universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„Correctly counting molecules using unique molecular identifiers“

verfasst von / submitted by

Dipl.-Ing. Florian Pflug

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2019 / Vienna 2019

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 794 685 490
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Molekulare Biologie / Molecular Biology
Betreut von / Supervisor:	Univ.-Prof. Dr. Arndt von Haeseler

This page intentionally left blank.

Acknowledgements

First of all, I want to thank my supervisor Arndt for offering me the chance to write this thesis, for allowing me to choose the direction of my research, and for his guidance and support throughout the project. The last 4 years at the *CIBIV* have been a blast and have taught me a lot, both scientifically and otherwise.

Much of my research presented here would not have been possible without Armin and Simon U.; I am grateful our inspiring and fruitful collaboration.

At times, navigating the intricacies of my research felt simple next to navigating the intricacies of university bureaucracy; without Iris, Gerlinde and Zahra, I would surely have failed at the latter.

Despite the interesting research, life at CIBIV would have been dull without my colleagues and friends there; many thanks to all present and former members; but in particular to Milica, Celine, Olga, Moritz, Luis, Connie, Tung and Philipp for making the CIBIV a welcoming and enjoyable place to be at from the first day on; and to Simon H., Veronika, Alina, Martin, Sebastian and Tamara for continuing to do so to the present day.

Thank you Johanna for teaching me biology and getting me interested in bioinformatics; Anna for many good talks during our quest for the “Goldene Wiener Wandernadel“; Shivi and Bernie for being good friends for more than half of my life; Judith, Vinka and Mathieu for many evenings filled with fun and interesting discussions; Uschi and Peter for all the time spent on snowy slopes and rocky walls; and all other people in my life for making life worth living no matter whether my research was progressing well or not.

Last but not least, I am grateful to my parents for always supporting my curiosity about science, math and computing, and to my sister Verena for showing me that science and math are not everything.

This page intentionally left blank.

Abstract

Precise measurement of the state of biological systems is a fundamental requirement in *quantitative biology* – the field dealing with the deduction and validation of quantitative models of biological phenomena. Such measurement techniques, for example of mRNA transcript or microbial species abundances, are often based on *next generation sequencing* (NGS).

A major factor limiting the precision of such *quantitative* NGS applications is *amplification bias* – the preferential amplification of some sequences by the *polymerase chain reaction* (PCR) step required before sequencing. To overcome this bias, *unique molecular identifiers* (UMIs) are added to each molecule prior to amplification, and abundances are then estimated from the number of distinct UMIs, not the number of sequencing reads. But while more faithful than read counts, such UMI counts are still *indirectly* biased by preferential amplification; stronger amplification of particular sequences translates into a lower risk of such molecules being overlooked during sequencing entirely, and thus of their UMI not being counted. This indirect bias is not amenable to improved *in vitro* techniques and must thus be tackled *in silico*.

Towards this goal, this work introduces the computational method TRUmiCount, based on a stochastic Galton-Watson branching process model of the PCR and a model of sequencing as Poissonian sampling. TRUmiCount combines these models to predict the number of unobserved from the number of observed molecules, and by employing a statistical denoising technique called *shrinkage estimation*, can do so on the level of individual mRNA transcripts or microbial species. These predictions thus allow *in silico* corrections for indirect amplification bias, and for UMI-based RNA-Seq experiments, TRUmiCount is demonstrated to produce unbiased mRNA transcript counts from raw bias-afflicted data.

Insertion pool sequencing (iPool-Seq) is another example of an experimental technique that benefits from TRUmiCount. iPool-Seq was developed to study host-

pathogen systems such as maize infected by *Ustilago maydis*, the fungus that causes corn smut. iPool-Seq compares the abundances of different mutants of the pathogen before and after infection of the host plant, with the goal of identifying mutants whose virulence – their ability to proliferate on the host – differs from the wild-type. For iPool-Seq data, TRUmiCount is shown to improve the fidelity of the abundance estimates, and thus ensures that amplification bias does not affect the measured virulences.

The statistical analysis of TRUmiCount-corrected NGS data poses unique challenges. After correction, UMI counts are fractional instead of integral, and can no longer be expected to obey the Poissonian mean-variance relationship often assumed for count data. To show how statistical inference methods can take TRUmiCount corrections into account while evading these issues, a statistical model of iPool-Seq incorporating these corrections is derived, and shown to provide a means of separating significant from insignificant differences between the virulences of mutants and wild-type.

Finally, a user-friendly data analysis pipeline for iPool-Seq data is presented. It includes all steps necessary to transform raw sequencing reads into TRUmiCount-corrected mutant abundances, and assesses the virulence of mutants, and the significance of their deviation from the wild-type. Detailed step-by-step description of both the wet-lab and the data-analysis parts of iPool-Seq are meant to provide easy access to the iPool-Seq method for as many potential users as possible.

Parts of this thesis have been published in the following articles:

Pflug F. G., & von Haeseler A. (2018). TRUmiCount: correctly counting absolute numbers of molecules using unique molecular identifiers. *Bioinformatics, 34*(18), 3137–3144. DOI:10.1093/bioinformatics/bty283

Uhse S., **Pflug F. G.**, Stirnberg A., Ehrlinger K., von Haeseler A., & Djamei A. (2018). In vivo insertion pool sequencing identifies virulence factors in a complex fungal-host interaction. *PLoS Biology, 16*(4), e2005129. DOI:10.1371/journal.pbio.2005129

Uhse S.[e], **Pflug F. G.**[e], von Haeseler A., & Djamei A. (2019). Insertion pool sequencing for insertional mutant analysis in complex host-microbe interactions. *Current Protocols in Plant Biology, 4*, e20097. DOI:10.1002/cppb.20097

[e]these authors contributed equally to this work

Kurzfassung

Präzise Messungen des Zustandes eines biologischen Systems bilden eine Grundlage der *quantitativen Biologie* – also jenes wissenschaftlichen Gebietes welches sich mit der Erstellung und Validierung quantitativer Modelle biologischer Systeme beschäftigt. Solche Techniken, beispielsweise zur Messung der Abundanz von Transkripten oder verschiedener mikrobieller Spezies, basieren oft auf Sequenzierungsmethoden hohen Durchsatzes (eng. next-generation sequencing; NGS).

Einer der hauptsächlichen die Genauigkeit limitierenden Faktoren in *quantitativen* NGS-Anwendungen sind Verzerrungen der gemessenen Abundanzen wegen der bevorzugten Amplifikation mancher Sequenzen durch die *Polymerasekettenreaktion* (eng. polymerase chain reaction; PCR). Um das zu vermeiden wird vor der Amplifikation jedes Molekül mit einem *eindeutigen molekularen Identifikator* (eng. unique molecular Identifier; UMI) versehen, und Abundanzen werden dann aus der Anzahl an verschiedenen UMIs geschätzt, nicht aus der Anzahl an sequenzierten Kopien. Aber auch die Anzahl an UMIs wird, wenngleich schwächer, trotzdem *indirekt* durch bevorzugte Amplifikation verzerrt; stärkere Amplifikation eines Moleküles reduziert das Risiko, dass keine seiner Kopien sequenziert und das Molekül damit nicht gezählt wird. Dieser indirekte Effekt kann nur durch Korrekturen *in silico* in Angriff genommen werden.

Dazu stellt diese Arbeit die Methode *TRUmiCount* vor, welche auf einem Modell der PCR als stochastischer Galton-Watson Verzweigungsprozess und von Sequenzierung als Poisson'scher Stichprobennahme basiert. Damit schätzt TRUmiCount die Anzahl an nicht beobachteten Molekülen aus jener der beobachteten, und verwendete eine statistische Rauschunterdrückungstechnik um diese Schätzung auf die Ebene einzelner Transkripte oder mikrobieller Spezies zu erweitern. Mit Hilfe von TRUmiCount können damit indirekte PCR-bedingte Verzerrungen korrigiert werden, und für UMI-basierte RNA-Sequenzierung wird gezeigt, dass TRUmiCount aus verzerrten Rohdaten unverzerrte Transkriptabundanzen ermitteln kann.

iPool-Seq (eng. *insertion pool sequencing*) ist eine experimentelle Methode zum Studium von Wirt-Pathogen-Systemen, z.B. Mais und *Ustilago maydis* (Verursacher des Maisbeulenbrandes), und profitiert ebenfalls von TRUmiCount. iPool-Seq vergleicht die Abundanz von Mutanten des Pathogens vor und nach der Infektion des Wirts, und identifiziert Mutanten mit veränderter *Virulenz* – ihrer Fähigkeit, sich auf dem Wirt zu vermehren. Es wird gezeigt, dass TRUmiCount die Genauigkeit dieser Messungen erhöht, und damit verhindert, dass PCR-bedingte Verzerrungen die gemessenen Virulenzen beeinflussen.

Die statistische Analyse von TRUmiCount-korrigierten NGS-Daten muss berücksichtigen, dass korrigierte Abundanzen nicht mehr ganzzahlig sind, und typischen Poisson'schen Annahmen über das Mittelwert-Varianz-Verhältnis nicht mehr entsprechen. Um zu zeigen wie diese Probleme umgangen werden können, leiten wir ein statistisches Modell für iPool-Seq her in welches Korrekturen für indirekte PCR-Verzerrungen einfließen, und zeigen, dass damit signifikante von insignifikanten Virulenzänderungen unterschieden werden können.

Abschließend wird ein benutzerfreundliches Programm zur Analyse von iPool-Seq-Daten präsentiert. Es inkludiert alle Schritte um aus rohen Sequenzdaten (korrigierte) Abundanzen zu ermitteln, und bestimmt für alle Mutanten Virulenz sowie die Signifikanz ihrer Abweichung vom Wildtyp. Eine schrittweise Beschreibung aller notwendigen Labor- und Datenauswertungsschritte, soll die Methode einem möglichst großen Kreis an Benutzern zugänglich machen.

Teile dieser Arbeit wurden in folgenden Artikeln publiziert:

Pflug F. G., & von Haeseler A. (2018). TRUmiCount: correctly counting absolute numbers of molecules using unique molecular identifiers. *Bioinformatics, 34*(18), 3137–3144. DOI:10.1093/bioinformatics/bty283

Uhse S., **Pflug F. G.**, Stirnberg A., Ehrlinger K., von Haeseler A., & Djamei A. (2018). In vivo insertion pool sequencing identifies virulence factors in a complex fungal-host interaction. *PLoS Biology, 16*(4), e2005129. DOI:10.1371/journal.pbio.2005129

Uhse S.[e], **Pflug F. G.**[e], von Haeseler A., & Djamei A. (2019). Insertion pool sequencing for insertional mutant analysis in complex host-microbe interactions. *Current Protocols in Plant Biology, 4*, e20097. DOI:10.1002/cppb.20097

[e] diese Autoren haben zu gleichen Teilen zu dieser Arbeit beigetragen

Contents

List of figures and tables

Chapter 6. Statistically modelling iPool-Seq

This page intentionally left blank.

Chapter 1

Introduction

Even a single living cell is a marvellously sophisticated system, sustained by a vast number of concurrent chemical and physical processes. Yet the whole is even more than the sum of its parts – it is not from these processes' individual complexities, but from their *interactions* that the almost infinite complexity of the living world around us emerges. To understand such a system we must thus not only catalogue all the individual parts, but also consider their effects on one another. Do, for example, the effects of two opposing processes cancel, or does one prevail? How do the effects of two processes working in conjunction add up? How sensitive is one process to changes in another? In short, we must understand the processes *quantitatively*.

Gaining such a quantitative understanding of the individual biological processes in a complex system, and through that an understanding of their possible interactions and thus of the system as a whole is the fundamental goal of *quantitative biology*. Such a quantitative understanding is inseparably linked with the creation of a mathematical model of the *state* of the system in question, of its temporal evolution, and of its spatial pattern – or in the language of physics, to the creation of a *theory* of the system (Howard, 2014).

As in physics, good theories are firmly rooted in and regularly challenged with experimental data, making high-fidelity measurements of the state of a system crucial to its understanding; as Howard (2014) puts it *measurement focuses the mind*.

Yet obtaining such high-fidelity measurements is notoriously difficult – in particular when we deal with complex systems, where we have to measure many parameters in parallel to capture the system's true state well, and where some of the chemical and physical processes and properties that sustain the system may interfere with measurements. Measurements of a complex system's state are thus always *incomplete*, and afflicted by *errors*[1] consisting of both *bias* (systematic deviations of measured from true values) and *noise* (random fluctuations of measured around true values). As Hastings *et al.* (2005) put it,

> "*Quantitative biologists must come to grips with a complex, stochastic, poorly observed system.*"

When we form theories and challenge them with data, we must account for our observations being *poor*, lest our theory confuse the properties of the system with the deficits of our method of observation. But to account for such effects when we form and challenge *quantitative* theories of a system with data we require *quantitative* insight into how *poor* our observations really are – into how strongly our observations are biased towards one value or another, how much they fluctuate randomly, and which experimental conditions affect these measurement errors.

In other words, we require a *theory* of our chosen method of observation. Only with such a quantitative theory of our methods as a basis can we then proceed as Hastings *et al.* (2005) suggests, and *couple* that theory with possible theories of the system to see whether they are consistent with observed data.

This thesis will present a theory for a particular class of methods based on *unique molecular identifiers* (UMIs) and *next-generation sequencing* (NGS) for observing the state of a single cell, a complex tissue, or a microbial community.

Before we describe these concepts and the class of methods considered in more detail, we take a step back and consider why these methods offer a powerful way of gaining insights into the state of many different types of complex biological systems.

[1] An insight succinctly rendered in german as: *Wer misst, misst Mist*

DNA or RNA abundances as a proxy of system state

When we attempt to measure the state of a biological system, the first difficulty to overcome is the staggering size of the full state space of even the smallest such system. Already for a single cell, and ignoring spatial configuration completely, measuring the full state seems intractable; it would require us to measure the precise abundance of every type of chemical compound in a cell simultaneously. To arrive at a tractable problem, we thus have to limit our attention further, to a smaller class of molecules; yet whose combined abundance data is still informative about the current full state of the cell and its future behaviour.

Cells contain four major types of small organic molecules, *sugars*, *fatty acids*, *amino acids* and *nucleotides*, each of which form the building blocks of one of the major classes of larger chemical compounds, namely *polysaccharids*, *lipids*, *proteins* and the *nucleic acids*, consisting of *deoxyribonucleic acid* (DNA) and *ribonucleic acid* (RNA) (Alberts *et al.*, 2002). While the abundance of the different types of chemical compounds in all of these classes could be argued to be relevant components of the current cell state, simultaneous measurement of all of these abundances is currently only tractable for DNA and RNA; this has been made possible through the advent of *next-generation sequencing* (NGS; Mardis, 2008, 2013; Metzker, 2010).

And there are also plenty of biological reasons why DNA and RNA abundances offer insight into the state of many different types of biological systems. By counting DNA we count *genomes*, and thus the abundance of (single-celled) species or mutants of a species; by counting mRNAs we can infer which genes are actively expressed; by immuno-chemically selecting DNA or RNA bound to a particular protein prior to counting, we can observe the intra-cellular machinery that controls DNA replication, transcription and translation in action; just to name a few examples.

Measuring DNA or RNA abundances through NGS

In the last two decades, DNA sequencing throughput has improved by more than six orders of magnitude, from being able to sequence about one hundred molecules in parallel to being able to sequence hundreds of millions of molecules. The sequencing methods that enable this are collectively referred to as *next-generation sequencing*

(NGS; Mardis, 2008, 2013; Metzker, 2010), with the most widely used method being the one offered commercially by *Illumina, Inc.*

The large number of molecules sequenced simultaneously allows these methods to be used *quantitatively*. An early such application of NGS is *RNA sequencing* (RNA-seq; Z. Wang *et al.*, 2009), where all *messenger RNAs* (mRNAs) are extracted from a biological sample (traditionally consisting of many similar cells), converted back to *complementary DNA* (cDNA) by reverse transcription, and subjected to next-generation sequencing (figure 1 in chapter 1 on page 21). Each of the millions of *sequencing reads* is then aligned to the (presumably known) reference genome of the organism in question, and by counting the number of reads overlapping each gene, we obtain simultaneous estimates of the (relative) abundances of all mRNA transcripts of the genes of the organism.

But quantitative applications of NGS are not limited to RNA-seq. The general principle of sequencing all DNA or RNA molecules in a sample and then counting the abundance of specific sequences can be used to measure a large range of different types of *state* of a biological system; a few (quite random and certainly non-exhaustive) examples are:

Metagenomic analyses to detect all present taxa and their abundances in a microbial community by sequencing, assignment of reads to taxa, and counting the number of reads per taxa (Escobar-Zepeda *et al.*, 2015);

ChIP-Seq to study DNA-protein interactions by immuno-precipitation of the chromatin using a specific antibody, sequencing, mapping of reads to the genome, and looking for coverage peaks (Barski *et al.*, 2007; Johnson *et al.*, 2007);

Ribo-Seq to study the location of the ribosomes on the transcripts by digestion of mRNAs except for the parts protected by a bound ribosome, sequencing, mapping of reads to the genome, and looking for coverage peaks (Ingolia *et al.*, 2009);

Bacteraemia diagnosis to enable faster personalised treatment of septic patients by sequencing of the free DNA in the patient's blood, assignment of reads to possible causative bacteria, and statistical analysis to separate pathogen(s) from background (Grumaz *et al.*, 2016).

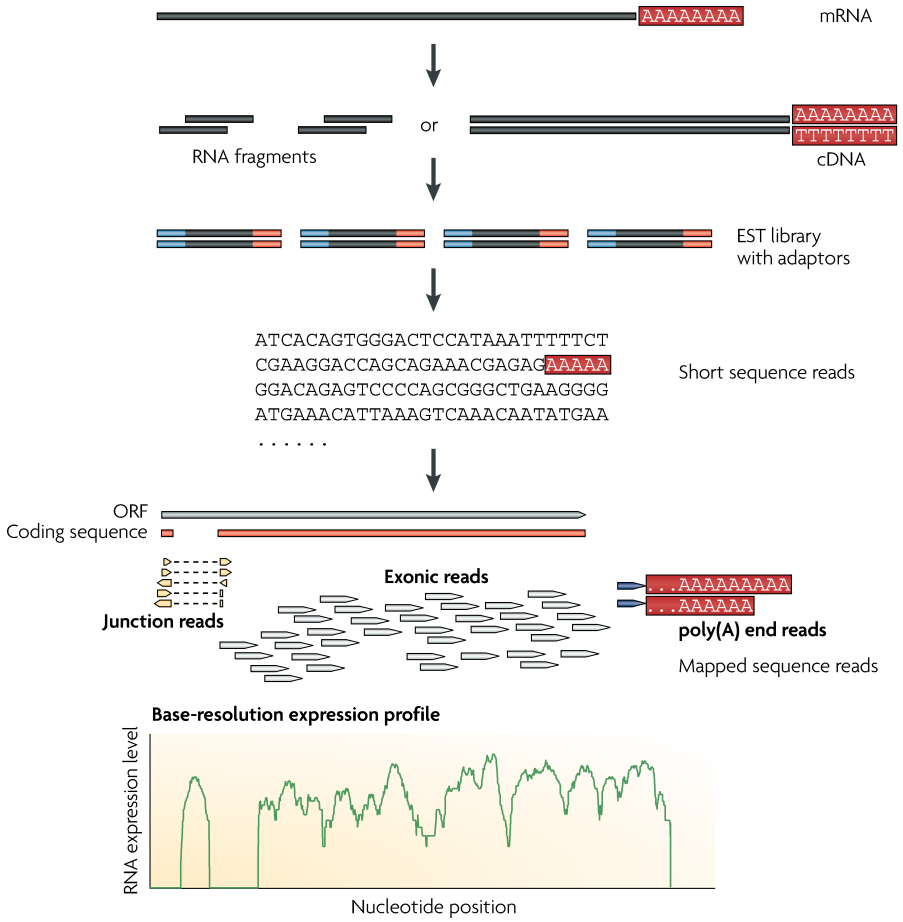

Figure 1: **A typical RNA-seq experiment.** Briefly, long RNAs are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation. Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using next-generation sequencing technology. The resulting sequence reads are aligned with the reference genome or transcriptome, and classified as three types: exonic reads, junction reads and poly(A) end-reads. These three types are used to generate a base-resolution expression profile for each gene, as illustrated at the bottom; a yeast ORF with one intron is shown (Z. Wang *et al.*, 2009; both figure and legend).

Amplification biases DNA and RNA abundances

While new applications of NGS continue to be developed, and while the technology itself has progressed far over the last decade, with both the number of reads produced by a single sequencing run, as well as their length, having increased greatly, one aspect has changed only little – the method is quite *lossy*. Typically, to sequence a sample on a Illumina HiSeq lane, a volume of at least 10 μl with a DNA concentration of 10 nM is required (Illumina Inc., 2012), which amounts to $\approx 10\,\text{nM} \cdot 10\,\mu\text{l} = 100$ femtomole or ≈ 60 billion DNA molecules[2]. As a single HiSeq lane can be expected to yield ≈300 million reads (Illumina Inc., 2019), of of the 60 billion input molecules thus only about 0.5% eventually get sequenced and the other 99.5% are lost. Others estimate the loss to be even higher, one estimate puts the loss for RNA-seq experiments at 99.9% of molecules (McIntyre *et al.*, 2011).

To compensate for this large loss of input molecules, and to reach the required DNA concentration and purity, DNA libraries are therefore almost always *amplified* before sequencing. After ensuring that the original DNA fragments carry appropriate adapters with fixed sequences at both ends, a *polymerase chain reaction* (PCR; Mullis *et al.*, 1986) is used to repeatedly duplicate each DNA molecule, which (over multiple cycles) produces thousands of copies from a single original template molecule.

But while the PCR can amplify (almost) any sequence of nucleotides, it does not generally do so with uniform *efficiency*. The reaction tends to systematically prefer some sequences over others, and over multiple cycles these preferences accumulate, considerably biasing the final library (Kanagawa, 2003; Suzuki and Giovannoni, 1996). While some factors like GC content that influence efficiency are known, the *magnitude* of their effect depends heavily experimental details such as the precise PCR protocol and enzyme used (Aird *et al.*, 2011; figure 2 in chapter 1 on page 23).

For quantitative NGS, these biases translate to miss-quantification of molecule abundances (Aird *et al.*, 2011; Dabney and Meyer, 2012; Kebschull and Zador, 2015; Pawluczyk *et al.*, 2015; Pinto and Raskin, 2012). While models exists to correct for some types of bias *in silico* (Love, Huber, *et al.*, 2014), a more general solution was sought that applies equally well to all types of quantitative NGS applications.

[2] One *mole* of a substance is defined as exactly $6.02214076 \cdot 10^{23}$ molecules.

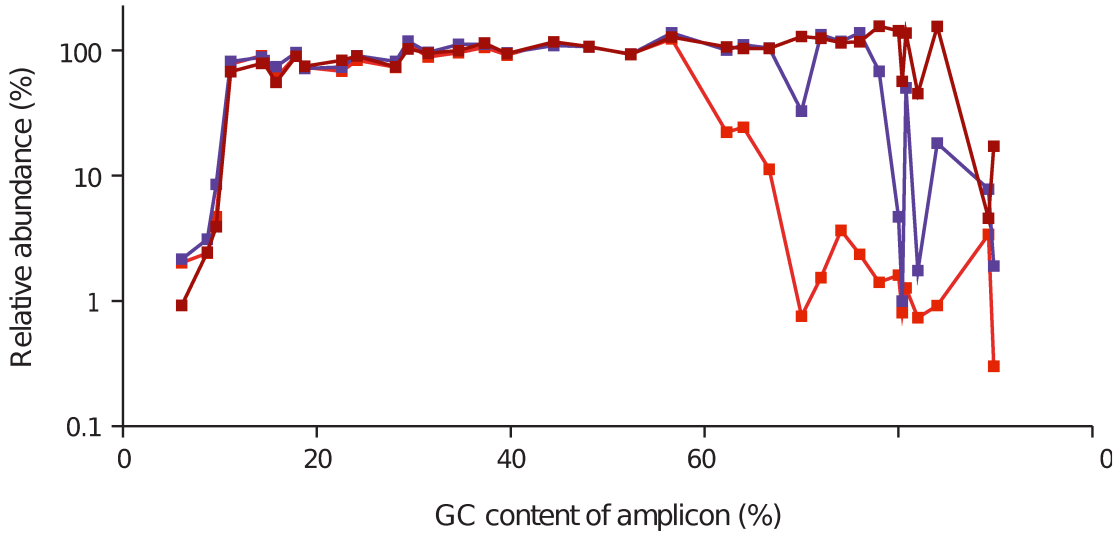

Figure 2: **Effect of temperature ramp rates.** The standard PCR protocol with Phusion HF DNA polymerase and short initial (30 s) and in-cycle (10 s) denaturation times was performed on three different thermo-cyclers at their respective default temperature ramp settings. Heating and cooling rates were 6°C/s and 4.5°C/s on thermo-cycler 1 (bright red line), 4°C/s and 3°C/s on thermo-cycler 2 (purple line) and 2.2°C/s and 2.2°C/s on thermo-cycler 3 (dark red line) (Aird *et al.*, 2011; both figure and legend).

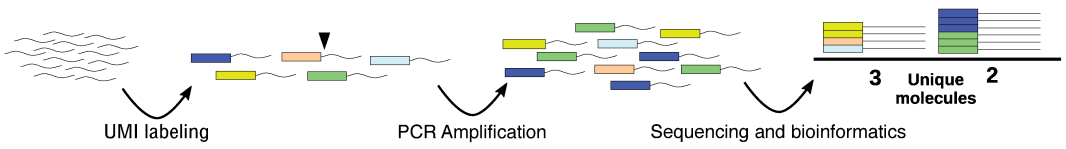

Figure 3: **Schematic representation of how UMIs are used to count unique molecules.** Fragmented DNA is labeled with a random UMI sequence (short oligonucleotide; represented as coloured blocks). Following PCR amplification, sequencing, and bioinformatics steps, the sequence read alignment coordinates and UMI sequences are used to identify sequence reads originating from the same initial DNA fragment (PCR duplicates) and so count the unique molecules. (Smith *et al.*, 2017; both figure and legend, figure slightly modified).

Unique molecular identifiers (UMIs)

Unique molecular identifiers (UMIs; Hug and Schuler, 2003; Kivioja *et al.*, 2011) were introduced to overcome these PCR-induced amplification biases. The idea is to make molecules distinguishable before sequencing through some sort of unique identifiers (the UMI) – typically by appending a short, random oligonucleotide to each molecule before amplification. After PCR amplification and sequencing, molecule abundances are determined by counting the number of distinct UMIs, amongst all reads with a particular sequence (figure 3 in chapter 1 on page 23). Compared to read counts, UMI counts show better concordance to true abundances due to being less sensitive to PCR biases (Kivioja *et al.*, 2011), and they provide absolute quantification of molecules numbers instead of only relative abundances as read counts do.

But while UMI counts reduce the effect of PCR bias and improve concordance, an *indirect* form of PCR bias remains. Of molecules that are amplified less efficiently, fewer copies will tend to exist post-amplification. This in turn increases those molecules' risk of all of their copies (and thus the original molecule) remaining *unobserved* by the sequencing process entirely – sequencing is, after all, not exhaustive but rather very lossy.

It may seem that this indirect form of PCR bias can be avoided by increasing the sequencing depth to a point where essentially no molecules are overlooked. Yet while doing so will indeed avoid indirect PCR bias, a thought experiment shows that doing so may actually *reduce* the overall accuracy! Let us imagine we sequence a given sample again and again to increase the total sequencing depth further and further. We should then expect each UMI to be represented by an ever larger number of reads; yet because some of those reads will inevitably contain errors, we will also detect an ever growing number of distinct UMIs as the total sequencing depth increases! In addition to sequencing errors, at higher sequencing depths (due to the growing number of PCR cycles necessary to produce enough material for sequencing) we must expect to also see a higher number of chimeric molecules (Haas *et al.*, 2011) and other amplification artefacts, which further increases the number of *phantom UMIs* that do not represent a molecule in the original sample.

To summarise, increasing the sequencing depth to counter-act indirect PCR

bias requires somehow filtering out phantom UMIs. Yet any such filter will inevitably remove some *true UMIs* (stemming from actual molecules) as well, in turn increasing PCR bias.

To chose optimally in this *Catch-22* (Heller, 1961) situation, we must understand the consequence of particular choices of sequencing depth and filtering on the accuracy of our measurements *quantitatively*. Or put in the language of the introductory paragraphs, we required a (quantitative) *theory*.

Towards a *theory* of UMI-based quantitative NGS

Since the introduction of the UMI method, some progress has been made to understand the method, and to provide tools to correct for its shortcomings *in silico*.

Algorithms have been developed to detect and merge erroneous versions of UMIs produced by sequencing errors (Liu, 2019; Smith *et al.*, 2017). Models of different types of PCR-induced distortions such as stochastic behaviour, biases, and template switching have been studied (Best *et al.*, 2015; Kebschull and Zador, 2015). Different variations of the UMI method have been developed, in particular some with UMIs designed to be more resilient to sequencing errors (Shiroguchi *et al.*, 2012).

Yet while these works offer some insight into the behaviour of UMI-based NGS experiments, it remains to form a coherent mathematical theory. Such a theory should enable us to predict the magnitude of indirect PCR biases under different experimental conditions and phantom UMI filters, and be precise enough to yield *corrections* which, when applied to observed UMI counts, increase the fidelity of the results.

This thesis presents such a theory and studies the accuracy of its predictions for different types of UMI-based NGS experiments and different conditions (Pflug & von Haeseler, 2018; chapters 2 and 5 of this work). With our theory of UMI-based NGS experiments as a basis, we then present a statistical model of a particular such method called *insertion pool sequencing* (iPool-Seq; Uhse et al., 2018; chapters 3, 4 and 6 of this work).

References in this chapter

Aird D., Ross M. G., Chen W.-S. S., Danielsson M., Fennell T., Russ C., … Gnirke A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, *12*(2), R18. DOI:10.1186/gb-2011-12-2-r18

Alberts B., Johnson A., & Lewis J. (2002). Molecular Biology of the Cell. (4th). New York (NY): Garland Science.

Barski A., Cuddapah S., Cui K., Roh T.-Y., Schones D. E., Wang Z., … Zhao K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, *129*(4), 823–837. DOI:10.1016/j.cell.2007.05.009

Best K., Oakes T., Heather J. M., Shawe-Taylor J., & Chain B. (2015). Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. *Nature Publishing Group*, 1–13. DOI:10.1038/srep14629

Dabney J., & Meyer M. (2012). Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *BioTechniques*, *52*(2). DOI:10.2144/000113809

Escobar-Zepeda A., Vera-Ponce de León A., & Sanchez-Flores A. (2015). The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. *Frontiers in genetics*, *6*, 348. DOI:10.3389/fgene.2015.00348

Grumaz S., Stevens P., Grumaz C., Decker S. O., Weigand M. A., Hofer S., … Sohn K. (2016). Next-generation sequencing diagnostics of bacteremia in septic patients. *Genome Medicine*, *8*(1), 73. DOI:10.1186/s13073-016-0326-8

Haas B. J., Gevers D., Earl A. M., Feldgarden M., Ward D. V., Giannoukos G., … Birren B. W. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome research*, *21*(3), 494–504. DOI:10.1101/gr.112730.110

Heller J. (1961). *Catch-22*. New York (NY): Simon and Schuster.

Howard J. (2014). Quantitative cell biology: the essential role of theory. *Molecular Biology of the Cell*, *25*(22), 3438–3440. DOI:10.1091/mbc.e14-02-0715

Hug H., & Schuler R. (2003). Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. *Journal of theoretical biology*, *221*(4), 615–624. DOI:10.1006/jtbi.2003.3211

Illumina Inc. (2012). TruSeq DNASample Preparation Guide. Retrieved August 6, 2019, from http://emea.support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_truseq/truseqdna/TruSeq_DNA_SamplePrep_Guide_15026486_C.pdf

Illumina Inc. (2019). HiSeq 3000/HiSeq 4000 System quality and performance. Retrieved August 6, 2019, from https://www.illumina.com/systems/sequencing-platforms/hiseq-3000-4000/specifications.html

Ingolia N. T., Ghaemmaghami S., Newman J. R. S., & Weissman J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, *324*(5924), 218–23. DOI:10.1126/science.1168978

Johnson D. S., Mortazavi A., Myers R. M., & Wold B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science*, *316*(5830), 1497–1502. DOI:10.1126/science.1141319

Kanagawa T. (2003). Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of Bioscience and Bioengineering*, *96*(4), 317–323. DOI:10.1016/S1389-1723(03)90130-7

Kebschull J. M., & Zador A. M. (2015). Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Research*, *43*(21), gkv717. DOI:10.1093/nar/gkv717

Kivioja T., Vähärautio A., Karlsson K., Bonke M., Enge M., Linnarsson S., & Taipale J. (2011). Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, *9*(1), 72–74. DOI:10.1038/nmeth.1778

Liu D. (2019). Algorithms for efficiently collapsing reads with Unique Molecular Identifiers. *bioRxiv*. DOI:10.1101/648683

Love M. I., Huber W., & Anders S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. DOI:10.1186/s13059-014-0550-8

Mardis E. R. (2008). Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*, *9*(1), 387–402. DOI:10.1146/annurev.genom.9.081307.164359

Mardis E. R. (2013). Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry*, *6*(1), 287–303. DOI:10.1146/annurev-anchem-062012-092628

McIntyre L. M., Lopiano K. K., Morse A. M., Amin V., Oberg A. L., Young L. J., & Nuzhdin S. V. (2011). RNA-seq: Technical variability and sampling. *BMC Genomics*, *12*(1), 293. DOI:10.1186/1471-2164-12-293

Metzker M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, *11*(1), 31–46. DOI:10.1038/nrg2626

Mullis K., Faloona F., Scharf S., Saiki R., Horn G., & Erlich H. (1986). Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, *51*, 263–273. DOI:10.1101/SQB.1986.051.01.032

Pawluczyk M., Weiss J., Links M. G., Egaña Aranguren M., Wilkinson M. D., & Egea-Cortines M. (2015). Quantitative evaluation of bias in PCR amplification and next-generation sequencing derived from metabarcoding samples. *Analytical and Bioanalytical Chemistry*, *407*(7), 1841–1848. DOI:10.1007/s00216-014-8435-y

Pinto A. J., & Raskin L. (2012). PCR Biases Distort Bacterial and Archaeal Community Structure in Pyrosequencing Datasets. *PLoS ONE*, *7*(8), e43093. DOI:10.1371/journal.pone.0043093

Shiroguchi K., Jia T. Z., Sims P. A., & Xie X. S. (2012). Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proceedings of the National Academy of Sciences*, *109*(4), 1347–1352. DOI:10.1073/pnas.1118018109

Smith T., Heger A., & Sudbery I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome research*, *27*(3), 491–499. DOI:10.1101/gr.209601.116

Suzuki M. T., & Giovannoni S. J. (1996). Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and environmental microbiology*, *62*(2), 625–30.

Wang Z., Gerstein M., & Snyder M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. DOI:10.1038/nrg2484

Chapter 2

TRUmiCount: Correctly counting absolute numbers of molecules using unique molecular identifiers (Pflug & von Haeseler, 2018. *Bioinformatics*, *34*(18), 3137-3144.)

Preamble

In this publication we introduce a *theory* of *unique molecular identifier* (UMI) - based *next-generation sequencing* (NGS) experiments. This theory provides the basis for the computational *TRUmiCount* method which removes PCR artefacts and (indirect) PCR bias from the abundances of particular DNA or DNA fragments measured in such experiments.

For the sake of simplicity, we present our method in the typical terms of *RNA-Sequencing* (RNA-Seq), i.e. we assume the quantities of interest are the copy numbers of *messenger RNA* (mRNA) molecules transcribed from the different genes of an organism. Despite this choice of presentation, the method itself is not restricted to RNA-Seq settings. In fact, it applies equally well to any experimental design where all individual DNA or RNA fragments are (made) *distinguishable*

before being amplified by a *polymerase chain reaction* (PCR), and where each sequencing read can thus be traced back to a single pre-amplification molecule.

This is in particular the case for *insertion pool sequencing* (iPool-Seq; Uhse *et al.*, 2018; chapter 3 of this work), developed to measure of the virulence of particular mutants of the maize pathogen *U. maydis*. In fact, the need to analyse iPool-Seq data provided the initial motivation for development of the TRUmiCount method; the generalisation to RNA-Seq and other UMI-based NGS experiments happened later. For iPool-Seq data, TRUmiCount is crucial in ensuring that the measured virulences are not biased by referential PCR amplification of particular mutant sequences; chapter 5 discusses the performance of the method for iPool-Seq data.

The author considers TRUmiCount to be the main contribution of this thesis to the field of NGS data analysis.

Author contributions

Mathematical modelling, software development and writing of the manuscript were done by the author of this thesis, under the guidance and supervision of Arndt von Haeseler.

Publication history and status

Submitted on February 2nd 2018 to *Bioinformatics*, accepted for publication on April 12th 2019, published online on April 16th 2018.

Pflug F. G., & von Haeseler A. (2018). TRUmiCount: correctly counting absolute numbers of molecules using unique molecular identifiers. *Bioinformatics, 34*(18), 3137–3144. DOI:10.1093/bioinformatics/bty283

Bioinformatics, 34(18), 2018, 3137–3144
doi: 10.1093/bioinformatics/bty283
Advance Access Publication Date: 16 April 2018
Original Paper

OXFORD

Gene expression

TRUmiCount: correctly counting absolute numbers of molecules using unique molecular identifiers

Florian G. Pflug[1,*] and Arndt von Haeseler[1,2]

[1]Center for Integrative Bioinformatics Vienna (CIBIV), Joint Institute of the University of Vienna and Medicial University of Vienna, Max F. Perutz Laboratories (MFPL), A-1030 Vienna, Austria and [2]Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, A-1090 Vienna, Austria

*To whom correspondence should be addressed.
Associate Editor: Bonnie Berger

Received on February 2, 2018; revised on March 23, 2018; editorial decision on April 4, 2018; accepted on April 12, 2018

Abstract

Motivation: Counting molecules using *next-generation sequencing* (NGS) suffers from PCR amplification bias, which reduces the accuracy of many quantitative NGS-based experimental methods such as RNA-Seq. This is true even if molecules are made distinguishable using *unique molecular identifiers* (UMIs) before PCR amplification, and distinct UMIs are counted instead of reads: Molecules that are lost entirely during the sequencing process will still cause underestimation of the molecule count, and amplification artifacts like PCR chimeras create phantom UMIs and thus cause over-estimation.
Results: We introduce the TRUmiCount algorithm to correct for both types of errors. The TRUmiCount algorithm is based on a mechanistic model of PCR amplification and sequencing, whose two parameters have an immediate physical interpretation as PCR efficiency and sequencing depth and can be estimated from experimental data without requiring calibration experiments or spike-ins. We show that our model captures the main stochastic properties of amplification and sequencing, and that it allows us to filter out phantom UMIs and to estimate the number of molecules lost during the sequencing process. Finally, we demonstrate that the phantom-filtered and loss-corrected molecule counts computed by TRUmiCount measure the true number of molecules with considerably higher accuracy than the raw number of distinct UMIs, even if most UMIs are sequenced only once as is typical for single-cell RNA-Seq.
Availability and implementation: TRUmiCount is available at http://www.cibiv.at/software/trumicount and through Bioconda (http://bioconda.github.io).
Contact: florian.pflug@univie.ac.at
Supplementary information: Supplementary information is available at *Bioinformatics* online.

1 Introduction

Experimental methods like RNA-Seq, ChIP-Seq and many others depend on *next-generation sequencing* (NGS) to measure the abundance of DNA or RNA molecules in a sample. The PCR amplification step necessary before sequencing often amplifies different molecules with different efficiencies, thereby biasing the measured abundances (Aird *et al.*, 2011). This problem can be alleviated by ensuring that all molecules are distinguishable before amplification by some combination of factors comprising a *unique molecular identifier* (UMI) (Hug and Schuler, 2003; Kivioja *et al.*, 2012), which usually includes a distinct molecular barcode ligated to each molecule before amplification (Fig. 1A, colored dots; see Smith *et al.* (2017) for a more extensive history of the UMI method). After amplification and sequencing, instead of counting reads, reads are grouped by UMI, and each distinct UMI is taken to reflect a distinct molecule in the original sample (Fig. 1A). But while the number

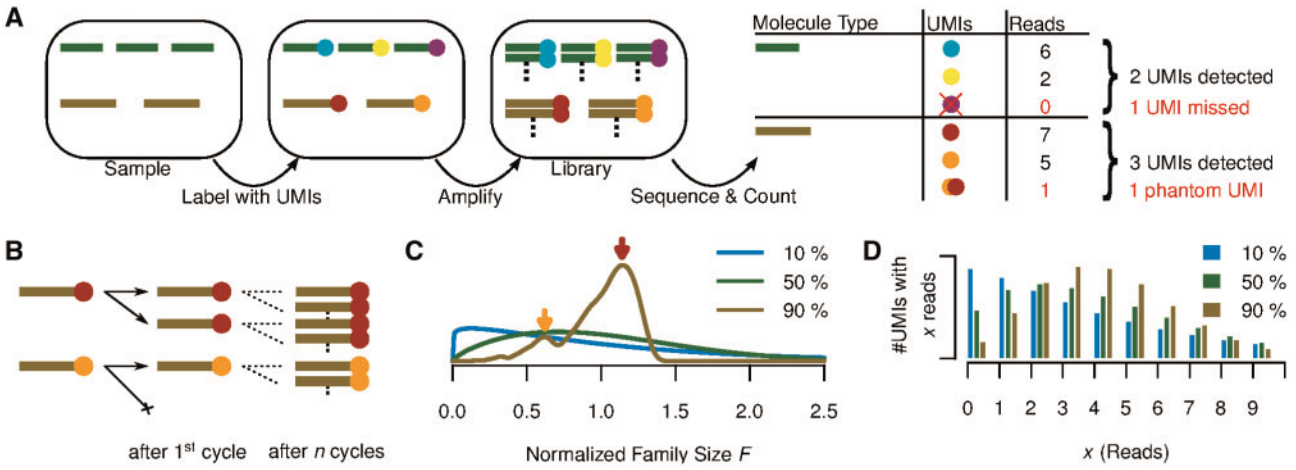

Fig. 1. (**A**) The relevant steps of library preparation when the UMI method is used. The sample initially contains three copies of molecule ▬ and two copies of ▬, which are made unique by labelling with UMIs (●, ●, ●, ●, ●). Each of those molecules is expanded into a molecular family during amplification, and a random selection of molecules from these families is sequenced. Counting unique UMIs then counts unique molecules, unless UMIs have read-count zero (●) or phantom UMIs are produced (●). **(B)** PCR as a Galton-Watson branching process. Molecule ▬● failed to be copied during the first PCR cycle and the final family size is thus reduced compared with ▬●. **(C)** Normalized family size distribution for efficiency 10, 50 and 90%. The arrows mark the most likely normalized family sizes for the two molecules from (B), assuming a reaction efficiency of 90%, and taking their distinct fates during the first PCR cycle into account. **(D)** Distribution of reads per UMI for efficiency 10, 50 and 90% assuming $D = 4$ Reads per UMI on average

of distinct UMIs may be a better proxy for the molecule count, it is still biased, for two reasons:

- Molecules that are amplified with low efficiency will have fewer copies made, hence fewer reads per UMI, and thus a higher chance of being left entirely unsequenced (Fig. 1A, green transcript, violet UMI).
- Sequencing errors, PCR chimeras, and index miss-assignment (Sinha *et al.*, http://www.biorxiv.org/content/early/2017/04/09/125724) in multiplexed sequencing runs can produce *phantom UMIs* which do not correspond to any molecule in the original sample (Fig. 1A, orange/red phantom UMI).

Various methods have been proposed to counter-act these effects: Smith *et al.* (2017) proposed an algorithm for merging highly similar, erroneous versions of the same original UMI to correct for sequencing errors and single-nucleotide PCR amplification errors. To filter out more complex PCR artifacts, strand-specific UMI-labeling protocols were introducted (Shiroguchi *et al.*, 2012; Schmitt *et al.*, 2012) that allow filtering out artifacts based on whether UMIs for both strands of a template molecule were detected. A correction for molecules left entirely unsequenced is mentioned by Kivioja *et al.* (2012), but being based on the Poission distribution, it severely under-estimates the amount of affected molecules; for their data by about an order of magnitude.

Instead of relying on sequence similary or complicated strand-specific UMI-labeling protocols, we rely on the per-UMI read count to separate *true UMIs* (i.e. UMIs of actual molecules in the original sample) from *phantom UMIs*. Chimeric PCR products are typically produced during later reaction cycles, and can therefore be expected to have smaller copy numbers and hence a lower read-count than non-chimeric PCR products. Index miss-assignment and sequencing errors typically happen randomly, and are unlikely to produce a larger number of reads showing the same phantom UMI. For these reasons, phantom UMIs can be expected to have a markedly lower read count than most *true UMIs*, i.e. UMIs of actual molecules in the original sample.

Our bias-correction and phantom-removal algorithm TRUmiCount exploits this difference in expected read counts between phantoms and true UMIs. It removes UMIs likely to be phantoms based on a read-count threshold, and then estimates and corrects the (gene-specific) loss, i.e. the fraction of molecules that were not sequenced or whose UMIs were mistaken for phantoms. For this correction TRUmiCount employs a model of PCR amplification that accounts for the stochasticity inherent to this amplification reaction.

2 Materials and methods

2.1 The TRUmiCount algorithm

The TRUmiCount algorithm consists of the following three steps:

1. We first filter out phantom UMIs by removing any UMI whose read count lies below a suitably chosen *error-correction threshold* (T).
2. We then estimate the loss (ℓ), i.e. the fraction of molecules that were not sequenced at all, or whose UMIs were removed by the error-correction threshold. This estimate is computed using a stochastic model of the amplification and sequencing process whose parameters are the PCR efficiency (E), and the sequencing depth (D), expressed as the average number of reads per UMI in the initial sample. From the observed distribution of reads per UMI, we estimate both (raw) gene-specific as well as library-wide values for these parameters, and compute corresponding estimates of the loss (see Section 2.2 for details).
3. Finally, we add the estimated number of lost UMIs back to the the observed number of true UMIs (those UMIs with $\geq$ threshold reads) to find the total number of molecules in the original sample. Since the loss can vary between genes, to yield unbiased counts, the correction must be based on gene-specific loss estimates. Due to the noise inherent to raw gene-specific estimates for genes with only few observed true UMIs, we employ a James-Stein-type (James and Stein, 1961) *shrinkage estimator*, adjusting the raw gene-specific parameter and loss estimates towards the library-wide ones (thus *shrinking* their difference). We choose the amount of shrinkage based on each estimate's precision, in such a way that the expected overall error is minimized (Carter and Rolph, 1974) (see Section 2.3).

2.2 Estimating the fraction of lost molecules

To estimate the loss, i.e. the fraction of molecules whose UMIs had a read count below the error-correction threshold, we model the distribution of per-UMI read counts by combining a stochastic model of PCR amplification with a model of NGS as random sampling.

2.2.1 A stochastic model of PCR amplification

To model PCR amplification, we use the *single-stranded* model of Krawczak *et al.* (1989), meaning we view PCR as a stochastic process that during each cycle duplicates each molecule independently with a particular probability E, called the reaction's *efficiency*. We further assume that a molecule is copied perfectly or not at all, i.e. that neither partial copies nor copies with a slightly different base-pair sequence are produced, that no molecules are destroyed or lost, and that the efficiency E stays constant throughout the reaction. Although this model has been extended by Weiss and von Haeseler (1997) to include the possibility of substitution errors during amplification, exhaustively modeling *all* possible sources of phantom UMIs seems futile. We therefore pursue a different approach, and model only the error-free case, trusting the error-correction threshold to remove phantoms. Over multiple cycles, each molecule is thus assumed to be expanded into a *molecular family* of identical copies. Since we use the single-stranded model, *molecule* for us always means a single-stranded piece of DNA, and we do not distinguish between a strand and its reverse complement. For our purposes, a piece of double-stranded DNA thus consists of two indistinguishable molecules.

Before amplification, we assume all molecules in the sample to be distinguishable by some UMI. During amplification, each of those molecules gives rise to a *molecular family* of (indistinguishable) copies. The initial size of such a family (i.e. the number of copies it is comprised of) is 1. During the first PCR cycle, the size increases to 2 if the single initial molecule is copied successfully, i.e. with probability E. Continuation of this process, always using all existing molecules as potential templates that are copied with probability E, produces a random sequence $M_0, M_1, M_2, \ldots$ of molecular family sizes after the 0th, 1st, 2nd, ... cycle. This sequence forms a Galton-Watson branching process (Weiss and von Haeseler, 1995), and follows the recursion

$$\begin{aligned} M_0 = 1, \quad M_i = M_{i-1} + \Delta_i \quad \text{where} \\ \Delta_i \sim \text{Binom}\,(M_{i-1}, E). \end{aligned} \tag{1}$$

Although we are not aware of a way to obtain an explicit formula for the distribution of the family size M_i after i PCR cycles, the expected value and variance of M_i can be computed explicitly. According to Harris (1989, Ch. 1), Equation (5.3), $\mathbb{V}M_i = \frac{\sigma^2 m^i (m^i - 1)}{m^2 - m}$ where m and σ are the mean and SD of M_1. In our case these are $m = 1 + E$ and $\sigma^2 = E \cdot (1 - E)$, thus we find

$$\mathbb{E}M_i = (1 + E)^i \tag{2}$$

$$\mathbb{V}M_i = \frac{1 - E}{1 + E} \cdot (1 + E)^i ((1 + E)^i - 1) \tag{3}$$

Equation 2 shows the well-known exponential growth of expected family sizes during PCR. But apart from recovering this well-known property of PCR, the Galton-Watson model also predicts the likelihood of *deviations* from this expectation due to random failures of copy operations, and by simulation allows us to find the actual distribution of M_i.

2.2.2 The normalized family size F

Due to the exponential growth of the expectation of M_i, the distribution of M_i depends heavily on the PCR cycle count i. That dependency, however, affects mostly the *scale*, not the *shape* of the distribution of M_i. To see the effect on the shape more clearly, the effect on the scale is removed by replacing M_i with a re-scaled version which has an expected value of one,

$$\tilde{M}_i = \frac{M_i}{\mathbb{E}M_i} = \frac{M_i}{(1 + E)^i}. \tag{4}$$

These re-scaled family sizes can be sensibly compared across PCR cycles. We observe that with growing cycle counts, the additional stochasticity introduced by each additional cycle drops rapidly. The re-scaled family size after the first cycle varies by a factor of two depending on whether the (single) copy operation during the first cycle succeeds or fails. Later on there are more templates to copy from, and thus the success or failure to copy any particular molecule averages out, making the behavior of the reaction more deterministic. Finally, $\tilde{M}_i \approx \tilde{M}_{i+1}$, because the family size M_i increases during each cycle almost exactly by a factor of $1 + E$, which matches the decrease of the re-scaling factor in $\tilde{M}_i$. This informal argument can be turned into a formal proof (see Harris (1989), Ch. 1, Th. 8.1) of the convergence of the re-scaled family size as i tends towards ∞, which allows us to remove the cycle count as a parameter entirely from what we call the *normalized family size*

$$F = \lim_{i \to \infty} \tilde{M}_i. \tag{5}$$

Although there is again no explicit formula known for the distribution of the normalized family size F, we find its variance from Equations (3–5) to be

$$\mathbb{V}F = \frac{1 - E}{1 + E}. \tag{6}$$

To quickly evaluate the density $f_F(x; E)$ of the distribution of F for a particular normalized family size x given reaction efficiency E, we interpolate using 2D polynomial interpolation (Akima, 1996) between pre-computed densities for different reaction efficiencies between 0 and 100% at different family sizes between 0 and 50 (see Supplementary Section S1.1 for details).

2.2.3 Modeling the sequencing process

The normalized family size distribution models the abundance of molecules with a particular UMI. To model the read count of a particular UMI after sequencing (i.e. the number of reads stemming from a particular pre-amplification molecule), we model NGS with a Poissonian sampling model (Marioni *et al.*, 2008). This amounts to assuming that (i) each individual copy has the same probability of being sequenced, (ii) this probability is small compared to the sequencing depth and (iii) there were many (distinguishable) original molecules. We further assume that a UMI is *on average* represented by D reads. Then the read count C of a UMI with known normalized molecular family size F is Poisson distributed,

$$\begin{aligned} C \mid F \sim \text{Poisson}\,(F \cdot D), \\ \mathbb{P}(C = k \mid F) = e^{-F \cdot D} \frac{(F \cdot D)^k}{k!}. \end{aligned} \tag{7}$$

In general, however, the exact family size F of any particular UMI is unknown—we only know the *distribution* of F. To compute the probability of a UMI having k reads, we average over all possible family sizes $x \in [0, \infty)$, weighting them with their respective density $f_F(x; E)$ in the distribution of the normalized family size F,

$$\mathbb{P}(C=k)=\int_0^\infty \mathbb{P}(C=k\,|\,F=x)\cdot f_F(x;\,E)\,dx. \tag{8}$$

We note that while $\mathbb{P}(C=k)$ depends on D and E, we omit these dependencies for brevity of notation. To compute the probabilities $\mathbb{P}(C=k)$, we integrate numerically using the midpoint rule on the grid of family sizes x for which $f_F(x;\,E)$ was pre-computed. For the mean and variance of C we find the explicit expressions

$$\mathbb{E}(C)=D,\quad \mathbb{V}(C)=D+D^2\frac{1-E}{1+E}. \tag{9}$$

Since we impose an error-correction threshold T and drop UMIs with fewer than T reads, the read-count distribution we actually observe is a *censored* version of C where the possible outcomes $C<T$ are removed. For the mean and variance of this censored distribution with threshold T we write

$$\mathbb{E}(C\,|\,C\geq T)=\frac{\sum_{k=T}^\infty k\cdot\mathbb{P}(C=k)}{\mathbb{P}(C\geq T)}, \tag{10}$$

$$\mathbb{V}(C\,|\,C\geq T)=\frac{\sum_{k=T}^\infty (k-\mathbb{E}(C\,|\,C\geq T))^2\cdot\mathbb{P}(C=k)}{\mathbb{P}(C\geq T)}. \tag{11}$$

To compute $\mathbb{E}(C\,|\,C\geq T)$, we rewrite the infinite sum in Equation (10) to $\mathbb{E}(C)-\sum_{k<T}k\cdot\mathbb{P}(C=k)$, and similarly for $\mathbb{V}(C\,|\,C\geq T)$.

2.2.4 Computing the loss

The expected loss ℓ is the expected fraction of true UMIs that either remain completely unsequenced, or that are removed by the error-correction threshold. Since we treat each per-UMI read count, and hence each UMI's fate (to be filtered or not) as independent stochastic quantities, this expected fraction is simply the probability that a single UMI has a read-count below the threshold T, i.e.

$$\ell=\mathbb{P}(C<T). \tag{12}$$

2.3 Correcting for lost molecules

Given n^{obs} experimentally observed UMIs (after applying the error-correction threshold T to filter out phantoms) and their read count vector $\mathbf{c}=(c_1,\ldots,c_{n^{\text{obs}}})$, we estimate the reaction efficiency E and the mean number of reads per UMI D. We use the *method of moments*, i.e. we find E and D such that the predicted mean equals the sample mean $\widehat{m}$ of $\mathbf{c}$, and the predicted variance its sample variance $\widehat{v}$. Since we only take observed UMIs with at least T reads into account, we must compute the predictions using the censored distribution, i.e. find E, D such that $\widehat{m}=\mathbb{E}(C\,|\,C\geq T)$ and $\widehat{v}=\mathbb{V}(C\,|\,C\geq T)$

If $T=0$, i.e. if $\widehat{m}$ and $\widehat{v}$ reflect the *uncensored* mean respectively variance, these equations can be solved explicitly by inverting Equation (9), which yields the method of moments estimates $\widehat{D}=\widehat{m}$ and $\widehat{E}=\frac{1-\nu'}{1+\nu'}$, where $\nu'=\frac{\widehat{v}-\widehat{m}}{\widehat{m}^2}$ limited to the interval $[0,1]$.

If $T>0$, we solve the system of equations numerically to find E and D (see Supplementary Section S1.2). With these parameter estimates, we then compute an estimate $\widehat{\ell}$ of the loss ℓ using Equation (12), and use it to correct for the expected number of lost molecules. Assuming that we observed n^{obs} UMIs and given $\widehat{\ell}$, we estimate the total number of molecules in the original sample to have been

$$\widehat{n}^{\text{tot}}=\frac{n^{\text{obs}}}{1-\widehat{\ell}}. \tag{13}$$

2.3.1 Gene-specific estimates and corrections

Since the reaction efficiency E and depth D, and hence also the loss, will usually vary between individual genes (or other genomic features of interest), to correct the observed number of transcripts of some gene $g\in 1,\ldots,K$ for the loss, a *gene-specific* loss estimate $\widehat{\ell}_g$ should be used. In principle, such estimates are found by applying the described estimation procedure to only the UMIs found for transcripts of gene g, i.e. by computing a gene-specific mean $\widehat{m}_g$ and variance $\widehat{v}_g$ of the number of reads per UMI, and solving for parameters E and D to find a gene-specific $\widehat{E}_g^{\text{raw}}$ and $\widehat{D}_g^{\text{raw}}$, and computing $\widehat{\ell}_g^{\text{raw}}$ using Equation (12). If the number n_g^{obs} of observed UMIs (i.e. transcripts) stemming from gene g is large, a correction based on $\widehat{\ell}_g^{\text{raw}}$ yields an (approximately) unbiased and accurate estimate of the total number of transcripts of that gene. But if n_g^{obs} is small, the error of the estimator $\widehat{\ell}_g^{\text{raw}}$ easily exceeds the variability of the true gene-specific value ℓ_g between genes. In such cases, correcting using the *library-wide* estimate $\widehat{\ell}^{\text{all}}$ computed from *all* UMIs found in the library will yield a more accurate (although biased) estimate of the total number transcripts of gene g.

Interestingly, by combining these two flawed estimators of the true gene-specific loss ℓ_g, we obtain a *shrinkage estimator* $\widehat{\ell}_g^{\text{shr}}$ that improves upon both in terms of *mean squared error* (MSE), see Carter and Rolph (1974) Equation (2.4),

$$\widehat{\ell}_g^{\text{shr}}=\lambda_g\cdot\widehat{\ell}_g^{\text{raw}}+(1-\lambda_g)\cdot\widehat{\ell}^{\text{all}}. \tag{14}$$

The gene-specific coefficient λ_g determines how much the raw gene-specific estimate is *shrunk* towards the global estimate, and its optimal choice (with respect to the MSE) depends on the variances the two constituent estimators. To determine the optimal λ_g we make the following assumptions about these estimators:

i. the library-wide estimate $\widehat{\ell}^{\text{all}}$ is a good proxy for the true *average* loss taken over all genes $1,\ldots,K$. This seems reasonable given the size of a typical library, comprising millions of UMIs.
ii. the estimator variance of the raw gene-specific estimator $\widehat{\ell}_g^{\text{raw}}$ depends only on the number n_g^{obs} of observed UMIs for gene g, and does so in an inversely proportional manner. This is certainly true asymptotically for large numbers of observations, for small numbers Supplementary Figure S4 shows this approximation to be reasonable.

We write s for the variance of the true loss between genes (i.e. for the mean squared difference of ℓ_g and $\widehat{\ell}^{\text{all}}$), and u for the proportionality constant between the estimator variance of $\widehat{\ell}_g^{\text{raw}}$ and $1/n_g^{\text{obs}}$. According to Carter and Rolph (1974) Equation (2.4ff) the optimal choice for λ_g is then

$$\lambda_g=\frac{s}{s+u/n_g^{\text{obs}}} \tag{15}$$

To compute the gene-specific shrinkage estimators $\widehat{\ell}_g^{\text{shr}}$, it remains to find constants u and s. Towards that end, we observe that the expected squared deviation of the raw gene-specific loss estimate $\widehat{\ell}_g^{\text{raw}}$ from its average $\bar{\ell}=\frac{1}{n}\sum_{g=1}^K\widehat{\ell}_g^{\text{raw}}$ is the total variance of $\widehat{\ell}_g^{\text{raw}}$, which is comprised of the between-gene variance s and the estimator variance u/n_g^{obs}, or in other words $\mathbb{E}\,(\widehat{\ell}_g^{\text{raw}}-\bar{\ell})^2=s+u/n_g^{\text{obs}}$.

This allows us to estimate s and u using *least squares regression*, i.e. by minimizing

$$\sum_{g=1}^K((\widehat{\ell}_g^{\text{raw}}-\bar{\ell})^2-s-u/n_g^{\text{obs}})^2\cdot w(n_g^{\text{obs}}). \tag{16}$$

Without weighting (i.e. for $w(n)=1$), the considerable drop in magnitude of $(\widehat{\ell}_g^{\text{raw}}-\bar{\ell})^2$ as n_g^{obs} increases would allow genes with small

number of observations to yield an unduly large influence over the estimates. Since it is the genes with a low to moderate number of observations that benefit from shrinking, some modest bias of this sort is actually desired—but not as strong a bias as $w(n) = 1$ exhibits, and one not so purely focused on genes with very few observations. We therefore use the weights $w(n) = \frac{n}{1+n/100}$, which initially increase linearly with the number of observations, but eventually converge to 100 instead of increasing further. This has the desired effect of shifting the focus away from rarely observed genes, and concentrating it on genes with a moderate number of observations.

3 Results

3.1 PCR stochasticity versus efficiency

During PCR amplification, each uniquely labeled molecule is amplified into a molecular family of indistinguishable copies. Random successes or failures to copy molecules during early reaction cycles lead to a variation in the final family sizes (Fig. 1B), even between identical (expect for their molecular barcode) molecules. As the family size of each initial molecule grows, the proportion of successful copy operations approaches the efficiency E, therefore reducing the amount of noise added by each additional cycle. The total number of cycles thus has little influence on the final family size distribution, and is therefore not a parameter of our model. For the same reason, a plateau effect (i.e. diminishing reaction efficiency during later cycles) has little effect on the final family size distribution, and is thus not included in the model. The final distribution does, however, depend strongly on the reaction efficiency, with fluctuations in family size decreasing as the efficiency grows towards 100% (Fig. 1C).

For efficiencies close to 100%, most molecular families are thus of about average size, except for those ($\sim 100 - E$ percent) families for which the first copy failed. These are about half the average size, and form a distinct secondary peak in the family size distribution (Fig. 1C, brown curve). We emphasize that due to this, even at efficiencies close to 100%, the distribution still shows considerable dispersion, meaning that even at high efficiencies stochastic PCR effects are not negligible. At lower efficiencies, the family sizes vary even more wildly, as extreme family sizes (on both ends of the scale) become more likely (Fig. 1C, blue and green curves).

If we add sequencing to the picture, i.e. combine the stochastic PCR model outlined above with a model of sequencing as random Poissonian sampling (Marioni *et al.*, 2008), the variability of per-UMI read counts (Fig. 1D) then has two sources—the variability of molecular family sizes and the Poissonian sampling introduced by sequencing. Although the latter is reduced by increasing the sequencing depth, the former is independent of the sequencing depth but is reduced by increasing the reaction efficiency. For all reasonable error-correction thresholds T the predicted fraction of true UMIs filtered out by the error-correction step thus grows with diminishing efficiency E.

3.2 Model validation and phantom UMI removal

To validate our model of amplification and sequencing, we compared the predicted distribution of per-UMI read counts to the distribution observed in two published RNA-Seq datasets. Kivioja *et al.* (2012) labeled and sequenced transcripts in *Drosophila melanogaster* S2 cells using 10 bp random molecular barcodes from the 5′ end. Shiroguchi *et al.* (2012) labeled and sequenced transcript fragments in *E.coli* cells on both ends, using (on each end) one of 145 molecular barcodes carefully selected to have large pairwise edit distances. The Y-shaped sequencing adapters used in the *E.coli* experiment were designed such that each strand of a labeled double-stranded cDNA molecule produces a related but distinguishable molecular family.

To see whether our algorithm offers an advantage over existing UMI error-correction strategies, we pre-filtered the observed UMIs in each of the two replicates of these datasets using the following existing algorithms: We first merged UMIs likely to be erroneously sequenced versions of the same molecule, using the algorithm proposed by Smith *et al.* (2017). For the *E.coli* experiment we also removed UMIs for which the complementary UMI corresponding to the second strand of the same initial template molecule was not detected, as proposed by Shiroguchi *et al.* (2012). See Supplementary Section S1.4 for details on the analysis pipeline we used.

To this pre-filtered set of UMIs we then applied our algorithm. For each dataset, we manually chose an error-correction threshold by visually comparing read-count distribution and model prediction for different thresholds, and picking the lowest threshold that yielded a reasonably good fit. Above the error-correction threshold (Fig. 2a, black bars), the observed library-wide distribution of reads per UMI closely follows the model prediction, and the *E.coli* data even shows traces of the secondary peak that represents molecules not duplicated in the first reaction PCR cycle. Choosing a different threshold will change the number of UMIs surviving the error-correction filter, but has little influence on the estimated reaction efficiency and on the estimated total number of UMIs after loss correction (Supplementary Fig. S1). We thus conclude that our model captures the main stochastic behavior of the amplification and sequencing processes, and accurately models the read-count distribution of true UMIs.

The UMIs removed by our filter, i.e. those with fewer reads than the error-correction threshold demands, (Fig. 2A, gray bars) are over-abundant compared to our prediction. This over-abundance increases further as per-UMI read counts drop, indicating the existence of a group of UMIs with significantly reduced molecular family sizes. While we may expect some systematic variation of family sizes between true UMIs (on top of the stochastic variations that our PCR model predicts), we would expect these to be gradual and not form distinct groups. We conclude that the extra UMIs causing the observed over-abundance are indeed phantoms that are rightly removed by our algorithm. We note that none of these phantoms were removed by either the UMI merging algorithm of Smith *et al.* (2017), or (for the *E.coli* data) by filtering UMIs for which the complementary UMI (representing the second strand of the template molecule) was not detected.

For the *D.melanogaster* data, our loss estimates of 9% (R1) and 8.8% (R2) are about a magnitude higher than the 1% (R1) and 2% (R2) estimated using the (truncated) Poisson distribution suggested by Kivioja *et al.* (2012). Given that using a Poisson model amounts to assuming a 100% efficient duplication of molecules during each PCR cycle, this severe underestimation by the Poisson model shows that the inherent stochasticity of the PCR cannot be neglected.

3.3 Gene-specific quantification bias

The gene-specific (shrunken) estimates for amplification efficiency, average reads per UMI, and loss that our algorithm produces, vary between genes to different degrees (Fig. 2B). We observe the smallest amount of variation for the average number of reads per UMI (Fig. 2B, left)—the estimates of this parameter are virtually identical for a large majority of genes, and differs only for a few outliers.

The estimated amplification efficiencies on the other hand can vary substantially between genes (Fig. 2B, middle). For the two

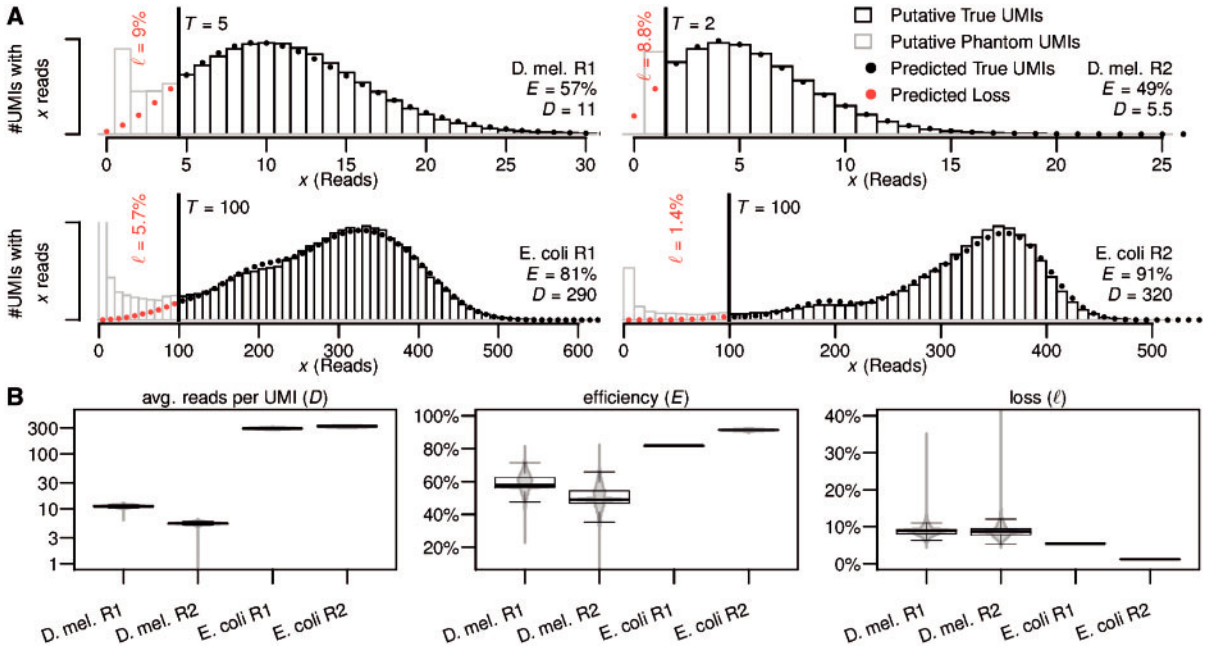

Fig. 2. (**A**) Observed and predicted library-wide distribution of reads per UMI and parameter and loss estimates. Filtered UMIs (grey bars, left of threshold *T*) are over-abundant and thus assumed to contain both phantom and true UMIs (red dots). UMIs surviving the filter (black bars) closely follow the predicted distribution (black dots) and are assumed to be true UMIs. **(B)** Variability of the (shrunken) model parameters and resulting loss between genes. Includes parameter for 7481 detected genes in *D.melanogaster* R1, 8001 genes in R2, 2380 genes in *E.coli* R1 and 2308 genes in R2. (Color version of this figure is available at *Bioinformatics* online.)

D.melanogaster replicates the range is 22–81% (R1) and 1–83% (R2). Considering that in this experiment only the 3' ends of transcripts were sequenced, and all fragments contributing to a gene hence share a similar sequence composition, this is not unexpected. These differences in efficiency cause the loss to vary heavily between genes as well (Fig. 2B, right), between 4 and 35% for R1 and between 4 and 89% for R2 (which has a much lower overall sequencing depth). Without gene-specific loss corrections, abundance comparisons between genes will thus suffer from systematic quantification bias against certain genes of up to $\approx 35–4\% = 31\%$ for R1 and up to $\approx 85\%$ for R2. The larger amount of systematic bias in *D.melanogaster* R2 is caused by the two-fold reduction of the number of reads per molecules in R2 compared with R1—due to the lower number of reads per molecule in R2, the same difference in amplification efficiencies between two genes translates into a larger difference of the number of lost molecules in R2 compared with R1.

In contrast, fragments from all parts of the transcript were sequenced in the *E.coli* experiments, and together with the high sequencing depth (≈300 reads per UMI), we now expect little variations of efficiency, and small and highly uniform losses across genes. Our efficiency and loss estimates reflects this (Fig. 2B, middle and right), and as the lack of outliers shows, they do so even for genes with only few UMIs. Yet for these genes, the raw (unshrunken) gene-specific estimates are noisy (Supplementary Fig. S3), proving that shrinking the raw estimates successfully reduces the noise to acceptable levels.

3.4 Bias-corrected transcript counts

To further verify the accuracy of the corrected transcript counts computed by our algorithm, we conducted a simulation study. We use the (loss-corrected) estimated total transcript abundances of *D.melanogaster* replicate 1, rounded to 10, 30, 100, 300, 1000, 3000 or 10 000 molecules as the true transcript abundances. We then simulated amplification and sequencing of these transcripts, using for each gene the previously estimated gene-specific efficiency and average number of reads per UMI (Fig. 2B). To the resulting list of UMIs and their read-counts for each gene we applied our algorithm to recover the true transcript abundances (threshold $T = 5$ as before), and determined for each gene the relative error of the recovered abundances compared with the simulation input.

Figure 3 shows these relative errors (i) if no correction is done (ii) if the correction is based soley on the raw gene-specific loss estimates (i.e. no shrinkage) and (iii) for the full TRUmiCount algorithm (i.e. using shrunken loss estimates). The uncorrected counts systematically under-estimate the true transcript counts, in 50% of the cases by at least ≈10%, independent of the true number of transcripts per gene. And even at high transcript abundances, the relative error still varies *between* genes, biasing not only absolute transcript quantification, but also relative comparisons between different genes. The counts corrected using raw gene-specific estimates are unbiased and virtually error-free for strongly expressed genes, but exhibit a large amount of additional noise for weakly expressed genes. The full TRUmiCount algorithm successfully controls the amount of added noise, and shows no additional noise for weakly expressed genes, while still being unbiased and virtually error-free for more strongly expressed genes.

3.5 Peformance for low sequencing depth

To assess the performance of the TRUmiCount algorithm at low sequencing depths such as are common for single-cell RNA-Seq experiments, we ran a second simulation with gene-specific depth parameters scaled such that the average across all genes was $D = 1$ read per molecule (Fig. 4). Under these conditions, the most likely outcome for a single molecule in the initial sample is to remain unsequenced (39% of molecules), and only 27% of molecules are found in more than one read. The library-wide efficiency estimate of 57%

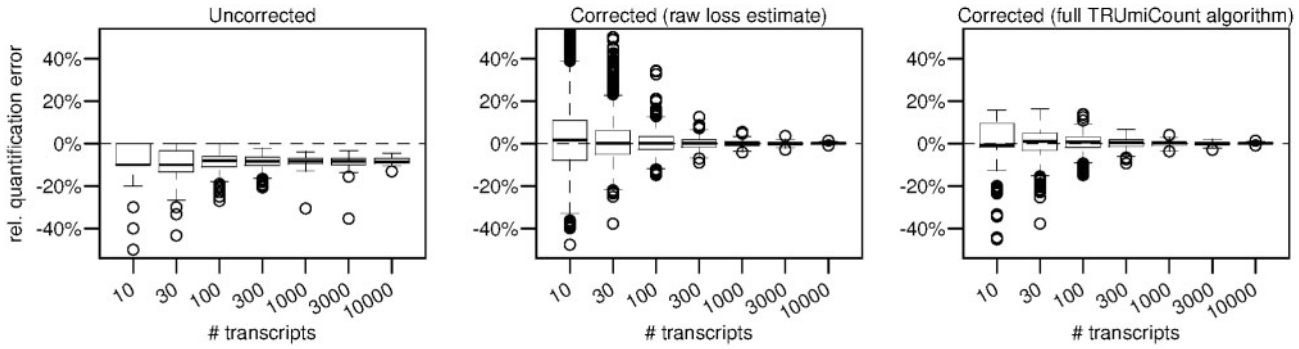

Fig. 3. Relative error of estimated total number of transcripts depending on the true number of transcripts. Left panel uses the observed number of UMIs without any correction. Middle panel uses the raw gene-specific loss estimates to correct for lost UMIs. Right panel uses the full TRUMmiCount algorithm employing shrunken gene-specific loss estimates to correct for lost UMIs

(Fig. 4A) is nevertheless accurate, and identical to the one computed for the full dataset (*D.melanogaster* R1) that the simulation was based on (Fig. 2A).

For the relative error of the corrected transcript counts we observed a roughly 2-fold increase at low-sequencing depth (Fig. 4B) compared with the situation at original sequencing depth (Fig. 3, right), but still no systematic over- or under-estimation. We estimate that Poissonian sampling effects account for about a $\sqrt{1-0.09}/\sqrt{1-0.39} \approx 1.2$-fold increase of the relative errors. The rest is probably due to the parameter estimation problem becoming harder at lower sequencing depths, particularly for weakly expressed genes. For more strongly expressed genes, the relative quantification error again drops towards zero, similar to the behavior at original sequencing depth.

4 Discussion

The TRUmiCount algorithm we presented successfully removes the biases inherent in raw UMI counts, and produces unbiased and low-noise measurements of transcript abundance, allowing for unbiased comparisons between different genes, exons and other genomic features. It does so even in the presence of various types of phantom UMIs and varying amplification efficiencies, both between samples and along the genome. Compared to other error-correction techniques, it is not restricted to particular types of phantom UMIs, or to a special Y-shaped design of the sequencing adapters.

Our model of the amplification and sequencing process is mechanistic, and its two parameters have an immediate physical interpretation. They can both be determined from the experimental data without the need for either guesses or separate calibration experiments. The TRUmiCount algorithm thus does not require any changes to library preparation over the basic UMI method. By inspecting the estimated parameters—in particular the amplification efficiency, the amplification reaction itself can be studied. For example, by estimating model parameters separately for sequenced fragments of different lengths, the drop of reaction efficiency with increasing fragment lengths can be quantified (Supplementary Fig. S2).

Although TRUmiCount requires that libraries are sequenced sufficiently deeply to detect at least some UMIs more than once, it can also deal with cases where a molecule is on average detected only by a single read, which is common e.g. for single-cell RNA-Seq. The performance of TRUmiCount is reduced a bit in such situations, but it still offers an improvement over uncorrected counts by removing

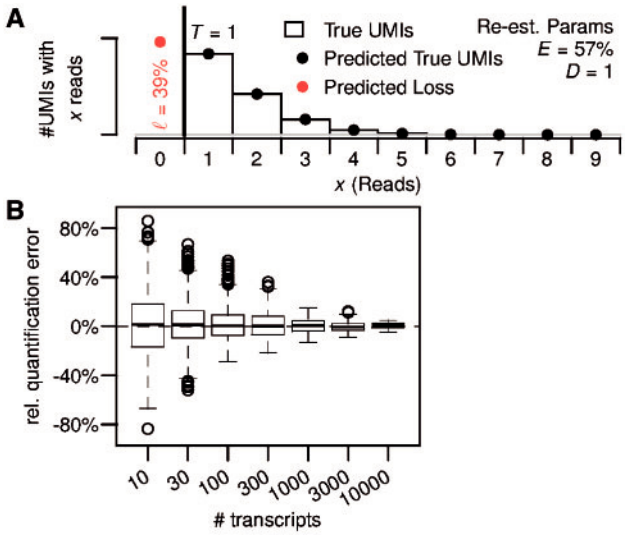

Fig. 4. TRUmiCount performance for low sequencing depth. **(A)** Overall distribution of observed and predicted reads per UMI for an average of $D = 1$ read per molecule. **(B)** Relative error of estimated total number of transcripts for different true numbers of transcripts and $D = 1$ read per molecules on average. (Color version of this figure is available at *Bioinformatics* online.)

systematic biases. For even lower read counts, where gene-specific bias correction becomes infeasible, we expect that TRUmiCount could still be used to correct for cell-specific (instead of gene-specific) biases, thus reducing the amount of technical noise when comparing absolute transcript counts of the same gene between individual cells.

The TRUmiCount algorithm can thus help to increase the accuracy of many quantitative applications of NGS, and by removing biases from comparisons between genes can aid in the quantitative unraveling of complex gene interaction networks. To make our method as easily accessible as possible to a wide range of researchers, we provide two readily usable implementations of our algorithm. Our R package *gwpcR* enables a flexible integration into existing R-based data analysis workflows. In addition, we offer the command-line tool TRUmiCount which is designed to work in conjunction with the *UMI-Tools* of Smith *et al.* (2017). Together they provide a complete analysis pipeline which produces unbiased transcript counts from the raw reads produced by a UMI-based RNA-Seq experiment (http://www.cibiv.at/software/trumicount).

Acknowledgements

We would like to thank Armin Djamei and Simon Uhse for introducing us to the concept of UMIs, all members of the *Center for Integrative Bioinformatics Vienna* (CIBIV), in particular Luis Paulin-Paz, Celine Prakash and Philipp Rescheneder for their valuable feedback throughout all stages of this project, and Olga Chernomor for sharing her insights into shrinkage estimators. We also thank the "DoktoratsKolleg RNA Biology" for funding this work.

Funding

This work was supported by the Austrian Science Fund (FWF): W1207-B09.

Conflict of Interest: none declared.

References

Aird,D. *et al.* (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, R18.

Akima,H. (1996) Algorithm 760; rectangular-grid-data surface fitting that has the accuracy of a bicubic polynomial. *ACM Trans. Math Softw.*, **22**, 357–361.

Carter,G.M. and Rolph,J.E. (1974) Empirical bayes methods applied to estimating fire alarm probabilities. *J. Am. Stat. Assoc.*, **69**, 880.

Harris,T.E. (1989). *The Theory of Branching Processes*. Dover, New York.

Hug,H. and Schuler,R. (2003) Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. *J. Theor. Biol.*, **221**, 615–624.

James,W. and Stein,C. (1961) Estimation with quadratic loss. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, vol. **1**, pp. 361–379. University of California Press, Berkeley, CA.

Kivioja,T. *et al.* (2012) Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, **9**, 72–74.

Krawczak,M. *et al.* (1989) Polymerase chain reaction: replication errors and reliability of gene diagnosis. *Nucleic Acids Res.*, **17**, 2197–2201.

Marioni,J.C. *et al.* (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.

Schmitt,M.W. *et al.* (2012) Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 14508–14513.

Shiroguchi,K. *et al.* (2012) Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 1347–1352.

Smith,T. *et al.* (2017) UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.*, **27**, 491–499.

Weiss,G. and von Haeseler,A. (1995) Modeling the polymerase chain reaction. *J. Comput. Biol.*, **2**, 49–61.

Weiss,G. and von Haeseler,A. (1997) A coalescent approach to the polymerase chain reaction. *Nucleic Acids Res.*, **25**, 3082–3087.

TRUmiCount: Correctly counting absolute numbers of molecules using unique molecular identifiers

Supplementary Information

Florian G. Pflug and Arndt von Haeseler

S1 Supplementary Methods

S1.1 Computing the distribution of F

To find the actual distribution (in terms its density $f_F(\cdot\,;\,E)$ with the reaction efficiency E as a parameter) of the normalized family size F for a particular efficiency E we resorted to simulation. We simulated the PCR process for efficiencies from 0.01 to 0.99 (steps of 0.01 up to 0.90, steps of 0.005 up to 0.94, steps of 0.002 up to 0.99). Each time, we simulated 10^9 independent trajectories, and ran each simulation until the expected family size was 10^7 molecules (i.e. for $n = 7/\log_{10}(1+E)$ cycles). At that point the stochasticity further cycles would introduce is negligible and we may thus assume $\tilde{M}_n \approx \tilde{M}_{n+1} \approx F$.

For each efficiency E, we normalized the simulated raw family sizes using Equation (4) to obtain 10^9 independent samples of F. Using kernel density estimation, we then estimated values of the density function $f_F(\lambda\,;\,E)$ of the normalized family size distribution on a grid of 318 values of λ between 0 and 50. The grid points are spaced non-uniformly, being finest (distance 0.0025) around 0 and 1 and getting coarser elsewhere.

This procedure resulted in a 123×318 matrix of densities, i.e. $f_F(\lambda\,;\,E)$ evaluated for each combination of one of the 123 simulated efficiencies E, and one of the 318 normalized family sizes λ. Using this (pre-computed and stored) matrix, the density function $f_F(\lambda\,;\,E)$ can be evaluated quickly for arbitrary values of E and λ by two-dimensional polynomial interpolation (Akima, 1996).

S1.2 Numerical method of moments estimates

To obtain method of moments estimates for model parameters D (reads per molecules) and E (reaction efficiency) in the general case $T \geq 0$ from the observed mean $\hat{m}$ and observed variance $\hat{v}$ of the number of reads per UMI, we must find D and E such that

$$\begin{aligned}\hat{m} &= \mathbb{E}(C \,|\, C \geq T),\\ \hat{v} &= \mathbb{V}(C \,|\, C \geq T)\end{aligned}$$

We solve this system of equations with an iterative method that starts with initialization step I and then repeats update step U until the estimates $\hat{D}$, $\hat{E}$ and $\mathbb{P}(C \geq T)$ converge (absolute or relative change less than 10^{-4}).

I: We start by pretending that $T = 0$, and set

$$\begin{aligned}\hat{D} &:= \hat{m},\\ \hat{E} &:= \frac{1-r}{1+r} \quad \text{where} \quad r = \frac{\hat{v}-\hat{m}}{\hat{m}^2} \quad \text{limited to } [0,1],\end{aligned}$$

U: Using the current model parameter estimates $\hat{D}$ and $\hat{E}$, we compute

$$\begin{aligned}\mathbb{P}(C = k) \quad &\text{for} \quad k = 0,\dots,T-1,\\ \mathbb{P}(C \geq T) &= 1 - \sum_{k=0}^{T-1} \mathbb{P}(C = k).\end{aligned}$$

We then exploit that the uncensored mean (and similarly the variance) can be partitioned into a sum of the (scaled) censored mean and the mean (or variance) terms "missing" from the censored mean, i.e. we compute updated estimates $\hat{D}'$ of the uncensored mean and $\hat{v}'_u$ of the uncensored variance,

$$\begin{aligned}\hat{D}' &:= \mathbb{P}(C \geq T)\cdot\hat{m} + \sum_{k<T} k\cdot\mathbb{P}(C=k),\\ \hat{v}'_u &:= \mathbb{P}(C \geq T)\cdot(\hat{v}+\hat{m}^2) - \hat{D}'^2 + \sum_{k<T} k^2\cdot\mathbb{P}(C=k).\end{aligned}$$

Given the updated estimates of the uncensored moments, the updated reaction efficiency estimate $\hat{E}'$ is computed as in the case $T = 0$ as

$$\hat{E}' := \frac{1-r}{1+r} \quad \text{where} \quad r = \frac{\hat{v}'_u - \hat{D}'}{\hat{D}'^2} \quad \text{limited to } [0,1].$$

S1.3 Multiple initial copies

If each distinct molecules the sample initially contains $R > 1$ identical copies (e.g. $R = 2$ if the initial molecules are double-stranded), each of these copies can be imagined to be amplified by a separate and independent PCR processes. But since the molecules are indistinguishable, these processes cannot be observed individually – we can observe only the (re-normalized) sum of the resulting family sizes. These observed normalized family size distribution is thus the average of R independent versions of F, and its variance is thus one R-th of the variance in Equation (6), i.e.

$$\mathbb{V}F = \frac{1-E}{1+E}\cdot\frac{1}{R}, \quad \mathbb{V}C = D + D^2\cdot\frac{1-E}{1+E}\cdot\frac{1}{R}. \tag{S1}$$

The density of distribution of F for $R > 1$ is the R-fold self-convolution of the density of F with itself (re-scaled to again have expected value one), and can thus be computed from the pre-computed matrix for the single-molecule case without performing additional simulations.

Parameter estimation proceeds just as for $R = 1$, except that when computing estimate v' of $\mathbb{V}F$, we must now account for the reduction of the observed variance of F by a factor of $\frac{1}{R}$, i.e. we set $v' = R\cdot\frac{\hat{v}-\hat{m}}{\hat{m}^2}$.

S1.4 Data Analysis

The reads from each of the downloaded sequenced libraries, were mapped (ignoring the barcode part) with *NGM* v0.5.2 (Sedlazeck *et al.*, 2013) to the

reference transcriptome of *D. melanogaster* (R6.08) respectively *E. coli* (strain K-12 MG1655). To avoid ambiguities during mapping for genes with multiple isoforms, we filtered the *D. melanogaster* transcriptome to contain only a single transcript per gene before mapping. For each gene, we picked either the single transcript with a FlyBase score of at least "moderately supported", or the longest transcript (if multiple ones had score "moderately supported" or higher). After mapping the reads, we used the combination of mapping coordinates (both start and end for the paired-end *E. coli* data, only start for the single-end *D. melanogaster* data) and barcode (on both ends in the case of *E. coli*) as UMI. To account for sequencing errors, we merged similar UMIs (barcodes differing at most in one position, mapping coordinates by at most 30 bases for paired-end, 5 for sing-end libraries) using the graph-based algorithm of Smith *et al.* (2017). For the *E. coli* data we additionally combined reciprocal UMIs stemming from the two strands of a single template molecule, but stored the read counts for plus- and minus-strand separately (see Shiroguchi *et al.* (2012)).

This yielded, for each of the libraries, a table comprising the gene id, start- end end position, barcode and read-count(s) of each detected UMI. Based on this table, the error-correction thresholds ($T = 5$ for *E. coli*, $T = 5$ for *D. melanogaster* R1, $T = 2$ for *D. melanogaster* R2), and the initial number of molecules (actually, strands) for each UMI ($R = 1$ for *E. coli* due to the Y-shaped adapters, $R = 2$ for *D. melanogaster* due to secondary strand synthesis before amplification) our algorithm computed library-wide and raw as well as shrunken gene-specific estimates of the reaction efficiency, of the average number of reads per UMI, and of the loss. For the *E. coli* data, the error-correction threshold was applied to the plus- and minus-strand read counts separately, filtering out UMIs if either count lay below the chosen threshold. This increased the loss of true UMIs, and we modified the definition of the loss accordingly to $\ell = 1 - (1 - \mathbb{P}(C < T))^2$ (compare to Equation (12)). (Note that in the histograms in Fig. 2A, for a lack of other options, we show plus- and minus-strand counts separately, but omit UMIs where one of the strands is not detected at all). In addition to the gene-specific parameter and loss estimates, our algorithm output the observed number of UMIs n_g^{obs} and the estimated total number of UMIs (i.e. transcript molecules) n_g^{tot}.

S1.5 Simulation

We determined the residual error of the corrected transcript counts using a simulation approach. We started from the (loss-corrected) estimated transcript counts n_g^{tot} and (shrunken) gene-specific estimates for reaction efficiency $\hat{E}_g$ and sequencing depth $\hat{D}_g$ of gene $g \in \{1, \dots, K\}$ that we computed for replicate 1 of the *D. melanogaster* dataset. First we rounded n^{tot} to the next number in the series $10, 30, 100, 300, \dots$ and used the resulting number as the *true* number n_g^{true} of transcripts of gene g. For each gene g, we then used the amplification+sequencing model (with parameters E_g, D_g and $R = 2$ meaning double-stranded molecules) to simulate the sequencing of n_g^{true} UMIs, which yielded for each gene n_g^{true} read counts, one for each UMI. To this list comprising gene id and (for each gene) n_g^{true} read counts, we applied our algorithm, using $T = 5$ and $R = 2$ as before (but passing along no other information from the first run of the algorithm). The algorithm thus dropped all UMIs with fewer than $T = 5$ reads, treated the remaining UMIs for each gene g as the *observed* number of UMIs n_g^{obs}, re-estimated the (shrunken) gene-specific losses, and used them to correct n_g^{obs} for these losses to arrive at an estimated total transcript count n_g^{tot}. Finally, we computed for each gene the *relative quantification error* as

$$\frac{\left|n_g^{\text{tot}} - n_g^{\text{true}}\right|}{n_g^{\text{true}}}. \tag{S2}$$

References

Akima, H. (1996). Algorithm 760; rectangular-grid-data surface fitting that has the accuracy of a bicubic polynomial. *ACM Transactions on Mathematical Software*, **22**(3), 357–361.

Sedlazeck, F. J., Rescheneder, P., and von Haeseler, A. (2013). NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics*, **29**(21), 2790–2791.

Shiroguchi, K., Jia, T. Z., Sims, P. A., and Xie, X. S. (2012). Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(4), 1347–1352.

Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome research*, **27**(3), 491–499.

S2 Supplementary Figures

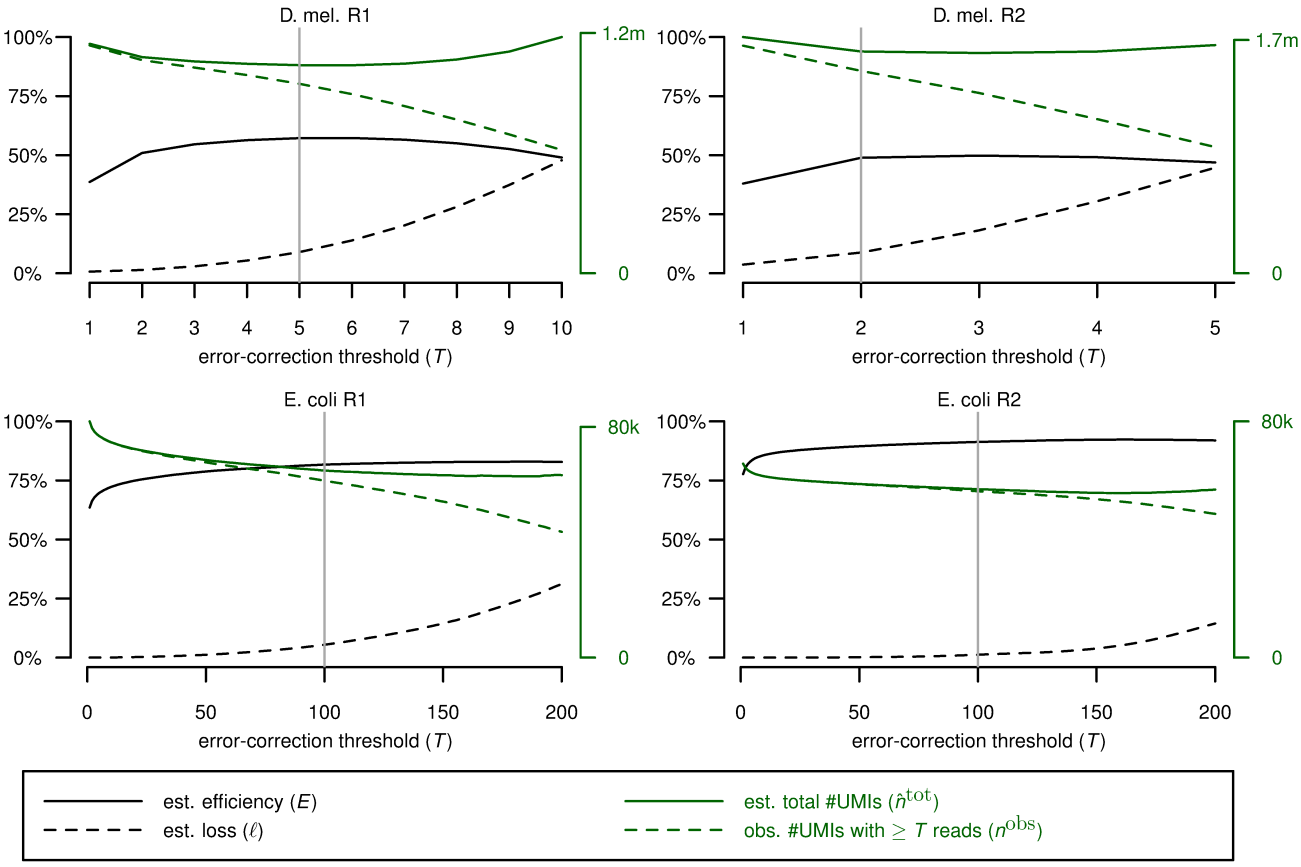

Fig. S1. Sensitivity of estimates to choice of threshold T. Shows the estimated efficiency (E, left y-axis), loss (ℓ, left y-axis), number of putative true UMIs (n^{obs}, right y-axis) and estimated total number of molecules ($\hat{n}^{\mathrm{tot}}$, right y-axis) for different choices for the error-correction threshold T.

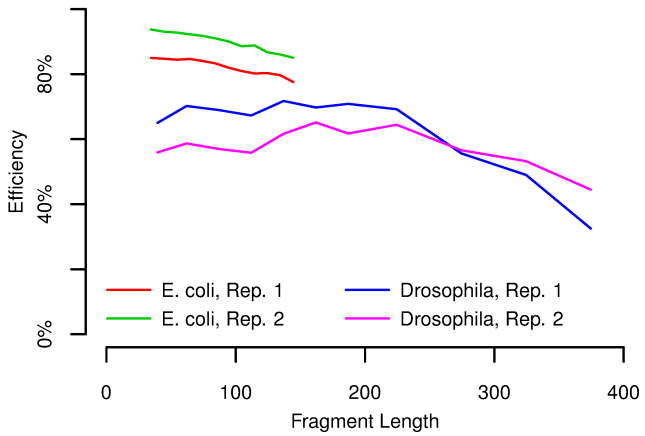

Fig. S2. Length dependence of PCR efficiency. For each experiment, the detected UMIs were binned according to fragment length, and the PCR efficiency estimated independently for each bin.

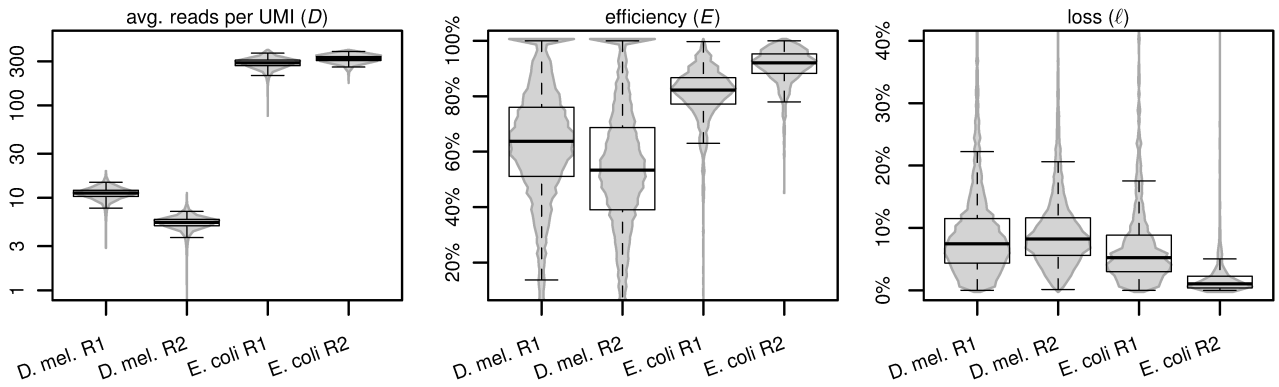

Fig. S3. Variability of the raw (unshrunken) model parameters and resulting loss between genes. Includes parameter for 7481 detected genes in D. mel. R1, 8001 genes in R2, 2380 genes in E. coli R1 and 2308 genes in R2.

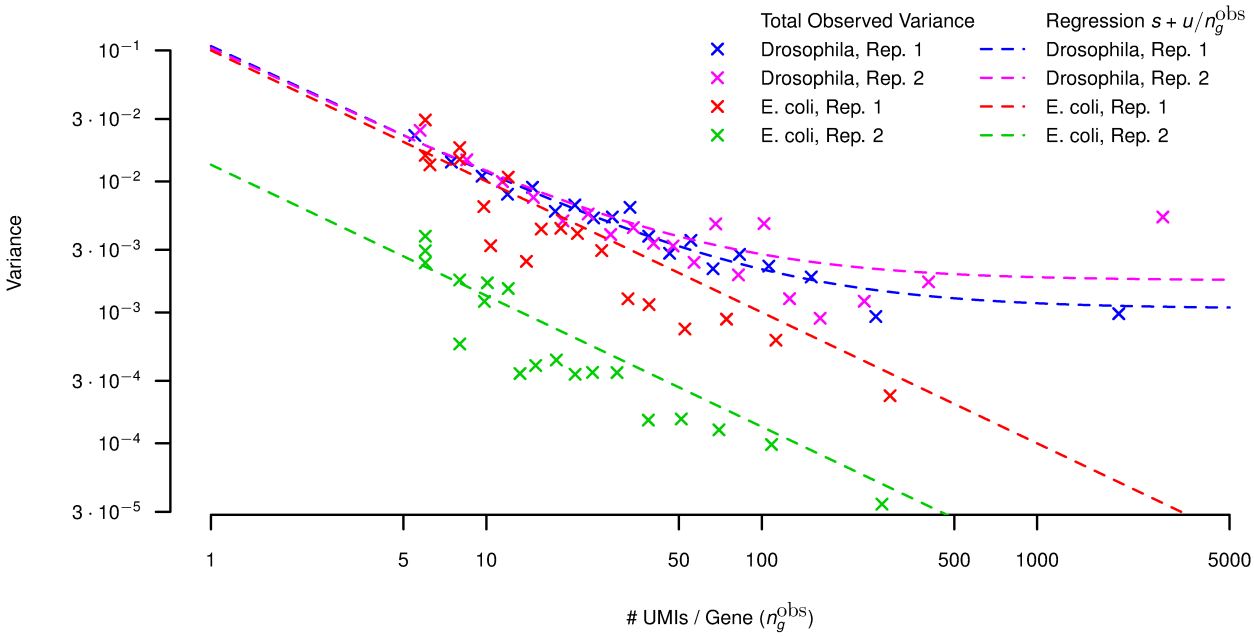

Fig. S4. Total variance of the raw gene-specific loss estimates. Total observed variance was computed for bins containing 20 genes with a similar number n_g^{obs} of observed true UMIs. The regression curve $s + u/n^{\mathrm{obs}}$ used to infer the optimal gene-specific shrinkage factors λ_g comprises two components, the variance s of the loss between genes, and the n_g^{obs}-dependent error of the (raw) gene-specific loss estimates u/n_g^{obs}. See also Gene-specific estimates & corrections.

Chapter 3

In vivo insertion pool sequencing identifies virulence factors in a complex fungal-host interaction (Uhse *et al.*, 2018. *PLoS Biology, 16*(4): e2005129)

Preamble

Here, we present a method to accurately measure the differences in virulence of different mutant strains of the maize pathogen *Ustilago maydis* by accurately measuring pre- and post infection abundances of the mutants using *next-generation sequencing* (NGS) and *unique molecular identifiers* (UMIs). The hundred-fold size difference between pathogen host genomes (20Mb vs. ≈ 2Gb), the small genetic differences between mutants (200b of 20Mb) and limited mechanical separability of host and pathogen make such measurements challenging.

To nevertheless successfully quantify mutant virulences using this approach, *theoretical* insight into the properties of iPool-Seq is necessary. The foundation for such insight is formed by the TRUmiCount algorithm and its underlying theory of UMI-based NGS experiments (Pflug & von Haeseler, 2018; chapter 2 of this

work). This theory is applicable to iPool-Seq as well (see chapter 5), allowing us to construct a statistical model of iPool-Seq experiments on top of it. This model provides a quantitative definition of *virulence* and a framework for detecting differences between mutants and wild-type; the model is described briefly in this chapter and in more detail in chapter 6).

Author contributions

Mathematical modelling, software development and data curation were done by the author of this thesis, under the guidance and supervision of Arndt von Haeseler. Data validation and visualisation were done in conjunction with Simon Uhse from Armin Djamei's lab.

Publication history and status

Submitted on December 14th 2017 to *PLoS Biology*, accepted for publication on March 20th 2018, published online on April 23rd 2018.

Uhse S., **Pflug F. G.**, Stirnberg A., Ehrlinger K., von Haeseler A., & Djamei A. (2018). In vivo insertion pool sequencing identifies virulence factors in a complex fungal-host interaction. *PLoS Biology, 16*(4), e2005129. DOI:10.1371/journal.pbio.2005129

PLOS BIOLOGY

METHODS AND RESOURCES

In vivo insertion pool sequencing identifies virulence factors in a complex fungal–host interaction

Simon Uhse[1], Florian G. Pflug[2], Alexandra Stirnberg[1], Klaus Ehrlinger[1], Arndt von Haeseler[2,3], Armin Djamei[1]*

1 Gregor Mendel Institute (GMI), Austrian Academy of Sciences, Vienna BioCenter (VBC), Vienna, Austria, **2** Center for Integrative Bioinformatics Vienna (CIBIV), Max F Perutz Laboratories (MFPL), University of Vienna, Medical University Vienna, Vienna, Austria, **3** Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria

* armin.djamei@gmi.oeaw.ac.at

OPEN ACCESS

Citation: Uhse S, Pflug FG, Stirnberg A, Ehrlinger K, von Haeseler A, Djamei A (2018) In vivo insertion pool sequencing identifies virulence factors in a complex fungal–host interaction. PLoS Biol 16(4): e2005129. https://doi.org/10.1371/journal.pbio.2005129

Academic Editor: Joseph Heitman, Duke University Medical Center, United States of America

Received: December 14, 2017

Accepted: March 20, 2018

Published: April 23, 2018

Data Availability Statement: Raw sequencing data has been deposited to ENA under accession number PRJEB23309 and processed data can be found in S1 Data and S2 Data.

Funding: European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013) (grant number GA335691 „Effectomics"). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Austrian

Abstract

Large-scale insertional mutagenesis screens can be powerful genome-wide tools if they are streamlined with efficient downstream analysis, which is a serious bottleneck in complex biological systems. A major impediment to the success of next-generation sequencing (NGS)-based screens for virulence factors is that the genetic material of pathogens is often underrepresented within the eukaryotic host, making detection extremely challenging. We therefore established insertion Pool-Sequencing (iPool-Seq) on maize infected with the biotrophic fungus *U. maydis*. iPool-Seq features tagmentation, unique molecular barcodes, and affinity purification of pathogen insertion mutant DNA from in vivo-infected tissues. In a proof of concept using iPool-Seq, we identified 28 virulence factors, including 23 that were previously uncharacterized, from an initial pool of 195 candidate effector mutants. Because of its sensitivity and quantitative nature, iPool-Seq can be applied to any insertional mutagenesis library and is especially suitable for genetically complex setups like pooled infections of eukaryotic hosts.

Author summary

Insertion mutant screens are widely used to identify genotype–phenotype relationships. In negative selection screens, a major limitation is the efficient identification of mutants that are lost or retained after selection. To identify these mutants, the two genomic sequences flanking the insertion cassette must be found. However, pinpointing these insertion flanks within a genome is like looking for a needle in a haystack; a problem that becomes even worse when several organisms form a biotrophic interaction. To overcome this hurdle, we developed insertion Pool-Sequencing (iPool-Seq). With iPool-Seq, we were able to efficiently amplify and enrich insertion flanks from complex genomic DNA samples. This technique allows for the quantification of relative insertion mutant abundance before and after selection by next-generation sequencing (NGS). We demonstrate the power of iPool-Seq with a negative selection screen by infecting maize with 195 candidate effector mutants of

Science Fund (FWF) (grant number P27429-B22, P27818-B22, I 3033-B22). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: ARS, autonomous replication sequence; Cm-medium, control Complete medium; dpi, days post infection; EGB, Early Golden Bantam; FDR, false discovery rate; gDNA, genomic DNA; hpt, hygromycin phosphotransferase; HITS, high-throughput insertion tracking by deep sequencing; iPool-Seq, insertion Pool-Sequencing; LB, left border; ME, mosaic end; NGS, next-generation sequencing; PE, paired-end; RB, right border; ROI, region of interest; SBS, sequencing primer binding site; UMI, unique molecular identifier; UPS, unique primer binding site; wt, wild-type.

the fungal pathogen *Ustilago maydis*. We identified 28 virulence factors, of which 23 have not been previously described. iPool-Seq is extremely sensitive, cost- and time-efficient, and promises to be a powerful tool for identifying target genes in large selection screens.

Introduction

Virulence factors are key for successful infections by pathogens. Their identification is of major interest because of the necessity to develop effective counter strategies. For instance, fungal virulence factors are typically identified by mutating single loci in fungi, followed by individual fungal mutant infections of host tissue and subsequent assessment of pathogen fitness [1–4]. Individual infection assays are not ideal for the genetic screening of a large number of pathogen mutants because they are laborious, cost-intensive, and—most importantly—assessment of infections is often subjective and qualitative rather than quantitative. An attractive alternative is infection with a pool of pathogen mutants allowing direct assessment of individual pathogen fitness in the same host tissue. However, using a pooled pathogen infection creates the challenge of identifying pathogens with reduced virulence within a complex mixture of genetic material extracted from infected host tissue.

Mutant collections can be efficiently generated using insertional mutagenesis. Insertional mutagenesis employs gene cassettes that commonly comprise a selectable marker under the control of a strong constitutive promoter. The detection of genome–cassette junctions can serve as a molecular identifier for each insertion mutant. During screening, insertional mutants before selection in the host are defined as the genetic input, whereas surviving insertional mutants after selection comprise the genetic output. Insertional mutagenesis can be achieved randomly through transposon insertion [5–8] or *Agrobacterium tumefaciens*-mediated transformation [3, 9], or specifically through site-specific insertion by homologous recombination [10, 11].

Over the last decade, several approaches were established that use massive parallel sequencing for the detection of inserted gene cassettes. These approaches were successfully used to track mutants from the small genomes of prokaryotic pathogens and allowed the identification of bacterial genes involved in virulence or host colonization after pooled infections [12–16]. However, only a few attempts were reported that identified virulence factors using pools of eukaryotic pathogens [17]. The main factors limiting the successful insertional mutagenesis of eukaryotic pathogens by pooled infections in complex host-pathogen systems are variable infection rates of individually mutated pathogens, the size ratio of host/pathogen genomes, the inability to sufficiently detect inserted gene cassettes from pathogenic material, and biases that arise through PCR-based amplification steps.

To enable successful and quantitative insertion mutant screen-based identification of virulence factors in complex biological systems, we developed insertion Pool-Sequencing (iPool-Seq). We determined the sensitivity and efficiency of iPool-Seq using an insertion mutant collection of 195 predicted virulence factors encoded by the maize pathogen *U. maydis*. The haploid *U. maydis* genome consists of approximately 20.5 megabases [18, 19], whereas the diploid genome of maize is 2.3 gigabases large [20]. This represents a 100-fold genome size difference, which is beside the proportion between fungal and host plant genome abundance as a limiting factor, making the robust detection of *U. maydis* sequence information in infected maize tissue necessary. The iPool-Seq workflow consists of Tn5 Transposase-mediated tagmentation of complex genomic DNA (gDNA) allowing efficient library preparation from low-input material [21, 22]. This is followed by the efficient enrichment of extremely rare insertion cassettes from fungal genomes using biotin-streptavidin affinity purification of PCR products. Amplification

biases are monitored through incorporated unique molecular identifiers (UMIs). Insertional mutant fitness within host tissues is directly measured through quantification of UMI counts present in infected output material compared to UMI counts from the input library.

iPool-Seq on *U. maydis* infections of maize confirmed the identity of 5 known fungal virulence factors that were included as positive controls in the screen. Importantly, 23 previously unreported virulence factors encoded by *U. maydis* were uncovered. Three of these factors were confirmed to be novel virulence factors of *U. maydis* after testing by individual infection. The combination of pooled insertion mutant infections and iPool-Seq technology represents a straightforward and cost-effective approach to map insertion mutants in complex host–pathogen systems with the potential to generate genome-wide virulence maps of relevant crop pathogens and beyond.

Results

iPool-sequencing design and library generation

We employed the smut fungus *U. maydis* as a model to establish iPool-Seq. We generated a Golden Gate cloning-compatible plasmid, which allows for recombination of multiple fragments in a single reaction [23]. To this end, we combined a hygromycin resistance cassette that is flanked by unique primer binding sites (UPSs) with the chromosomal up- and downstream regions (1,000 bp) of 195 predicted *U. maydis* effector genes (Fig 1A; S1 Table). Plasmids were linearized and transformed into *U. maydis* SG200 protoplasts for deletion of the putative virulence factors by homologous recombination (Fig 1B). For each of the insertion mutant constructs, we isolated 3 independent transformants and analyzed deletion events using PCR primers directed against the effector genes sequences. Absence of PCR products

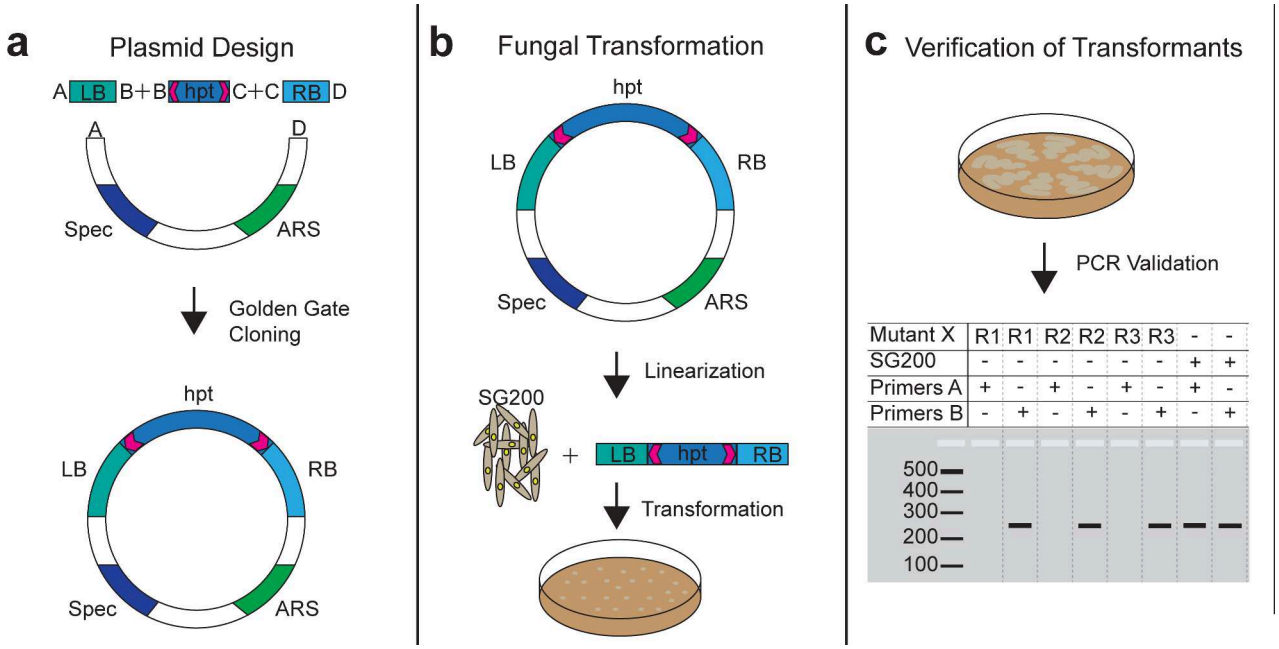

Fig 1. Design of deletion constructs and *U. maydis* insertional mutants. (**a**) Plasmid backbones containing a Spec and an ARS were combined with an hpt resistance cassette and specific borders (LB & RB) via Golden Gate Cloning [23]. The hpt resistance cassette is flanked by UPSs (magenta arrows). (**b**) Plasmids were linearized with *Asc*I and combined with haploid SG200 protoplasts. Transformants were selected on plates supplemented with hygromycin. (**c**) Schematic overview of PCR verification of transformants. Three independent fungal transformants were verified for each mutant locus via PCR. PCR products from primer-pair A targeting insertional mutant X was absent in positive transformants and detectable in SG200 control strains. A control primer-pair B gave a product in both insertional mutant X and SG200. ARS, autonomous replication sequence; hpt, hygromycin phosphotransferase; LB, left border; RB, right border; Spec, Spectinomycin resistance cassette; UPS, unique primer binding site.

https://doi.org/10.1371/journal.pbio.2005129.g001

indicated successful deletions (Fig 1C). For each successful deletion, 3 independent transformant replicates were verified and stored separately, allowing for individual propagation to avoid growth competition prior to pooled infections. We performed 2 independent infections with pools containing the entire collection of 195 insertional mutants and established the iPool-Seq library preparation protocol (S2 Fig).

For later comparison of mutant material abundance within the collection, iPool-Seq libraries were prepared from gDNA representing the mutant pool before infection (the input) and from infected tissues containing both maize and *U. maydis* genomes (the output, Fig 2A). To minimize the number of library preparation steps and conserve material, we replaced mechanical shearing of gDNA (requiring DNA-end repair, tailing, and adapter ligation steps) with Tn5-mediated tagmentation (Fig 2B) [21]. Although this approach yields a wider size range of DNA fragments, simultaneous DNA fragmentation and adapter ligation makes Tn5-mediated tagmentation preferable to DNA shearing approaches. We produced recombinant Tn5 transposase and adapted the published protocol to large gDNA inputs (S3 Fig) [21]. Furthermore, customized adapters for Tn5-mediated tagmentation were designed containing 12 bp unique molecular identifiers (UMIs) followed by a sequencing primer binding site (SBS; Fig 2B; S2 Table), which enables sequencing of UMIs using a custom-made first strand sequencing primer. Fragmented gDNA from pooled fungal infections of maize are not only highly diverse but fungal DNA content will certainly be underrepresented, making it necessary to efficiently enrich for insertion cassette junctions with genomic regions. To enrich for such junctions, the tagmentation-derived DNA fragments were amplified using specific adapter primers and biotinylated primers that bind to

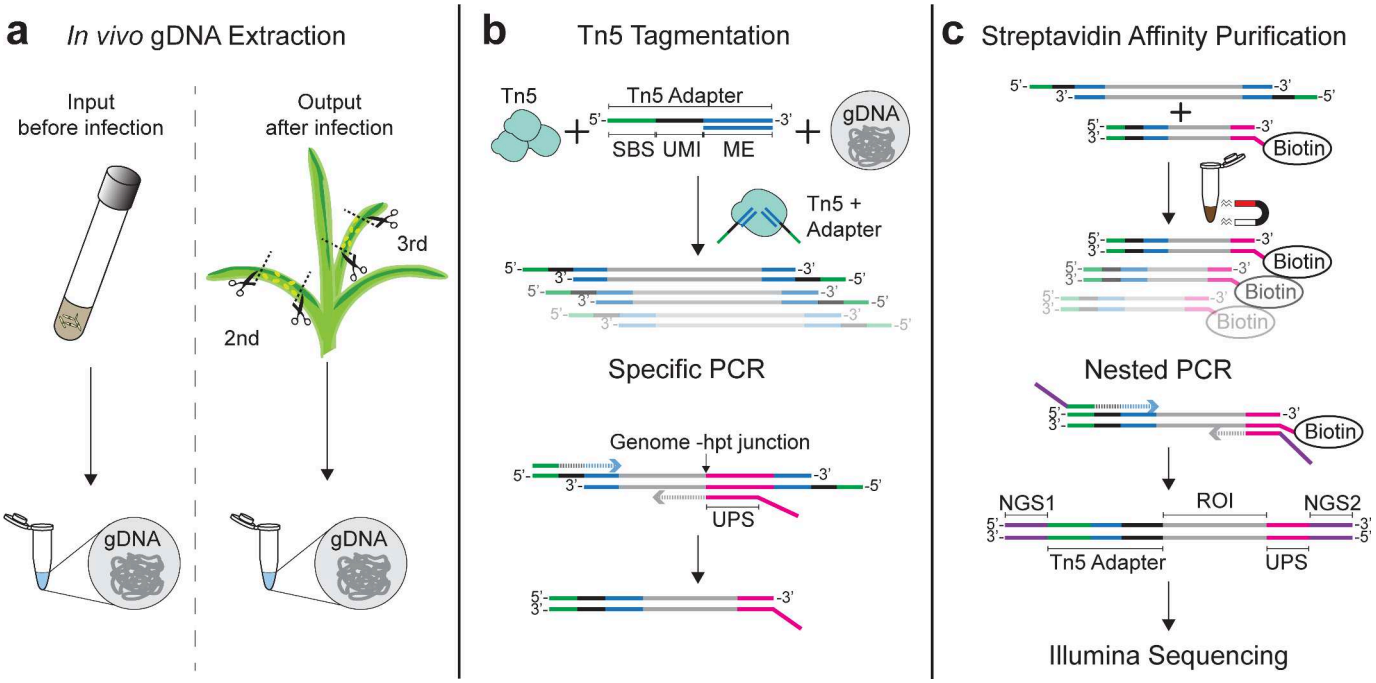

Fig 2. iPool-Seq library preparation workflow features tagmentation and UMIs. (**a**) Library preparation was carried out for the input mutant collection and for the output after infection. For the output, we harvested infected areas of the second and third maize leaves and isolated gDNA. (**b**) Extracted gDNA was fragmented with Tn5 Transposase loaded with custom adapters containing an SBS (green), 12-bp UMI, and Tn5 hyperactive MEs (blue). Genome–hpt resistance cassette junctions were PCR-amplified with biotinylated primers directed against UPSs (magenta) and adapter-specific primers directed at the SBS. (**c**) Biotinylated PCR products were streptavidin-affinity–purified and Illumina-compatible P5 (purple; NGS1) and P7 (purple; NGS2) ends were introduced by nested PCR. Final products were subjected to Illumina PE sequencing on a MiSeq platform. gDNA, genomic DNA; hpt, hygromycin phosphotransferase; iPool-Seq, insertion Pool-Sequencing ME, mosaic end; PE, paired-end; ROI, region of interest; SBS, sequencing primer binding site; UMI, unique molecular identifier; UPS, unique primer binding site.

https://doi.org/10.1371/journal.pbio.2005129.g002

unique sequences at the distal ends of deletion cassettes (Fig 2B; S2 Table). Consequently, both genomic junctions of individual insertion cassettes were amplified, yielding biotinylated PCR products from all insertional mutants. Biotinylated PCR products were isolated using streptavidin-based affinity purification (Fig 2C) and Illumina-compatible adapters were introduced via nested PCR (S2 Table). Sequencing was performed on an Illumina MiSeq platform. In conclusion, we designed iPool-Seq to benefit from tagmentation, specific amplification, and streptavidin purification for efficient enrichment of ultra-rare genome deletion cassette junctions out of a highly diverse gDNA mixture.

iPool-Seq facilitates the identification of fungal virulence factors

We infected maize in two independent experiments with three biological replicates of a pool of 195 verified insertional *U. maydis* mutants (S1 Table), resulting in six input and output libraries. The libraries were prepared as described above and sequenced on an Illumina MiSeq platform with paired-end (PE) sequencing. After read validation and read mapping, 87.7% ± 1.7% and 85.3% ± 1.6% of the obtained sequencing reads (input versus output, respectively) were mapped to *U. maydis* insertional mutation loci (Fig 3A; S1 Supporting methods).

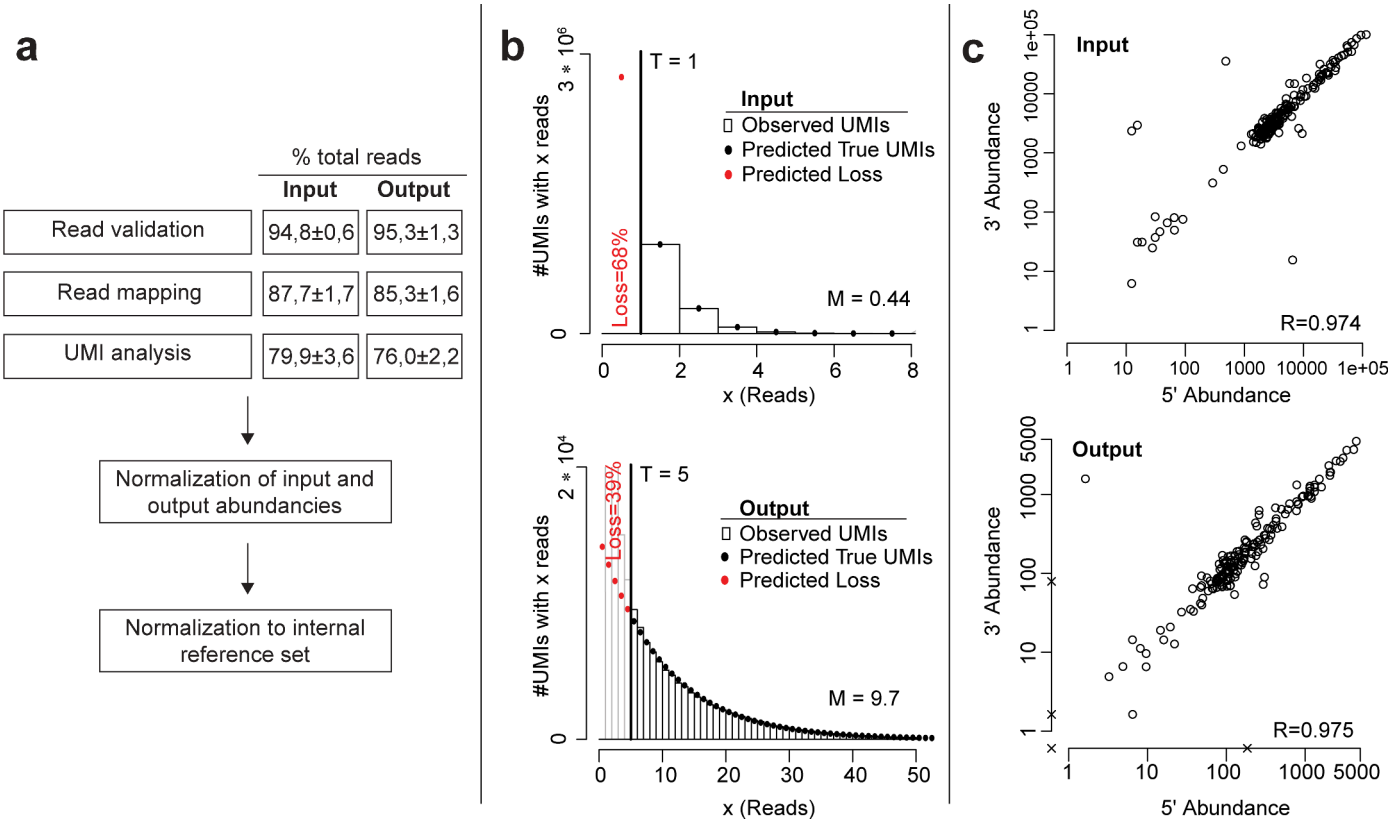

Fig 3. Quality control of iPool-Seq library. (**a**) Bioinformatic workflow of iPool-Seq analysis. Input and output read percentage after validation, mapping, and UMI analysis shows the mean ± SEM of 3 biological replicates and 2 independent infections. (**b**) Distribution of reads per individual UMI (bars) and model prediction (dots) over all insertional mutants of 1 representative replicate for input and output. Here, the error-correction threshold was set to 1 for the input and 5 for the output. Predicted true and lost UMIs are indicated. (**c**) Correlation plot of UMI counts for 5′- and 3′- genomic junctions of the hpt resistance cassette. One representative replicate of input and output is depicted. Each circle represents an insertional mutant. Missing up or downstream reads are marked with x. hpt, hygromycin phosphotransferase; iPool-Seq, insertion Pool-Sequencing; M, mean number of reads per UMI in the predicted distribution; R, correlation value; T, threshold; UMI, unique molecular identifier.

https://doi.org/10.1371/journal.pbio.2005129.g003

To remove reads produced by PCR bias and which would affect quantitative evaluation of input and output reads, we collapsed all reads with highly similar UMIs to a single UMI count after sequencing. Based on the observed distribution of reads per UMI and comparison to a model prediction, we then set a library-specific read count threshold, removed UMIs with fewer reads than the threshold as likely PCR and sequencing artifacts, and corrected the number of remaining UMIs for the estimated loss of real UMIs (Fig 3B, S1 Supporting methods). After this UMI analysis, we retained 79.9% ± 3.6% and 76.0% ± 2.2% of initial reads from input and output for downstream analyses, respectively (Fig 3A).

The sequencing results indicated that three-fourths of all iPool-Seq reads were informational for insertion mutant abundance. Moreover, iPool-Seq generated similar amounts of valid reads from input- and output-derived gDNAs, indicating that yield performance was not diminished using gDNA derived from two organisms.

Since each inserted mutagenesis cassette has two junctions with neighbouring genomic regions, an unbiased library preparation should produce similar read numbers for up- and downstream junctions. We observed high correlation values (R) for all insertion mutants for the input and output samples, indicating that iPool-Seq is not suffering from considerable PCR biases during exponential amplification of DNA fragments containing mutagenesis cassette–genome junctions (Fig 3C).

To identify *U. maydis* virulence factors, we analyzed input and output reads for significantly depleted sequences from the pool of 195 insertion mutants. First, the read output of all insertional mutants was normalized to the corresponding input reads. Second, we defined an internal reference set of *U. maydis* mutant strains that do not have virulence phenotypes [18, 24] and whose output and input reads showed a neutral and linear relationship (Fig 4A, neutral; Fig 4B; S3 Table). Our collection contains additional mutants that were previously reported to be neutral. In these communications, neutral mutants formed symptoms with the same severity as the progenitor strain SG200. However, these observations did not provide any distinct information about quantitative growth defects of these mutants. Therefore, we constrained the neutral reference set to five mutants that displayed a reproducible neutral behavior in the iPool-Seq data (S1 Supporting methods).

We then calculated, for each mutant, the level of depletion from the output sample compared to the input and determined significance through normalization to the internal reference set. This resulted in the identification of a substantial proportion of sequences that were significantly depleted from the output libraries (Fig 4B, red circles; S1 Data). We analyzed this depleted sequence set for known virulence factors and identified Pep1, Pit2, and Stp1 (UMAG_01987, UMAG_01375, and UMAG_02475) [25–27] as known essential virulence factors of *U. maydis* (Fig 4A, lost virulence). In addition, we found the previously described virulence factors ApB73 (UMAG_02011) [28] and Fer1 (UMAG_00105) [29] among the less depleted and reduced candidate sequences (Fig 4A, reduced). Two other mutants (UMAG_06223 and UMAG_02239), for which minor defects in disease symptom induction had been reported previously, were not significantly depleted in the iPool-seq results and one mutant (UMAG_12313) previously reported to be unaffected in virulence showed a weak but significant reduction in our iPool-seq approach (S4 Table) [24]. In summary, iPool-Seq results largely overlap with previously reported symptom scoring data for characterized virulence factors (S4 Table). It is also sensitive, as not only apathogenic but also reduced virulence factor mutants were identified. Importantly, analysis of the depleted sequence set yielded 23 fungal mutants that are potential novel virulence factors of *U. maydis* (Fig 4C; S4 Table).

In contrast to the identification of depleted mutant sequences, we did not identify sequences that were reproducibly enriched in all biological replicates, indicating that none of the fungal mutants tested conferred enhanced virulence to *U. maydis* on the tested host accession Early Golden Bantam (EGB; Fig 4C).

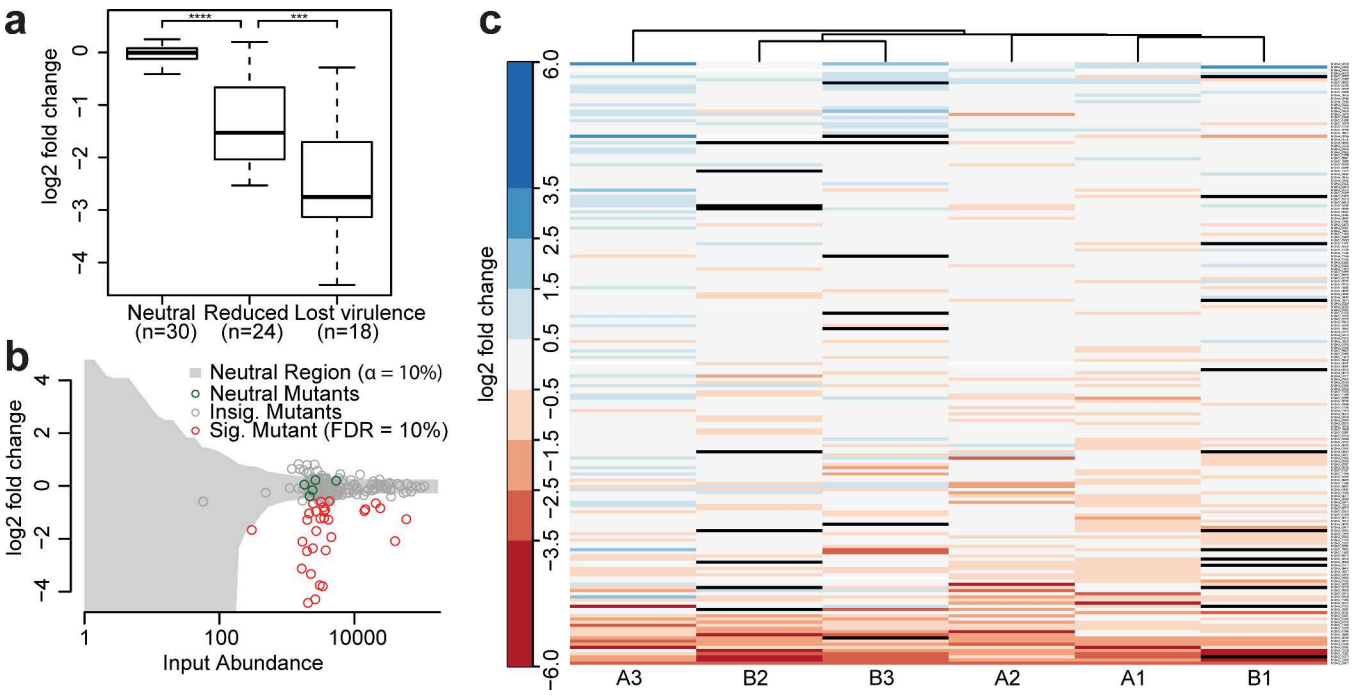

Fig 4. iPool-Seq identifies significantly depleted mutants after pooled infection. (**a**) Log_2-fold changes between normalized output abundances and internal reference set for mutants with known phenotypes. *p*-Values were calculated with Mann–Whitney U tests. $p = 5e^{-9}$ for neutral versus reduced and $p = 3e^{-4}$ for reduced versus lost virulence with $^{***}p < 0.001$; $^{****}p < 0.0001$ (S3 Table). (**b**) Log_2-fold change of output over input abundances for 1 representative replicate. Each circle represents 1 insertion mutant. Internal references are marked in green, significantly depleted in red (tested against reference set using negative binomial test; S1 Data; S1 Supporting methods), unaffected mutants in gray; Insig. area is also highlighted in gray. (**c**) Heatmap of log_2-fold changes of input normalized UMI counts of all insertional mutants sorted by mean level of abundance. Infection A and B are two independent experiments and 1, 2, and 3 are three biological replicates, which were clustered according to similarity. Mutants without detectable reads in output libraries are displayed in black (S1 Data; S1 Supporting methods). FDR, false discovery rate; Insig., insignificant; iPool-Seq, insertion Pool-Sequencing; Sig., significant; UMI, unique molecular identifier.

https://doi.org/10.1371/journal.pbio.2005129.g004

We next modeled the performance of iPool-Seq on a high-throughput mutant library of *U. maydis* (S9 Fig, S1 Supporting methods). To this end, we used the following parameters: 1) 20,000 insertion mutants were chosen cover the approximately 20-MB genome of *U. maydis* with approximately 1,000 bp average distance of insertion sites. 2) During maize colonization, approximately 1,500 of the approximately 6,900 *U. maydis* genes are transcriptionally up-regulated—and we showed that about 14% of all mutants from our library contributed to virulence (Fig 4C; S4 Table) [18, 30]. Based on these observations, we extrapolate that approximately 3% of all *U. maydis* genes are likely to be involved in virulence. 3) We showed with iPool-Seq that known reduced virulence factors of *U. maydis* had a mean logarithmic fold change of −1.53 and known essential virulence factors of −2.75 in comparison to the neutral reference set, respectively (Fig 4A). Due to a lack of data, the model does not take into account the number of unsuccessful infection events on the host plant but assumes 100% infection rate for each individual of a neutral mutant strain.

The model resulted in 40 (for essential virulence factors) and, respectively, 100 (for weak virulence factors) detected individuals necessary for each mutant in the input samples to identify virulence factors with 99% sensitivity. Based on observed average of approximately 10 reads per UMI (Fig 3B) and due to the insertion flank sequencing efficiency of at least 75% (Fig 3A), the required sequencing depth would be 26 Mio reads (20,000·100·10·1,33 = 26,600,000) per library. This suggests that the iPool-Seq technology can be used for large scale mutant screens in *U. maydis* and similar systems.

Validation of novel essential *U. maydis* virulence factors

To validate the 23 potential virulence factors identified by iPool-Seq, we chose three top candidates and tested their effects on virulence using individual infection assays. We observed a severe loss of *U. maydis* virulence upon infection of plants with fungi carrying these mutations. Whereas the wild-type progenitor strain SG200 produced galls on infected maize, all three mutant strains failed to form galls, indicating that they are essential for fungal virulence (Fig 5A). This effect was specifically due to virulence, as growth assays under stress-inducing conditions showed no difference between these mutant strains and SG2000 (Fig 5B). Using confocal microscopy on infected plants, we observed that mutant strains were severely impaired in colonizing maize leaf tissues (Fig 5C). Our combined results show that iPool-Seq facilitates the identification of essential factors for *U. maydis* virulence. Furthermore, the streamlined library preparation of iPool-Seq should make the method widely applicable for identifying unknown virulence factors in complex biological systems, such as in vivo infected tissues.

Discussion

Pooled mutant screens have proven to be very powerful tools to uncover individual genes affecting particular phenotypes in a time- and cost-effective fashion. Positive selection screens usually lead to limited numbers of individual surviving cells that are easily identifiable by a

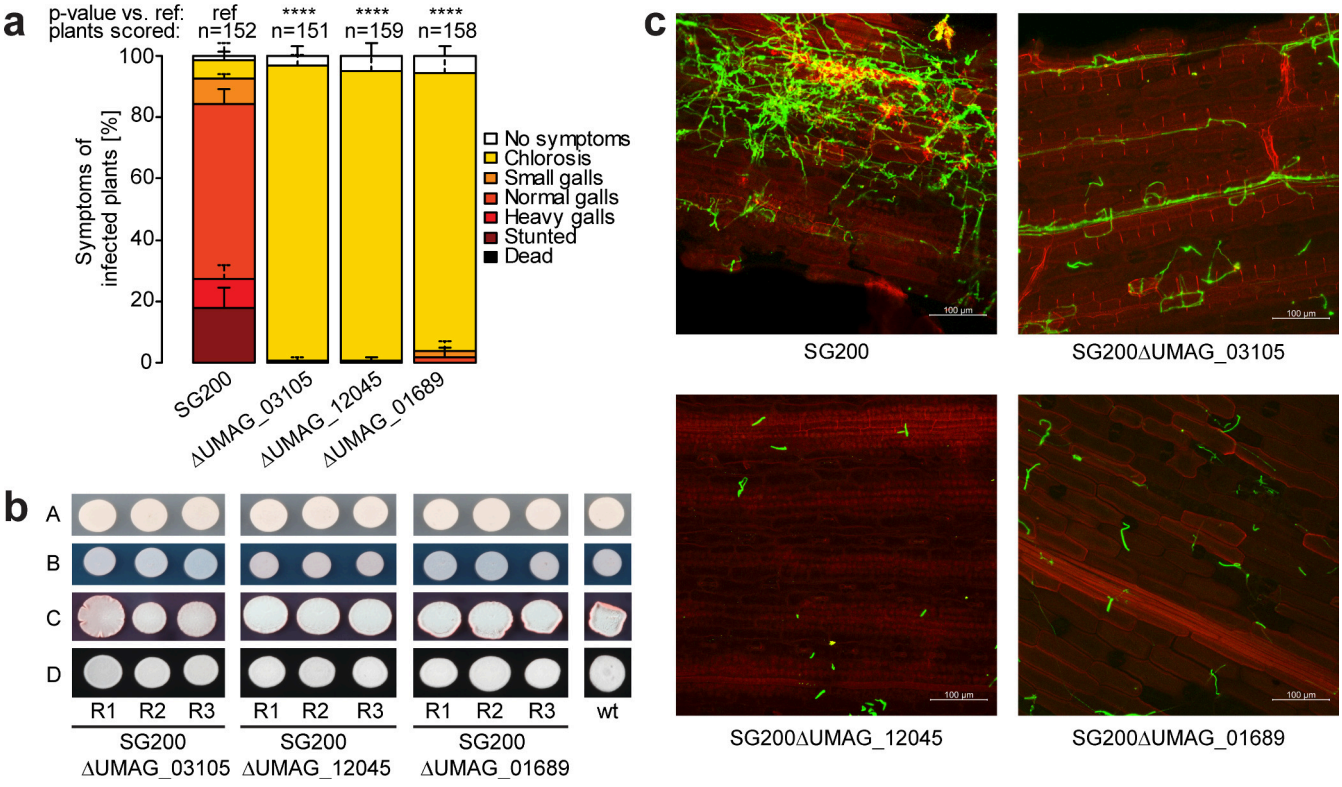

Fig 5. Virulence factor mutants identified by iPool-Seq cause reduced disease symptoms on maize. (**a**) Disease rating of insertional mutant strains 7 dpi. Mean standard deviation of relative counts from 3 replicates are displayed. Only positive error bars are shown. *p*-Values were calculated by Fishers exact test. Multiple testing correction was done by Benjamini-Hochberg algorithm. **** $p < 0.0001$. (S2 Data) (**b**) Growth assay of insertion mutants on (A) Cm-medium, or Cm-medium supplemented with (B) 75 μg/mL Calcofluor (cell wall stress), (C) 45 μg/mL Congo red (cell wall stress), and (D) Charcoal (b-filament inducing). (**c**) Confocal microscopy of maize infected with indicated insertional mutant strains 7 dpi. Infected plant tissue was stained with propidium iodide (red) and fungal hyphae with lectin binding WGA-AF488 (green). One representative picture of 9 infected plants is shown. Cm-medium, control Complete medium; dpi, days post infection; iPool-Seq, insertion Pool-Sequencing; ref, reference; wt, wild-type.

https://doi.org/10.1371/journal.pbio.2005129.g005

combination of restriction enzyme digests, inverse PCR, and sequencing. Negative selection screens rely on the survival of most analyzed cells, making it necessary to devise methodology that allows comparing the presence/absence of genetic information before and after selection. To tackle the later challenge, several insertional mutagenesis approaches have been developed [31]. Although successful in bacterial systems for the elucidation of virulence factors [5, 13, 32], such insertion mutant approaches were not widely used in eukaryotic systems, mainly because of unresolved technical issues such as low sensitivity and system-intrinsic limitations (for example, genome ploidy, lifestyle of the investigated model system).

Here, we introduce iPool-Seq as a versatile and highly sensitive method for the analysis of insertion mutant pools before and after selection, enabling both negative and positive mutant selection screens in complex eukaryotic systems including the analysis of host–pathogen interactions. We used iPool-Seq to examine virulence factors from a defined set of mutants of the crop fungus *U. maydis*, both confirming known factors and identifying novel ones. From the predicted mutant collection we used, most mutants were not significantly depleted from the output reads, indicating no function in virulence for the underlying genes. However, the role of some factors could be difficult to decipher, for example, because their action could be covered by functional redundancy of other virulence factors. Although we infected insertion mutants in dense pools, depleted insertional mutants appeared not to be affected by in trans complementation, by using the secreted factors of neighbouring fungal cells for example. Nevertheless, it cannot be excluded that, for certain gene products, in trans complementation could occur and mask the virulence defect of the respective mutant in a pooled infection setup. In conclusion, negative depletion screens have limitations to decipher redundancy and potential in trans complementation of virulence factors. In addition, we did not identify significantly enriched mutants in the iPool-Seq analysis of the mutant collection. A significant enrichment of output reads would indicate the loss of a negative regulator of virulence. A possible reason that we did not find enrichment could be our choice of the maize accession, EGB, which is highly susceptible to the *U. maydis* strain SG200.

Microscopy of *U. maydis* strain SG200 infecting maize tissue implies that many cells fail to penetrate the host [28]. In very complex insertion mutant libraries, this large individual failure rate could lead to the loss of mutants that lack any real defect in virulence. Therefore, for genome-wide virulence maps of *U. maydis* and similar biotrophic pathogens, the size of the insertion mutant pool must be individually adapted to the infection rate of the respective pathogen. To overcome this problem, genome-wide screens might need to be performed in subpools, as it has been done in a previous study with the fungal pathogen *Cryptococcus neoformans* [11].

iPool-Seq uses insertion cassette–specific primers to amplify the genomic insertion junctions from a mutant pool [17]. Additionally, iPool-Seq enriches for PCR products by using biotin/streptavidin interaction, an approach that has previously been used in bacterial transposon integration site identification methods such as high-throughput insertion tracking by deep sequencing (HITS) [5]. Importantly, UMIs in the adapter primer allow in silico elimination of PCR biases. The unique barcode identifiers additionally overcome cluster position identification problems during Illumina sequencing that would otherwise occur when the first bases from the insertion flank would otherwise be identical for all mutant loci. Dark cycle sequencing, as used in Quantitative insertion-site sequencing (QIseq) for example, is therefore unnecessary [17].

iPool-Seq was established using a defined insertion mutant collection of *U. maydis*. However, the technology can be adapted to any insertion mutant collection, such as transposon or *A. tumefaciens*-derived T-DNA libraries [33, 34]. The modeling of the iPool-Seq sensitivity indicates that iPool-Seq meets all premises to work for high-throughput. Therefore, iPool-Seq

promises to be a versatile technology for reanalysis of existing knock-in, activation-tagging, or transposon-insertion libraries, dramatically reducing labor costs for selection screens when compared to classical scoring approaches. Additionally, the relatively low costs of iPool-Seq for broad screens could also foster research in less funded emerging model systems. Due to the strong enrichment of insertion gene cassettes, the sequencing depth and costs of iPool-Seq are low. Thus, this technology will enable researchers to test diverse new selection criteria to efficiently build genotype–phenotype relationships. This will help to fill the knowledge gap that is currently still hampering research as exemplified for the well annotated *U. maydis* genome with 6,786 protein-encoding genes, of which 41.5% are in the category unknown [35]. Moreover, even if genes are annotated, their involvement in various biological processes might, simply, not yet be known.

From the candidate virulence factors that we identified with iPool-Seq, we chose 3 for verification and confirmed their virulence defect by classical scoring of disease symptoms. However, the assessment of disease symptoms is indirect, and discrepancies between the two methods might occur for other novel virulence factors. We speculate that the *U. maydis* genome encodes virulence factors whose mutants show reduced proliferation but still cause full disease symptoms based on qualitative measures. In line with this, the iPool-Seq data did not show significant depletions for two mutants that were previously reported with mild defects in symptom induction [24]. In contrast to these disease ratings, iPool-Seq has the potential to identify virulence factors that do not have an obvious effect on symptom formation on a genome-wide level.

In summary, we have demonstrated the functional genomic technology iPool-Seq by identifying both known and novel virulence factors from pooled infection assays of a biotrophic fungus within a complex host background. iPool-Seq is therefore a sensitive in vivo tool for researchers to help fill the genotype–phenotype gap in the post-genomic era.

Methods

Vector construction and insertional mutant generation

For all DNA manipulation we used *Escherichia coli* Mach1 (Thermo Fisher Scientific). The vector backbone for the generation of the mutant collection is based on pGBKT7 (Clontech Laboratories). We replaced kanamycin resistance with a spectinomycin resistance cassette and removed internal *Sap*I, *Bsa*I, *BsmB*I, and *Bbs*I restriction sites by direct mutagenesis from a derivative of the original vector, respectively [36]. The hygromycin resistance marker originates from vector pHwtFRT [37]; and *Sap*I, *Bsa*I, *BsmB*I, and *Bbs*I restriction sites were removed by site-directed mutagenesis. Moreover, we elongated the hygromycin cassette with a UPS on the 5′- and 3′-end (5′-TCGCCACAGGATACCACAGGACATCTGGGATATC and 3′-GCCACTCACGCCACAGGATACCACAGGACATCTGGGATATC; UPS is underlined). In detail, for each mutant locus we amplified 1,000 bp up- and downstream borders from *U. maydis* gDNA with standard molecular cloning procedures [38] and combined them with the modified hygromycin-selectable marker cassette flanked with UPS (Fig 2; S2 Table) and the plasmid backbone. Depending on the occurrence of internal restriction sites, we used either *Sap*I, *Bsa*I, *BsmB*I, or *Bbs*I restriction sites (ordered by priority of choice) for Golden Gate cloning [23]. Constructs were verified by Sanger sequencing and subsequently transformed into the haploid solopathogenic strain SG200 of *U. maydis* as previously described [18, 39, 40]. Transformants were verified by direct PCR: single mutants were grown in YepsLight (0.4% yeast extract, 0.4% peptone and 2% sucrose) liquid medium at 28˚C with shaking at 200 rpm in 48-well plates overnight. The next day, 100 μL overnight culture was pelleted and resuspended in 20 μL 0.02 M NaOH. 1 μL was then utilized for a direct PCR reaction with a primer pair directed against the replaced gene. As a positive control, a primer pair binding to another mutant locus was used.

Subsequently, we isolated gDNA from at least 4 PCR positive strains and repeated the direct PCR using 1 μL of 1:10 diluted gDNA as a template. All primer pairs used for the verification of deletion strains produced PCR products from a gDNA template from the progenitor strain SG200. Three independently verified *U. maydis* insertional mutants were preserved at −80˚C in PD liquid supplemented with 50% glycerol.

Growth conditions and pooled infection

For each mutant collection pool replicate we infected at least 100 plants of maize variety EGB (Olds Seeds, Madison, WI, USA). Seedlings were grown under a 14-hour/10-hour light/dark cycle at 28˚C/20˚C in plant growth chambers and infected 7 days after potting. *U. maydis* mutant strains were grown individually on selective PD plates supplemented with 200 μg/mL hygromycin for 2–3 days at 28˚C. Subsequently, for each mutant strain, 1 mL YepsLight (0.4% yeast extract, 0.4% peptone and 2% sucrose) liquid preculture was inoculated in 48-well plates and grown at 28˚C overnight with shaking at 200 rpm. For main cultures, precultures were diluted 1:2,000 in 3 mL YepsLight in test tubes and grown at 28˚C with shaking at 200 rpm overnight. After 14–16 hours, the main cultures of all mutants were adjusted to an OD_{600} of 3 and mixed in equal amounts. The mutant pool was pelleted at 2,000 x g for 10 minutes and resuspended in sterile water. 250 μL of the mutant pool was infected in each maize seedling with a syringe. After 7 days, infected areas from the second and third leaves were harvested, ground to a fine powder in liquid nitrogen, and preserved at −80˚C until iPool-Seq library preparation.

iPool-Seq library preparation

For output gDNA extraction, 0.75–1 g of infected plant powder was supplemented with 2 mLLysis buffer (10 mM Tris, pH 8; 100 mM NaCl; 1 mM EDTA; 2% Triton X 100 [v/v]; 1% SDS [w/v]), 2.5 mL TE-buffer equilibrated phenol, chloroform, and isoamyl alcohol (25:24:1, pH 7.5–8, Carl Roth) and 100 μL sterile glass beads (450–600 μM, B.Braun) in a 7-mL Precellys tube. The material was processed for 20 seconds at 4,500 rpm with a Precellys evolution bead mill (Bertin). The debris was pelleted at 17,000 x g for 15 minutes, and 2 mL supernatant was added to 2.2 mL Isopropanol. The precipitated gDNA was washed with 1 mL 80% EtOH and eluted in 150 μL or 200 μL TE supplemented with RNAse A (20 μg/mL, Thermo Fisher Scientific). For input gDNA extraction, gDNA was extracted from 2 mL of insertional mutant pool as previously described [41]. gDNA concentrations were determined with PicoGreen (Thermo Fisher Scientific). Tn5 fragmentation of a total of 10 μg gDNA for output and 1 μg gDNA for the input was adapted from [20], and performed as follows [21]: We combined 1 μg gDNA per 20 μL reaction with Tn5 transposase (150 ng/μL f.c.) preloaded with 25-μM adapters in 1x TAPS buffer (50 mM TAPS-NaOH, 25 mM MgCl2, 50% v/v DMF, pH 8.5 at 25˚C) and incubated the reaction mix in a thermocycler at 55˚C for 10 minutes. We purified each reaction mix with a 1:1 ratio of Agencourt AMPure XP beads (Beckman Coulter) according to the manufacturer's protocol and performed PCR with Phusion polymerase (New England Biolabs) using an adapter specific forward primer and a biotinylated insertion specific primer from 250 ng fragmented gDNA (denaturation for 15 seconds at 95˚C, annealing for 15 seconds at 65˚C, elongation for 30 seconds at 72˚C; repeated for 15 cycles; 1 minute final elongation). We pooled all PCRs of the same sample and purified 1/5 with Agencourt AMPure XP beads (ratio 1:1; Beckman Coulter). The PCR amplicons eluted from each sample were split into 4 PCR reactions and amplified with nested primers to add Illumina compatible P5 and P7 ends (15 cycles, with 65˚C annealing temperature and 30 seconds elongation at 72˚C). The final PCR products were purified with Agencourt AMPure XP beads in a 1:1 ratio. The average fragment

size was measured on a fragment analyzer (Advanced Analytical Technologies, Inc.) and library quality was controlled with qPCR. Illumina Sequencing was performed on a MiSeq platform with 75 PE conditions. We used a custom designed forward sequencing primer and the standard Illumina primers for reverse and index sequencing (S2 Table).

Confirmation of iPool-Seq candidate virulence factors

We confirmed the results of iPool-Seq for 3 candidate genes with individual infection assays, microscopy, and in vitro growth assays. The infection assay was performed as previously described [18]. In summary, for each insertional mutant, 3 replicates of *U. maydis* were grown overnight in YepsLight liquid medium (0.4% yeast extract, 0.4% peptone and 2% sucrose) with 200 rpm agitation to an OD_{600} of 0.6–1 and adjusted to an OD_{600} of 1 in sterile water. We syringe-infected 7-day-old maize seedlings of the variety EGB with approximately 250 μL fungal suspension per plant. Symptoms were scored 7 days post infection (dpi) according to the published protocol [18]. Fungal leaf colonization was assessed 7 dpi via microscopy. Fungal hyphae were stained with WGA-AF488 (Thermo Fisher Scientific) and plant cell walls with propidium iodide (Sigma-Aldrich) as previously described [28]. Confocal microscopy was performed with the following settings: We utilized an LSM780 Axio Observer confocal laser scanning microscope with an LD LCI Plan-Apochromat 25x/0.8 Imm Corr DIC M27 objective (Zeiss, Jena, Germany). WGA-AF488 was excited at 488 nm and detected at 500–540 nm; propidium iodide was excited at 561 nm and detected at 580–660 nm.

Bioinformatic analysis

For each sequenced library, adapter read-throughs were removed from the raw Illumina reads, UMIs were extracted and stored separately, and the reads (lacking UMIs) were mapped to the *U. maydis* reference genome [18] using NextGenMap [42]. The reads mapping to each flank (5' and 3') of each insertional mutant were grouped by UMI, and highly similar UMIs were merged to correct for sequencing errors [43]. UMIs with fewer reads than the error-correction threshold were removed as likely artifacts, and the number of surviving (and thus likely true) UMIs for each gene and flank were counted. To correct for biases caused by different detection losses (i.e., # undetected genomes/# total genomes) between mutants and flanks, the mutant- and flank-specific losses were estimated from the observed mutant- and flank-specific distributions of reads per UMI (S1 Supporting methods) using the TRUmiCount algorithm (see S1 Supporting methods for details) [44]. To discern stochastic fluctuations from knockout phenotypes, the number of true UMIs detected in the output pool for neutral insertional mutants were assumed to follow a negative binomial distribution with mean $\mu_m = \lambda \cdot n_m^{\text{in}} \cdot \left(1 - \ell_m^{\text{out}}\right) / \left(1 - \ell_m^{\text{in}}\right)$ and (inverse) overdispersion parameter $r_m = n_m^{\text{in}} / \left(1 + d \cdot n_m^{\text{in}}\right)$. Briefly, a neutral mutant m's expected UMI count in the output pool thus depends on (1) the number n_m^{in} of detected UMIs in the input pool, (2) the estimated losses ℓ_m^{out} and ℓ_m^{in} for the output and input pool, and (3) a mutant-independent normalization factor λ to account for differences in total genome count between input and output samples. The sources of overdispersion of the output counts are (4) the (Poissonian) sampling uncertainty of the input pool counts n_m^{in}, and (5) random fluctuations of fungus proliferation accounted for by the mutant-independent parameter d. For each output pool, parameters λ and d were estimated (see S1 Supporting methods for details) by fitting the model to a reference set of presumed neutral mutants (S3 Table), 2 one-sided p-values for the significance of depletion (respectively, enrichment) compared to the reference set were computed for each insertional mutant and transformed to q-values to control for the false discovery rate (FDR) [45]. Undetected insertional mutants (i.e., with zero UMIs) in input pools were

excluded from the analysis of the corresponding output pools. Undetected insertional mutants in output pools were not assigned *p*- or q-values.

To quantify the change in virulence of an insertional mutant, its abundance in the output was first normalized to its abundance in the input (thus assuming independent fates of the individuals in the input). Then, the $\log_2$-fold change between its normalized output abundance and the normalized output abundance of the internal reference set was computed (see S1 Supporting methods for details). Further details on the modeling can be found in S1 Supporting methods.

Supporting information

S1 Data. q-Values of *U. maydis* mutant strains.
(XLSX)

S2 Data. Symptom rating of mutant strains.
(XLSX)

S1 Fig. Workflow of pooled infection of maize. For each replicate of the *U. maydis* mutant collection, at least 100 maize plants of the accession EGB were potted. Mutants were grown on selective plates for 2–3 days. From plates, precultures were inoculated and grown ON. The precultures were used for inoculation of the main cultures to avoid dead material in the infection pool. All main cultures were pooled with equal amounts that were adjusted to the same optical density and infected in 7-day old maize seedlings with a syringe. Infected areas of the second and third leaf of each plant were harvested 7 days after the infection. All 3 biological replicates of the mutant collection were processed in 14 days. EGB, Early Golden Bantam; ON, overnight.
(TIF)

S2 Fig. Tn5 fragmentation of gDNA with modified adapters. Recombinantly produced hyperactive Tn5 was tested with standard Tn5-ME-A and custom UMI-ME-A on 1 μg gDNA of *U. maydis*-infected maize tissue with indicated concentrations. gDNA; genomic DNA; In, Input; M, Marker 1 kb-ladder (Thermo Scientific); ME, mosaic end; Tn5-ME-A, Tn5-ME-A-dapter; UMI-ME-A, UMI-ME-adapter.
(TIF)

S3 Fig. Sensitivity of iPool-Seq. Estimated sensitivity of iPool-Seq for a genome-wide library of *U. maydis* mutants. Model shows for different (1 up to 100) mutant copies detected in the input sample for the sensitivity of virulence factor detection. Depicted model curves are given assuming 3% of all mutants have a reduced virulence of log2(FC) −1.53 and log2(FC) of −2.75, respectively, and the other 97% are neutral in respect to virulence. The sensitivity reaches 99% at 40 detected mutants (lost virulence) and 100 detected mutants (reduced virulence), respectively. FC, fold change; iPool-Seq, insertion Pool-Sequence.
(TIF)

S1 Software. iPool-Seq analysis pipeline. iPool-Seq, insertion Pool-Sequencing.
(TGZ)

S1 Supporting methods. iPool-Seq analysis pipeline description. iPool-Seq, insertion Pool-Sequencing.
(PDF)

S1 Table. *U. maydis* genes targeted for insertional mutagenesis.
(XLSX)

S2 Table. Key primers used in this study.
(XLSX)

S3 Table. *U. maydis* mutants used for the internal reference set.
(XLSX)

S4 Table. Significantly depleted *U. maydis* mutants identified by iPool-Seq. iPool-Seq, insertion Pool-Sequencing.
(XLSX)

Acknowledgments

We thank the Molecular Biology service at the Vienna BioCenter for their excellent Sanger Sequencing support. Special thanks go to the NGS Core Facility (VBCF) for the adaptation of sequencing protocols. Our acknowledgement goes to the Vienna BioCenter Core Facility (VBCF) plant facilities for their support with plant growth. We thank Dr. Kirsten Senti for technical advice. We would like to thank Dr. Matthias Schäfer for helpful advice on the manuscript. Special thanks go to Dr. Youssef Belkhadir and Dr. Yasin Dagdas for critical reading of the manuscript and constructive suggestions. We would also like to thank all the members of the Djamei lab and of the CIBIV (Center for Integrative Bioinformatics Vienna), particularly Dr. Angelika Czedik-Eysenberg, Denise Seitner, Luis Paulin-Paz and Celine Prakash, for valuable feedback on the project and manuscript, and Dr. J. Matthew Watson for proofreading.

Author Contributions

Conceptualization: Armin Djamei.

Data curation: Florian G. Pflug.

Funding acquisition: Arndt von Haeseler, Armin Djamei.

Investigation: Simon Uhse, Alexandra Stirnberg, Klaus Ehrlinger.

Methodology: Simon Uhse, Florian G. Pflug, Armin Djamei.

Project administration: Arndt von Haeseler, Armin Djamei.

Resources: Simon Uhse, Alexandra Stirnberg, Klaus Ehrlinger, Armin Djamei.

Software: Florian G. Pflug.

Supervision: Arndt von Haeseler, Armin Djamei.

Validation: Simon Uhse, Florian G. Pflug.

Visualization: Simon Uhse, Florian G. Pflug.

Writing – original draft: Simon Uhse.

Writing – review & editing: Armin Djamei.

References

1. Idnurm A, Reedy JL, Nussbaum JC, Heitman J. Cryptococcus neoformans virulence gene discovery through insertional mutagenesis. Eukaryot Cell. 2004; 3(2):420–9. https://doi.org/10.1128/EC.3.2.420-429.2004 PubMed PMID: WOS:000220945300018. PMID: 15075272
2. Wamatu J, White D, Chen W. Insertional mutagenesis of Sclerotinia sclerotiorum through Agrobacterium tumefaciens-mediated transformation. Phytopathology. 2005; 95(6):S108–S. PubMed PMID: WOS:000202991401123.

3. Jeon J, Park SY, Chi MH, Choi J, Park J, Rho HS, et al. Genome-wide functional analysis of pathogenicity genes in the rice blast fungus. Nat Genet. 2007; 39(4):561–5. https://doi.org/10.1038/ng2002 PMID: 17353894.
4. Michielse CB, van Wijk R, Reijnen L, Cornelissen BJC, Rep M. Insight into the molecular requirements for pathogenicity of Fusarium oxysporum f. sp lycopersici through large-scale insertional mutagenesis. Genome Biol. 2009; 10(1). doi: ARTN R4 https://doi.org/10.1186/gb-2009-10-1-r4 PubMed PMID: WOS:000263823200010. PMID: 19134172
5. Gawronski JD, Wong SMS, Giannoukos G, Ward DV, Akerley BJ. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung. P Natl Acad Sci USA. 2009; 106(38):16422–7. https://doi.org/10.1073/pnas.0906627106 PubMed PMID: WOS:000270071600076. PMID: 19805314
6. van Opijnen T, Bodi KL, Camilli A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat Methods. 2009; 6(10):767–U21. https://doi.org/10.1038/nmeth.1377 PubMed PMID: WOS:000270355200023. PMID: 19767758
7. Langridge GC, Phan MD, Turner DJ, Perkins TT, Parts L, Haase J, et al. Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. Genome Res. 2009; 19(12):2308–16. https://doi.org/10.1101/gr.097097.109 PubMed PMID: WOS:000272273400015. PMID: 19826075
8. Goodman AL, McNulty NP, Zhao Y, Leip D, Mitra RD, Lozupone CA, et al. Identifying Genetic Determinants Needed to Establish a Human Gut Symbiont in Its Habitat. Cell Host Microbe. 2009; 6(3):279–89. https://doi.org/10.1016/j.chom.2009.08.003 PubMed PMID: WOS:000270290700011. PMID: 19748469
9. Michielse CB, Hooykaas PJJ, van den Hondel CAMJJ, Ram AFJ. Agrobacterium-mediated transformation as a tool for functional genomics in fungi. Curr Genet. 2005; 48(1):1–17. https://doi.org/10.1007/s00294-005-0578-0 PubMed PMID: WOS:000230624600001. PMID: 15889258
10. Colot HV, Park G, Turner GE, Ringelberg C, Crew CM, Litvinkova L, et al. A high-throughput gene knockout procedure for Neurospora reveals functions for multiple transcription factors. P Natl Acad Sci USA. 2006; 103(27):10352–7. https://doi.org/10.1073/pnas.0601456103 PubMed PMID: WOS:000239069400037. PMID: 16801547
11. Liu OW, Chun CD, Chow ED, Chen C, Madhani HD, Noble SM. Systematic genetic analysis of virulence in the human fungal pathogen Cryptococcus neoformans. Cell. 2008; 135(1):174–88. https://doi.org/10.1016/j.cell.2008.07.046 PMID: 18854164; PubMed Central PMCID: PMCPMC2628477.
12. Crimmins GT, Mohammadi S, Green ER, Bergman MA, Isberg RR, Mecsas J. Identification of MrtAB, an ABC Transporter Specifically Required for Yersinia pseudotuberculosis to Colonize the Mesenteric Lymph Nodes. PLoS Pathog. 2012; 8(8). doi: ARTN e1002828 https://doi.org/10.1371/journal.ppat.1002828 PubMed PMID: WOS:000308558000009. PMID: 22876175
13. van Opijnen T, Camilli A. A fine scale phenotype-genotype virulence map of a bacterial pathogen. Genome Res. 2012; 22(12):2541–51. https://doi.org/10.1101/gr.137430.112 PubMed PMID: WOS:000311895500022. PMID: 22826510
14. Wang ND, Ozer EA, Mandel MJ, Hauser AR. Genome-Wide Identification of Acinetobacter baumannii Genes Necessary for Persistence in the Lung. Mbio. 2014; 5(3). doi: ARTN e01163-14 https://doi.org/10.1128/mBio.01163-14 PubMed PMID: WOS:000338875900069. PMID: 24895306
15. Cole BJ, Feltcher ME, Waters RJ, Wetmore KM, Mucyn TS, Ryan EM, et al. Genome-wide identification of bacterial plant colonization genes. PLoS Biol. 2017; 15(9):e2002860. https://doi.org/10.1371/journal.pbio.2002860 PMID: 28938018.
16. Le Breton Y, Belew AT, Freiberg JA, Sundar GS, Islam E, Lieberman J, et al. Genome-wide discovery of novel M1T1 group A streptococcal determinants important for fitness and virulence during soft-tissue infection. PLoS Pathog. 2017; 13(8):e1006584. https://doi.org/10.1371/journal.ppat.1006584 PMID: 28832676.
17. Bronner IF, Otto TD, Zhang M, Udenze K, Wang CQ, Quail MA, et al. Quantitative insertion-site sequencing (QIseq) for high throughput phenotyping of transposon mutants. Genome Res. 2016; 26(7):980–9. https://doi.org/10.1101/gr.200279.115 PubMed PMID: WOS:000378986000011. PMID: 27197223
18. Kamper J, Kahmann R, Bolker M, Ma LJ, Brefort T, Saville BJ, et al. Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis. Nature. 2006; 444(7115):97–101. https://doi.org/10.1038/nature05248 PubMed PMID: WOS:000241701500053. PMID: 17080091
19. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, et al. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomic. 2015; 15(2):141–61. https://doi.org/10.1007/s10142-015-0433-4 PubMed PMID: WOS:000351397700003. PMID: 25722247
20. Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, et al. The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science. 2009; 326(5956):1112–5. https://doi.org/10.1126/science.1178534 PubMed PMID: WOS:000271951000044. PMID: 19965430

21. Picelli S, Bjorklund AK, Reinius B, Sagasser S, Winberg G, Sandberg R. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. 2014; 24(12):2033–40. https://doi.org/10.1101/gr.177881.114 PubMed PMID: WOS:000345810600011. PMID: 25079858

22. Stern DL. Tagmentation-Based Mapping (TagMap) of Mobile DNA Genomic Insertion Sites. bioRxiv. 2017. https://doi.org/10.1101/037762

23. Engler C, Gruetzner R, Kandzia R, Marillonnet S. Golden Gate Shuffling: A One-Pot DNA Shuffling Method Based on Type IIs Restriction Enzymes. PLoS ONE. 2009; 4(5). doi: ARTN e5553 https://doi.org/10.1371/journal.pone.0005553 PubMed PMID: WOS:000266107300017. PMID: 19436741

24. Schilling L, Matei A, Redkar A, Walbot V, Doehlemann G. Virulence of the maize smut Ustilago maydis is shaped by organ-specific effectors. Mol Plant Pathol. 2014; 15(8):780–9. https://doi.org/10.1111/mpp.12133 PubMed PMID: WOS:000342131900003. PMID: 25346968

25. Schipper K, Brefort T, Doehlemann G, Djamei A, Muench K, Kahmann R. The secreted protein Stp1 is cruical for establishment of the biotrophic interaction of the smut fungus Ustilago maydis with its host plant maize. Eur J Cell Biol. 2008; 87:29–. PubMed PMID: WOS:000255316100068.

26. Doehlemann G, van der Linde K, Amann D, Schwammbach D, Hof A, Mohanty A, et al. Pep1, a Secreted Effector Protein of Ustilago maydis, Is Required for Successful Invasion of Plant Cells. PLoS Pathog. 2009; 5(2). doi: ARTN e1000290 https://doi.org/10.1371/journal.ppat.1000290 PubMed PMID: WOS:000263928000034. PMID: 19197359

27. Mueller AN, Ziemann S, Treitschke S, Assmann D, Doehlemann G. Compatibility in the Ustilago maydis-Maize Interaction Requires Inhibition of Host Cysteine Proteases by the Fungal Effector Pit2. PLoS Pathog. 2013; 9(2). doi: ARTN e1003177 https://doi.org/10.1371/journal.ppat.1003177 PubMed PMID: WOS:000315648900027. PMID: 23459172

28. Stirnberg A, Djamei A. Characterization of ApB73, a virulence factor important for colonization of Zea mays by the smut Ustilago maydis. Mol Plant Pathol. 2016; 17(9):1467–79. https://doi.org/10.1111/mpp.12442 PubMed PMID: WOS:000389134900013. PMID: 27279632

29. Eichhorn H, Lessing F, Winterberg B, Schirawski J, Kamper J, Muller P, et al. A ferroxidation/permeation iron uptake system is required for virulence in Ustilago maydis. Plant Cell. 2006; 18(11):3332–45. https://doi.org/10.1105/tpc.106.043588 PubMed PMID: WOS:000243093700035. PMID: 17138696

30. Lanver D, Muller AN, Happel P, Schweizer G, Haas FB, Franitza M, et al. The biotrophic development of Ustilago maydis studied by RNAseq analysis. Plant Cell. 2018. https://doi.org/10.1105/tpc.17.00764 PMID: 29371439.

31. Gray AN, Koo BM, Shiver AL, Peters JM, Osadnik H, Gross CA. High-throughput bacterial functional genomics in the sequencing era. Curr Opin Microbiol. 2015; 27:86–95. https://doi.org/10.1016/j.mib.2015.07.012 PubMed PMID: WOS:000365065400014. PMID: 26336012

32. Armbruster CE, Forsyth-DeOrnellas V, Johnson AO, Smith SN, Zhao L, Wu W, et al. Genome-wide transposon mutagenesis of Proteus mirabilis: Essential genes, fitness factors for catheter-associated urinary tract infection, and the impact of polymicrobial infection on fitness requirements. PLoS Pathog. 2017; 13(6):e1006434. https://doi.org/10.1371/journal.ppat.1006434 PMID: 28614382; PubMed Central PMCID: PMCPMC5484520.

33. Kemppainen M, Duplessis S, Martin F, Pardo AG. T-DNA insertion, plasmid rescue and integration analysis in the model mycorrhizal fungus Laccaria bicolor. Microb Biotechnol. 2008; 1(3):258–69. https://doi.org/10.1111/j.1751-7915.2008.00029.x PMID: 21261845; PubMed Central PMCID: PMCPMC3815887.

34. Honkanen S, Jones VAS, Morieri G, Champion C, Hetherington AJ, Kelly S, et al. The Mechanism Forming the Cell Surface of Tip-Growing Rooting Cells Is Conserved among Land Plants. Current Biology. 2016; 26(23):3238–44. https://doi.org/10.1016/j.cub.2016.09.062 PubMed PMID: WOS:000389590500032. PMID: 27866889

35. Walter MC, Rattei T, Arnold R, Guldener U, Munsterkotter M, Nenova K, et al. PEDANT covers all complete RefSeq genomes. Nucleic Acids Research. 2009; 37:D408–D11. https://doi.org/10.1093/nar/gkn749 PubMed PMID: WOS:000261906200074. PMID: 18940859

36. Rabe F, Seitner D, Bauer L, Navarrete F, Czedik-Eysenberg A, Rabanal FA, et al. Phytohormone sensing in the biotrophic fungus Ustilago maydis—the dual role of the transcription factor Rss1. Mol Microbiol. 2016; 102(2):290–305. https://doi.org/10.1111/mmi.13460 PubMed PMID: WOS:000386101000009. PMID: 27387604

37. Khrunyk Y, Munch K, Schipper K, Lupas AN, Kahmann R. The use of FLP-mediated recombination for the functional analysis of an effector gene family in the biotrophic smut fungus Ustilago maydis. New Phytol. 2010; 187(4):957–68. https://doi.org/10.1111/j.1469-8137.2010.03413.x PubMed PMID: WOS:000280998600012. PMID: 20673282

38. Sambrook J, Russell DW, Sambrook J. The condensed protocols from Molecular cloning: a laboratory manual. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press; 2006. v, 800 p. p.

39. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977; 74(12):5463–7. PMID: 271968; PubMed Central PMCID: PMCPMC431765.

40. Kamper J. A PCR-based system for highly efficient generation of gene replacement mutants in Ustilago maydis. Mol Genet Genomics. 2004; 271(1):103–10. https://doi.org/10.1007/s00438-003-0962-8 PMID: 14673645.

41. Hoffman CS, Winston F. A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of Escherichia coli. Gene. 1987; 57(2–3):267–72. PMID: 3319781.

42. Sedlazeck FJ, Rescheneder P, von Haeseler A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. Bioinformatics. 2013; 29(21):2790–1. https://doi.org/10.1093/bioinformatics/btt468 PMID: 23975764.

43. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. 2017; 27(3):491–9. https://doi.org/10.1101/gr.209601.116 PubMed PMID: WOS:000395694000015. PMID: 28100584

44. Pflug FG, von Haeseler A. TRUmiCount: Correctly counting absolute numbers of molecules using unique molecular identifiers. bioRxiv. 2017. https://doi.org/10.1101/217778

45. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995; 57(1):289–300. http://www.jstor.org/stable/2346101?seq=1#page_scan_tab_contents

This page intentionally left blank.

Chapter 4

Insertion Pool Sequencing for Insertional Mutant Analysis in Complex Host-Microbe Interactions (Uhse *et al.*, 2019. *Current Protocols in Plant Biology*, *4*: e20097)

Preamble

The iPool-Seq protocol (Uhse *et al.*, 2018; chapter 3 of this work) depends on the TRUmiCount algorithm (Pflug & von Haeseler, 2018; chapter 2 of this work) to ensure unbiased quantification of mutants abundances, and on a tailored statistical model to distinguish significant differences between mutants' virulences (chapter 6). By providing a polished and easy-to-use data analysis pipeline and step-by-step instructions, the publication presented here aims to make both the iPool-Seq wet-lab protocol and the computational analysis of iPool-Seq data as easily accessible to potential users as possible.

The instructions also include guidance on how to select TRUmiCount's *error-*

correction threshold T; see figure 2 and section *Adjusting the TRUmiCount phantom rejection threshold* (Uhse *et al.*, 2019; chapter 4 of this work; pages 78 and 82).

Note: Basic Protocols 1 through 3 pertain to the *wet-lab* parts of iPool-Seq and are neither the work of the author of this thesis, nor are they relevant to the rest of the work (except maybe Basic Protocol 3, which describes the Streptavidin purification step, and thus has some bearing on the applicability of TRUmiCount; see the discussion in chapter 5). These parts are included only for completeness' sake.

Author contributions

Design, development and testing of the iPool-Seq pipeline, writing of *Basic Protocol 4. Bioinformatic analysis*, and writing of the pipeline-related discussion in the *Commentary* were done by the author of this thesis, under the guidance and supervision of Arndt von Haeseler.

Publication history and status

Submitted on March 4th 2019 to *Current Protocols in Plant Biology*, accepted for publication on May 31st 2019, published online on July 3rd 2019.

Uhse S.[e], **Pflug F. G.**[e], von Haeseler A., & Djamei A. (2019). Insertion pool sequencing for insertional mutant analysis in complex host-microbe interactions. *Current Protocols in Plant Biology, 4*, e20097. DOI:10.1002/cppb.20097

[e]these authors contributed equally to this work

Insertion Pool Sequencing for Insertional Mutant Analysis in Complex Host-Microbe Interactions

Simon Uhse,[1,5] Florian G. Pflug,[2,5] Arndt von Haeseler,[2,3] and Armin Djamei[1,4,6]

[1]Gregor Mendel Institute (GMI), Austrian Academy of Sciences, Vienna BioCenter (VBC), Vienna, Austria
[2]Center for Integrative Bioinformatics Vienna (CIBIV), Joint Institute of the University of Vienna and Medical University of Vienna, Max F. Perutz Laboratories (MFPL), Vienna, Austria
[3]Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria
[4]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany
[5]These authors contributed equally to this work
[6]Corresponding author: *djamei@ipk-gatersleben.de*

Insertional mutant libraries of microorganisms can be applied in negative depletion screens to decipher gene functions. Because of underrepresentation in colonized tissue, one major bottleneck is analysis of species that colonize hosts. To overcome this, we developed insertion pool sequencing (iPool-Seq). iPool-Seq allows direct analysis of colonized tissue due to high specificity for insertional mutant cassettes. Here, we describe detailed protocols for infection as well as genomic DNA extraction to study the interaction between the corn smut fungus *Ustilago maydis* and its host maize. In addition, we provide protocols for library preparation and bioinformatic data analysis that are applicable to any host-microbe interaction system.

Keywords: fungal genomics • maize • plant genomics • plant-microbe interactions • unique molecular identifiers • *Ustilago maydis* • virulence factors

How to cite this article:
Uhse, S., Pflug, F. G., von Haeseler, A., & Djamei, A. (2019). Insertion pool sequencing for insertional mutant analysis in complex host-microbe interactions. *Current Protocols in Plant Biology*, *4,* e20097. doi: 10.1002/cppb.20097

INTRODUCTION

Insertional mutagenesis has been used in fungi, including baker's yeast and *Cryptococcus neoformans*, to decipher gene functions (Giaever et al., 2002; Idnurm, Reedy, Nussbaum, & Heitman, 2004; Winzeler et al., 1999). Negative depletion screening in combination with insertional mutagenesis libraries is a powerful approach to decipher gene functions

CURRENT PROTOCOLS
A Wiley Brand

Current Protocols in Plant Biology e20097, Volume 4
Published in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cppb.20097

under a certain condition, e.g., host infection (Jeon et al., 2007). Recent advances in massive parallel sequencing allow for large-scale approaches in bacteria that permit analysis of larger pools of insertional mutants (Gawronski, Wong, Giannoukos, Ward, & Akerley, 2009; van Opijnen, Bodi, & Camilli, 2009). Here, we describe the insertion pool sequencing (iPool-Seq) pipeline that we recently established with the *Ustilago maydis–Zea mays* interaction (Uhse et al., 2018).

U. maydis is a smut fungus that colonizes and overcomes the immunity of the crop plant maize (Kamper et al., 2006). Many molecular and genetic tools are available for *U. maydis*, and therefore, the fungus is an important model organism in the field of plant-microbe interactions (Lanver et al., 2017). Especially useful for generation of an insertional mutagenesis library is the availability of a solopathogenic *U. maydis* strain that is haploid and capable of infecting (Kamper et al., 2006). Many plant pathogens, including *U. maydis,* rely on the versatile repertoire of effector genes that mediate and shape the interaction with the host plant. The majority of predicted *U. maydis* effector genes are unstudied, and it is unclear if and to what extent they contribute to virulence (Kamper et al., 2006). To gain insights about the effector repertoire of *U. maydis*, we generated an insertional mutagenesis library and employed it in a negative depletion screen during infection of maize to elucidate novel virulence factors in the fungus.

Transformation of insertional cassettes via homologous recombination is well established in *U. maydis* and has been used successfully to delete clusters of predicted effector genes (Kamper et al., 2006). We created an insertional mutant library for *U. maydis* via homologous recombination of a selectable marker conferring resistance to hygromycin. Next, we established the iPool-Seq workflow based on this library, allowing for controlled insertional mutagenesis at loci of predicted effectors that are likely to contribute to the virulence of *U. maydis* (Uhse et al., 2018). All newly generated *U. maydis* mutants were verified for deletion of the targeted effector genes via PCR on cultures and on extracted genomic DNA (gDNA). Eventually, the library comprised three independent replicates for each insertional mutant, with 195 putative virulence factor mutants for *U. maydis* in total. We used this library to conduct a negative depletion screen by infection of the host plant maize and subsequent analysis of the input and output compositions. The input, i.e., the gDNA of the mutant pool before the infection, and the output, i.e., the gDNA of the infected host material, were prepared, deep-sequenced, and bioinformatically analyzed.

iPool-Seq was performed on the entire library of 195 insertional mutants. The method was highly selective for reads from *U. maydis* insertional mutant loci, yielding >75% informative reads for input and output samples. Moreover, we identified 28 reproducibly and significantly depleted mutants from next-generation sequencing (NGS) reads after infection in the maize Early Golden Bantam. Several of the mutants that we found with iPool-Seq were previously shown to be virulence factors. For instance, mutants of the *U. maydis* effectors Pep1, ApB73, and, more recently, Cce1 displayed severe virulence defects based on classical disease ratings (Doehlemann et al., 2009; Seitner, Uhse, Gallei, & Djamei, 2018; Stirnberg & Djamei, 2016). The confirmation of these known virulence factors by iPool-Seq indicates that iPool-Seq yields reliable results for depleted mutants. To further strengthen this finding, we tested three potential virulence factors identified by iPool-Seq and were able to show a virulence phenotype for these mutants based on classical disease ratings in comparison to the wildtype solopathogenic strain SG200 (Uhse et al., 2018).

The protocols described here aim to make the full potential of iPool-Seq accessible to the larger scientific community. The iPool-Seq workflow, from infection to sequence analysis, is divided into four parts (Fig. 1): Basic Protocol 1 describes the process of infection of the host plant maize with insertional mutant pools. Basic Protocol 2 describes the extraction of gDNA from input samples before infection and from output samples

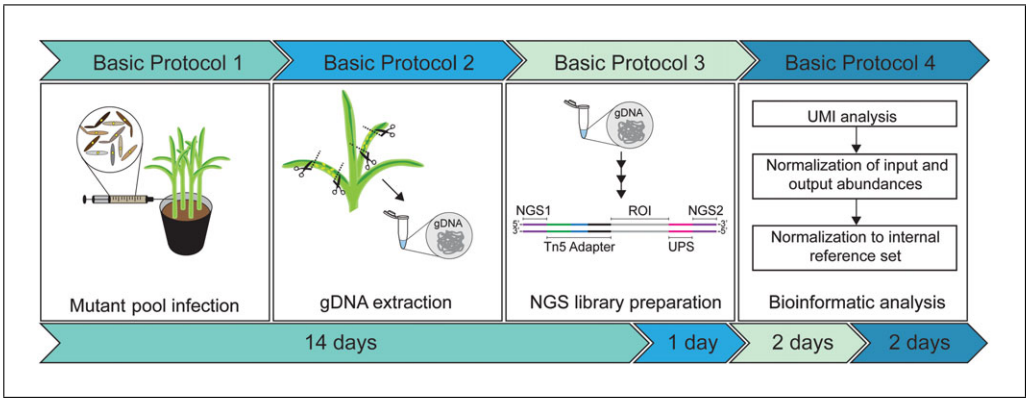

Figure 1 Overview of the iPool-Seq pipeline. The pipeline contains four parts, which can be finished sequentially in ~20 days. gDNA, genomic DNA; NGS, next-generation sequencing; UMI, unique molecular identifier.

after infection. Both Basic Protocol 1 and Basic Protocol 2 were established for *the U. maydis–Z. mays* pathosystem and might require adaptation when applied to other host-microbial interaction systems. In contrast, Basic Protocols 3 and 4 are applicable to any host-microbial interaction system: Basic Protocol 3 describes the NGS library preparation in detail, and Basic Protocol 4 details bioinformatic analysis of the sequencing results for the input and output libraries, with the goal of detecting changes in the virulence of particular insertional mutants compared to a reference set of neutral controls.

U. MAYDIS INSERTIONAL MUTANT POOL INFECTION IN MAIZE

BASIC PROTOCOL 1

For generation of a negatively depleted output, the insertional mutant library of *U. maydis* must be raised, pooled, and infected into its host maize. The following protocol describes the processes of infection and of harvest of the infected maize tissue.

NOTE: Repeat the procedure for a total of three biological replicates.

Materials

Soil [4:1 mixture of standard potting soil (Einheitserde Werkverband e.V.) with perlite (Granuflor)]
Early Golden Bantam (EGB) maize seeds
Nematode (Biohelp) solution (1 g in 3 L water)
Cryopreserved *U. maydis* insertional mutant library (strain SG200 background genotype; transformed via homologous recombination; Kamper et al., 2006; Uhse et al., 2018)
Potato dextrose–agar plates containing 200 µg/ml hygromycin (see recipe)
YepsLight liquid medium (see recipe)
Double-distilled water
0.05% (v/v) Tween-20 (Sigma-Aldrich) in double-distilled water
Liquid nitrogen

12-cm-diameter round pots
Phytochamber
5-ml glass pipets
48-deep-well liquid culture plates (5-ml well volume, UCT)
28°C incubator-shaker
15- and 50-ml Falcon tubes
Rotator
Photometer/plate reader
500-ml centrifuge tube

Standard tabletop centrifuge
1-ml syringe (B. Braun)
0.45-mm-diameter needle (25-mm long, B. Braun)
Scissors
1-L beaker (Duran)
Magnetic stirrer (IKA) and stir bar
Mortar and pestle
Metal spatula
Laboratory Mixer Mill MM 200 (Retsch) and compatible container
8-mm-diameter metal balls (Kugel-Pompel)

NOTE: All reagents, consumables, and equipment coming into contact with living *U. maydis* axenic culture cells must be sterile. Working in a laminar flow hood is recommended, if possible.

Potting of maize

1. Distribute soil in 12-cm-diameter round pots and water pots sufficiently. Seed five EGB maize seeds per pot for a total of >100 seeds per insertional mutant pool replicate. Treat each pot with 100 ml nematode solution for pest control.
2. Grow maize in a phytochamber under the following conditions: 14-hr/10-hr light/dark cycle at 28°C/20°C with a total light intensity of 183.21 µmol m^{-} s^{-1}.

After 7 days, maize seedlings are ready for infection with the insertional mutant pool.

Growth of U. maydis insertional mutant library

3. Distribute cryopreserved *U. maydis* insertional mutant library strains on potato dextrose–agar plates containing 200 µg/ml hygromycin using 5-ml glass pipets. Keep individual mutant strains separate by streaking ≤8 strains on designated sectors per plate. Grow at 28°C for 2 to 3 days in the dark.
4. Inoculate each strain in 2 ml YepsLight liquid medium per well in 48-deep-well liquid culture plates. Grow infection pre-culture overnight in a 28°C incubator-shaker with agitation at 180 rpm.
5. Inoculate 5 ml YepsLight per 15-ml Falcon tube with pre-culture at a 1:3000 ratio to form infection main cultures. Grow overnight (for 15 hr) at 28°C in a rotator at 20 rpm.
6. Measure optical density at 600 nm in a photometer/plate reader for each individual mutant strain infection main culture. Adjust the amount of culture to achieve an optical density between 0.6 and 1 for each strain and pool equal volumes of cultures of all mutant strains in a 500-ml centrifuge tube.
7. Centrifuge 10 min at 2000 × *g* and discard supernatant by decanting, ensuring removal of all supernatant. Resuspend pellet in double-distilled water to an optical density at 600 nm of 1 by pipetting up and down.

Infection of U. maydis insertional mutant library in maize seedlings

8. Using a 1-ml syringe and a 0.45-mm-diameter needle, inject ∼250 µl pooled infection culture into 7-day-old EGB maize seedlings (see step 2) that display three juvenile leaves. Make sure to infect maize seedlings in the center of the leaf whirl by piercing the stem halfway. Infect a total of >100 maize plants with the pool.
9. For the input control, centrifuge 10 ml pooled infection culture in a 15-ml Falcon tube for 1 min at 10,000 × *g*. Discard supernatant and store pellet at −70°C until isolation of gDNA from the input sample (see Basic Protocol 2).

Harvest of infected maize tissue

10. Grow infected maize seedlings from step 8 for another 7 days in a phytochamber with a 14-hr/10-hr light/dark cycle at 28°C/20°C with a total light intensity of 183.21 µmol m^{-2} s^{-1}.

11. Harvest infected second and third maize leaves at a 1-cm distance from the infection site with scissors, making sure to restrict the harvested material to infected tissue. Wash harvested tissue in 0.05% Tween-20 in double-distilled water in a 1-L beaker by stirring on a magnetic stirrer with a magnetic stir bar at 200 rpm for 5 min. Subsequently, wash leaves twice in double-distilled water. Air-dry wet leaves at room temperature before cryopreservation (see steps 12 to 14).

    The wash steps facilitate removal of any remaining dead insertional mutants located on the leaf epidermis.

12. Crush dry infected maize tissue in liquid nitrogen with a mortar and pestle.

    From now on, cool down all consumables and equipment with liquid nitrogen and avoid thawing of the maize tissue to ensure gDNA integrity.

13. Using a metal spatula, transfer crushed material into a container compatible with the Laboratory Mixer Mill MM 200 and add three 8-mm-diameter metal balls. Mill samples at 25 Hz for 90 sec. Cool mill container in liquid nitrogen for 60 sec and repeat milling step.

14. Using a metal spatula, transfer milled maize powder into a 50-ml Falcon tube and store at −70°C until gDNA extraction from the output sample (see Basic Protocol 2).

gDNA EXTRACTION FROM MUTANT POOL BEFORE AND AFTER INFECTION OF MAIZE

BASIC PROTOCOL 2

The starting material for the iPool-Seq Illumina library preparation is gDNA (Basic Protocol 1). Firstly, gDNA from the input library is required to analyze the composition of the initial insertional mutants. Secondly, the gDNA of the infected material is required to obtain insights about the insertional mutant pool composition after infection. *U. maydis* gDNA extraction is based on a protocol established in yeast that was adapted for iPool-Seq (Hoffman & Winston, 1987).

Materials

Glass beads (450 to 600 µM, B. Braun)
Input pellet (10-ml pellet of pooled mutants before infection, stored at −70°C; see Basic Protocol 1, step 9)
TE-equilibrated phenol/chloroform/isoamyl alcohol [25:24:1 (v/v/v), pH 7.5 to 8, Carl Roth; store at 4°C in the dark]
Ustilago lysis buffer (see recipe)
80% (v/v) and 100% ethanol (p.a.)
1× TE containing 20 µg/ml RNase A (Thermo Fisher Scientific, EN0531; store at −20°C)
Homogenized infected maize tissue (output, stored at −70°C; see Basic Protocol 1, step 14)
Isopropanol (p.a.)

VXR basic Vibrax (IKA) or equivalent
Refrigerated tabletop centrifuge, 4°C
ThermoMixer C (Eppendorf)
Scale

7-ml Precellys tube (Bertin)
Precellys Evolution bead mill (Bertin)
5-ml Eppendorf tubes

NOTE: Conduct all steps on ice and pre-cool consumables and equipment to 4°C.

CAUTION: Phenol/chloroform is highly toxic and volatile. Take protective measures when working with phenol/chloroform, including working in a laminar flow hood. Handle phenol-contaminated waste appropriately.

CAUTION: Store the ethanol and isopropanol at 5°C to 30°C in a safety cabinet for flammable liquids.

gDNA extraction from input

1. Add ~200 µl glass beads to the input pellet (see Basic Protocol 1, step 9). Add 500 µl TE-equilibrated phenol/chloroform/isoamyl alcohol and 400 µl Ustilago lysis buffer.
2. Vortex mixture at 1500 rpm for 15 min at room temperature in a VXR basic Vibrax or equivalent device.
3. Centrifuge 30 min at 13,000 × *g*, 4°C.
4. In the meantime, prepare a 1.5-ml Eppendorf tube containing 1 ml of 100% ethanol.
5. Add 400 µl of upper, aqueous layer from step 3 to the tube from step 4 and mix vigorously by vortexing for 30 sec.
6. Incubate mixture for >1 hr at −20°C to improve gDNA precipitation.

   Stopping point: Store the mixture overnight at −20°C.
7. Centrifuge 5 min at 13,000 × *g*, 4°C.
8. Wash gDNA pellet with 1 ml of 80% ethanol. Invert tube several times and centrifuge 5 min at 13,000 × *g*, 4°C.
9. Remove supernatant carefully by pipetting, briefly spin down tube, and remove residual supernatant.
10. Add 30 µl of 1× TE containing 20 µg/ml RNase A.
11. Incubate for 15 min at 55°C on a ThermoMixer C with agitation at 800 rpm and with an open lid.

    This step allows for evaporation of residual ethanol.
12. Store extracted input gDNA at −20°C until library preparation (see Basic Protocol 3).

gDNA extraction from output

13. Weigh 1 g homogenized infected maize tissue (output; see Basic Protocol 1, step 14) and transfer into a 7-ml Precellys tube.
14. Add ~1000 µl glass beads to output. Add 2.5 ml TE-equilibrated phenol/chloroform/isoamyl alcohol and 2 ml Ustilago lysis buffer.
15. Vortex mixture at 5000 rpm for 30 sec at room temperature in a Precellys Evolution bead mill. Transfer mixture into 5-ml Eppendorf tubes.
16. Centrifuge 30 min at 13,000 × *g*, 4°C.

17. In the meantime, prepare a 5-ml Eppendorf tube containing 2.2 ml isopropanol.
18. Add 1.5 ml of upper, aqueous layer from step 16 to the tube from step 17 and mix vigorously by vortexing for 30 sec.
19. Incubate mixture for >1 hr at −20°C to improve gDNA precipitation.

    Stopping point: Store the mixture overnight at −20°C.
20. Centrifuge 5 min at 13,000 × *g*, 4°C.
21. Wash gDNA pellet with 1 ml of 80% ethanol. Invert tube several times and centrifuge 5 min at 13,000 × *g*, 4°C.
22. Remove supernatant carefully, briefly spin down tube, and remove residual supernatant.
23. Add 150 µl of 1× TE containing 20 µg/ml RNase A.
24. Incubate for 15 min at 55°C on a ThermoMixer C with agitation at 800 rpm and with an open lid.

    As in step 11, this step allows for evaporation of residual ethanol.
25. Store extracted output gDNA at −20°C until library preparation (see Basic Protocol 3).

ILLUMINA SEQUENCING LIBRARY PREPARATION USING gDNA FROM INSERTIONAL MUTANT POOLS

BASIC PROTOCOL 3

The purified gDNA from the insertional mutant library input and output (Basic Protocol 2) is further processed via an iPool-Seq library preparation protocol to obtain Illumina sequencing–compatible libraries. The protocol is optimized for specific enrichment of insertion mutant flanks and high double-stranded DNA (dsDNA) yields, enabling library preparation directly from infected host tissue.

NOTE: Check the dsDNA concentration after each step to ensure successful preparation. We recommend quantification via a fluorescence assay, e.g., PicoGreen.

Materials

Oligonucleotides: 100 µM Adapters P1 and P2 and 5 µM PCR1-F, PCR1-R, PCR2-F, and PCR2-R (see Table 1)
Reassociation buffer (see recipe)
Purified Tn5 transposase [1 µg/µl, prepared according to prior protocol (Picelli et al., 2014); store at −70°C in aliquots for ≤12 months]
100% glycerol
Input and output gDNA (see Basic Protocol 2, steps 12 and 25, respectively)
5× TAPS buffer (see recipe)
Nuclease-free water
SPRI magnetic beads (e.g., Agencourt AMPure XP beads, Beckman Coulter; store at 4°C)
Phusion High-Fidelity DNA Polymerase, with 5× Phusion HF Buffer (New England Biolabs, M0530; store at −20°C)
10 mM dNTPs (New England Biolabs, N04472; store at −20°C)
Dynabeads M-270 Streptavidin
1× and 2× B&W buffers (see recipe)

PCR tubes
Thermocycler

Gel electrophoresis chamber or fragment analyzer
1.5-ml DNA LoBind tubes (Eppendorf)
Magnetic stand
Rotator

Additional reagents and equipment for gel electrophoresis (see Current Protocols article; Gallagher, 2012)

Tn5 fragmentation of input and output gDNA

1. Combine the following in a PCR tube to 100 µl total volume and mix well by pipetting up and down:

   - 25 µl of 100 µM Adapter P1
   - 25 µl of 100 µM Adapter P2
   - 50 µl reassociation buffer.

   Perform primer annealing in a thermocycler starting from 90°C, with a 1°C decrement per minute.

2. Combine the following in a PCR tube to 100 µl total volume and mix well by pipetting up and down:

   - 25 µl purified Tn5 transposase
   - 25 µl annealed adapters (see step 1)
   - 50 µl of 100% glycerol.

3. Incubate for 30 min at 37°C in a thermocycler.

4. Combine the following in separate PCR tubes and mix well by pipetting up and down:

   - 250 ng input or output gDNA
   - Tn5 transposase loaded with adapters (see step 3) to a final concentration of 150 ng/µl

Table 1 Oligonucleotides Used for Adapters, Specific PCR1 and PCR2, and Illumina Sequencing in Basic Protocol 3

Oligonucleotide	Sequence[a]
Adapter P1[b]	5′-CACGACGCTCTTCCGATCTNNNNNNNNNNNNAGATGTGTATAAGAGACAG-3′
Adapter P2	5′-[phos]CTGTCTCTTATACACATC[3InvdT]-3′
PCR1-R[c,d]	5′-[BioTEG]CCAGATGTCCTGTGGTATCCTGTG-3′
PCR1-F	5′-GAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC-3′
PCR2-F	5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACAC-3′
PCR2-R[d]	5′-CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCTGTGGTATCCTGTGGCG-3′
Illumina Rd1[e]	5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′
Illumina Rd2[e]	5′-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3′

[a]Depending on the library of insertional mutants that is screened, these sequences may potentially need to be adjusted.
[b]The 12 Ns in Adapter P1 constitute the UMI and should be random.
[c]The 6 Ns in PCR2-R constitute the (single-index) library multiplexing barcode.
[d]With PCR1-R and PCR2-R as listed, enrichment of insertion mutant flanks is specific for the sequence 5′-CCAGATGTCCTGTGGTATCCTGTGGCG-3′, and read2 will start with the underlined part of PCR2-R.
[e]Illumina Rd1 and Rd2 are the standard Illumina TruSeq sequencing primers.

- 4 µl of 5× TAPS buffer and
- Nuclease-free water to 20 µl.

CAUTION: *TAPS buffer contains dimethylformamide (DMF), which is toxic and volatile. Take appropriate safety measures and work in a laminar flow hood.*

Prepare separate reactions for the input and output and use 250 ng per PCR reaction. For the Tn5 fragmentation, we recommend total amounts of 500 ng for the input gDNA and 10 × 1 µg for the output gDNA for a first trial. Upscaling is recommended if final read coverage of insertional mutants and unique molecular identifier (UMI) diversity is low.

5. Incubate reaction for 10 min at 55°C in a thermocycler.
6. Clean fragmentations with SPRI magnetic beads, e.g., Agencourt AMPure XP beads, according to the manufacturer's protocol.

We recommend gDNA purification with SPRI beads to minimize DNA loss. A ratio of 1.5× SPRI beads to DNA yields good results. Size selection is possible but not recommended. Column-based purification systems can be used instead of SPRI magnetic beads but will most likely provide lower DNA recovery yields and thus less diverse sequencing libraries.

7. Ensure fragmentation success via gel electrophoresis or using a fragment analyzer.
8. Determine dsDNA concentration for subsequent PCR.

We recommend quantification via a fluorescence assay, e.g., PicoGreen, and a minimal final concentration of 20 ng/µl.

Stopping point: Store the fragmented DNA at −20°C until specific PCR.

Specific PCR of mutant cassette genome junctions

9. Combine the following in a PCR tube and mix well by pipetting up and down.

- 5 µl of 5× Phusion HF Buffer
- 1 µl of 10 mM dNTPs
- 1 µl of 5 µM PCR1-F
- 1 µl of 5 µM PCR1-R
- 0.5 µl Phusion High-Fidelity DNA Polymerase
- 250 ng fragmented and cleaned gDNA (see step 6)
- Nuclease-free water to 25 µl.

10. Run the following PCR program in a thermocycler:

Initial step:	1 min	95°C	(initial denaturation)
15 cycles:	15 sec	95°C	(denaturation)
	15 sec	65°C	(annealing)
	30 sec	72°C	(elongation)
Final step:	1 min	72°C	(final elongation).

11. Clean PCR reactions with SPRI magnetic beads according to the manufacturer's protocol.
12. Pool clean eluates. Optional: Check dsDNA concentration prior to streptavidin enrichment (see steps 13 to 19).

Stopping point: Store the cleaned specific PCR fragments at −20°C until streptavidin enrichment.

Streptavidin enrichment

13. For each input and output sample, wash 30 µl Dynabeads M-270 Streptavidin in 200 µl of 1× B&W buffer in a 1.5-ml DNA LoBind tube on a magnetic stand.

14. Repeat washing step three additional times.
15. Resuspend beads in 2× B&W buffer, matching the volume of the clean input and output PCR1 eluates.
16. Pool eluates and beads and allow for enrichment of biotinylated PCR amplicons by rotation at room temperature for 15 min.
17. Place tubes in a magnetic stand for 1 min and discard supernatant.
18. Wash beads with 200 µl of 1× B&W buffer while the tubes remain in the magnetic stand.
19. Repeat washing step three additional times. Resuspend beads in 34 µl nuclease-free water and proceed with nested PCR (see steps 20 and 21).

Stopping point: Store the enriched PCR fragments at −20°C until nested PCR.

Nested PCR of enriched fragments

20. Combine the following in PCR tubes to 50 µl total volume and mix well by pipetting up and down:

- 34 µl nuclease-free water containing beads (see step 19)
- 10 µl of 5× Phusion HF Buffer
- 1 µl of 10 mM dNTPs
- 2 µl of 5 µM PCR2-F
- 2 µl of 5 µM PCR2-R
- 1 µl Phusion High-Fidelity DNA Polymerase.

21. Run the following PCR program in a thermocycler:

Initial step:	1 min	95°C	(initial denaturation)
15 cycles:	15 sec	95°C	(denaturation)
	15 sec	65°C	(annealing)
	30 sec	72°C	(elongation)
Final step:	1 min	72°C	(final elongation).

22. Place tubes in a magnetic stand for 1 min and transfer supernatant containing PCR2 amplicons into fresh PCR tubes.
23. Clean PCR reactions with SPRI magnetic beads according to the manufacturer's protocol.
24. Check dsDNA concentration and proceed with Illumina sequencing (see Basic Protocol 4).

We recommend a minimal final dsDNA concentration of 0.1 ng/µl to enable Illumina sequencing. Moreover, we suggest quality control of iPool-Seq library preparation via qPCR and fragment length analysis prior to Illumina sequencing. We sequence the iPool-Seq libraries on a MiSeq Illumina platform with 75PE conditions.

IMPORTANT NOTE: *The standard Illumina Nextera sequencing primers are not compatible with our Tn5 adapter and will interfere with sequencing if present. Instead, Illumina TruSeq sequencing primers (Table 1, Illumina Rd1 and Rd2) must be used.*

BASIC PROTOCOL 4

BIOINFORMATIC ANALYSIS

This protocol describes how insertion pool data are analyzed, using the pipeline that we developed (available at *http://www.cibiv.at/software/ipoolseq-pipeline*), to find insertional mutants with significantly increased or decreased virulence. Virulence is measured as the abundance of a deletional mutant in the post-infection output pool relative to the

pre-infection input and is compared to the virulence of a reference set of known neutral mutants to find significant deviations from neutral behavior.

NOTE: In the following, commands intended to be entered on a Unix-style terminal, either directly on a Linux machine or via SSH, are printed in a `monospaced font`. Outside of such commands, filenames are printed `monospaced and underlined.` Placeholders for names of experiments or libraries that have to be supplied by the user are printed in *italic*.

NOTE: iPool-Seq pipeline is based on the workflow engine "snakemake" (Köster & Rahmann, 2012). It is thus not strictly necessary to proceed step by step; in particular, jumping to step 13 directly after adding all required data in step 6 will cause all intermediate steps to be executed automatically. However, given that it is good practice to check the results of key intermediate steps (like mapping and KO assignment) for validity before proceeding, we recommend following the steps outlined below and also performing the checks and validations suggested in the Troubleshooting section for each individual step.

Materials

Workstation running Linux or Windows 10 with Windows Subsystem for Linux (WSL), with 64-bit CPU and 8 GB or more of RAM and with free disk space ≥5 times size of raw sequencing data
iPool-Seq analysis pipeline (*http://www.cibiv.at/software/ipoolseq-pipeline*)
Reference genome of *U. maydis* in FASTA format
FASTA file containing sequences at 5′ end (named "5p") and 3′ end (named "3p") of knockout (KO) cassette
List of deletional mutants of *U. maydis* as GFF2 file listing KO cassette insertion positions
Single BAM file containing unmapped sequencing reads or two separate compressed FASTQ files (one for read1 and one for read2) for each sequenced library (prepared according to Basic Protocol 3)
Web browser
PDF viewer

NOTE: For the FASTA file describing the KO cassette, the sequences must reflect the part of read2 that overlaps with the cassette, i.e., start with the underlined part of PCR2-R (Table 1) and extend up to the end of the cassette. See the (included) list of deletional mutants used by Uhse et al. (2018) `cfg/Uhse_et_al.2018/cassette.fa` for an example.

NOTE: For the list of deletional mutants, each entry must carry ≥2 two tags: a unique "Name" and a flag "Neutral" with value 0 or 1 that decides whether a particular deletion strain is included in the reference set of assumed neutral deletions. See the (included) list of deletional mutants used by Uhse et al. (2018) `cfg/Uhse_et_al.2018/knockouts.gff` for an example.

Installing the iPool-Seq pipeline

1. On a workstation running Linux or Windows 10 with WSL, download and install iPool-Seq analysis pipeline by executing

```
VER=latest-release
URL=http://github.com/Cibiv/ipoolseq-pipeline
curl -L -O $URL/archive/$VER.tar.gz
tar xzf $VER.tar.gz
cd ipoolseq-pipeline-$VER
```

You can also use a web browser to download the latest release from http://github.com/Cibiv/ipoolseq-pipeline/releases. Then, change into the directory containing the downloaded file in a terminal window and continue with the tar command to unpack the archive.

2. Install and activate our Bioconda (*http://bioconda.github.io*; Grüning et al., 2018) environment, containing all software packages required by pipeline, with

```
./install-environment.sh
source ./environment/bin/activate
```

Should environment.tar.gz fail to download, download that file from http://github.com/Cibiv/ipoolseq-pipeline manually and place it in the folder containing the pipeline.

IMPORTANT NOTE: *The environment must be re-activated (but not re-created) whenever you open a new terminal window.*

3. Optional: Test for successful installation of library by re-analyzing one of the replicates of Uhse et al. (2018). To download the raw sequencing data for replicate A1 (data/Uhse_et_al.2018/expA.r1-in.bam and expA.r1-out.bam) and to run the differential virulence analysis, do

```
snakemake data/Uhse_et_al.2018/expA.r1.dv.tab
```

Afterward, data/Uhse_et_al.2018/expA.r1.dv.tab contains a table listing the differential virulence analysis results for all deletional mutants (see Table 3 for a list of columns), and an accompanying HTML report is written to data/Uhse_et_al.2018/expA.r1.dv.html. Note that although the pipeline discussed here uses the same approach as the pipeline used by Uhse et al., it differs in some details and does not cover combining data from multiple replicates. The re-analysis thus cannot be expected to reproduce the published results exactly.

Adding the reference genome, cassette file, KO list, and libraries

4. Pick a name (e.g., *your_design*) for the experimental design (including the reference genome and the list of KO cassette insertions) and create two folders with

```
mkdir -p data/your_design
mkdir -p cfg/your_design
```

5. Copy the reference genome of *U. maydis* in FASTA format to cfg/*your_design*/reference.fa, the FASTA file containing the sequences at 5′ end (named "5p") and 3′ end (named "3p") of the KO cassette to cfg/*your_design*/cassette.fa, and the list of deletional mutants of *U. maydis* as a GFF2 file listing KO cassette insertion positions to cfg/*your_design*/knockouts.gff.

6. For a single replicate (named *your_replicate* here, but it can have an arbitrary name), which always consists of two libraries, name pre-infection input pool *your_replicate*-in and post-infection output pool *your_replicate*-out. For each of those two libraries (in the following, referred to as *your_lib*, which thus stands for either *your_replicate*-in or *your_replicate*-out), place raw paired-end sequencing data either into data/*your_design*/*your_lib*.bam as a single BAM file containing unmapped sequencing reads or into data/*your_design*/*your_lib*.1.fq.gz (read1) and data/*your_design*/*your_lib*.2.fq.gz (read2) as two separate compressed FASTQ files.

This naming scheme ensures that the pipeline knows which reference genome and KO list belong to a particular library and (in step 13) which input and output libraries constitute one replicate.

Table 2 Columns in the Knockout Abundance Table Generated by TRUmiCount in Basic Protocol 4, step 10

Column name	Type	Description
gene	String	Combination of knockout name and flank, "*knockout*:*flank*"
n.umis	Integer	Observed UMI count (after filtering)
n.tot	Numeric	Est. number of total UMI count, *n.umis/(1-loss)*
efficiency	Numeric	Estimated PCR efficiency
depth	Numeric	Avg. number of reads per UMI (including lost UMIs)
loss	Numeric	Est. fraction of lost (unobserved) UMIs
n.obs	Integer	Same as for *n.umis*
n.reads	Integer	Observed total read count (after filtering)
n.umis.prefilter	Integer	Observed UMI count before read-count filter
n.reads.prefilter	Integer	Observed total read count before read-count filter

Steps 7 to 12 must be executed for each library, i.e., twice per replicate, replacing your_lib first with your_replicate-in and then with your_replicate-out.

Trimming and UMI extraction (per library)

7. To trim the technical sequences (Fig. 2A) from the raw sequencing reads for library *your_lib*, do

```
snakemake data/your_design/your_lib.trim.1.fq.gz
```

Executing the trimming step for either read1 (in the command above) or read2 automatically trims the other read as well. If you started from an unmapped BAM file, the file is also automatically converted into two FASTQ files, with one for each read, during this step.

8. Optional (but recommended): To verify the trimming, generate FastQC (Andrews, 2010) reports for trimmed first and second reads with

```
snakemake data/your_design/your_lib.fastqc.1.html
snakemake data/your_design/your_lib.fastqc.2.html
```

Mapping and assignment to insertional KOs (per library)

9. To map and assign trimmed reads to KOs (Fig. 2B) for the sequencing library *your_lib*, do

```
snakemake data/your_design/your_lib.assign.bam
```

Determining KO abundances (per library)

10. To produce a table of genome abundance estimates for sequencing library *your_lib*, do

```
snakemake data/your_design/your_lib.count.tab
```

For a description of the columns of `data/your_design/your_lib.count.tab`, see Table 2. During this step, the TRUmiCount algorithm (Pflug & von Haeseler, 2018) also produces the accompanying PDF report `data/your_design/your_lib.count.pdf`.

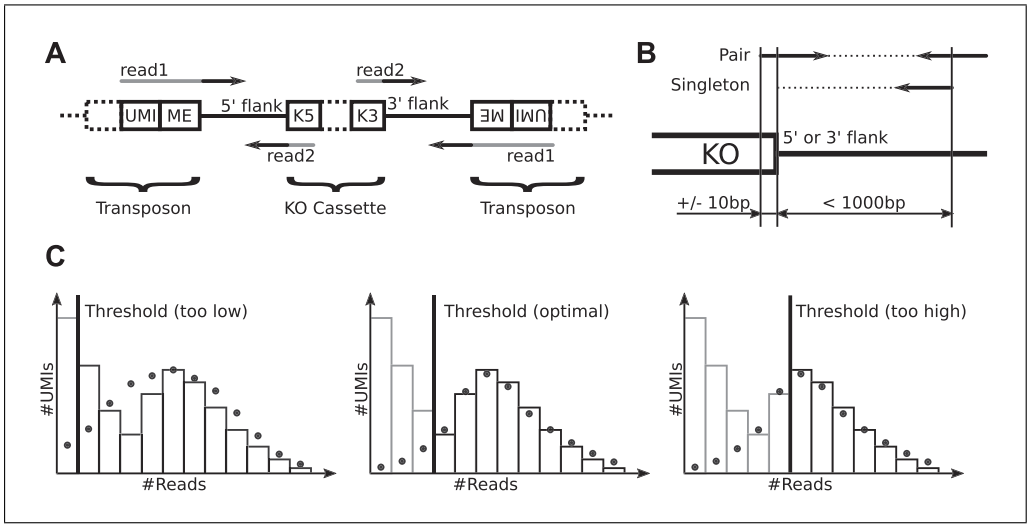

Figure 2 (**A**) Layout of sequenced fragments on both sides of knockout (KO) cassette insertions. The gray parts of the reads are non-genomic and are trimmed before mapping. The unique molecular identifier (UMI) consists of 12 random bases, ME=5′-AGATGTGTATAAGAGACAG-3′. (**B**) Required mapping locations and directions for read pairs with either both mates or one mate mapped to be assigned to a specific KO. (**C**) Data indicating that the chosen TRUmiCount threshold is too low (left), optimal (middle), or too high (right).

Adjusting the TRUmiCount phantom rejection threshold (per library)

11. To see if adjustment of TRUmiCount's read-count threshold is necessary, check read-count distribution plot in the TRUmiCount report `data/your_design/your_lib.count.pdf`.

12. Optional: If there is a clear overabundance of observed vs. predicted UMIs for read counts slightly larger than the threshold, increase threshold. If the predicted and observed numbers of UMIs agree well for read counts below the threshold, decrease threshold (Fig. 2C). To set a library-specific threshold *T* for library *your_lib*, add the following lines to the "trumicount" block in `cfg/config.yaml` (be sure to match the indentation of the existing lines in that block):

```
- file: 'data/your_design/your_lib.*'
opts: '--threshold T'
```

IMPORTANT NOTE: *After changing the setting, remove `data/your_design/your_lib.count.tab` and re-run step 10.*

Finding differentially virulent KOs (per replicate)

13. To compute the log fold changes of KO virulence for *your_replicate* and p-values for how significantly these log fold changes deviate from zero (i.e., no change compared to the neutral reference set), do

```
snakemake data/your_design/your_replicate.dv.tab
```

The differential virulence analysis is based on the KO abundances and loss correction factors from the tables `your_replicate-in.count.tab` and `your_replicate-out.count.tab`, both created in the folder `data/your_design` during step 10. Step 13 also produces an accompanying HTML report in `data/your_design/your_replicate.dv.html`.

Downstream analysis

14. To produce plots and to combine data from multiple replicates, load output table `data/your_design/your_replicate.dv.tab` from step 13 into a

Table 3 Columns in the Differential Virulence Table Generated in Basic Protocol 4, step 13

Column name	Type	Description
knockout	String	Name of the knockout as in the knockout list GFF file
is.neutral	Flag	1 if the knockout is part of the reference set, 0 otherwise
n.out	Integer	Sum of 5′ and 3′ UMI counts after filtering (output pool)
loss.out	Numeric	Average of 5′ and 3′ est. fraction of lost UMIs (output pool)
abundance.out	Numeric	Est. number of genomes, 0.5**n.out/(1-loss.out)* (output pool)
n.in	Integer	Sum of 5′ and 3′ UMI counts after filtering (input pool)
loss.in	Numeric	Average of 5′ and 3′ est. fraction of lost UMIs (input pool)
abundance.in	Numeric	Est. number of genomes, 0.5**n.in/(1-loss.in)* (input pool)
log2fc	Numeric	Virulence log fold change compared to the reference set ($\log_2 \Delta$v)
low.pval	Numeric	p-value for *log2fc* being significantly low
high.pval	Numeric	p-value for *log2fc* being significantly high
low.qval	Numeric	FDR-corrected p-value for log2fc being significantly low
high.qval	Numeric	FDR-corrected p-value for log2fc being significantly high

table-oriented tool such as Microsoft Excel, GraphPad Prism, R, or SPSS. Then, filter KOs based on fold change, p-value, q-value, number of genomes, etc. Combine data from multiple replicates either by filtering based on criteria from multiple replicates (e.g., significance) or by computing a combined p-value (e.g., with Fisher's method).

See Table 3 for a description of the columns of `data/your_design/your_replicate.dv.tab`, *which contains the results of the differential virulence analysis step.*

REAGENTS AND SOLUTIONS

B&W buffer, 1×

5 mM Tris, pH 7.5
1 M NaCl
0.5 mM EDTA
Store ≤6 months at room temperature

B&W buffer, 2×

10 mM Tris, pH 7.5
2 M NaCl
1 mM EDTA
Store ≤6 months at room temperature

Potato dextrose–agar plates containing 200 µg/ml hygromycin

2.4% (w/v) potato dextrose broth (Difco)
2.0% (w/v) agar (Difco)
Add sterile deionized water and autoclave at 121°C
Cool to 50°C, add 50 mg/ml hygromycin (Roche) to 200 µg/ml, and mix by stirring
Pour 20 ml per 9-cm petri dish
Store ≤2 weeks at 4°C in the dark

Reassociation buffer

10 mM Tris, pH 8.0
50 mM NaCl

1 mM EDTA
Store ≤6 months at room temperature

TAPS buffer, 5×

250 mM TAPS-NaOH
125 mM $MgCl_2$
50% (v/v) DMF
Adjust to pH 8.5 at 25°C
Store ≤4 weeks at room temperature in the dark in a cabinet for flammable liquids
CAUTION: *DMF is toxic; work in a laminar flow hood with appropriate protective measures.*

Ustilago lysis buffer

10 mM Tris, pH 8.0
100 mM NaCl
1 mM EDTA
2% (v/v) Triton X-100
1% (w/v) SDS
Store ≤6 months at room temperature

YepsLight liquid medium

1.0% (w/v) yeast extract (Difco)
0.4% (w/v) Bacto Peptone (Difco)
0.4% (w/v) sucrose
Add sterile deionized water and autoclave at 121°C
Store ≤4 weeks at room temperature

COMMENTARY

Background Information

The analysis of insertional mutant libraries is well established for bacteria (Gawronski et al., 2009; Goodman et al., 2009; Langridge et al., 2009; van Opijnen et al., 2009) but has not been applied extensively to eukaryotic microorganisms. Genome-wide insertional mutant libraries were generated successfully for baker's yeast (*Saccharomyces cerevisiae*) by homologous recombination (Winzeler et al., 1999) and, more recently, by transposition (Michel et al., 2017) and for the rice pathogenic fungus *Magnaporthe oryzae* by the kinase ATM (Jeon et al., 2007). However, tools that allow for efficient high-throughput analysis of a negative depletion screen in the context of a host, for instance in the case of *M. oryzae* and rice (Jeon et al., 2007), were not available until recently. Therefore, we developed iPool-Seq, which, due to its high selectivity and sensitivity, allows for analysis of pooled infections of insertional mutants directly from the infected host tissue (Uhse et al., 2018). iPool-Seq provides a powerful tool for scientists who want to analyze mutant pool composition after colonization of the host, without any biases that could arise due to separation of the output fraction.

Outlook

We previously demonstrated (Uhse et al., 2018) that iPool-Seq offers an elegant possibility to analyze a *U. maydis* insertional mutant pool after colonization of its host maize. We further suggest that iPool-Seq could be applied to genome-wide insertional mutant pools generated by high-throughput techniques, e.g., transposon-mediated mutagenesis. Due to its high selectivity and sensitivity, iPool-Seq enables analysis of large insertional mutant pools directly from the colonized host tissue and obviates the need for separation of the host tissue and colonizing microbes. We propose that iPool-Seq not only is suitable for analysis of mutant pools of microorganisms in the context of a plant host but also may be applied to animal-microbe or microbe-microbe interaction systems.

Trimming and UMI extraction

During this step, all non-genomic sequences (with UMI and KO cassette overlap) are removed from the raw sequencing read pairs, as produced by (paired-end) sequencing (Fig. 2A), so as to not interfere with the mapping process. The UMIs are instead stored as

part of the pairs' read names to ensure that the UMIs are passed alongside the reads through the following processing steps. Reads that do not overlap with the KO cassette or that do not contain a UMI are removed.

Mapping and assignment to insertional KOs

During this step, the trimmed reads are mapped to the reference genome using NextGenMap (Sedlazeck, Rescheneder, & von Haeseler, 2013) and assigned to the individual insertional KOs (Fig. 2B). In short, proper read pairs (pairs with both mates mapped and correctly oriented) are assigned to a specific flank (5′ or 3′) of a KO cassette insertion (i.e., a KO strain) if one read starts close (±10 bp) to the respective end of the cassette and extends away from the cassette. Singleton reads (reads whose mate could not be mapped) must map ≤1000 bp away from the respective end of the cassette and extend toward the cassette. Reads that cannot be assigned unambiguously are not assigned at all.

Determining KO abundances (counting genomes)

This step consists of error correction, phantom removal, and loss estimation for the UMIs detected for a particular combination of flank and KO strain. The UMIs are first corrected for sequencing errors with UMI-Tools (Smith, Heger, & Sudbery, 2017), which merges similar UMIs found within the same flank of a KO cassette insertion. The merged UMIs are then processed further with TRUmiCount (Pflug & von Haeseler, 2018), which filters based on per-UMI read counts to remove additional *phantom UMIs* (mostly amplification artifacts) and then, for each flank of each KO cassette insertion, estimates and corrects for the percentage of lost (i.e., unobserved) true UMIs. This ensures that the estimated KO abundances are unaffected by PCR amplification bias. The output comprises, per combination of KO and flank, the filtered UMI count (number of observed genomes), the estimated loss, and the loss-corrected genome count (Table 2).

Critical Parameters

Generating insertion libraries suitable for iPool-Seq

There are different techniques available for generation of insertion libraries. We based the generation of the *U. maydis* insertional mutant library on homologous recombination. Insertional mutagenesis via homologous recombination has the advantage of not being prone to multiple insertions per individual, which is more likely to happen with untargeted, random insertional mutagenesis approaches, like Agrobacterium-mediated transformation (Michielse, Hooykaas, van den Hondel, & Ram, 2005). However, homologous recombination is more laborious than high-throughput methods. Moreover, the choice of the fungal model system is critical, and advantages and drawbacks of systems should be evaluated prior to insertional mutant library generation. Here, we use *U. maydis*, a fungal model that is genetically accessible to homologous recombination and well suited for the application of iPool-Seq in the context of a host infection due to the availability of a solo-pathogenic strain (Bölker, Genin, Lehmler, & Kahmann, 1995).

For the final analysis (Basic Protocol 4), an internal reference set of mutants with unaffected virulence is essential to find mutants that are comparatively depleted or enriched mutants. The reference set can be defined either by known unaffected mutants that have been published before or by individual mutants identified via infection tests.

Mutant pool infection

The iPool-Seq protocol provides a high specificity for the inserted sequences and thus can be applied directly to infected host material (Basic Protocol 1). However, the number of pathogens that infect an individual host can constitute a bottleneck. We suggest overcoming this bottleneck by increasing the number of infected host plants or by reducing the complexity of the insertion mutant library.

Sequencing

We recommend sequencing the finished library (Basic Protocol 3) on an Illumina MiSeq platform and aiming for 2 to 3 million reads per library. However, it is also possible to use a different platform and to sequence more deeply to improve individual mutant UMI counts.

For complex libraries, it is possible to increase the amount of gDNA for the infected output sample and to proportionally increase the sequencing depth, which will likely yield a higher coverage of individual mutants. For a given mutant library and amount of extracted gDNA, the average number of reads per UMI (found in the TRUmiCount report) is an indicator of whether increasing the sequencing depth would be beneficial. For libraries with <1 read per UMI on average, deeper sequencing can be expected to improve the accuracy of abundance measurements; after that, the benefit drops gradually, and more than ∼10

reads per UMI will not provide any additional benefit.

Troubleshooting

Trimming and UMI extraction

The optional FastQC (Andrews, 2010) reports created for the first and second reads after trimming offer a first quality check of library preparation and sequencing (Basic Protocol 4). Aspects to check for are as follows: (a) under "Basic Statistics," that most (>90%) reads survived the trimming step; (b) under "Per base sequence quality" and "Per base N content," that the sequenced bases are high quality and do not contain many Ns; and (c) under "Adapter Content," that trimming indeed removed all sequencing adapters from the reads. If many reads are lost during the trimming step, they either were contaminants or did not match the sequence pattern that the library preparation should produce. In this case, we recommend blasting a few random reads to check for contamination and to manually compare their sequence composition to the expected pattern (Fig. 2A) and to the sequences in `data/your_design/cassette.fa`. Should most reads survive but show either a strong drop-off of base qualities toward the end or many Ns, it may be necessary to include an additional quality-based trimming step in the Trimmomatic (Bolger, Lohse, & Usadel, 2014) command in `cfg/config.yaml` (see the Trimmomatic manual for details). If there are still adapter sequences detected in the trimmed reads, add any custom adapter sequences that differ from the adapters mentioned in Basic Protocol 3 to `cfg/Uhse_et_al.2018.adapters.fa` (again, see the Trimmomatic manual for details).

Mapping and assignment to insertional KOs

We recommend checking the results of the mapping and KO assignment process visually for a few libraries and KO cassette insertions in a genomic viewer like the Integrated Genomics Viewer (IGV; Robinson et al., 2011). You should find most of the reads mapped to the two flanks (3′ and 5′) of KO cassette insertions, carrying the name and flank of the insertional KO in the form "*name*:*flank*" in the XT tag, and not extending more than a few base pairs into the regions replaced by the KO cassette. You may find some spurious reads mapped to arbitrary locations in the genome; these will later be ignored by the pipeline and thus are not a cause for concern (unless overly abundant).

If reads are mapped correctly but not assigned to the correct KO cassette insertion, check that the genomic coordinates in your GFF2 files listing the KO cassette insertions are correct. If a substantial fraction of the sequenced fragments are longer than 1000 bp or if many reads extends more than a few base pairs into the regions replaced by the KO cassette due to mapping imprecisions, adjust the "mapping fuzzyness" or "max fragment length" parameter of the "knockout_assignment" step in `cfg/config.yaml`.

Adjusting the TRUmiCount phantom rejection threshold

For optimal separation of phantom UMIs (i.e., amplification artifacts) from true UMIs by the TRUmiCount algorithm (Pflug & von Haeseler, 2018), it can be necessary to adjust the automatically chosen read-count threshold T (Fig. 2C). This is true in particular at higher sequencing depths, where *phantom UMIs* can make up a large proportion of UMIs (but not of reads, due to the phantom's lower read counts). Setting the threshold too low (Fig. 2C, left) will cause phantom UMIs to be mistaken for true UMIs, which has the potential to distort the results. It will also distort TRUmiCount's parameter estimates (the PCR efficiency in particular), resulting in a bad fit of model and data. Choosing a threshold that is too high (Fig. 2C, right), in contrast, will cause more *true UMIs* (i.e., UMIs that reflect actual genomes in the sequenced pool) to be filtered out. However, given that TRUmiCount estimates and corrects for this loss, the net effect is only a reduced precision of the measured abundances (due to the lower absolute genome counts), and not an introduction of systematic biases. When choosing a threshold value, too high is thus preferable over too low.

Understanding Results

The results contain several types of information. Firstly, the KO abundance tables (produced in Basic Protocol 4, step 10) list the number of genomes per mutant found in the input and output libraries, i.e., they contain information about the absolute abundances of the individual KOs. These tables in particular also provide information about which mutants are not detected at all in either input or output (i.e., that show zero detected genomes), which is possibly due to very slow growth or not being viable at all. Gradual changes in a mutant's

virulence are detected by comparison of the mutant's input and output abundances to a reference set of neutral mutants and summarized in the differential virulence report (produced in Basic Protocol 4, step 13).

Finding differentially virulent KOs (statistical analysis of abundances)

Given that KO abundances in the input can be spread across multiple orders of magnitude, the dominant factor that determines the abundance of a KO in the output pool is its abundance in the input pool; the effects due to different genotypes that we want to detect is typically subordinate to that. The iPool-Seq pipeline accounts for this by assuming that a KO's loss-corrected abundance A_{out} in the output pool depends (linearly) on its loss-corrected abundance A_{in} in the input pool, in addition to depending on the virulence factor Δv relative to neutral KOs ($\Delta v = 1$ for neutral KOs). To account for differences in genome capture efficiency and total genome count between the two libraries, the pipeline also includes a scaling factor λ (which is replicate-specific, but not KO-specific). The loss-corrected input and output abundances of a KO are computed from the (3′ and 5′ summed) filtered UMI counts N_{in} and N_{out}, which are corrected for unobserved UMIs by dividing by $1 - \ell_{in}$ and $1 - \ell_{out}$, where ℓ_{in} and ℓ_{out} are the (3′ and 5′ averaged) loss estimates computed by TRUmiCount. The pipeline thus computes the $\log_2$ fold change of a KO's virulence as

$$\log_2 \Delta v = \log_2 \frac{A_{out}}{\lambda \cdot A_{in}} = \log_2 \frac{N_{out} \cdot (1 - \ell_{in})}{\lambda \cdot N_{in} \cdot (1 - \ell_{out})}$$

To test for significant deviations of Δv from neutrality (i.e., $\Delta v = 1$), the pipeline assumes a negative binomial model for N_{out} given N_{in}, which accounts for the uncertainty in A_{out} and A_{in} due to sampling noise, as well as for an additional amount of dispersion d due to random fluctuations of the virulence of neutral KOs:

$$N_{out}|N_{in} \sim \text{NegBin}\left(\mu = \lambda \cdot N_{in} \cdot \frac{1 - \ell_{out}}{1 - \ell_{in}}, \; r = \frac{N_{in}}{1 + d \cdot N_{in}}\right)$$

The two parameters λ and d are estimated by fitting the model to the set of neutral KOs. Parameter λ measures the relative total size (i.e., the loss-corrected number of UMIs) of the output library relative to the input library, and d the squared coefficient of variation of output abundances due to biological noise (i.e., due to differences in the growth of neutral mutants).

Interpretation of the differential virulence report

The differential virulence report created in Basic Protocol 4, step 13, contains diagnostic statistics and plots that serve as quality checks and as verification that the assumptions of the statistical model are fulfilled to a reasonable degree.

The "Quality Control / Read and UMI count statistics" provide an overview of how many usable reads and UMIs remain after each step of the analysis pipeline. Discarding up to about one third of reads during the course of the analysis should be considered normal; if considerably fewer reads than that remain after the "TRUmiCount" step, the steps that remove the largest percentages of reads should be checked carefully for problems. For the number of UMIs, at high sequencing depths (on average ~10 reads per UMI or more), it is normal for the removed percentage to be considerably higher because of TRUmiCount filtering of UMIs with a low read count.

The precision of the KO abundances determined for the input and output pools is reflected by the correlation (found under "Quality Control / Correlation of 3′ and 5′ Flank Abundances") of the abundances computed for each flank of the KO cassette insertions. After loss correction by TRUmiCount, a correlation of ~0.9 or more can be expected.

For neutral KOs, our statistical model also assumes a linear relationship between pre-infection input abundance and post-infection output abundance, which can be verified in the input vs. output plots and correlations (found under "Quality Control/Correlation of Input and Output Abundances"). There exists no generally applicable lower bound for the expected correlation of input and output because the expected correlation depends on the percentage of non-neutral KOs among the KOs in the experiment and on how much faster or slower these KOs proliferate. More important than the correlation is the qualitative behavior seen in the input vs. output plots: the relationship should be linear across the full range of observed input abundances, without any plateau effect for highly abundant KOs. If such a plateau effect is observed, it is likely that the carrying capacity of the host plants

has been reached, and either the number of mutants that each plant is infected with should be reduced or the statistical model should be modified to account for the carrying capacity of the host plant.

Time Considerations

The iPool-Seq protocol can be finished in ∼3 weeks for the *U. maydis–Zea mays* pathosystem (Fig. 1). For other pathosystems, we recommend first determining critical parameters and generating an insertional mutant library for infection. In comparison to *U. maydis*, variations in time considerations for other pathogens will mainly depend on the infection protocol.

Acknowledgements

We thank Alexandra Stirnberg and Klaus Ehrlinger for their contribution to the generation of the *Ustilago maydis* deletion strain library. We would like to thank the whole Djamei Lab and the Center for Integrative Bioinformatics Vienna (CIBIV) for their critical advice and help regarding this manuscript. The research leading to these protocols received funding from the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013)/ERC grant agreement n° GA335691 ("Effectomics"), the Austrian Science Fund (FWF; P27429-B22, P27818-B22, I 3033-B22, W1207-B09), the "DoktoratsKolleg RNA Biology," and the Austrian Academy of Science (OEAW).

Literature Cited

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. doi: 10.1093/bioinformatics/btu170.

Bölker, M., Genin, S., Lehmler, C., & Kahmann, R. (1995). Genetic regulation of mating and dimorphism in Ustilago maydis. *Canadian Journal of Botany*, *73*(S1), 320–325. doi: 10.1139/b95-262.

Doehlemann, G., van der Linde, K., Amann, D., Schwammbach, D., Hof, A., Mohanty, A., ... Kahmann, R. (2009). Pep1, a secreted effector protein of Ustilago maydis, is required for successful invasion of plant cells. *PLoS Pathogens*, *5*(2), e1000290. doi: 10.1371/journal.ppat.1000290.

Gallagher, S. R. (2012). One-dimensional SDS gel electrophoresis of proteins. *Current Protocols in Protein Science*, *68*, 10.1.1–10.1.44. doi: 10.1002/0471140864.ps1001s68.

Gawronski, J. D., Wong, S. M. S., Giannoukos, G., Ward, D. V., & Akerley, B. J. (2009). Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(38), 16422–16427. doi: 10.1073/pnas.0906627106.

Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., ... Johnston, M. (2002). Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, *418*(6896), 387–391. doi: 10.1038/nature00935.

Goodman, A. L., McNulty, N. P., Zhao, Y., Leip, D., Mitra, R. D., Lozupone, C. A., ... Gordon, J. I. (2009). Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host & Microbe*, *6*(3), 279–289. doi: 10.1016/j.chom.2009.08.003.

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., ... Bioconda, T. (2018). Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, *15*(7), 475–476. doi: 10.1038/s41592-018-0046-7.

Hoffman, C. S., & Winston, F. (1987). A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of Escherichia coli. *Gene*, *57*(2–3), 267–272. doi: 10.1016/0378-1119(87)90131-4.

Idnurm, A., Reedy, J. L., Nussbaum, J. C., & Heitman, J. (2004). Cryptococcus neoformans virulence gene discovery through insertional mutagenesis. *Eukaryotic Cell*, *3*(2), 420–429. doi: 10.1128/EC.3.2.420-429.2004.

Jeon, J., Park, S. Y., Chi, M. H., Choi, J., Park, J., Rho, H. S., ... Lee, Y. H. (2007). Genome-wide functional analysis of pathogenicity genes in the rice blast fungus. *Nature Genetics*, *39*(4), 561–565. doi: 10.1038/ng2002.

Kamper, J., Kahmann, R., Bolker, M., Ma, L. J., Brefort, T., Saville, B. J., ... Birren, B. W. (2006). Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis. *Nature*, *444*(7115), 97–101. doi: 10.1038/nature05248.

Köster, J., & Rahmann, S. (2012). Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. doi: 10.1093/bioinformatics/bts480.

Langridge, G. C., Phan, M. D., Turner, D. J., Perkins, T. T., Parts, L., Haase, J., ... Turner, A. K. (2009). Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Research*, *19*(12), 2308–2316. doi: 10.1101/gr.097097.109.

Lanver, D., Tollot, M., Schweizer, G., Lo Presti, L., Reissmann, S., Ma, L. S. ... Kahmann, R. (2017). Ustilago maydis effectors and their impact on virulence. *Nature Reviews Microbiology*, *15*(7), 409–421. doi: 10.1038/nrmicro.2017.33.

Michel, A. H., Hatakeyama, R., Kimmig, P., Arter, M., Peter, M., Matos, J., ... Kornmann, B. (2017). Functional mapping of yeast genomes

by saturated transposition. *eLife*, *6*, e23570. doi: 10.7554/eLife.23570.

Michielse, C. B., Hooykaas, P. J. J., van den Hondel, C. A. M. J. J., & Ram, A. F. J. (2005). Agrobacterium-mediated transformation as a tool for functional genomics in fungi. *Current Genetics*, *48*(1), 1–17. doi: 10.1007/s00294-005-0578-0.

Pflug, F. G., & von Haeseler, A. (2018). TRUmiCount: Correctly counting absolute numbers of molecules using unique molecular identifiers. *Bioinformatics*, *34*(18), 3137–3144. doi: 10.1093/bioinformatics/bty283.

Picelli, S., Bjorklund, A. K., Reinius, B., Sagasser, S., Winberg, G., & Sandberg, R. (2014). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Research*, *24*(12), 2033–2040. doi: 10.1101/gr.177881.114.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. doi: 10.1038/nbt.1754.

Sedlazeck, F. J., Rescheneder, P., & von Haeseler, A. (2013). NextGenMap: Fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics*, *29*(21), 2790–2791. doi: 10.1093/bioinformatics/btt468.

Seitner, D., Uhse, S., Gallei, M., & Djamei, A. (2018). The core effector Cce1 is required for early infection of maize by Ustilago maydis. *Molecular Plant Pathology*, *19*(10), 2277–2287. doi: 10.1111/mpp.12698.

Smith, T., Heger, A., & Sudbery, I. (2017). UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*, *27*(3), 491–499. doi: 10.1101/gr.209601.116.

Stirnberg, A., & Djamei, A. (2016). Characterization of ApB73, a virulence factor important for colonization of Zea mays by the smut Ustilago maydis. *Molecular Plant Pathology*, *17*(9), 1467–1479. doi: 10.1111/mpp.12442.

Uhse, S., Pflug, F. G., Stirnberg, A., Ehrlinger, K., von Haeseler, A., & Djamei, A. (2018). In vivo insertion pool sequencing identifies virulence factors in a complex fungal-host interaction. *PLoS Biology*, *16*(4), e2005129. doi: 10.1371/journal.pbio.2005129.

van Opijnen, T., Bodi, K. L., & Camilli, A. (2009). Tn-seq: High-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nature Methods*, *6*(10), 767–U721. doi: 10.1038/nmeth.1377.

Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., ... Davis, R. W. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science*, *285*(5429), 901–906. doi: 10.1126/science.285.5429.901.

This page intentionally left blank.

Chapter 5

The efficacy of TRUmiCount

Preamble

In our original publication describing the TRUmiCount method (Pflug & von Haeseler, 2018; chapter 2 of this work), we decided against including any data on the performance of the method for iPool-Seq data; mostly due to space constraints. This was despite the fact that iPool-Seq data is well-suited for assessing the performance of the method for real-world data; since iPool-Seq provides independent abundance estimates for the 5'-flank and 3'-flank of each knocked-out gene, the variation between the two estimates may be used as a gauge of the fidelity of an abundance estimate

This chapter explores this idea and assesses the efficacy of TRUmiCount by comparing the 5'- and 3'-derived abundance estimates before and after application of the algorithm. It thus complements the simulation studies of Pflug & von Haeseler (2018; chapter 2 of this work) on the performance of TRUmiCount.

We then link the results we find for iPool-Seq back to RNA sequencing, the context in which TRUmiCount was originally presented by testing it with a dataset obtained by sequencing a defined mixture of synthetic mRNAs defined by the *external RNA control consortium* (ERCC).

Finally, we cover another omission in the original paper, and discuss the relationship of the two model parameters that TRUmiCount estimates, PCR efficiency and per-UMI sequencing depth.

This page intentionally left blank.

Introduction

Together with the initial presentation of the TRUmiCount algorithm (Pflug & von Haeseler, 2018; chapter 2 of this work), we showed that TRUmiCount's underlying model of *unique molecular identifier* (UMI) - based *next generations sequencing* (NGS) experiments fits observed data well, and allows us to use the observed per-UMI read count distribution to estimate the *loss*, i.e. the fraction of molecules that remained unobserved.

The main purpose of TRUmiCount, however, is to *correct* for this loss; and thus to remove *indirect PCR bias* from UMI-based quantitative NGS experiments. While counting distinct UMIs instead of reads avoids *direct PCR bias*, preferential amplification (i.e. higher PCR efficiency) of molecules with particular sequences can decrease those molecule's risk of remaining undetected, and thus indirectly bias the results in favour of those molecules. While Pflug & von Haeseler (2018; chapter 2 of this work) show the efficacy of TRUmiCount in removing indirect PCR bias for *simulated* data, it remains to do so for experimental data. For this, two general strategies are possible, *internal validation* where we asses the internal concordance of the results, and *external validation* where we compare the results to the known truth.

Internal validation

By *internal validation* we mean assessing the performance of a measurement technique by testing the concordance of different measurements of the same quantity. Since the indirect PCR biases we hope to remove with TRUmiCount are sequence-dependent and thus likely persist between replicates, simply repeating the same experiment and comparing results is not sufficient. Instead, we need an experimental design where the same underlying abundance can be measured by counting molecules with different sequences, and therefore different indirect PCR biases.

By lucky coincidence, iPool-Seq data allows for this type of internal validation. Briefly, the iPool-Seq protocol (Uhse et al., 2018; chapter 3 of this work) measures the abundances of particular mutant strains of a single-cellular organism, originally the fungus *Ustilago maydis*. Each mutant strain has a single gene knocked out by

replacement with a particular sequence called *knockout cassette* (KO cassette). To measure the abundances of the strains in a pool of knockouts, iPool-Seq fragments the genomic DNA, labels each fragment with an UMI, amplifies the fragments that overlap either *flank*[1] of the KO cassette (5' or 3'; relative to a specific strand), and sequences the resulting library. Under ideal conditions, the number of detected UMIs for the 5' and the 3' flank of the KO cassette should therefore agree and reflect the number of genomes of each mutant in the mutant pool.

The sequence content of the amplified 5'- and 3'-flank fragments mostly consists of the (non-coding) genomic sequence that flank the knocked-out gene, and thus differs both between mutants and between the two flanks of a single mutants. It follows that PCR bias (both direct and indirect) can be expected to not only vary between mutants, but also to affect the 5'- and 3'-flank-derived UMI counts of a single mutant in different ways; this makes the two counts two independently biased measurements of the same true mutant genome count.

By comparing the concordance of the 5' and 3'-flank-derived genome counts before and after applying the TRUmiCount algorithm, we can therefore assess the algorithm's performance.

External validation

External validation refers to an alternative strategy where the measurement technique to be assessed is applied to a particular *standard* – an object or situation in which the *true* value of the measured quantity is known with high accuracy – and the resulting measurements are tested for concordance with the expected results. For RNA abundances, a suitable standard was developed by the *External RNA Controls Consortium* (ERCC), who defined different mixtures containing 92 synthetic mRNAs in known concentrations (External RNA Controls Consortium, 2005a,b).

However, the large difference of about 7 orders of magnitude between the concentration of the most and the least abundant mRNA in these ERCC mixtures

[1] With fragments that overlap either flank we mean fragments whose one end lies within the KO cassette and whose other end lies within the original genome. See figure 2 of Uhse et al. (2018; chapter 4 of this work on page 78) for a more detailed description on the sequence content of these fragments

makes applying TRUmiCount somewhat difficult[2]. For these large differences in mRNA abundance (which exceed the abundance differences in living cells by about 2 or 3 magnitudes), the typical number of a few hundred thousand different UMIs used for labelling is too small to prevent identical molecules from sometimes being labelled with identical UMIs. These *UMI collisions* break TRUmiCount's assumption of distinguishable molecules before PCR amplification, and thus alter the expected per-UMI read count distribution. This technical difficulty prevented us from include any such performance assessment in the original presentation of the TRUmiCount algorithm; at that time all UMI-based datasets available to us either failed to include any synthetic RNA mixtures, or were affected by UMI collisions within most of the synthetic RNAs they included.

The single-cell RNA sequencing method developed by 10x Genomics provides a way to overcome these difficulties. Their method encapsulates each individual cell within a single droplet of a water-based buffer in an oil suspension, and uses that droplet as a nanoliter reaction chamber to reverse transcribe the cell's mRNA into cDNA. During reverse transcription, each mRNA is labelled not only with an UMI, but also with a droplet-specific *unique cell identifier* (CID).

To create the *ERCC 1k* (10x Genomics, 2016; Zheng *et al.*, 2017) dataset, Zheng *et al.* input a synthetic ERCC mRNA mixture instead of actual cells. This makes the "transcriptome" of all "cells" identical, and lets the CIDs act simply as a second-level identifier of molecules – the UMIs themselves only need to be long enough to distinguish all identical mRNAs within a single droplet. The resulting sequencing data shows no signs of UMI collisions, and thus allows us to apply TRUmiCount and to assess its performance.

PCR efficiency vs. mean number of reads per molecule

TRUmiCount estimates two model parameters for each *molecular species* that is being counted (e.g. transcripts of a particular gene in the case of RNA-sequencing; genomics fragments overlapping one of the flanks of the KO cassette for iPool-Seq). These parameters are the *PCR efficiency* E and the *mean number of reads per*

[2]For *ERCC* Spike-In Mix 2, the most abundance mRNA is ERCC-00002 with $3 \cdot 10^4$ attomole/μl and the least abundant one ERCC-00083 with $7 \cdot 10^{-3}$ attomole/μl.

molecule D. Because TRUmiCount works with a normalised version of the (post-amplification) family size distribution (Pflug & von Haeseler, 2018; chapter 2 of this work, see section *2.2.2 The normalized family size F* on page 33), TRUmiCount treats these two parameters as independent, meaning that smaller values for the PCR efficiency E will increase the *variance* of the predicted per-UMI read counts, but not decrease their mean (which is equal to D per definition).

Yet while this normalisation and the ensuing independence of the parameters is convenient mathematically, there does not seem to be a *physical* reasons for it – clearly, under any model of the PCR, we would expect the differences in PCR efficiencies to strongly affect post-amplification copy numbers.

From the presence or absence of the expected relationship between efficiency E and sequencing depth D in experimental data, we can thus learn more about how well TRUmiCount's model captures PCR behaviour, and which types of effects it might fail to consider.

Methods

Internal validation

The iPool-Seq data, more specifically the raw sequencing reads from the input and output pools of replicates 1-3 of experiments A and B of Uhse *et al.* (2018; chapter 3 of this work) were processed (until step 10) of the data analysis pipeline of Uhse et al. (2019) to produce 12 (knockout) mutant abundance tables (see table 2 of Uhse et al., 2019; chapter 4 of this work on page 77), one for each sequenced pool. The phantom UMI rejection threshold was set based on manual inspection of the observed UMI count distributions to $T = 1$ for pools $A1^{(in)}$, $A1^{(out)}$, $A2^{(in)}$, $A2^{(out)}$, $A3^{(in)}$, $A3^{(out)}$, $B2^{(in)}$, to $T = 2$ for pool $B1^{(in)}$, to $T = 5$ for pools $B1^{(out)}$, $B2^{(out)}$ and to $T = 9$ for pool $B3^{(out)}$. Each of the 12 mutant abundance tables for the 12 pools contain, for the each flank f (5' or 3') of each mutant m (encoded as "m:f" in the column *gene*), the number $N_{m,f}$ of observed UMIs (after applying the phantom UMI rejection threshold; column *n.umis*), the estimated loss $\ell_{m,f}$ (column *loss*) and the loss-corrected abundance $A_{m,f} = N_{m,f}/(1 - \ell_{m,f})$ (column *n.tot*).

From these tables, we then asses the concordance of 5'- and 3'-derived abundance

measures by computing (for each of the 12 pools) two correlation coefficients, R_{raw} for the correlation of the (phantom-filtered but not loss-corrected) UMI counts $N_{m,5'}$ and $N_{m,3'}$, and R_{corr} for the correlation of the loss-corrected abundances $A_{m,5'}$ and $A_{m,3'}$. Both correlation coefficients are computed as the Pearson correlation coefficient of the *square root* of the counts or abundances; under a Poissonian sampling model, the square root reduces heteroscedasticity and avoids overdue influence of highly abundant mutants.

External validation

We first filtered the raw sequencing read pairs of the *ERCC 1k* dataset (10x Genomics, 2016; Zheng *et al.*, 2017) with UMI-Tools' *whitelist* and *extract* tools (Smith *et al.*, 2017) to remove reads with bogus cell identifiers (CIDs), then mapped them (their non-technical parts, i.e. read 2) against the 92 ERCC sequences, and grouped reads with similar UMIs with UMI-Tools' *group* tool. Finally, we ran TRUmiCount with phantom UMI rejection threshold $T = 10$ (chosen again based on visual inspection) to produce an mRNA abundance table (similar to the mutant abundance tables produced for iPool-Seq data; see table 2 of chapter 4 on page 77). The mRNA abundance table lists for each mRNA g its UMI count (phantom-filtered by thresholding) N_g (column *n.umis*), estimated PCR efficiency E_g (column *efficiency*), estimated mean number of reads per molecule D_g (column *depth*), loss estimate ℓ_g (column *loss*) and loss-corrected abundance $A_g = N_g/(1-\ell_g)$ (column *n.tot*). We then repeated these steps with subsamples containing 50%, 20%, 10%, 5%, 2% and 1% of the raw sequencing read pairs; the phantom UMI rejection threshold was adjusted accordingly to 5, 2 and 1 (for the 10% subsample and all smaller ones).

For each subsample (including the full sample), we then computed using these mRNA abundance tables the Pearson correlation of the *square root*[3] of measured abundances and known concentrations (Thermo Fischer, n.d.) of the 92 synthetic mRNAs to asses their concordance. We did this twice, once to compute the correlation R_{raw} for the uncorrected UMI counts N_g, and once to compute R_{corr} for the loss-corrected abundances A_g instead.

[3]The square root again stabilises the variances of the observed abundances under a Poissonian sampling model, and avoid an overdue influence of highly abundance mRNAs on the results

Fraction of variance unexplained

Given a correlation coefficient R (either R_{raw} for uncorrected UMI counts or R_{corr} for loss-corrected abundances), we call R^2 the *fraction of variance explained* and $1-R^2$ the *fraction of variance unexplained*[4] (FVU).

To quantify the improvement in concordance due to TRUmiCount, either between 5'- and 3'-derived abundance estimates for internal validation or between estimated and known abundances for external validation, we compute the relative reduction of the FVU as

$$\begin{aligned}\Delta\text{FVU} &= \frac{\text{FVU}_{\text{raw}} - \text{FVU}_{\text{corr}}}{\text{FVU}_{\text{raw}}} \\ &= \frac{R^2_{\text{corr}} - R^2_{\text{raw}}}{1 - R^2_{\text{raw}}},\end{aligned} \tag{1}$$

and express it in percent (%).

Any $\Delta\text{FVU} > 0$ then reflects an improved concordance after applying TRUmiCount, and $\Delta\text{FVU} = 50\%$ for example means that the amount of unexplained variance was halved.

Modelling the relationship of E and D

Under an exponential growth model of PCR, a reaction that runs for n cycles with efficiency E amplifies molecule copy numbers by a factor of $(1+E)^n$. We then assume that we may treat *copy numbers*, *concentration* and *average number of reads per UMI* as the same abundance-related quantity expressed in different units[5]to arrive at the following model for the relationship between a molecular species' PCR efficiency E and average number of reads per UMI D,

$$D = B \cdot (1+E)^{n_{\text{eff}}}. \tag{2}$$

[4]For external validation, this is the typical nomenclature in regression analysis – the known concentrations take the role of a *predictor* variable, which then explains some fraction of the variance of the *response* variable. For internal validation, either the 5' or the 3' measurement can arbitrarily be chose as a predictor, and the other as a response.

[5]This is not true exactly, but the difference is negligible. We ignore that the *concentration* or *average number of reads per UMI* does not only depend on a species' own copy number, but also the sum of all other species' copy numbers.

The number of "cycles" n_{eff} in this relationship does not have to correspond to the actual number of PCR cycles, and it is not necessarily integral. Instead, it reflects the *effective* number of cycles where the reaction runs with efficiency E – after that, we assume the efficiencies to drop to a common value for all molecular species. The *baseline abundance* B represents the expected number of reads per UMI for molecules that are not amplified at all over those n_{eff} cycles.

For each subsample (including the full sample) of the ERCC 1k dataset, we fit the log-transformed[6] model from equation 2 using *ordinary least squares* (OLS) to the observed pairs (E_g, D_g) for the 92 synthetic mRNAs g to estimate n_{eff} and B.

Results

Model fit

For the 10x Genomics ERCC data (figure 3 in chapter 5 on page 98), we observe a good agreement between observed and predicted read count distributions, and reasonable model parameters, in particular for the PCR efficiency.

For iPool-Seq data (figure 1 in chapter 5 on page 96), the situation is more ambiguous. For experiment B, the input pools show a good agreement between model and observation, and reasonable (yet low) efficiency estimates. For the corresponding output pools, the efficiency estimate is zero, and obtaining a good fit required relatively high phantom-rejection thresholds ($T = 5$ for $\text{B1}^{(\text{out})}$ and $\text{B2}^{(\text{out})}$, $T = 10$ for $\text{B3}^{(\text{out})}$). These thresholds filter out 51% ($\text{B1}^{(\text{out})}$), 57% ($\text{B2}^{(\text{out})}$) respectively 72% ($\text{B3}^{(\text{out})}$) of the UMIS; with these UMIs making up for 9% ($\text{B1}^{(\text{out})}$), 13% ($\text{B2}^{(\text{out})}$) respectively 10% ($\text{B3}^{(\text{out})}$) of reads. For the 6 pools in experiment A, the estimated efficiencies are zero as well; apart from that, the sequencing depth is too low to make a reliable statement about the goodness of fit[7].

[6] The log transform yields the linear model $\log D = \log B + n \cdot \log(1 + E)$

[7] Essentially, only reads counts 1 and 2 are observed which any two-parameter model can likely model well

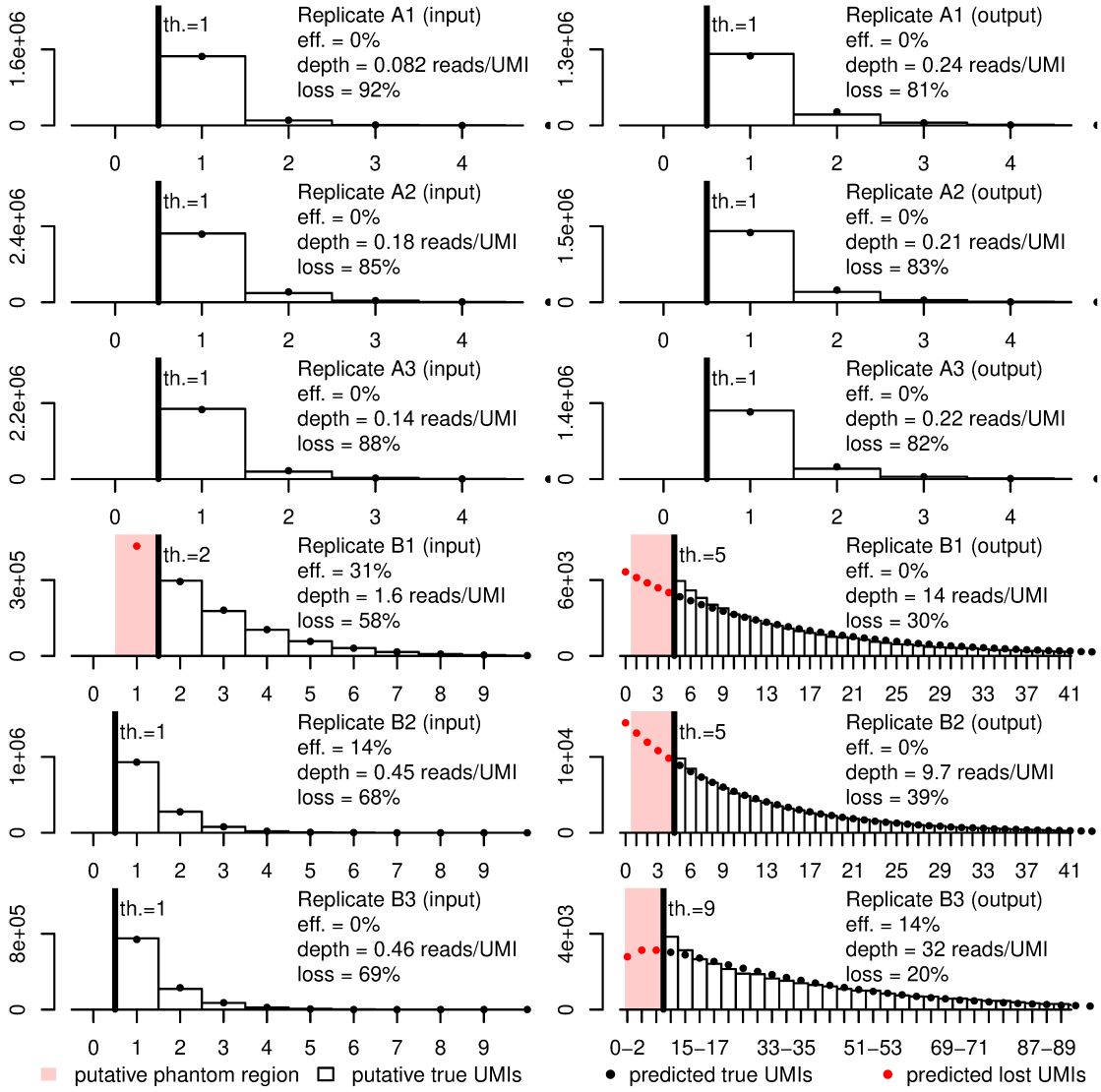

Figure 1: **Observed and predicted per-UMI read count distributions for iPool-Seq data.** The x-axis shows per-UMI read counts, the y-axis the observed (black bars) respectively predicted (black and red dots) frequencies of these read counts amongst UMIs. The UMIs that fall into the (red) area below the phantom-rejection threshold are not shown. Data stems from input and output pools of replicates A1-A3, B1-B3 of Uhse *et al.* (2018; chapter 3 of this work)

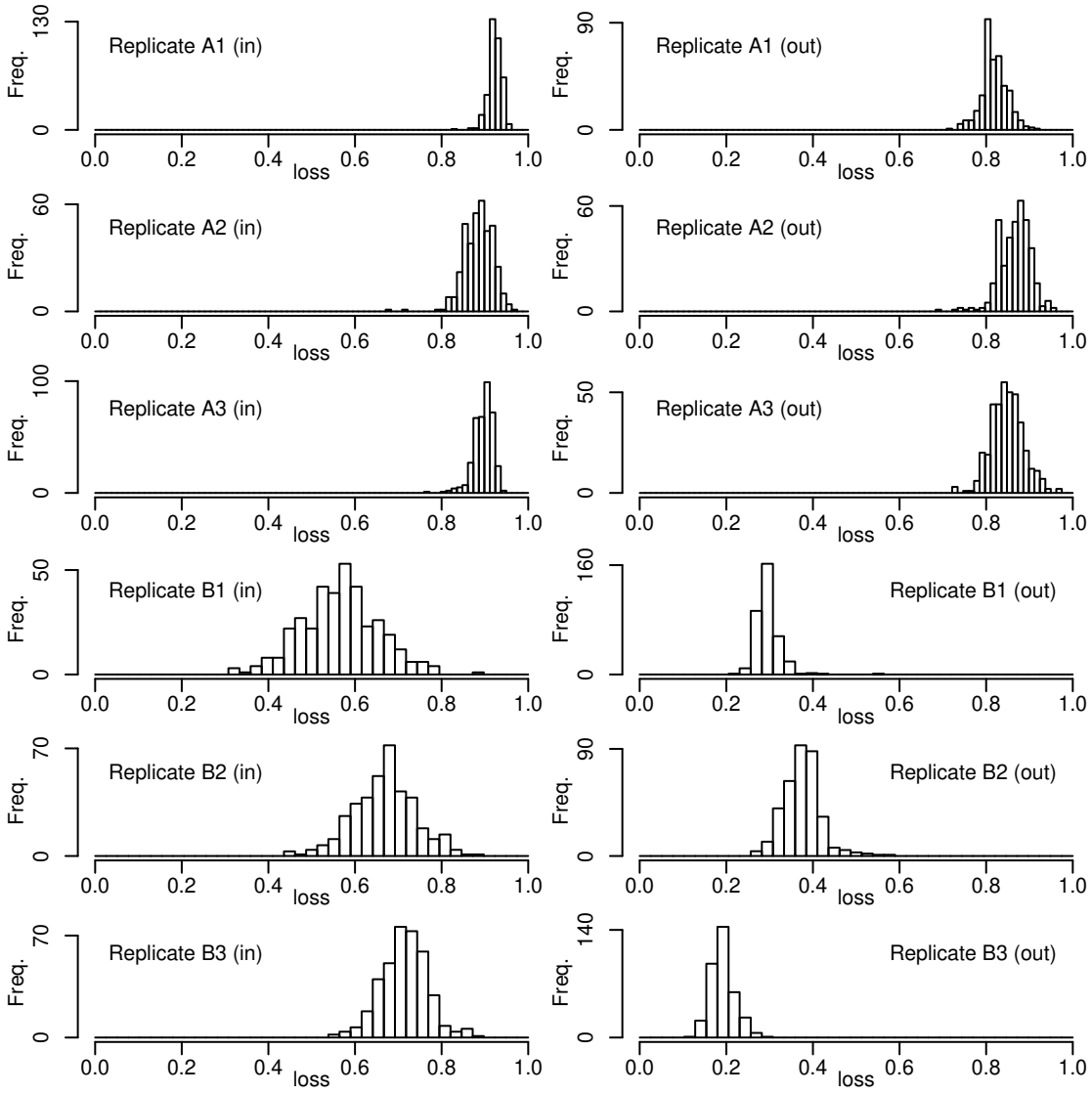

Figure 2: **Estimated mutant- and flank-specific UMI losses for iPool-Seq data.** Histograms showing the distribution of the mutant- (m) and flank- (f; 5' or 3') specific TRUmiCount UMI loss estimates $\ell_{m,f}$ for the input and output pools of replicates 1-3 of experiments A and B of Uhse *et al.* (2018; chapter 3 of this work).

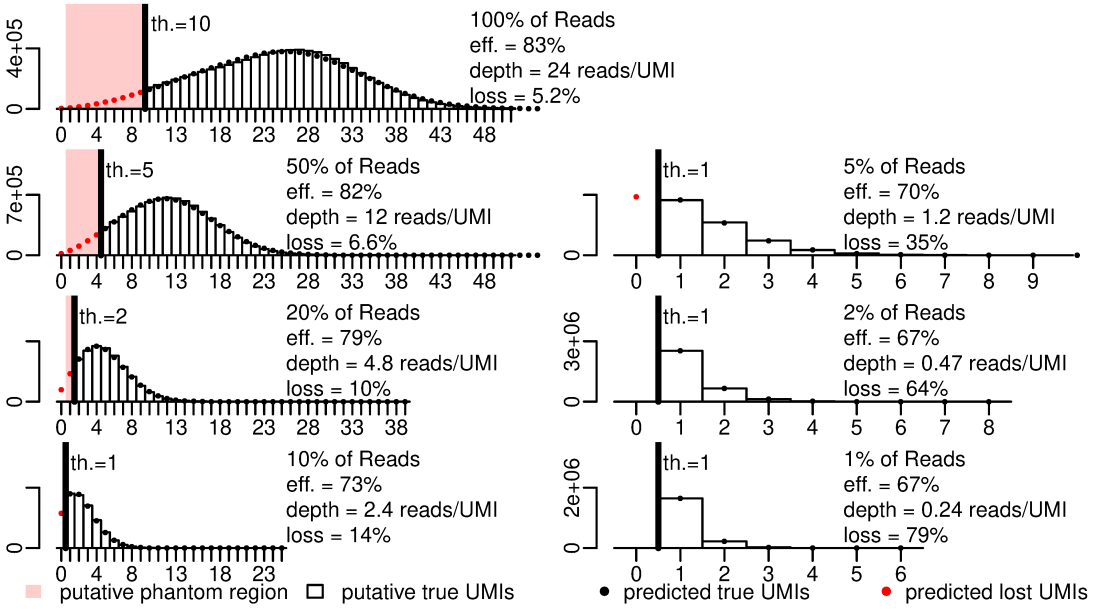

Figure 3: **Observed and predicted UMI read count distributions for 10x ERCC data.** The x-axis shows per-UMI read counts, the y-axis the observed (black bars) respectively predicted (black and red dots) frequency of these read counts amongst UMIs. The UMIs that fall into the (red) area below the phantom-rejection threshold are not shown. Shown are 100%, 50%, …, 1% subsamples of the ERCC 1k dataset (10x Genomics, 2016; Zheng *et al.*, 2017).

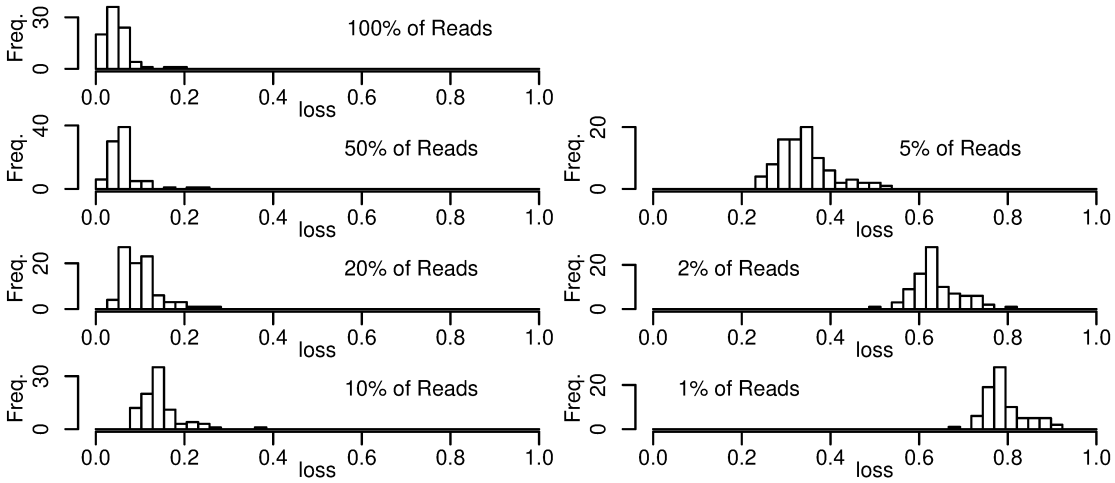

Figure 4: **Estimated mRNA-specific UMI losses for 10x ERCC data.** The distribution of the mRNA-specific TRUmiCount UMI loss estimates ℓ_g across the synthetic mRNAs for 100%, 50%, …, 1% subsamples of the ERCC 1k dataset (10x Genomics, 2016; Zheng *et al.*, 2017).

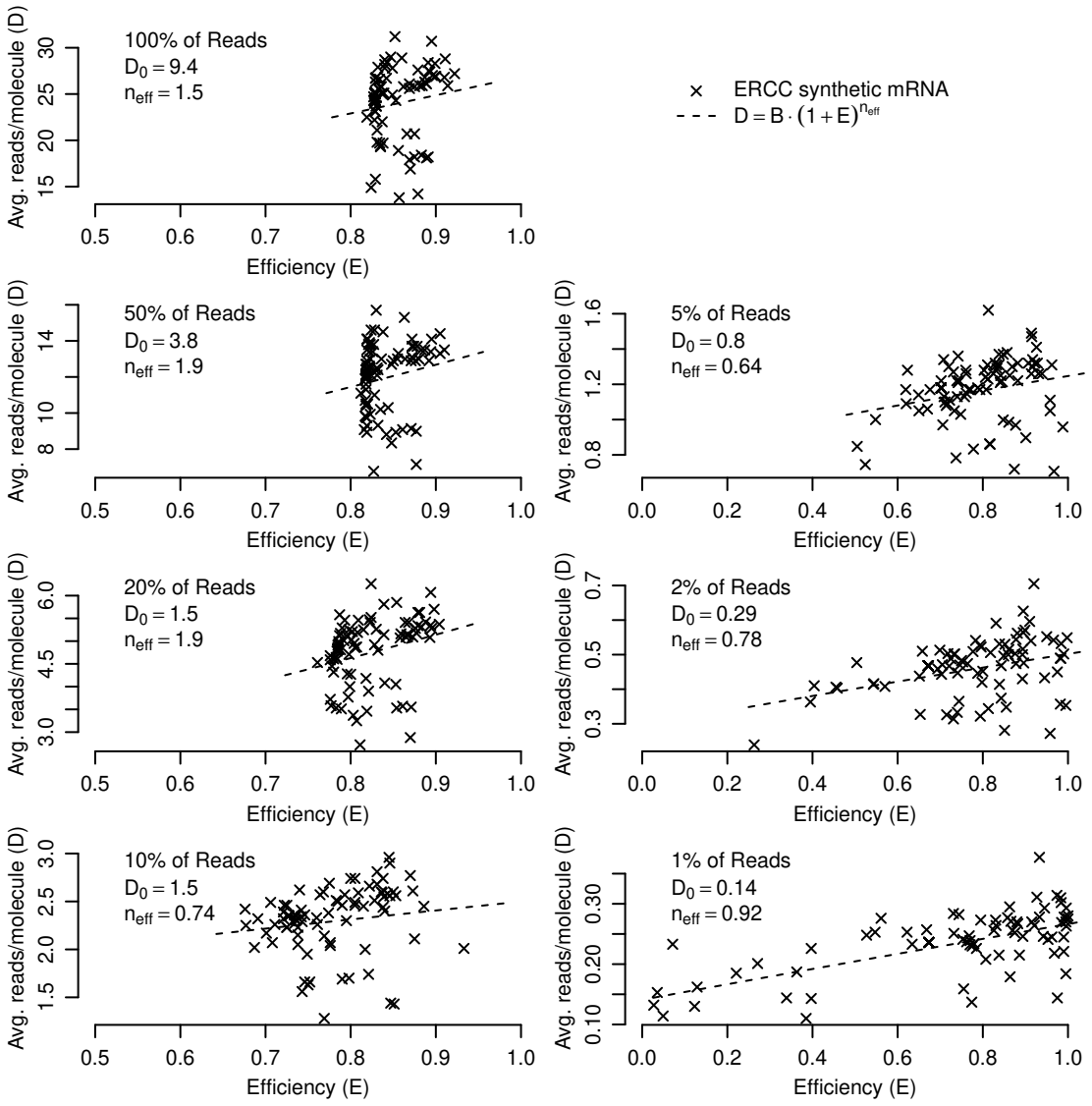

Figure 5: **The relationship of PCR efficiency and mean number of reads per molecule.** Estimated mRNA-specific efficiencies E_g vs. mean number of reads per molecule D_g across the 92 synthetic mRNAs g for 100%, 50%, …, 1% subsamples of the ERCC 1k dataset (10x Genomics, 2016; Zheng *et al.*, 2017).

Experiment	input			output		
	R_{raw}	R_{corr}	ΔFVU (% of raw)	R_{raw}	R_{corr}	ΔFVU (% of raw)
A1	0.902	0.931	29%	0.882	0.931	40%
A2	0.936	0.949	21%	0.919	0.945	30%
A3	0.959	0.960	3%	0.935	0.950	22%
B1	0.911	0.938	29%	0.980	0.987	34%
B2	0.947	0.961	25%	0.950	0.952	3%
B3	0.927	0.941	19%	0.938	0.942	6%

Table 1: **Reduction of unexplained variance between 5' and 3' mutant abundances for iPool-Seq data.** Columns R_{raw} (before correction) and R_{corr} (after loss correction by TRUmiCount) show the correlation of the UMI counts (Pearson correlation of the count's *square roots* to stabilise the variance) for the 5' and 3' flank of each knockout in the *input* respectively *output* pools of the 6 experiments of Uhse *et al.* (2018; 3 of this work). Column ΔFVU shows the reduction of the *fraction of variance unexplained* (FVU = $1 - R^2$) due to the loss correction, relative to the FVU before correction.

% Reads	R_{raw}	R_{corr}	ΔFVU (% of raw)	% Reads	R_{raw}	R_{corr}	ΔFVU (% of raw)
100%	0.970	0.973	11%				
50%	0.970	0.974	15%	5%	0.966	0.977	31%
20%	0.970	0.976	20%	2%	0.961	0.976	39%
10%	0.971	0.977	20%	1%	0.959	0.975	38%

Table 2: **Reduction of unexplained variance of observed mRNA abundances for 10x ERCC data.** Columns R_{raw} (before correction) and R_{corr} (after loss correction by TRUmiCount) show the correlation of the UMI counts for the 92 artificial mRNAs in ERCC Mix 2 with their known abundances (Pearson correlation of the count's and abundance's *square roots* to stabilise the variances) for 1% to 100% subsamples of 10X Genomics' *ERCC 1k* sample dataset (10x Genomics, 2016). Column ΔFVU shows the reduction of the *fraction of variance unexplained* (FVU = $1 - R^2$) due to the loss correction, relative to the FVU before correction.

Model parameters & estimated losses

The estimated losses follow a similar pattern across the input and output pools of replicates A1-A3, B1-B3 (figure 2 in chapter 5 on page 97), but are shifted towards one for lower and towards zero for higher overall sequencing depths (i.e. average number of reads per UMI), as expected. For the subsamples of the 10x ERCC data (figure 4 in chapter 5 on page 98), the estimated losses also shift depending on the sequencing depth, with their distribution mostly retainings its shape as it shifts.

For subsamples of the 10x ERCC data with low sequencing depth (10% of all reads and lower), the relationship between the mRNA-specific parameter estimates for PCR efficiency E_g and average number of reads per molecule D_g follows the exponential growth model from equation 2, but with only between $n_{\text{err}} = 0.64$ and $n_{\text{eff}} = 0.92$ effective PCR cycles (figure 5 in chapter 5 on page 99) At high sequencing depths (100% down to 20% of all reads), the efficiency estimates never drop below ≈ 0.8, which "squashes" the points together and breaks the predicted relationship of E_g and D_g.

Internal validation

Despite the ambiguity about the goodness of fit for iPool-seq data, we observed an increased concordance between 5'- and 3'-derived abundances estimates across all 12 sequenced mutant pools; the *reduction of the fraction of unexplained variance* (ΔFVU; equation 1 on page 94) ranges from 3% to 40%, with most pools showing a ΔFVU of around 20%-30% (table 1 in chapter 5 on page 100).

External validation

For the 10x ERCC data, we observe a ΔFVU of 11% for the full dataset (table 2 in chapter 5 on page 100). The reason for this rather modest improvement of concordance is the rather small overall UMI loss of 5% on average, with a maximum of 19% for ERCC-00171 (figure 4 in chapter 5 on page 98). In contrast, for subsamples of the 10x ERCC data, the ΔFVU increases to up to 39%. It is also noteworthy that while the raw correlation decreases with decreasing sequencing depth from 0.97 to 0.959, the correlation after applying TRUmiCount stays constant

at ≈ 0.975. The loss-corrected abundances thus contain about 5% of unexplained variance irrespective of sequencing depth which ranges from ≈10 reads/UMI down to ≈0.25 reads/UMI.

Discussion

Model fit

10x Genomics' single-cell RNA-seq method (Zheng *et al.*, 2017) is considerably more complex than the typical RNA-seq protocol. Yet, all of the added complexity deals with labelling mRNAs (or rather the corresponding cDNA produced by reverse transcription) with unique cell- and molecule identifiers *before* amplification. After that, amplification and sequencing of the cDNA library proceeds normally, and in particular without any interspersed purification steps. Since the model underlying TRUmiCount focusses on the amplification and sequencing steps, it still closely resembles experimental reality for droplet-based single-cell RNA-seq methods such as the one offered by 10x Genomics.

This close resemblance is reflected by the good agreement between observed and prediction per-UMI read count distributions (figure 3 in chapter 5 on page 98) for the 10x ERCC dataset (10x Genomics, 2016); the model fits the data as well as it fit for the two RNA-seq datasets that TRUmiCount was originally tested against (Pflug & von Haeseler, 2018; chapter 2 of this work).

iPool-Seq (Uhse et al., 2018; chapter 3 of this work) initially labels all fragments with UMIs, but only those that *uniquely* identify a single mutant strain are then enriched by a combination of specific PCR and Streptavidin purification (figure 2 of Uhse et al., 2018; chapter 3 of this work on page 48). While the specific PCR matches TRUmiCount's model, the purification step interspersed between the two PCRs violates it, for the following reason:

As discussed in more detail by Pflug & von Haeseler (2018; chapter 2 of this work, see section *2.2.2 The normalized family size F* on page 33), as each molecule's copy number grows, the amount of additional stochasticity introduced by every additional PCR cycle drops, which allows TRUmiCount to disregard the exact

number of cycles and use the limit distribution instead. Bead-based purification steps such as the Streptavidin purification that iPool-Seq relies on, however, are generally quite lossy, meaning their output contains considerably fewer molecules than their input, even for molecules they actually select for. This random loss of copies directly introduces additional stochasticity into the post-purification family sizes, and it also indirectly introduces stochasticity by lowering the copy numbers that the subsequent PCR starts from.

This makes the model used by TRUmiCount a coarser approximation of reality for iPool-Seq than for RNA-seq. This miss-specification, however, appears to be slight enough to not be observable in the comparison of observed and expected read counts – unless possibly for pool $B3^{(out)}$ where it may be connected to the rather large phantom-rejection threshold $T = 10$ that was required to obtain a good fit. The miss-specification is, however, indirectly observable through the unrealistically low PCR efficiency estimates (zero for most pools); with the PCR efficiency being the only parameter of TRUmiCount's model that affects the variance independently of the mean, any non-modelled source of variance will generally skew the estimated efficiency towards lower values.

Model parameters & estimated losses

Given the exponential growth behaviour of the PCR, we expected to see a strong link between the estimated PCR efficiencies E_g and mean number of reads per molecule D_g over all molecular species g. We, however, do not consistently observe such a link – while it appears to be somewhat present for small subsamples of the ERCC data (fewer than 20% of all reads), for larger subsamples all efficiency estimates lie close together, but the average number of reads per molecule still varies considerably (figure 5 in chapter 5 on page 99). Yet the range of loss estimates seems reasonable in both cases (figure 4 in chapter 5 on page 98), and TRUmiCount consistently reduces the *fraction of variance unexplained* (table 2 in chapter 5 on page 100).

We cannot currently offer an explanation for this effect. We previously found strong evidence that our efficiency parameter E actually measure the (early cycle) efficiency of the PCR reaction; this was corroborated by the slightly secondary peak

we observed for the *E. coli* dataset we previously tested TRUmiCount with, and by the efficiency-length relationship we found (figures 2 and S2 of Pflug & von Haeseler, 2018; chapter 2 of this work on pages 36 and 41). Yet for the 10x ERCC data, the efficiency estimate seems to act more as nuisance parameter that we must estimate to compute the loss, but which does not have an immediate connection to experimental reality.

Disentangling the precise relationship between the PCR efficiency we estimate and the physical behaviour of the reaction would require more data; in particular, it would require data where reaction parameter are modified individually so that the effect of these reaction parameters the estimated efficiency and the mean number of reads per molecule can be studied.

Here, we can only conclude that in the light of the subtle relationship between efficiency and mean number of read per molecule, treating these as independent model parameters is sensible. Doing so allows TRUmiCount to adapt better to different experimental conditions, and to still produces usable loss estimates even if some modelling assumptions are violated.

Efficacy in correcting for indirect PCR bias

We see improved concordance, both *internally* between the 5'- and 3'-derived abundance estimates for iPool-Seq data, and *externally* with the known input concentrations of different synthetic mRNAs for 10x ERCC data (tables 1 and 2 in chapter 5 on page 100).

The improvement ranges from a very slight improvement in unexplained variance (3% reduction for pools $A2^{(in)}$ and $B2^{(out)}$ of Uhse *et al.*) to removing more than a third of the residual variance ($A1^{(out)}$ and $B1^{(out)}$ of Uhse *et al.*, 1% and 2% subsample of the 10x ERCC data), and large improvements can be observed even for samples where the fit of the model is not perfect and some parameter estimates are unrealistic.

We conclude that, despite the fact that the model does not describe experimental reality perfectly, TRUmiCount improves the fidelity of abundances measured via quantitative NGS experiments, and in doing so is pretty robust against violations of its modelling assumptions.

References in this chapter

10x Genomics. (2016). Single Sell Gene Expression Dataset ERCC 1k. Retrieved October 20, 2018, from https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/ercc

External RNA Controls Consortium. (2005a). Proposed methods for testing and selecting the ERCC external RNA controls. *BMC Genomics*, *6*(1), 150. DOI:10.1186/1471-2164-6-150

External RNA Controls Consortium. (2005b). The External RNA Controls Consortium: a progress report. *Nature Methods*, *2*(10), 731–734. DOI:10.1038/nmeth1005-731

Smith T., Heger A., & Sudbery I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome research*, *27*(3), 491–499. DOI:10.1101/gr.209601.116

Thermo Fischer. (n.d.). ERCC Controls Analysis: ERCC RNA Spike-In Control Mixes. Retrieved October 20, 2018, from https://assets.thermofisher.com/TFS-Assets/LSG/manuals/cms_095046.txt

Zheng G. X. Y., Terry J. M., Belgrader P., Ryvkin P., Bent Z. W., Wilson R., … Bielas J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*(1), 14049. DOI:10.1038/ncomms14049

This page intentionally left blank.

Chapter 6

Statistically modelling iPool-Seq

Preamble

This chapter augments the *Bioinformatic analysis* section (Uhse *et al.*, 2018; chapter 3 of this work; page 56 ff.) and the *Finding differentially virulent KOs* section (Uhse *et al.*, 2019; chapter 4 of this work; page 82 ff.). In these sections, space constraints prevented a proper definitions of what *virulence* means in the context of iPool-Seq, how it is computed, how the statistical model we employ to detect significant difference between different mutants is derived, and how the model is applied to experimental data.

Here, all of these points are addressed. This chapter thus does not introduce any new model, but instead provides background information and a rigorous derivation of the statistical model of iPool-Seq used by Uhse et al. (2018 and 2019; chapters 3 and 4 of this work).

The chapter is also intended to serve as a blueprint of how TRUmiCount is can be combined with other statistical methods – the naive approach of using loss-corrected counts in place of raw counts often fails. First because loss-corrected counts are non-integral, and second because they no longer obey the Poissonian "mean equal variance" relationship that many methods assume for count data.

This page intentionally left blank.

Introduction

The purpose of the *iPool-Seq* method (Uhse *et al.*, 2018; chapter 3 of this work) is to measure the *virulence* of particular mutants (strains) of a single-cellular pathogen and to detect mutants with a *virulence phenotype*, meaning a virulence that is significantly increased or decreased compared to wild-type strains. While in general many definitions of *virulence* are possible, our measure will reflect a strain's ability to proliferate on the host, and will not, for example, take into account the magnitude of any adverse effects on the host the pathogen might have.

The prerequisite for iPool-Seq is to have a *pool* (i.e. mixture) of mutants of the pathogen, each of which had a single gene knocked out by replacement with a particular sequence called the *KO* (knockout) cassette. The host is then infected with this (pre-infection) *input* pool of mutants, and after giving the pathogens time to proliferate on the host, the host (or its infected parts) are harvested, which yields the (post-infection) *output* pool of mutants. The DNA in both pools is then extracted, fragmented and labelled with *unique molecular identifiers* (UMIs; Hug and Schuler, 2003; Kivioja *et al.*, 2011). After labelling, the fragments that overlap either the 5' or the 3' flank of the KO cassette (and thus uniquely identify a particular mutant) are amplified and subjected to *next-generation sequencing* (NGS; Mardis, 2008, 2013; Metzker, 2010). Processing the resulting (many millions) sequencing reads (steps 7 to 10 of Basic Protocol 4 of Uhse *et al.*, 2019; chapter 4 of this work) then produces the following basic read-outs for a particular mutant m:

$N^{(\mathrm{in})}_{m,5'}$, $N^{(\mathrm{in})}_{m,3'}$, $N^{(\mathrm{out})}_{m,5'}$, $N^{(\mathrm{out})}_{m,3'}$**:** The numbers of detected unique genomic fragments (UMIs) in the input (in) respectively output (out) pool that overlap the 5' respectively 3' flank of the KO cassette in mutant m. We assume each copy of mutant m contains a single copy of its genome with exactly one KO cassette integration; consequently, exactly one 5' and one 3' fragment per mutant copy should be detected in each pool. For every mutant m, iPool-Seq thus yields two *independent* measurements, $N^{(\mathrm{in})}_{m,5'}$ and $N^{(\mathrm{in})}_{m,3'}$, related to the mutant's abundance (i.e. copy number) in the input pool, and similarly $N^{(\mathrm{out})}_{m,5'}$ and $N^{(\mathrm{out})}_{m,3'}$, for the mutant's abundance in the output pool.

$\ell^{(\mathrm{in})}_{m,5'}$, $\ell^{(\mathrm{in})}_{m,3'}$, $\ell^{(\mathrm{out})}_{m,5'}$, $\ell^{(\mathrm{out})}_{m,3'}$**:** The *loss* estimated by TRUmiCount (Pflug & von Haeseler,

2018; chapter 2 of this work) for the corresponding unique fragment counts, i.e. the estimated fraction of copies of mutant m in pool p (*in* or *out*) whose genomic fragment overlapping the flank f (*5'* or *3'*) of the KO cassette didn't contribute to the count $N^{(p)}_{m,f}$.

We also assume that we are given a particular *reference set* $\mathcal{R}$ of *reference mutants* against which the virulence of all other mutants is to be measured. Usually, this set will consist of mutants for which the knockout is known to have no effect on it's behaviour during host infections.

Our goal in this chapter is two-fold:

First, we will quantify the *virulence* of a particular mutant m by assigning it a numerical virulence value $v_m \in [0, \infty)$; for this virulence scale 0 will mean "cannot proliferate on the host at all" and 1 will mean "proliferates as fast as the reference mutants.

Second, we want to quantify the *significance* of a particular mutant's reduction or increase in virulence's compared to the reference mutants. This will entail the creation of a statistical model that captures the behaviour of reference mutants.

While *RNA Sequencing* (RNA-Seq; Z. Wang *et al.*, 2009) is superficially similar to iPool-Seq in that the abundance of individual of some particular *type* (genotypes for iPool-Seq, transcripts of a particular gene for RNA-Seq) is measured by quantitative *next-generation sequencing* (NGS), the models used by differential expression callers for RNA-Seq data like *edgeR* (M. D. Robinson, McCarthy, *et al.*, 2010), *DESeq* (Anders and Huber, 2010) and *DESeq2* (Love, Huber, *et al.*, 2014) are not directly applicable to iPool-Seq.

We will thus derive a statistical model tailored specifically to iPool-Seq experiments. It will rely on the same, proven ingredients (for example the *negative binomial distribution* as many existing models for RNA-Seq and other NGS-based experiments, but takes the specific design of iPool-Seq experiments into account.

From unique fragment counts to abundances

Given the iPool-seq read-out described above, i.e. unique fragment counts $N^{(p)}_{m,f}$ and loss estimates $\ell^{(p)}_{m,f}$ for every combination of mutant m, flank f (5' or 3') and pool p (in or out), we want to combine the flank-specific estimates to compute one abundance $A^{(p)}_m$ per mutant and pool. An obvious approach would be to correct each flank-specific fragment count $N^{(p)}_{m,f}$ separately for it's loss and average the resulting abundances, i.e. set $A^{(p)}_m$ to the average of $N^{(p)}_{m,f}/\big(1-\ell^{(p)}_{m,f}\big)$ for $f = 5'$ and $f = 3'$.

This combined abundance estimate, however, isn't well-suited given our goal of statistically modelling iPool-Seq. It doesn't seem to be associated with any simple statistical model that connects $A^{(p)}_m$, $N^{(p)}_{m,5'}$ and $N^{(p)}_{m,3'}$, and it does fail to take into account the *precision* of the two averaged abundances. The more unique fragments we detect for a particular flank, the lower the variance of $N^{(p)}_{m,f}/\big(1-\ell^{(p)}_{m,f}\big)$; and hence the higher should we *weight* that estimate.

We thus combine the two flank-specific abundance estimates of a mutant in a slightly different way. We start by assuming that there is a single true abundance $A^{(p)}_m$ of mutant m in pool p, and that the two losses $\ell^{(p)}_{m,5'}$ and $\ell^{(p)}_{m,3'}$ are fixed. We further assume that $N^{(p)}_{m,5'}$ and $N^{(p)}_{m,3'}$ are (conditionally on $A^{(p)}$ and the losses $\ell^{(p)}_{m,3'}$ and $\ell^{(p)}_{m,5'}$) independent, and that flank-specific fragment counts are modelled by

$$N^{(p)}_{m,f} \sim \text{Poisson}\Big(A^{(p)}_m \cdot \big(1-\ell^{(p)}_{m,f}\big)\Big) \tag{1}$$

(any other discrete stable distribution would work as well). We then set

$$\begin{aligned} N^{(p)}_m &= N^{(p)}_{m,5'} + N^{(p)}_{m,3'}, \\ \bar{\ell}^{(p)}_m &= \tfrac{1}{2}\ell^{(p)}_{m,5'} + \tfrac{1}{2}\ell^{(p)}_{m,3'} \end{aligned} \tag{2}$$

and find for the sum $N^{(p)}_m$ of the flank-specific counts the model

$$N^{(p)}_m \sim \text{Poisson}\Big(2 \cdot A^{(p)}_m \cdot \big(1-\bar{\ell}^{(p)}_m\big)\Big), \tag{3}$$

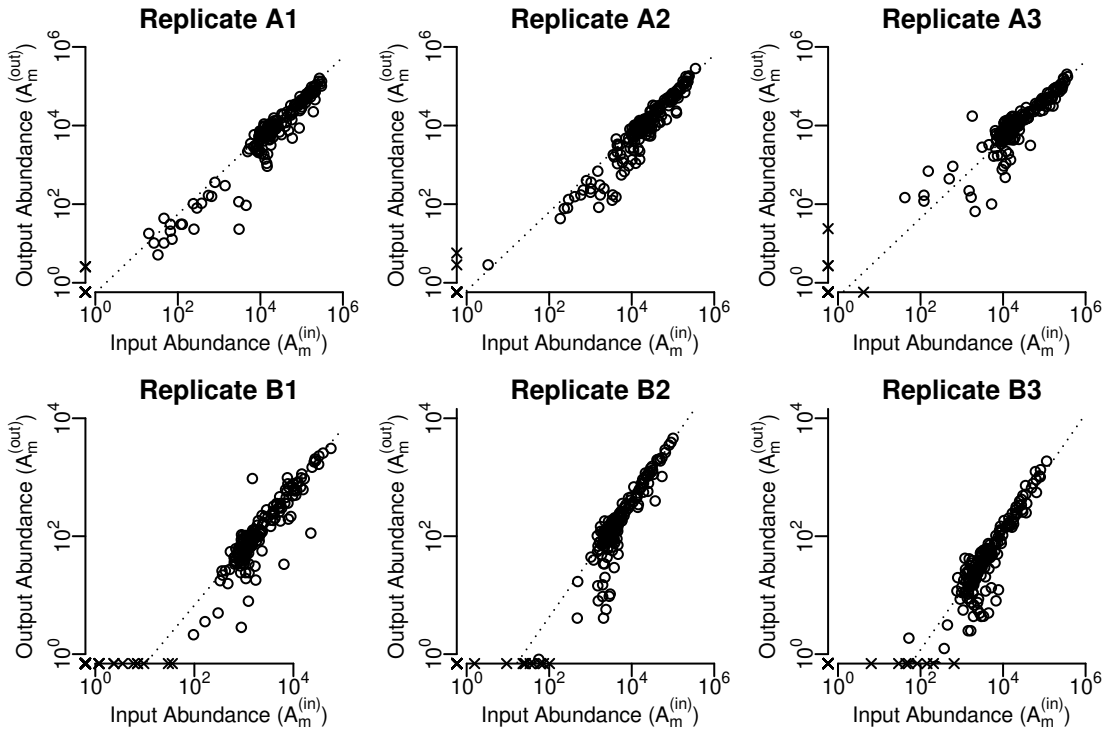

Figure 1: **Input pool vs. output pool abundances of each mutant.** A cross (×) on the x- or y-axis marks mutants undetected in the output or input pools. The dotted line reflects the output abundance prediction for reference mutants ($A_m^{(\mathrm{out})} = \lambda \cdot A_m^{(\mathrm{in})}$; see equation 6 for $v_m = 1$). Data from replicates 1 to 3 of experiments A and B of Uhse *et al.* (2018; chapter 3 of this work). Loss-corrected genome number and model predictions were computed with the pipeline of Uhse *et al.* (2019; chapter 4 of this work).

which yields the *maximum-likelihood* (ML) abundance estimate

$$A_m^{(p)} = \frac{1}{2} \cdot \frac{N_m^{(p)}}{1 - \bar{\ell}_m^{(p)}} \,. \tag{4}$$

This model addresses the issues raised above for a simple average of flank-specific abundances. In particular, it stems from an explicit statistical model of the unique detected fragment counts; we will return to and extend this Poissonian fragment sampling model when we construct the full statistical model of iPool-Seq.

Quantifying virulence

We return now to the goal of defining a measure of a mutant m's *virulence* v_m the reflects it's ability to proliferate on the host. As long as the total pathogen load is far from the hosts carrying capacity when the host is harvested, individual pathogen cells can be expected to proliferate approximately independently, which gives rise to a *linear* relationship between a mutant m's input abundance $A_m^{(\text{in})}$ and its output abundance $A_m^{(\text{out})}$ with a mutant-dependent proliferation factor λ_m, i.e.

$$A_m^{(\text{out})} = \lambda_m \cdot A_m^{(\text{in})}. \tag{5}$$

This linearity assumption is supported by the data as well, see figure 1 (chapter 6 on page 112).

For reference mutants $r \in \mathcal{R}$, we expect the proliferation factors λ_r to agree (at least approximately; we leave stochastic effects aside for now, these are considered when we construct the full stochastic model). We denote this common factor the *reference proliferation factor* λ, and call the *relative change* between a mutant's proliferation factor and the reference the mutant's *virulence* (v_m). Then,

$$A_m^{(\text{out})} = v_m \cdot \lambda \cdot A_m^{(\text{in})}, \tag{6}$$

and given λ a mutant's virulence is computed as

$$v_m = \frac{A_m^{(\text{out})}}{\lambda \cdot A_m^{(\text{in})}} = \frac{1}{\lambda} \cdot \frac{1 - \bar{\ell}_m^{(\text{in})}}{1 - \bar{\ell}_m^{(\text{out})}} \cdot \frac{N_m^{(\text{out})}}{N_m^{(\text{in})}}. \tag{7}$$

Given a suitable estimate for reference proliferation factor λ, this virulence scales satisfies the requirements listed earlier – in particular, it assign a virulence of (roughly) 1 to mutants in the reference set.

A suitable estimate for the reference proliferation factor λ could be found by averaging $A_m^{(\text{out})}/A_m^{(\text{in})}$ over all mutants $m \in \mathcal{R}$. However, in practice, λ is estimated together with the other parameters of the complete statistical model of iPool-Seq, see below.

A statistical model of iPool-Seq

We begin the construction of a full statistical model of iPool-Seq by returning to the idea of modelling the relationship between the abundance $A_m^{(p)}$ of mutant m in pool p (in or out), the average loss $\bar{\ell}_m^{(p)}$ and the observed unique fragment count $N_m^{(p)}$ with the Poissonian sampling model from equation 3, and combine it with the linear model for $A_m^{(\text{out})}$ given $A_m^{(\text{in})}$ from equation 6

$$\begin{aligned} N_m^{(\text{in})} &\sim \text{Poisson}\Big(2 \cdot A_m^{(\text{in})} \cdot \big(1 - \bar{\ell}_m^{(\text{in})}\big)\Big), \\ N_m^{(\text{out})} &\sim \text{Poisson}\Big(2 \cdot A_m^{(\text{out})} \cdot \big(1 - \bar{\ell}_m^{(\text{out})}\big)\Big), \\ \mathbb{E}A^{(\text{out})} &= v_m \cdot \lambda \cdot A_m^{(\text{in})}. \end{aligned} \tag{8}$$

Here, v_m is again a mutant m's virulence and λ the proliferation factor of reference mutants. By weakening the linear model for $A_m^{(\text{out})}$ given $A_m^{(\text{in})}$ to a statement about the expectation, we leave room for the inclusion of *biological noise*; by this we mean differences between mutant's true output abundances $A_m^{(\text{out})}$ that are neither caused by different input abundances, nor by their genetic background, by rather by random variations of how well mutants are able to proliferate.

We introduce an additional model parameter d for the amount of biological noise that afflicts $A_m^{(\text{out})}$; to make d easier to interpret, the parameter is defined to be the *squared biological coefficient of variation* (biological CV^2); it measures the additional variance of $A_m^{(\text{out})}$ in units of squared expected value of $A_m^{(\text{out})}$, which means the total $\mathbb{V}A_m^{(\text{out})}$ of $A_m^{(\text{out})}$ now has two components,

$$\mathbb{V}A_m^{(\text{out})} = \underbrace{v_m^2 \cdot \lambda^2 \cdot \mathbb{V}A_m^{(\text{in})}}_{\text{scaled variance of } A_m^{(\text{in})}} + \underbrace{d \cdot \big(\mathbb{E}A_m^{(\text{out})}\big)^2}_{\text{biological noise}} \tag{9}$$

To remove the unknown true input abundance $A_m^{(\text{in})}$ from our model, we view the Poissonian sampling model for $N_m^{(\text{in})}$ from equation 8 from a Bayesian angle, and exploit the well-known fact that the Gamma distribution is the conjugate prior of the Poisson distribution. More specifically, if we assume a Gamma(α, β) distribution[1](where α, β are hyper-parameters) as the *prior* distribution on the

[1]We refer to the parametrisation with shape α, rate β and density $f_\Gamma(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$.

true input abundance[2], the *posterior* distribution of $A_m^{(\text{in})}$ after observing $N_m^{(\text{in})}$ is[3,4]

$$A_m^{(\text{in})} \,\big|\, N_m^{(\text{in})} \;\sim\; \text{Gamma}\big(\alpha', \beta'\big), \tag{10}$$

where

$$\begin{aligned}\alpha' &= \alpha + N_m^{(\text{in})},\\ \beta' &= (\beta+1)\cdot\big(1-\bar{\ell}_m^{(\text{in})}\big)\cdot 2.\end{aligned}$$

From the posterior distribution of the true input abundance $A_m^{(\text{in})}$ we find the posterior of the true output abundance $A_m^{(\text{out})}$ by modifying the Gamma distribution parameters to scale the mean by $v_m \cdot \lambda$ (equation 8) and increase the variance by d times the squared mean (equation 9), which yields[5]

$$A_m^{(\text{out})} \,\big|\, N_m^{(\text{in})} \;\sim\; \text{Gamma}\big(\alpha'', \beta''\big), \tag{11}$$

where

$$\begin{aligned}\alpha'' &= \frac{\alpha'}{1+d\cdot\alpha'} = \frac{\alpha + N_m^{(\text{in})}}{1+d\cdot\big(\alpha+N_m^{(\text{in})}\big)},\\ \beta'' &= \frac{1}{v_m\cdot\lambda}\cdot\frac{\beta'}{1+d\cdot\alpha'} = 2\cdot\frac{1-\bar{\ell}_m^{(\text{in})}}{v_m\cdot\lambda}\cdot\frac{\beta+1}{1+d\cdot\big(\alpha+N_m^{(\text{in})}\big)}.\end{aligned}$$

Finally, we translate the conditional distribution of $A_m^{(\text{out})}$ given an observed input fragment count $N_m^{(\text{in})}$ into the condition distribution of $N_m^{(\text{out})}$. Under the Poissonian sampling model from equation 8, the conditional distribution of $N_m^{(\text{out})}$ is a *mixture* of Poisson distribution with a Gamma(α''', β''')-distributed rate,

$$\begin{aligned} N_m^{(\text{out})} \,\big|\, A_m^{(\text{out})} \;&\sim\; \text{Poisson}\Big(2\cdot\big(A_m^{(\text{out})}\cdot\big(1-\bar{\ell}_m^{(\text{out})}\big)\big)\Big),\\ 2\cdot A_m^{(\text{out})}\cdot\big(1-\bar{\ell}_m^{(\text{out})}\big) \;&\sim\; \text{Gamma}\big(\alpha''', \beta'''\big),\end{aligned} \tag{12}$$

[2]for technical reasons, we put the prior on $\mathbb{E}N_m^{(\text{in})} = 2\cdot A_m^{(\text{in})}\cdot\big(1-\bar{\ell}_m^{(\text{in})}\big)$, not on $A_m^{(\text{in})}$ (eq. 8).

[3]*Update rule*: If $N \mid A \sim \text{Poisson}(A)$ with prior distribution $A \sim \text{Gamma}(\alpha, \beta)$ before observing N, then the posterior distribution of A after having observed N is $A \mid N \sim \text{Gamma}(\alpha+N, \beta+1)$.

[4]*Scaling rule*: If $X \sim \text{Gamma}\big(\alpha, \beta\big)$ then $sX \sim \text{Gamma}\big(\alpha, \beta/s\big)$

[5]We use that if Gamma(α,β) has mean $\mu = \frac{\alpha}{\beta}$, variance $\sigma^2 = \alpha/\beta^2$ then Gamma$(\alpha/\nu, \beta/\nu)$ for $\nu = (1+d\alpha)$ has also mean $(\alpha/\nu)/(\beta/\nu) = \mu$ but variance $(\alpha/\nu)/(\beta^2/\nu^2) = \sigma^2(1+d\alpha) = \sigma^2 + d\mu^2$.

where

$$\alpha''' = \frac{\alpha + N_m^{(\text{in})}}{1 + d \cdot \left(\alpha + N_m^{(\text{in})}\right)},$$
$$\beta''' = \frac{1}{v_m \cdot \lambda} \cdot \frac{1 - \bar{\ell}_m^{(\text{in})}}{1 - \bar{\ell}_m^{(\text{out})}} \cdot \frac{\beta + 1}{1 + d \cdot \left(\alpha + N_m^{(\text{in})}\right)},$$

which makes the conditional distribution of $N_m^{(\text{out})}$ *negative binomial* (NB)[6],

$$N_m^{(\text{out})} \,\big|\, N_m^{(\text{in})} \sim \text{NegBin}(r, p), \tag{13}$$

where

$$r = \alpha''' = \frac{\alpha + N_m^{(\text{in})}}{1 + d \cdot \left(\alpha + N_m^{(\text{in})}\right)},$$
$$p = \frac{1}{1 + \beta'''}.$$

In our application, the common parametrisation of the negative binomial distribution with success probability p and number of failures r (which can be non-integral) is not particularly convenient, and it hides the connection between the (conditional) mean of $N_m^{(\text{out})}$ and our definition of the virulence v_m of a mutant from equation 5. The latter connection is more clearly visible if the negative binomial distribution is parametrised by its mean $\mu = \frac{rp}{1-p}$ and the number of failures r, which for the conditional distribution $N_m^{(\text{out})}$ gives

$$N_m^{(\text{out})} \,\big|\, N_m^{(\text{in})} \sim \text{NegBin}(\mu, r), \tag{14}$$

where

$$\mu = \frac{r \cdot p}{1 - p} = \frac{r}{\beta'''} = v_m \cdot \lambda \cdot \frac{1 - \bar{\ell}_m^{(\text{out})}}{1 - \bar{\ell}_m^{(\text{in})}} \cdot \frac{\alpha + N_m^{(\text{in})}}{\beta + 1},$$
$$r = \alpha''' = \frac{\alpha + N_m^{(\text{in})}}{1 + d \cdot \left(\alpha + N_m^{(\text{in})}\right)}.$$

[6] We use that if $X \sim \text{Poisson}(\Lambda)$ and $\Lambda \sim \text{Gamma}(\alpha, \beta)$, then $X \sim \text{NB}(r, p)$, $r = \alpha$, $p = 1/(1+\beta)$.

It remains to choose hyper-parameters α, β. Under a strictly Bayesian view of the model, only choices $\alpha, \beta > 0$ are possible – otherwise, our Gamma(α, β) prior on $A_m^{(\text{out})}$ is *improper* (meaning not a proper probability distribution). But any such choice introduces a bias towards a certain input abundances (no uniform proper prior on the whole positive real line is possible!), and causes disagreement between the negative binomial model in equation 14 and a natural re-interpretation of the virulence definition from equation 5 as a statement about the conditional expectation of $N_m^{(\text{out})}$ given $N_m^{(\text{in})}$. Both issues are resolved if the requirement that the prior is a proper probability distribution is dropped and α, β are set to zero. All posterior distribution, and in particular $N_m^{(\text{out})} \,|\, N_m^{(\text{in})}$, remain valid distributions as long as $N_m^{(\text{in})} > 0$, i.e. for all mutants that are detected as being present in the input pool. This choice of hyper-parameters yields the simplified negative binomial (NB) model for iPool-Seq

$$N_m^{(\text{out})} \,|\, N_m^{(\text{in})} \sim \text{NegBin}\left(\mu := v_m \cdot \lambda \cdot \frac{1-\bar{\ell}_m^{(\text{out})}}{1-\bar{\ell}_m^{(\text{in})}} \cdot N_m^{(\text{in})}, \quad r := \frac{N_m^{(\text{in})}}{1 + d \cdot N_m^{(\text{in})}}\right). \tag{15}$$

which for $v_m = 1$ models the behaviour of *reference mutants*.

Variance components of $N_m^{(\text{out})}$

In the (conditional) negative binomial model (equation 15) for $N_m^{(\text{out})}$, the mean and (total) variance of $N_m^{(\text{out})}$ given $N_m^{(\text{in})}$ are[7]

$$\begin{aligned} \mu_m &= \mathbb{E}\left(N_m^{(\text{out})} \,\middle|\, N_m^{(\text{in})}\right) = v_m \cdot \lambda \cdot \frac{1-\bar{\ell}_m^{(\text{out})}}{1-\bar{\ell}_m^{(\text{in})}} \cdot N_m^{(\text{in})} \\ \sigma_m^2 &= \mathbb{V}\left(N_m^{(\text{out})} \,\middle|\, N_m^{(\text{in})}\right) = \mu_m + \mu_m^2 \cdot \frac{1 + d \cdot N_m^{(\text{in})}}{N_m^{(\text{in})}}. \end{aligned} \tag{16}$$

[7] We use that if $X \sim \text{NB}(\mu, r)$, then $\mathbb{V}\, X = \mu + \mu^2/r$ (for NB parametrised with μ, r, see eq. 14)

and the squared *coefficient of variation* (CV) of $N_m^{(\text{out})}$ given $N_m^{(\text{in})}$, which reflects the variance but expressed in units of squared expectation, is thus

$$\text{CV}_m^2 = \frac{\sigma_m^2}{\mu_m^2} = \frac{1}{N_m^{(\text{in})}} \quad + \quad \frac{1}{\mu_m} \quad + \quad d. \tag{17}$$

This partitioning of CV_m^2 shows that there are three (independent) source of variance that together comprise the total variance of $N^{(\text{out})}$:

Non-exhaustive sampling of the input pool $\left(1/N_m^{(\text{in})}\right)$**.** Under our Poissonian sampling model, the fewer unique fragments from a particular mutant are observed in the input pool, the larger the (relative) uncertainty about its true input abundance $A_m^{(\text{in})}$.

Non-exhaustive sampling of the output pool $\left(1/\mu_m\right)$**.** Again due to our Poissonian sampling model, the smaller the expected number of unique fragments from a particular mutant is in the output pool, the larger are the expected (relative) variations of the observed fragment count.

Biological CV^2 $\left(d\right)$**.** Added explicitly to the model to account for *biological noise*, meaning random variations of mutant growth on the host.

Estimating parameters, computing virulences and testing significance

We start from the basic read-out of an iPool-Seq experiment in the form of unique fragment counts $N_{m,f}^{(p)}$ and TRUmiCount loss estimates $\ell_{m,f}^{(p)}$ for every combination of mutant m, pool p (in or out), and flank f (5' or 3'). We compute per-mutant virulences v_m and *false discovery rate* (FDR; Benjamini and Hochberg, 1995) - corrected p-values $q_m^{(\text{low})}$ and $q_m^{(\text{high})}$, indicating the significance of $v_m <$ (lower virulence than reference) and $v_m > 1$ (higher virulence than reference), by executing the following procedure:

1. Combine flank-specific (5' and 3') unique fragment counts and TRUmiCount loss estimates into a single count $N_m^{(p)}$ and a single loss estimate $\bar{\ell}_m^{(p)}$ per

mutant m and pool p (equation 2) and compute input ($A_m^{(\text{in})}$) and output ($A_m^{(\text{out})}$) - pool abundances (equation 4).

2. Estimate *maximum likelihood* (ML) parameters for the *reference proliferation factor* (λ) and *biological* CV^2 (d) by optimising the combined likelihood of all observed unique fragment counts $N_m^{(\text{out})}$ for *reference mutants* $m \in \mathcal{R}$ under the negative binomial model (equation 15; set $v_m = 1$ since $m \in \mathcal{R}$).

3. Compute each mutant's *virulence* v_m from its observed unique fragment counts and estimated losses, using the ML estimate for the reference proliferation factor λ (equation 7).

4. For each mutant m that was detected as present in both the input and the output pool, compute two p-values $p_m^{(\text{low})}$ and $p_m^{(\text{high})}$ for the significance (under the negative binomial model, equation 15), of a decreased (low; $v_m < 1$) respectively an increased (high; $v_m > 1$) virulence compared to the mutants in the reference set $\mathcal{R}$,

$$\begin{aligned} p_m^{(\text{low})} &= F_{\text{NB}}\big(N_m^{(\text{out})}; \mu_m, r_m\big), \\ p_m^{(\text{high})} &= 1 - F_{\text{NB}}\big(N_m^{(\text{out})} - 1; \mu_m, r_m\big). \end{aligned} \tag{18}$$

where

$$\mu_m = \lambda \cdot \frac{1 - \bar{\ell}_m^{(\text{out})}}{1 - \bar{\ell}_m^{(\text{in})}} \cdot N_m^{(\text{in})}, \qquad r_m = \frac{N_m^{(\text{in})}}{1 + d \cdot N_m^{(\text{in})}}$$

and $F_{\text{NB}}(x; \mu, r)$ denotes the *cumulative distribution function* (CDF) of the negative binomial distribution, parametrised with mean μ and (not necessarily integral) failure count r.

These p-values reflect the probability of detecting fewer or equally many ($p_m^{(\text{low})}$) respectively more or equally many ($p_m^{(\text{high})}$) than $N_m^{(\text{out})}$ unique fragments in the output pool for a mutant with $v_m = 1$ (reference virulence) assuming $N_m^{(\text{in})}$ detected fragment in the input pool, taking the estimates fractions $\bar{\ell}_m^{(\text{in})}$, $\bar{\ell}_m^{(\text{out})}$ of lost fragments into account. Note that for reasonably large $N_m^{(\text{out})}$, $p_m^{(\text{high})} \approx 1 - p_m^{(\text{low})}$ because the subtraction of 1 from $N_m^{(\text{out})}$ in the computation of $p_m^{(\text{high})}$ then has little effect.

5. Apply the *Benjamini-Hochberg* (BH; Benjamini and Hochberg, 1995) correction separately to the sets of p-values for reduced and increased virulence to obtain FDR-corrected p-values $q_m^{(\text{low})}$ and $q_m^{(\text{high})}$. Then filter the list of mutants according to $q_m^{(\text{low})} \leq \text{FDR}$ respectively $q_m^{(\text{high})} \leq \text{FDR}$ to find a list of genes whose removal impacts the pathogen's virulence negatively respectively positively, containing (on average) no more than a fraction of FDR *false positives*.

The asymptotic power of iPool-Seq

We consider the case of large input fragment counts $N_m^{(\text{in})}$, $\lambda = 1$, and negligible losses $\bar{\ell}_m^{(\text{in})} \approx \bar{\ell}_m^{(\text{out})} \approx 0$. The output fragment counts $N_m^{(\text{out})}$ are thus also large (which is the point of these assumptions; neither $\lambda = 1$ nor the negligible losses themselves are material). For the Gamma(α'', β'')-distribution of $A_m^{(\text{out})}$, equation 11 (where we again set $\alpha = \beta = 0$ for the hyper-parameters) then yields $\alpha'' \approx d^{-1}$ and $\beta'' \approx \left(d v_m N_m^{(\text{in})}\right)^{-1}$, and thus the following approximation[8]

$$A_m^{(\text{out})} \approx \text{Gamma}\left(\frac{1}{d}, \frac{2}{d \cdot v_m \cdot N_m^{(\text{in})}}\right). \tag{19}$$

This approximate distribution of $A_m^{(\text{out})}$ has, as one should expect, mean value $\mu_m'' = \alpha''/\beta'' = v_m N_m^{(\text{in})}/2$ and variance $(\sigma_m'')^2 = d(\mu_m'')^2$. The distribution of $N_m^{(\text{out})}$ is mixture of Poisson distributions whose mean is distributed according to $2 \cdot A_m^{(\text{out})}$ (equation 12), and whose variance is, being Poisson, equal to the mean. For large $N_m^{(\text{in})}$ (and fixed d), almost all of these Poisson distributions thus are, relative to the variance $d(\mu_m'')^2$ of $A_m^{(\text{out})}$, essentially concentrated at a single point, making the mixture moot and the distribution of $N_m^{(\text{out})}$ the same as that of of $2 \cdot A_m^{(\text{out})}$, meaning

$$N_m^{(\text{out})} \approx \text{Gamma}\left(\frac{1}{d}, \frac{1}{d \cdot v_m \cdot N_m^{(\text{in})}}\right), \tag{20}$$

with mean $\mu_m = v_m N_m^{(\text{in})}$ and $\sigma_m^2 = d\mu_m^2$.

[8] By *approximation* we here mean convergence in distribution for $N_m^{(\text{in})} \to \infty$; possibly after shifting and re-scaling to ensure that the (distributional) limit exists.

We can remove the dependency on $N_m^{(\text{in})}$ by considering the observed *virulence* $\hat{v}_m = N_m^{(\text{out})}/N_m^{(\text{in})}$ (eq. 7; remember $\lambda = 1$ and losses are negligible) instead,

$$\hat{v}_m = \frac{N_m^{(\text{out})}}{N_m^{(\text{in})}} \approx \text{Gamma}\left(\frac{1}{d}, \frac{1}{d \cdot v_m}\right). \tag{21}$$

Given a particular false-positive rate (or significance level)[9] φ, a virulence $v_m < 1$ and biological CV^2 parameter d, we can then proceed as follows to compute the (approximate) power κ for recognising a mutant's virulence to be reduced:

First, we compute the *critical value* t_φ, meaning the largest observed variance $\hat{v}_m$ that is still *significantly smaller than 1* at significance level φ,

$$t_\varphi = F_\Gamma^{-1}\left(\varphi\,;\, \frac{1}{d}, \frac{1}{d}\right), \tag{22}$$

where $F_\Gamma(\cdot\,; \alpha, \beta)$ is the CDF of the Gamma distribution with shape α and rate β and F_Γ^{-1} its inverse (i.e. the quantile function).

Then we compute the probability of a mutant with true v_m producing an observed virulence of at most t_φ,

$$\kappa = \mathbb{P}(\hat{v}_m \leq t_\varphi) = F_\Gamma\left(t_\varphi\,;\, \frac{1}{d}, \frac{1}{d \cdot v_m}\right). \tag{23}$$

Table 1 (chapter 6 on page 121) shows the power for a few combinations of virulence, biological CV^2 and significance level.

[9] Since α is already taken, we denote the significance level or false-positive rate φ

		Sig. level 0.05					Sig. level 0.01			
d	v_m=	0.75	0.5	0.25	0.125	v_m=	0.75	0.5	0.25	0.125
0.001		1.00	1.00	1.00	1.00		1.000	1.00	1.00	1.00
0.01		0.89	1.00	1.00	1.00		0.680	1.00	1.00	1.00
0.1		0.19	0.64	1.00	1.00		0.054	0.32	0.97	1.00

Table 1: **The power of iPool-Seq.**. Power to detect significant reductions of mutant m's virulence to v_m for different biological CV^2 (d) and significance levels.

Results

We ran the pipeline of Uhse *et al.* (2019; chapter 4 of this work) on the raw sequencing reads for the 6 replicates of an iPool-Seq experiment published by Uhse *et al.* (2018; chapter 3 of this work). The pipeline transforms the raw sequencing reads of each replicate into pool (p), mutant (m) and flank (f) - specific unique fragment counts $N^{(p)}_{m,f}$ and corresponding TRUmiCount loss estimates $\ell^{(p)}_{m,f}$, and then applies the procedure outlined in the section *Estimating parameters, computing virulences and testing significance* (page 118). The results of that procedure are replicate-specific estimates of the two model parameters λ (reference proliferation factor) and d (biological CV^2), as well as replicate- and mutant (m) - specific estimates of input abundance $A^{(\mathrm{in})}_m$, output abundance $A^{(\mathrm{out})}_m$, and virulence v_m, and finally p-values $p^{(\mathrm{high})}_m$ and $p^{(\mathrm{low})}_m$ and *false discovery rate* (FDR) - corrected q-values $q^{(\mathrm{high})}_m$ and $q^{(\mathrm{low})}_m$ that quantify the significance of $v_m > 1$ respectively $v_m < 1$. Virulence $v_m = 1$ reflects the behaviour of the mutants in the reference set; the reference set used by Uhse *et al.* (2018) contains 5 single-gene knockout mutants of the genes UMAG_01302, UMAG_02193, UMAG_03202, UMAG_10403, UMAG_10553. We note that since the pipeline used by Uhse *et al.* (2018) relied on an older version of the TRUmiCount algorithm, and also differed in some other details from the version used to generate the data shown here, the data shown here differs slightly from the results presented here.

As already briefly discussed when the linear model for the relationship of the

	Replicate					
Parameter	A1	A2	A3	B1	B2	B3
λ	0.55	0.63	0.44	0.067	0.043	0.012
d	$1.8 \cdot 10^{-2}$	$2.5 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$	$1.4 \cdot 10^{-2}$	$2.4 \cdot 10^{-2}$	$1.9 \cdot 10^{-9}$

Table 2: **Model parameter estimates for experimental data.** Maximum-likelihood (ML) estimates for reference proliferation factor (λ) and biological CV^2 (d). Shown are the results from replicates 1 to 3 of experiments A and B of Uhse *et al.* (2018; chapter 3 of this work). Results were computed with the pipeline of Uhse *et al.* (2019; chapter 4 of this work).

expected values of $A_m^{(\text{in})}$ and $A_m^{(\text{out})}$ was introduced (equations 5 and 6), the data of Uhse *et al.* supports the assumption of linearity well (figure 1 in chapter 6 on page 112). In particular, even for most abundant mutants in the input pool, the output pool abundances show no trace of a plateau effect that could be point to the carrying capacity of the host being reached. This is crucial – any plateau effect would cause the model to either systematically under-estimate the virulence of highly abundant mutants in the input pool, or over-estimate the virulence of rarer mutants.

While the linear relationship of the expectations of $A_m^{(\text{in})}$ and $A_m^{(\text{out})}$ is defined by the model parameter called *reference proliferation factor* (λ), the other model parameter (d) called *biological* CV^2 defines the relationship of their variances. This parameter measures the additional variance of $A_m^{(\text{out})}$ over $A_m^{(\text{in})}$ due to random growth differences between mutants; and does so in units of squared expectation. With the exception of replicate B3 (which appears to be an outlier), the estimates of d lie within a factor of 2 between the smallest ($1.4 \cdot 10^{-2}$, replicate A2) and the largest ($2.5 \cdot 10^{-2}$, replicate A3), see table 2 (chapter 6 on page 122). This is despite the fact that the range of output abundances varies by at least an order of magnitude between these replicates – for replicates A1-3, the bulk of output abundances range roughly from 10^3 to 10^5, for replicates B1-B3 they range from $10^2 - 10^4$ and for replicate B3 they range from $10^1 - 10^3$ (y-axis of figure 1 in chapter 6 on page 112).

The estimated virulences v_m show for most of the 6 replicates a clearly asym-

	Replicate					
Direction	A1	A2	A3	B1	B2	B3
Increase ($q_m^{(\text{high})} \leq \alpha$)	2	2	72	7	2	45
Decrease ($q_m^{(\text{low})} \leq \alpha$)	90	70	32	56	36	45

Table 3: **Number of mutants with significant virulence phenotype.** Number of mutants whose virulence deviation significantly from the reference set, meaning $q_m^{(\text{high})} \leq \alpha$ or $q_m^{(\text{low})} \leq \alpha$. Shown are the results for replicates 1 to 3 of experiments A and B of Uhse *et al.* (2018; chapter 3 of this work). Results were computed with the pipeline of Uhse *et al.* (2019; chapter 4 of this work).

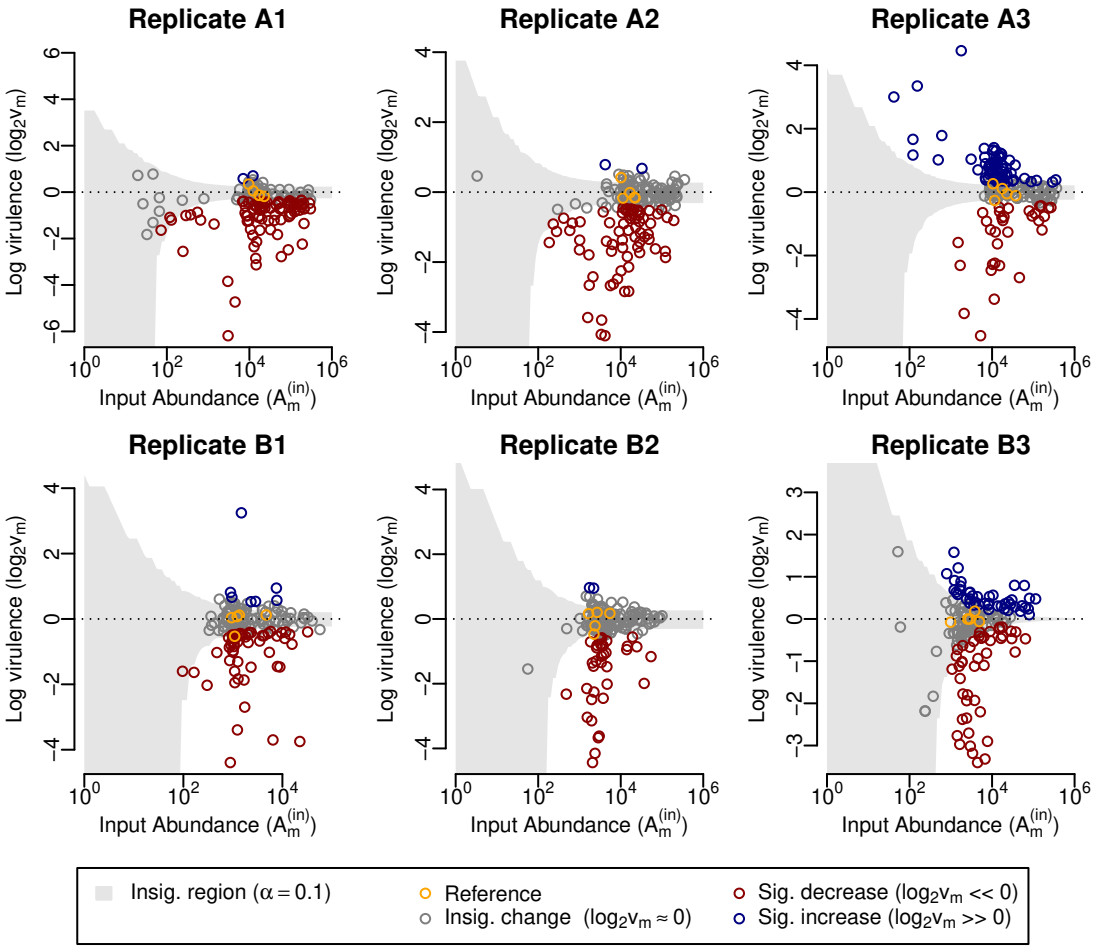

Figure 2: **Virulence of mutants compared to the reference set.** Virulence ($\log_2 v_m$) vs. input input abundance ($A_m^{(\text{in})}$). Grey area shows the (approximate; no FDR-correction and actual area depends on $\ell_m^{(\text{in})}$, $\ell_m^{(\text{out})}$) range of ($\log_2$) virulences for a particular $A_m^{(\text{in})}$ that do *not* significantly deviate from $v_m = 1$ (reference behaviour). Mutants are coloured according to their classification (classified with FDR threshold $\alpha = 0.1$). Shown are the results from replicates 1 to 3 of experiments A and B of Uhse *et al.* (2018; chapter 3 of this work). Results were computed with the pipeline of Uhse *et al.* (2019; chapter 4 of this work).

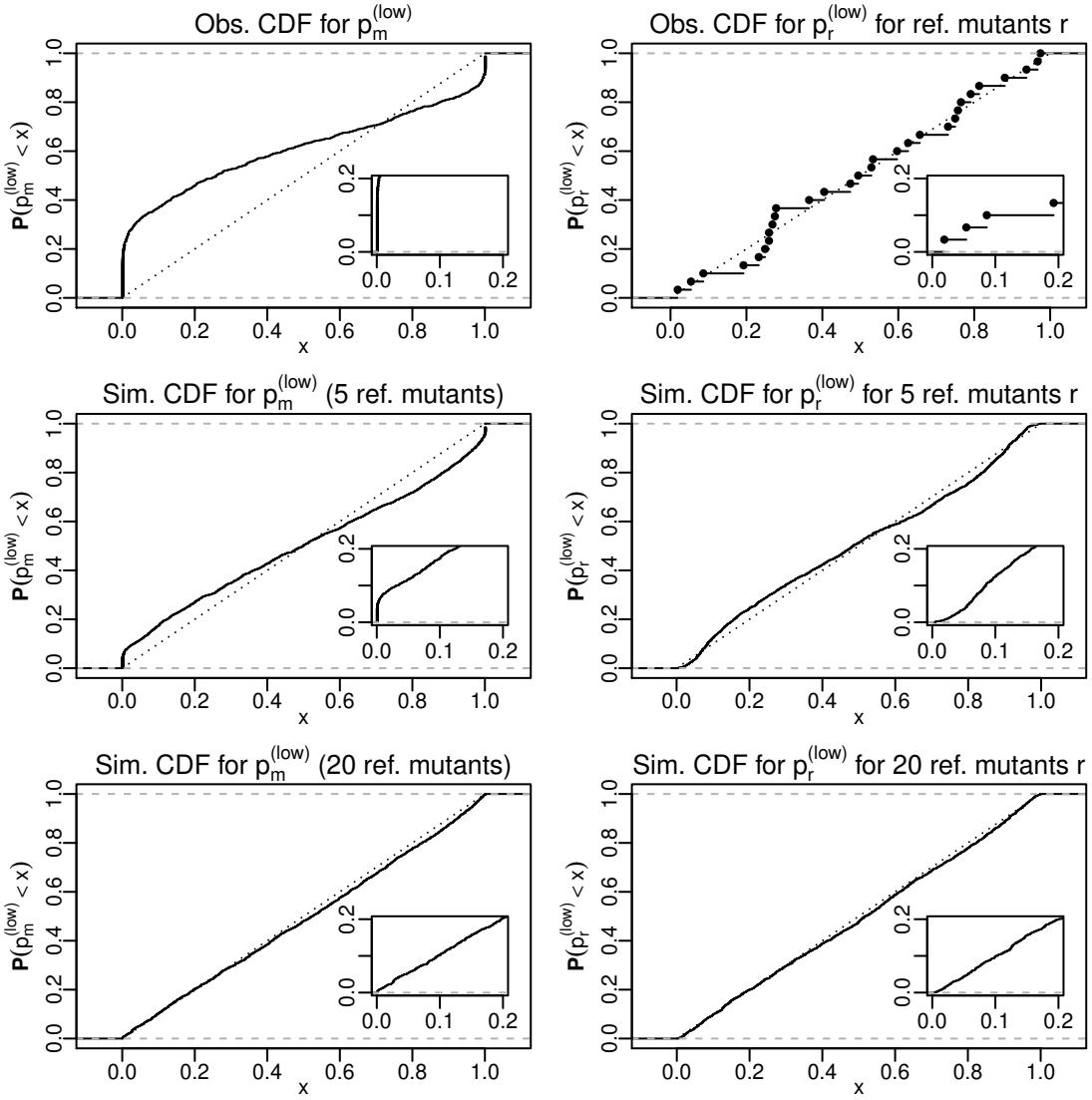

Figure 3: **Distribution of p-values.** First row shows the *cumulative distribution function* (CDF) of $p_m^{(\text{low})}$ across the 6 replicates of Uhse *et al.* (2018; chapter 3 of this work), model model fitted separately to each replicate. The CDFs in the left column include p-values for all ≈ 200 mutants, the CDFs on the right only p-values for the 5 reference mutants. Second and third row show the CDF obtained from 500 *simulated* iPool-Seq replicates with 5 (second row) respectively 20 (third row) reference mutants per replicate. Simulations were based on the model in equation 15; $v_m = 1$ for all simulated mutants, input abundances $A_m^{(\text{in})}$ log-uniformly distributed on $[10, 10^5]$, loss estimates beta-distributed with $\alpha = \beta = 10$, input fragment counts Poisson-distribution as in equation 3. Results for experimental data where computed with the pipeline of Uhse *et al.* (2019; chapter 4 of this work).

metric distribution; a decrease in virulence compared to the reference mutants is more common than an increase (figure 2 in chapter 6 on page 124; table 3 in chapter 6 on page 123). The exceptions are replicates A3 and B3, which classify 72 respectively 45 mutants as having an increased virulence, whereas the highest number of mutants with that classification across the other replicates is 7.

For replicate B3, the likely reason for the inflated number of mutants classified as having a significant increases in virulence is the atypically small biological CV^2 estimate $d = 1.9 \cdot 10^{-9}$ (table 2 in chapter 6); and while the estimate $d = 1.5 \cdot 10^{-2}$ for replicate A3 is not as low, it still lies on the low end of the scale.

Other replicates also show signs of an inflated false-discovery rate, though, both for detected virulence increases as well as decreases. Out of ≈ 200 mutants, 107 are classified as having an increased virulence, and 149 mutants as having a decreased virulence in at least one replicate at an FDR threshold of $\alpha = 0.1$; but if we require three or more replicates to classify a mutant as significant, no mutants remain whose virulence is classified as increased, and only 40 of the initial 149 remain classified as decreased (table 4 in chapter 6 on page 126).

The p-value distribution of $p_m^{(\text{low})}$ tells a similar tale – the observed CDF of $p_m^{(\text{low})}$ across all mutants and replicates A1-B3 shows a clear over-abundance of both p-values close to zero (about 20% ≈ 40 mutants) and close to one (about 10% ≈ 20 mutants) compared to a uniform distribution (left upper plot of figure 3 in chapter

	No. of mutants with sig. ($\alpha = 0.1$) virulence phenotype in $\geq k$ replicates					
Direction	$k =$ 1	2	3	4	5	6
Increase ($q_m^{(\text{high})} \leq \alpha$)	107	23	0	0	0	0
Decrease ($q_m^{(\text{low})} \leq \alpha$)	149	95	40	21	16	8

Table 4: **Reproducibility of significant virulence changes.** The number of mutates that showed a significant ($\alpha = 0.1$) virulence increase ($q_m^{(\text{high})} \leq \alpha$) respectively decrease ($q_m^{(\text{low})} \leq \alpha$) in at least $k = 1, 2, \ldots, 6$ replicates. Shown are the results from replicates 1 to 3 of experiments A and B of Uhse *et al.* (2018; chapter 3 of this work). Results were computed with the pipeline of Uhse *et al.* (2019; chapter 4 of this work).

6 on page 125). Roughly half of the ≈ 40 mutants with exceptionally small p- are likely mutants with an actual reduction in virulence – 21 mutants were classified as reduced virulence in 4 out of 6 replicates (FDR threshold $\alpha = 0.1$); requiring 5 replicates reduces the number to 16 (table 4 in chapter 6 on page 107). The ≈ 20 mutants with p-values $p_m^{(\mathrm{low})}$ exceptionally close to one (which corresponds to $p_m^{(\mathrm{high})}$ being exceptionally close to zero; $p_m^{(\mathrm{high})} \approx 1 - p_m^{(\mathrm{low})}$, see equation 18) on the other hand cannot be explained by their virulence being truly increased – no mutants are reproducibly classified as having increased virulence in more than 2 out of 6 replicates (table 4 in chapter 6 on page 107).

A possible source of the unexplained over-abundances of p-values close to zero and close to one is model over-fitting – in particular, because the reference set used by Uhse *et al.* (2018; chapter 3 of this work) contains only 5 mutants, which are used to estimate two model parameters (λ and d).

Simulations corroborate the small size of the reference set and the ensuing over-fitting as the reason behind the inflated false-discovery rate. For simulated iPool-Seq experiment with 5 reference mutants from which model parameters are inferred, the over-abundance of p-values $p_m^{(\mathrm{low})}$ close to one mimics the one observed for experimental data, and like for experimental data no (strong) deviation form a uniform distribution is observed within the p-values of the reference mutants. P-values close to zero are less over-abundant that for experimental data; this is to be expected since all simulated mutants have true virulence $v_m = 1$ (middle row of 3 in chapter 6 on page 125).

Finally, if the number of reference mutants per experiment is increased to 20, the p-value distribution of $p_m^{(\mathrm{low})}$ becomes uniform – indicating that for about 20 mutants in the reference set, the false-discovery rate would no longer be inflated compared to the chosen false discovery rate target.

Discussion

We have presented a model for count data that models *asymmetric* designs where the abundances of particular *species* (e.g. mutants of a pathogen, or transcripts of a particular gene) in a population are sampled before and after some event (such as infecting the host), and which allows classifying species by whether they cope *worse*, *equally well* or *better* with that event than some reference group of species.

This is different from common count data models in *next-generation sequencing* (NGS) applications like the models used by *edgeR* (M. D. Robinson, McCarthy, *et al.*, 2010), *DESeq* (Anders and Huber, 2010) and *DESeq2*(Love, Huber, *et al.*, 2014) to call differentially expressed genes in *RNA Sequencing* (RNA-seq) data. These models usually are *symmetric*, and compare species- or gene-specific counts between two samples for significant differences. These methods typically *do not* use a specific *reference group* to define a baseline – instead, they often use either the average, the majority response or a combination of both as a baseline (M. D. Robinson and Oshlack, 2010). For differential expression analysis of RNA-Seq experiments this usually works well because we often expect only a subset of genes to be differentially expressed between e.g. two tissues. But in mutant screens such as iPool-Seq experiments, we in many cases cannot now a priori whether just a few or almost all of the mutants will show a phenotype, and the referenced-based baseline setting of our model is thus advantageous (provided an appropriate reference group is included in the screen).

When we applied our model to the 6 replicates of Uhse *et al.* (2018; chapter 3 of this work), the individual replicates showed strong signs of over-fitting. This is not surprising given the small number of reference mutants (5) compared to the number of model parameters (2). The over-fitting could be alleviated by taking the variance of the estimates of d and λ into account and "broadening" the conditional distribution of $N_m^{(\mathrm{out})}$ further – similar in spirit to how a t-test uses Student's t-distribution in place of the normal distribution to account for the uncertainty of sample variance as an estimate of the true variance.

However, given the reduction in power such a modification would entail, it is not clear that it would be worth-while. Since a modest increase of the number of reference mutants from 5 to 20 virtually removes the observed over-fitting,

designing further experiments to include more reference mutants see preferential.

In fact, we must mention that the reference set used by Uhse *et al.* (2018; chapter 3 of this work) initially contained 13 mutants, selected based on their description as showing wild-type behaviour in the literature. But initial analysis of the data cast doubt on the neutral behaviour of many of them, and the set was reduced *after the fact* to only those 5 mutants that clearly showed no phenotype. While this clearly had an adverse effect on the false-positive rate, at the same time it ensured that we would not inflate the false-negative rate by using non-neutral mutants as a baseline. In the context of a screen, such trading of a higher false-positive rate in favour of a lower false-negative rate was deemed preferential, as it avoids erroneously removing mutants from the list of interesting candidates to be studied further.

Except for the over-fitting, the model seems to work well for its intended purpose, and managed to identify 21 mutants which, showing a significant deviation from reference behaviour in 4 out of 6 replicates, can be considered to have a virulence phenotype with high confidence. This was also independently confirmed by validation experiments – 3 novel genes with a virulence phenotype were tested in an independent virulence essay and found to cause much weaker symptoms of infection on the host plant than the wild-type (Uhse et al., 2018; figure 5 in chapter 3 of this work on page 52).

Outlook

The most delicate modelling assumption we make is the linear relationship between input and output abundances of mutants – while small violations of a model's assumptions often have little consequence on the fidelity of the results, linearity violations here translate to strong systematic biases against particular mutants. While the *U. maydis*-maize system seems to adhere quite strictly to the assumed linearity, other host-pathogen systems might not. To analyse experimental data from such host-pathogen systems, the model would then need to be extended to handle non-linear input-output abundance relationships.

The structure of the model makes such a modification relatively straight-forward. One would need to assume a transfer function ϑ parametrised by a mutant's

virulence v_m and additional (not mutant-specific) parameters Θ that maps input to output abundances,

$$A_m^{(\text{out})} = \vartheta\big(A_m^{(\text{in})}; \Theta, v_m\big). \tag{24}$$

To compute the virulence of a particular mutant, this equation would need to be solved for v_m, after having done the (straight-forward) translation of $N_m^{(\text{in})}$ into $A_m^{(\text{in})}$ and $N_m^{(\text{out})}$ into $A_m^{(\text{out})}$, and after having estimated parameters Θ. The estimation of Θ would again rely on likelihood maximisation – but of course for a modified stochastic model which also has the linear relationship between $A_m^{(\text{in})}$ and $A_m^{(\text{out})}$ replaced with equation 24.

Doing so exactly would require drastic changes to the model, as $A_m^{(\text{out})}$ would no longer follow a Gamma-distribution. But we can, as before, weaken equation 24 to be a statement about expectations, and thus include it into the stochastic model 8 in the form

$$\mathbb{E}A_m^{(\text{out})} = \vartheta\big(\mathbb{E}A_m^{(\text{in})}; \Theta, v_m\big). \tag{25}$$

For the variance, we modify the relationship from equation 9 by linearising ϑ,

$$\mathbb{V}A_m^{(\text{out})} = \Big(\vartheta'\big(\mathbb{E}A_m^{(\text{in})}; \Theta, v_m\big)\Big)^2 \cdot \mathbb{V}A_m^{(\text{in})} + d \cdot \big(\mathbb{E}A_m^{(\text{out})}\big)^2, \tag{26}$$

where ϑ' is the partial derivative of ϑ with respect to the input abundance, and d is an additional model parameter that, as before, represents *biological noise*.

Since variations in the input abundances typically cause much larger variations of output abundances than biological noise (i.e. random growth rate fluctuations), the linearisation in equation 26 is should not hamper this extended model's ability to accurately model host-pathogen systems that show a plateau effect due to the pathogen load of mutants highly abundance int he input reaching the hosts carrying capacity. We thus believe that the model presented here would be, with the extensions outlined above, also be applicable to such systems.

References in this chapter

Anders S., & Huber W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. DOI:10.1186/gb-2010-11-10-r106

Benjamini Y., & Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300. DOI:10.2307/2346101

Hug H., & Schuler R. (2003). Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. *Journal of theoretical biology*, *221*(4), 615–624. DOI:10.1006/jtbi.2003.3211

Kivioja T., Vähärautio A., Karlsson K., Bonke M., Enge M., Linnarsson S., & Taipale J. (2011). Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, *9*(1), 72–74. DOI:10.1038/nmeth.1778

Love M. I., Huber W., & Anders S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. DOI:10.1186/s13059-014-0550-8

Mardis E. R. (2008). Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*, *9*(1), 387–402. DOI:10.1146/annurev.genom.9.081307.164359

Mardis E. R. (2013). Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry*, *6*(1), 287–303. DOI:10.1146/annurev-anchem-062012-092628

Metzker M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, *11*(1), 31–46. DOI:10.1038/nrg2626

Robinson M. D., McCarthy D. J., & Smyth G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–40. DOI:10.1093/bioinformatics/btp616

Robinson M. D., & Oshlack A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. DOI:10.1186/gb-2010-11-3-r25

Wang Z., Gerstein M., & Snyder M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. DOI:10.1038/nrg2484

This page intentionally left blank.

Chapter 7

Conclusions

Quantitative *next-generation sequencing* (NGS; Mardis, 2008, 2013; Metzker, 2010) continues to be an important technique to gain insight into complex biological systems by separately measuring the abundances of many different types of DNA or RNA molecules; and by counting *unique molecular identifiers* (UMIs; Hug and Schuler, 2003; Kivioja *et al.*, 2011) instead of sequencing reads, the impact of *polymerase chain reaction* (PCR) bias on the method's accuracy can be reduced. However, counting UMIs instead of reads does not remove all PCR-related bias from the measured abundances – without *in silico* corrections, *indirect PCR bias* (see *Unique molecular identifies* in chapter 1 on page 24) through non-uniform risks of overlooking a molecule remains. To remove indirect PC bias *in silico*, we must understand it quantitatively – meaning we require a mathematical *theory* of how this bias arises, and on which experimental factors its magnitude depends.

A theory of UMI-based quantitative NGS

We presented such a theory, showed that it explains observed data reasonably well, and used it to develop the *TRUmiCount* algorithm which corrects UMI counts for indirect PCR bias (Pflug & von Haeseler, 2018; chapter 2 of this work). The corrected abundances reported by TRUmiCount show better concordance with known true abundances than raw UMI counts; both for simulations (Pflug & von Haeseler, 2018; chapter 2 of this work) as well as for experimental data (chapter 5).

The predictions of our theory remain valid even if its assumptions are (mildly) violated; in particular if instead of a single PCR step, multiple PCR steps with intermediate purification steps are used to enrich particular DNA fragments prior to sequencing to reduce the required sequencing depth, as *insertion pool sequencing* (iPool-Seq) does (Uhse *et al.*, 2018; chapter 3 of this work). While this model miss-specification affects the estimated model parameters, the corrections computed by our theory remain accurate enough to be usable. In particular, the correction is shown to still improve (internal) concordance over uncorrected abundance estimates (chapter 5).

While TRUmiCount improves the accuracy of molecule abundance estimates in both in a single-cell RNA-seq settings as well as for iPool-Seq data (chapter 5), some questions remain; in particular about the relationship between its two model parameters, the PCR efficiency E and the average number of reads per UMI D. Under TRUmiCount's model of PCR as a stochastic exponential growth process, we would except these two parameters to be highly correlated. Yet the data is ambiguous – while observe some correlation under some circumstances, under others we hardly observe any (figure 5 in chapter 5 on page 99 and related discussion in section *Model parameters & estimated losses* on page 103 ff.). How this fits the observation that the estimates efficiencies do anti-correlate with the amplified fragment's length (figure S2 in chapter 2 on page 41), and thus do seem to seem to measure the true efficiency of the PCR is unclear.

Possible extensions of the theory

Understanding why we don't consistently observe strong correlation of the estimated PCR efficiency E and the average number of reads per UMI D, and finding a realistic model of their relationship would benefit our theory of UMI-based quantitative NGS experiments. First, such a model might provide further insight into the reaction behaviour for different sequences. Second, it may reduce the number of gene-specific (other specific to any other genomic feature whose abundance is to be measured) parameters that have to be estimated; such a reduction would in turn reduce overfitting and thus increase the accuracy of the estimated fractions of lost molecules, in particular for genes where only little data is available.

Other approachs to reduce the number of independent parameters to estimate seem possible. Love *et al.* (2016), for example, in their *Alpine* method use a *generalised linear model* (GLM) to predict sequence- and length-specific bias across the whole genome. Through a similar model, the parameter estimates used by TRUmiCount for different genomic regions could be linked, thus once more reducing the number of independent parameters that require estimation.

A fragment bias model like the one used by Love *et al.* (2016) for Alpine, modified for TRUmiCount to predict fragment-specific PCR efficiencies and average number of reads per UMI from fragment parameters like length and sequence composition would also allow us to extend TRUmiCount to experimental methods like *chromatin-immunoprecipitation sequencing* (ChIP-Seq; Barski *et al.*, 2007; Johnson *et al.*, 2007) whose output is a *coverage graph* and not a *count table*. In such methods, reads respectively UMIs are not assigned to individual features like genes or mutants, but rather contribute to the coverage of the bases they contain. For TRUmiCount in its current form that poses a problem, because there are no natural groups of UMIs (like those assigned to a single gene) for which model parameters could be estimated individually; with an Alpine-inspired fragment bias model we could instead compute a loss estimate for each individual UMI which would indicate how many similar UMIs remained unobserved.

Any approach that avoids having to estimate independent PCR efficiencies for each gene would also make it possible to model more complex experimental situations without giving up identifiability of the parameters. For example, to model a protocol like iPool-Seq that consists of two PCR steps with an intermediate purification step we can proceed as follows:

To model each UMI's copy number after the first PCR and subsequent purification, we use the same mixture of Poisson distributions as does TRUmiCount; the Poisson distribution now reflects the (lossy) purification process instead of sequencing. Writing M_{pur} for the number of copies after purification, $P = \mathbb{E}\, M_{\text{pur}}$ for the average number of copies, and E_1 for the (first) PCR's efficiency, we thus have

$$F_1 \sim \text{PCR}(E_1, 1), \qquad M_{\text{pur}} \mid F_1 \sim \text{Poisson}(F_1 \cdot P) \tag{1}$$

where $\text{PCR}(E, I)$ stands for the (normalized) family size distribution after a PCR

with efficiency E starting from I initial copies. M_{pur} is then the initial copy number for the second PCR; after the second PCR (with efficiency E_2) and sequencing (with D reads per UMI on average), the read count C should then obey (c.f. Pflug & von Haeseler, 2018; chapter 2 of this work, see section *2.2.3 Modeling the sequencing process* on page 33)

$$F_2 \sim \text{PCR},(E_2, M_{\text{pur}}), \qquad C \mid F_2 \sim \text{Poisson}(F_2 \cdot D). \tag{2}$$

The difficulty of this model is not its mathematical treatment, but rather the estimation of E_1, E_2 and P – all these three parameters together define the observed variance of C, making robust independent estimation from limited data impossible. Parameters of such multi-stage models might, however, become identifiable if they are linked so that instead of having to estimate separate parameter for separate loci, all UMIs in a sequencing library become informative for every locus.

Considering such mixtures also allows us to model *UMI collisions*, i.e. situations where two identical molecules are sometimes labelled with identical UMIs, thus breaking the assumption of molecules being distinguishable before amplification. In place of M_{pur} in equation 2, a different distribution of initial copy numbers would be used that reflects the probability of a particular combination of UMI and molecule occurring once, twice, thrice, …. While non-ideal, such UMI collisions sometimes do sometimes occur if some mRNAs are more highly abundant than expected, or UMI length is restricted by experimental limitations; extending the theory to cover this case would thus be useful.

Finally, considering UMI collisions is interesting from a theoretical point of view, because it links the theory we developed for UMI-based NGS experiments to the much older idea of *PCR deduplication* (Ebbert *et al.*, 2016; Li, 2011; Li *et al.*, 2009). There, the sequence of a fragment itself (without artificially added random nucleotides) serves as its UMI, and reads with identical sequences are thus discarded as *PCR duplicates*. While these "UMIs" are prone to collisions and the removal of "PCR duplicates" can thus be over-zealous, the process can still be beneficial under some circumstances. By modifying our theoretical framework to include it, it may be possible to offer better advice, or even automated decision procedures, on whether reads with identical sequences should be collapsed or not.

Applying our theory of UMI-based quantitative NGS

The *raison d'être* of a quantitative theory for an experimental method is to *couple* it with a quantitative theory of the system being observed; by coupling two separate quantitative theories, confusing the behaviour of the system with its method of observation can be avoided.

We followed this approach when we developed a statistical model for the NGS and UM-based method iPool-Seq (Uhse *et al.*, 2018; chapter 3 of this work), which measures a particular pathogen's virulence by observing its abundance before and after infection of a host. The model exhibits the typical features of such models that are suitable for NGS-derived data (Anders and Huber, 2010; Love, Huber, *et al.*, 2014; M. D. Robinson, McCarthy, *et al.*, 2010): By assuming a (conditional) Poisson distribution, it takes the inherent statistical uncertainty of UMI counts obtained through non-exhaustive sampling into account. And by additionally assuming some uncertainty in the *true* abundances, additional variance is introduced that makes the (unconditional) distributions of observed counts a *Negative Binomial* (chapter 6).

Our iPool-Seq model extend the typical models in two ways: First, it combines counts observed *before* and *after* infection of the host, and second it takes the indirect PCR bias corrections supplied by TRUmiCount into account. For the latter, we consider our iPool-Seq model to be a *blueprint* of how to integrate these corrections into typical count data models. Simply applying indirect PCR bias corrections beforehand is not practical for such models, because doing so would yield non-integral counts, and break assumptions about the mean-variance relationship these models often rely on.

When used to find pathogen mutants with significantly different virulences, the iPool-Seq model we developed performs reasonable well, given experimental limitations such as the low number of reference mutants (chapter 6). Three of the mutants our model identifies as showing reduced virulence where independently tested in an infection assay, and showed a clear reduction of symptoms there as well (Uhse *et al.*, 2018; chapter 3 of this work, see figure 5 on page 52 and corresponding discussion on page 54), thus confirming the methods specificity.

References in this chapter

Anders S., & Huber W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. DOI:10.1186/gb-2010-11-10-r106

Barski A., Cuddapah S., Cui K., Roh T.-Y., Schones D. E., Wang Z., … Zhao K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, *129*(4), 823–837. DOI:10.1016/j.cell.2007.05.009

Ebbert M. T. W., Wadsworth M. E., Staley L. A., Hoyt K. L., Pickett B., Miller J., … Ridge P. G. (2016). Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC bioinformatics*, *17*(7), 239. DOI:10.1186/s12859-016-1097-3

Hug H., & Schuler R. (2003). Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. *Journal of theoretical biology*, *221*(4), 615–624. DOI:10.1006/jtbi.2003.3211

Johnson D. S., Mortazavi A., Myers R. M., & Wold B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science*, *316*(5830), 1497–1502. DOI:10.1126/science.1141319

Kivioja T., Vähärautio A., Karlsson K., Bonke M., Enge M., Linnarsson S., & Taipale J. (2011). Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, *9*(1), 72–74. DOI:10.1038/nmeth.1778

Li H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. DOI:10.1093/bioinformatics/btr509

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. DOI:10.1093/bioinformatics/btp352

Love M. I., Hogenesch J. B., & Irizarry R. A. (2016). Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nature biotechnology*, *34*(12), 1287–1291. DOI:10.1038/nbt.3682

Love M. I., Huber W., & Anders S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. DOI:10.1186/s13059-014-0550-8

Mardis E. R. (2008). Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*, *9*(1), 387–402. DOI:10.1146/annurev.genom.9.081307.164359

Mardis E. R. (2013). Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry*, *6*(1), 287–303. DOI:10.1146/annurev-anchem-062012-092628

Metzker M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, *11*(1), 31–46. DOI:10.1038/nrg2626

Robinson M. D., McCarthy D. J., & Smyth G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–40. DOI:10.1093/bioinformatics/btp616

This page intentionally left blank.

References (combined)

10x Genomics. (2016). Single Sell Gene Expression Dataset ERCC 1k. Retrieved October 20, 2018, from https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/ercc

Aird D., Ross M. G., Chen W.-S. S., Danielsson M., Fennell T., Russ C., … Gnirke A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, *12*(2), R18. DOI:10.1186/gb-2011-12-2-r18

Akima H. (1996). Algorithm 760: Rectangular-grid-data surface fitting that has the accuracy of a bicubic polynomial. *ACM Transactions on Mathematical Software*, *22*(3), 357–361. DOI:10.1145/232826.232854

Alberts B., Johnson A., & Lewis J. (2002). Molecular Biology of the Cell. (4th). New York (NY): Garland Science.

Anders S., & Huber W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. DOI:10.1186/gb-2010-11-10-r106

Andrews S. (2010). FastQC: A quality control tool for high throughput sequence data. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Armbruster C. E., Forsyth-DeOrnellas V., Johnson A. O., Smith S. N., Zhao L., Wu W., & Mobley H. L. T. (2017). Genome-wide transposon mutagenesis of *Proteus mirabilis*: Essential genes, fitness factors for catheter-associated urinary tract infection, and the impact of polymicrobial infection on fitness requirements. *PLoS pathogens*, *13*(6), e1006434. DOI:10.1371/journal.ppat.1006434

Barski A., Cuddapah S., Cui K., Roh T.-Y., Schones D. E., Wang Z., … Zhao K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, *129*(4), 823–837. DOI:10.1016/j.cell.2007.05.009

Benjamini Y., & Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300. DOI:10.2307/2346101

Best K., Oakes T., Heather J. M., Shawe-Taylor J., & Chain B. (2015). Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. *Nature Publishing Group*, 1–13. DOI:10.1038/srep14629

Bolger A. M., Lohse M., & Usadel B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. DOI:10.1093/bioinformatics/btu170

Bölker M., Genin S., Lehmler C., & Kahmann R. (1995). Genetic regulation of mating and dimorphism in *Ustilago maydis*. *Canadian Journal of Botany*, *73*(S1), 320–325. DOI:10.1139/b95-262

Breton Y. L., Belew A. T., Freiberg J. A., Sundar G. S., Islam E., Lieberman J., … McIver K. S. (2017). Genome-wide discovery of novel m1t1 group a streptococcal determinants important for fitness and virulence during soft-tissue infection. *PLoS pathogens*, *13*(8), e1006584. DOI:10.1371/journal.ppat.1006584

Bronner I. F., Otto T. D., Zhang M., Udenze K., Wang C., Quail M. A., … Rayner J. C. (2016). Quantitative insertion-site sequencing (qiseq) for high throughput phenotyping of transposon mutants. *Genome research*, *26*(7), 980–9. DOI:10.1101/gr.200279.115

Carter G. M., & Rolph J. E. (1974). Empirical bayes methods applied to estimating fire alarm probabilities. *Journal of the American Statistical Association*, *69*(348), 880–885. DOI:10.2307/2286157

Cole B. J., Feltcher M. E., Waters R. J., Wetmore K. M., Mucyn T. S., Ryan E. M., … Visel A. (2017). Genome-wide identification of bacterial plant colonization genes. *PLoS biology*, *15*(9), e2002860. DOI:10.1371/journal.pbio.2002860

Crimmins G. T., Mohammadi S., Green E. R., Bergman M. A., Isberg R. R., & Mecsas J. (2012). Identification of mrtab, an abc transporter specifically required for *Yersinia pseudotuberculosis* to colonize the mesenteric lymph nodes. *PLoS pathogens*, *8*(8), e1002828. DOI:10.1371/journal.ppat.1002828

Dabney J., & Meyer M. (2012). Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *BioTechniques*, *52*(2). DOI:10.2144/000113809

Doehlemann G., van der Linde K., Aßmann D., Schwammbach D., Hof A., Mohanty A., … Kahmann R. (2009). Pep1, a secreted effector protein of *Ustilago maydis*, is required for successful invasion of plant cells. *PLoS Pathogens*, *5*(2), e1000290. DOI:10.1371/journal.ppat.1000290

Ebbert M. T. W., Wadsworth M. E., Staley L. A., Hoyt K. L., Pickett B., Miller J., … Ridge P. G. (2016). Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC bioinformatics*, *17*(7), 239. DOI:10.1186/s12859-016-1097-3

Escobar-Zepeda A., Vera-Ponce de León A., & Sanchez-Flores A. (2015). The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. *Frontiers in genetics*, *6*, 348. DOI:10.3389/fgene.2015.00348

External RNA Controls Consortium. (2005a). Proposed methods for testing and selecting the ERCC external RNA controls. *BMC Genomics*, *6*(1), 150. DOI:10.1186/1471-2164-6-150

External RNA Controls Consortium. (2005b). The External RNA Controls Consortium: a progress report. *Nature Methods*, *2*(10), 731–734. DOI:10.1038/nmeth1005-731

Gallagher S. R. (2012). One-Dimensional SDS Gel Electrophoresis of Proteins. *Current Protocols in Protein Science*, *68*, 10.1.1–10.1.44. DOI:10.1002/0471140864.ps1001s68

Gawronski J. D., Wong S. M. S., Giannoukos G., Ward D. V., & Akerley B. J. (2009). Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung. *Proceedings of the National Academy of Sciences*, *106*(38), 16422–16427. DOI:10.1073/pnas.0906627106

Giaever G., Chu A. M., Ni L., Connelly C., Riles L., Véronneau S., … Johnston M. (2002). Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, *418*(6896), 387–391. DOI:10.1038/nature00935

Goodman A. L., McNulty N. P., Zhao Y., Leip D., Mitra R. D., Lozupone C. A., … Gordon J. I. (2009). Identifying Genetic Determinants Needed to Establish a Human Gut Symbiont in Its Habitat. *Cell Host & Microbe*, *6*(3), 279–289. DOI:10.1016/j.chom.2009.08.003

Gray A. N., Koo B.-M., Shiver A. L., Peters J. M., Osadnik H., & Gross C. A. (2015). High-throughput bacterial functional genomics in the sequencing era. *Current opinion in microbiology*, *27*, 86–95. DOI:10.1016/j.mib.2015.07.012

Grumaz S., Stevens P., Grumaz C., Decker S. O., Weigand M. A., Hofer S., … Sohn K. (2016). Next-generation sequencing diagnostics of bacteremia in septic patients. *Genome Medicine*, *8*(1), 73. DOI:10.1186/s13073-016-0326-8

Grüning B., Dale R., Sjödin A., Chapman B. A., Rowe J., Tomkins-Tinch C. H., … Köster J. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, *15*(7), 475–476. DOI:10.1038/s41592-018-0046-7

Haas B. J., Gevers D., Earl A. M., Feldgarden M., Ward D. V., Giannoukos G., … Birren B. W. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome research*, *21*(3), 494–504. DOI:10.1101/gr.112730.110

Harris T. E. (1989). *The Theory of Branching Processes*. New York (NY): Dover.

Heller J. (1961). *Catch-22*. New York (NY): Simon and Schuster.

Hoffman C. S., & Winston F. (1987). A ten-minute dna preparation from yeast efficiently releases autonomous plasmids for transformaion of *Escherichia coli*. *Gene*, *57*(2-3), 267–272. DOI:10.1016/0378-1119(87)90131-4

Honkanen S., Jones V. A. S., Morieri G., Champion C., Hetherington A. J., Kelly S., … Dolan L. (2016). The mechanism forming the cell surface of tip-growing rooting cells is conserved among land plants. *Current biology*, *26*(23), 3238–3244. DOI:10.1016/j.cub.2016.09.062

Howard J. (2014). Quantitative cell biology: the essential role of theory. *Molecular Biology of the Cell*, *25*(22), 3438–3440. DOI:10.1091/mbc.e14-02-0715

Hug H., & Schuler R. (2003). Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. *Journal of theoretical biology*, *221*(4), 615–624. DOI:10.1006/jtbi.2003.3211

Idnurm A., Reedy J. L., Nussbaum J. C., & Heitman J. (2004). *Cryptococcus neoformans* virulence gene discovery through insertional mutagenesis. *Eukaryotic Cell*, *3*(2), 420–429. DOI:10.1128/EC.3.2.420-429.2004

Illumina Inc. (2012). TruSeq DNASample Preparation Guide. Retrieved August 6, 2019, from http://emea.support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_truseq/truseqdna/TruSeq_DNA_SamplePrep_Guide_15026486_C.pdf

Illumina Inc. (2019). HiSeq 3000/HiSeq 4000 System quality and performance. Retrieved August 6, 2019, from https://www.illumina.com/systems/sequencing-platforms/hiseq-3000-4000/specifications.html

Ingolia N. T., Ghaemmaghami S., Newman J. R. S., & Weissman J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, *324*(5924), 218–23. DOI:10.1126/science.1168978

James W., & Stein C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, *vol. 1*, 361–379. Retrieved from http://projecteuclid.org/euclid.bsmsp/1200512173

Jeon J., Park S.-Y., Chi M.-H., Choi J., Park J., Rho H.-S., … Lee Y.-H. (2007). Genome-wide functional analysis of pathogenicity genes in the rice blast fungus. *Nature Genetics*, *39*(4), 561–565. DOI:10.1038/ng2002

Johnson D. S., Mortazavi A., Myers R. M., & Wold B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science*, *316*(5830), 1497–1502. DOI:10.1126/science.1141319

Kämper J., Kahmann R., Bölker M., Ma L.-J., Brefort T., Saville B. J., … Birren B. W. (2006). Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. *Nature*, *444*(7115), 97–101. DOI:10.1038/nature05248

Kanagawa T. (2003). Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of Bioscience and Bioengineering*, *96*(4), 317–323. DOI:10.1016/S1389-1723(03)90130-7

Kebschull J. M., & Zador A. M. (2015). Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Research*, *43*(21), gkv717. DOI:10.1093/nar/gkv717

Kemppainen M., Duplessis S., Martin F., & Pardo A. G. (2008). T-dna insertion, plasmid rescue and integration analysis in the model mycorrhizal fungus

Laccaria bicolor. *Microbial biotechnology*, *1*(3), 258–69. DOI:10.1111/j.1751-7915.2008.00029.x

Kivioja T., Vähärautio A., Karlsson K., Bonke M., Enge M., Linnarsson S., & Taipale J. (2011). Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, *9*(1), 72–74. DOI:10.1038/nmeth.1778

Köster J., & Rahmann S. (2012). Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. DOI:10.1093/bioinformatics/bts480

Krawczak M., Reiss J., Schmidtke J., & Rösler U. (1989). Polymerase chain reaction: Replication errors and reliability of gene diagnosis. *Nucleic acids research*, *17*(6), 2197–201. DOI:10.1093/nar/17.6.2197

Land M., Hauser L., Jun S.-R., Nookaew I., Leuze M. R., Ahn T.-H., … Ussery D. W. (2015). Insights from 20 years of bacterial genome sequencing. *Functional & integrative genomics*, *15*(2), 141–61. DOI:10.1007/s10142-015-0433-4

Langridge G. C., Phan M.-D., Turner D. J., Perkins T. T., Parts L., Haase J., … Turner A. K. (2009). Simultaneous assay of every *Salmonella Typhi* gene using one million transposon mutants. *Genome Research*, *19*(12), 2308–2316. DOI:10.1101/gr.097097.109

Lanver D., Müller A. N., Happel P., Schweizer G., Haas F. B., Franitza M., … Kahmann R. (2018). The biotrophic development of *Ustilago maydis* studied by RNA-seq analysis. *The Plant cell*, *30*(2), 300–323. DOI:10.1105/tpc.17.00764

Lanver D., Tollot M., Schweizer G., Lo Presti L., Reissmann S., Ma L.-S., … Kahmann R. (2017). *Ustilago maydis* effectors and their impact on virulence. *Nature Reviews Microbiology*, *15*(7), 409–421. DOI:10.1038/nrmicro.2017.33

Li H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. DOI:10.1093/bioinformatics/btr509

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. DOI:10.1093/bioinformatics/btp352

Liu D. (2019). Algorithms for efficiently collapsing reads with Unique Molecular Identifiers. *bioRxiv*. DOI:10.1101/648683

Love M. I., Hogenesch J. B., & Irizarry R. A. (2016). Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nature biotechnology*, *34*(12), 1287–1291. DOI:10.1038/nbt.3682

Love M. I., Huber W., & Anders S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. DOI:10.1186/s13059-014-0550-8

Mardis E. R. (2008). Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*, *9*(1), 387–402. DOI:10.1146/annurev.genom.9.081307.164359

Mardis E. R. (2013). Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry*, *6*(1), 287–303. DOI:10.1146/annurev-anchem-062012-092628

Marioni J. C., Mason C. E., Mane S. M., Stephens M., & Gilad Y. (2008). Rna-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, *18*(9), 1509–17. DOI:10.1101/gr.079558.108

McIntyre L. M., Lopiano K. K., Morse A. M., Amin V., Oberg A. L., Young L. J., & Nuzhdin S. V. (2011). RNA-seq: Technical variability and sampling. *BMC Genomics*, *12*(1), 293. DOI:10.1186/1471-2164-12-293

Metzker M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, *11*(1), 31–46. DOI:10.1038/nrg2626

Michel A. H., Hatakeyama R., Kimmig P., Arter M., Peter M., Matos J., … Kornmann B. (2017). Functional mapping of yeast genomes by saturated transposition. *eLife*, *6*. DOI:10.7554/eLife.23570

Michielse C. B., Hooykaas P. J. J., van den Hondel C. A. M. J. J., & Ram A. F. J. (2005). Agrobacterium-mediated transformation as a tool for functional genomics in fungi. *Current Genetics*, *48*(1), 1–17. DOI:10.1007/s00294-005-0578-0

Mueller A. N., Ziemann S., Treitschke S., Aßmann D., & Doehlemann G. (2013). Compatibility in the *Ustilago maydis*-maize interaction requires inhibition of host cysteine proteases by the fungal effector pit2. *PLoS pathogens*, *9*(2), e1003177. DOI:10.1371/journal.ppat.1003177

Mullis K., Faloona F., Scharf S., Saiki R., Horn G., & Erlich H. (1986). Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction.

Cold Spring Harbor Symposia on Quantitative Biology, *51*, 263–273. DOI:10.1101/SQB.1986.051.01.032

Pawluczyk M., Weiss J., Links M. G., Egaña Aranguren M., Wilkinson M. D., & Egea-Cortines M. (2015). Quantitative evaluation of bias in PCR amplification and next-generation sequencing derived from metabarcoding samples. *Analytical and Bioanalytical Chemistry*, *407*(7), 1841–1848. DOI:10.1007/s00216-014-8435-y

Picelli S., Björklund A. K., Reinius B., Sagasser S., Winberg G., & Sandberg R. (2014). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome research*, *24*(12), 2033–40. DOI:10.1101/gr.177881.114

Pinto A. J., & Raskin L. (2012). PCR Biases Distort Bacterial and Archaeal Community Structure in Pyrosequencing Datasets. *PLoS ONE*, *7*(8), e43093. DOI:10.1371/journal.pone.0043093

Rabe F., Seitner D., Bauer L., Navarrete F., Czedik-Eysenberg A., Rabanal F. A., & Djamei A. (2016). Phytohormone sensing in the biotrophic fungus *Ustilago maydis* - the dual role of the transcription factor rss1. *Molecular microbiology*, *102*(2), 290–305. DOI:10.1111/mmi.13460

Robinson J. T., Thorvaldsdóttir H., Winckler W., Guttman M., Lander E. S., Getz G., & Mesirov J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. DOI:10.1038/nbt.1754

Robinson M. D., McCarthy D. J., & Smyth G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–40. DOI:10.1093/bioinformatics/btp616

Robinson M. D., & Oshlack A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. DOI:10.1186/gb-2010-11-3-r25

Sambrook J. (2006). *The condensed protocols from molecular cloning: A laboratory manual*. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press.

Schilling L., Matei A., Redkar A., Walbot V., & Doehlemann G. (2014). Virulence of the maize smut *Ustilago maydis* is shaped by organ-specific effectors. *Molecular plant pathology*, *15*(8), 780–9. DOI:10.1111/mpp.12133

Schipper K., Brefort T., Doehlemann G., Djamei A., Münch K., & Kahmann R. (2008). The secreted protein stp1 is cruical for establishment of the biotrophic interaction of the smut fungus ustilago maydis with its host plant maize. *European Journal of Cell Biology*, *87*(supp. 1), 29–29.

Schmitt M. W., Kennedy S. R., Salk J. J., Fox E. J., Hiatt J. B., & Loeb L. A. (2012). Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(36), 14508–13. DOI:10.1073/pnas.1208715109

Sedlazeck F. J., Rescheneder P., & von Haeseler A. (2013). Nextgenmap: Fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics*, *29*(21), 2790–1. DOI:10.1093/bioinformatics/btt468

Seitner D., Uhse S., Gallei M., & Djamei A. (2018). The core effector cce1 is required for early infection of maize by *Ustilago maydis*. *Molecular Plant Pathology*, *19*(10), 2277–2287. DOI:10.1111/mpp.12698

Shiroguchi K., Jia T. Z., Sims P. A., & Xie X. S. (2012). Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proceedings of the National Academy of Sciences*, *109*(4), 1347–1352. DOI:10.1073/pnas.1118018109

Smith T., Heger A., & Sudbery I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome research*, *27*(3), 491–499. DOI:10.1101/gr.209601.116

Stern D. L. (2017). Tagmentation-based mapping (tagmap) of mobile dna genomic insertion sites. *bioRxiv*. DOI:10.1101/037762

Stirnberg A., & Djamei A. (2016). Characterization of apb73, a virulence factor important for colonization of *Zea mays* by the smut *Ustilago maydis*. *Molecular plant pathology*, *17*(9), 1467–1479. DOI:10.1111/mpp.12442

Suzuki M. T., & Giovannoni S. J. (1996). Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and environmental microbiology*, *62*(2), 625–30.

Thermo Fischer. (n.d.). ERCC Controls Analysis: ERCC RNA Spike-In Control Mixes. Retrieved October 20, 2018, from https://assets.thermofisher.com/TFS-Assets/LSG/manuals/cms_095046.txt

van Opijnen T., Bodi K. L., & Camilli A. (2009). Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nature Methods*, *6*(10), 767–772. DOI:10.1038/nmeth.1377

van Opijnen T., & Camilli A. (2012). A fine scale phenotype-genotype virulence map of a bacterial pathogen. *Genome research*, *22*(12), 2541–51. DOI:10.1101/gr.137430.112

Wamatu J., White D., & Chen W. (2005). Insertional mutagenesis of sclerotinia sclerotiorum through agrobacterium tumefaciens-mediated transformation. *Phytopathology*, *95*(6), S108.

Wang N., Ozer E. A., Mandel M. J., & Hauser A. R. (2014). Genome-wide identification of *Acinetobacter baumannii* genes necessary for persistence in the lung. *mBio*, *5*(3), e01163–14. DOI:10.1128/mBio.01163-14

Wang Z., Gerstein M., & Snyder M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. DOI:10.1038/nrg2484

Weiss G., & von Haeseler A. (1995). Modeling the polymerase chain reaction. *Journal of computational biology*, *2*(1), 49–61. DOI:10.1089/cmb.1995.2.49

Weiss G., & von Haeseler A. (1997). A coalescent approach to the polymerase chain reaction. *Nucleic acids research*, *25*(15), 3082–7. DOI:10.1093/nar/25.15.3082

Winzeler E. A. (1999). Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*, *285*(5429), 901–906. DOI:10.1126/science.285.5429.901

Zheng G. X. Y., Terry J. M., Belgrader P., Ryvkin P., Bent Z. W., Wilson R., … Bielas J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*(1), 14049. DOI:10.1038/ncomms14049