Dissertation zur Erlangung des Doktorgrades der Fakultät für Chemie und Pharmazie der Ludwig-Maximilians-Universität München

Consequences of DNA variation on gene regulation and human disease via RNA sequencing

Daniel Magnus Bader

from
Wertheim, Germany

2017

Erklärung
Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Herrn Prof. Dr. Klaus Förstemann betreut.

Eidesstattliche Versicherung
Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, den 11.05.2017 ....
Daniel M Bader

Dissertation eingereicht am: 11.05.2017

1. Gutachter: Prof. Dr. Klaus Förstemann

2. Gutachter: Prof. Dr. Julien Gagneur

Mündliche Prüfung am: 05.07.2017

Acknowledgments

I am very grateful to Julien Gagneur, who not only supervised me, but also listened to my problems as much as to my ideas. It was always fun to discuss projects as well as general questions about science and life with you. Thank you very much for being the group leader that you are, I know you will become even better.

I also want to thank Klaus Foerstemann for supervising this thesis together with the other members of my examination board: Eckhard Wolf, Dietmar Martin, Veit Hornung, and Franz Herzog. My thanks goes also to Christoph Klein who was part of my thesis advisory commitee.

My thanks goes to all the collaboration partners: Garwin Pichler from the Matthias Mann group, the groups of Lars Steinmetz and Holger Prokisch. Without you there would be no science here. In particular I want to thank Laura Kremer with whom I co-authored the second project of my PhD. I also want thank my first mentor Andre Altmann, who supervised my Bachelor thesis and introduced me to science. Without you I would not have started a PhD.

My gratitude goes also to the Quantitative Biosciences Munich graduate school. Foremost to Ulrike Gaul and Erwin Frey for installing this great melting pot of science. With their financial support I could visit inspiring conferences that helped me to grow as scientist and maybe even more as a person. Additionally, I thank the staff Mara Kieke, Julia Schlehe, Filiz Civril, Markus Hohle and Michael Mende who organized so many great lectures, workshops, and events for us.

I want to thank my group, the Gagneur lab, for their support and distraction in- and outside the lab. You were more than mere colleagues, you became true friends. In particular to Chriss, who is a constant in my life since I came to Munich; and to Juri, Vicente and Ziga for proofreading: Thank you! My thanks also goes to the Soeding group with whom we shared the office space at the LMU gene center for many years. Foremost Anja, Mark, Phillipp, Matthias, and Bjoern (Cramer group) thank you for creating this unique atmosphere google employees dream of. I thank the entire Rost group for their warm welcome in our new office at the TUM in Garching.

I want to thank my family for raising me curious and encouraging me in this career, where none of us went before.

Susann you are the love of my life.

Summary

With the complete knowledge about the DNA sequence of human and other model organisms like yeast, the foundation for a new, technology-driven era of biological research was laid at the beginning of this millennium. The availability of quantitative DNA and RNA information increased due to improved sequencing technology based on these reference genomes. This wealth of data has enabled us to ask biological questions of cause and consequence at nucleotide resolution genome-wide. In this work I investigated the consequences of DNA variation on RNA expression using sequencing data in two projects.

The first project distinguishes between controversial mechanisms that confer robustness to gene expression against regulatory variants. Previous studies suggested widespread buffering of RNA misexpression on protein levels during translation. We do not find evidence that translational buffering is common. Instead, we find extensive buffering at the level of RNA expression, exerted through negative feedback regulation acting in trans, which reduces the effect of regulatory variants. Our approach is based on a novel experimental design in which allelic differential expression in a yeast hybrid strain is compared to allelic differential expression in a pool of its spores. Allelic differential expression in the hybrid is due to cis-regulatory differences only. Instead, in the pool of spores allelic differential expression is not only due to cis-regulatory differences but also due to local trans effects that include negative feedback. We found that buffering through such local trans regulation is widespread, typically compensating for about 15% of cis-regulatory effects on individual genes. Negative feedback is stronger not only for essential genes, indicating its functional relevance, but also for genes with low to middle levels of expression, for which tight regulation matters most. We suggest that negative feedback is one mechanism of Waddington's canalization, facilitating the accumulation of genetic variants that might give selective advantage in different environments.

In the second project we develop a bioinformatic pipeline that improves the diagnosis of Mendelian disorders using RNA sequencing. Mendelian disorders can be caused by DNA variants in a single gene. However, the causal variants are hard to identify due to low sample numbers and often complex disease phenotypes. Accordingly, about 70% of patients with suspected Mendelian disorders remain undiagnosed after whole exome sequencing. This lack of diagnosis could be explained by disease-causing variants in non-coding regions. Whole genome sequencing facilitates the discovery of all genetic variants, but their sizeable number, coupled with a poor understanding of the non-coding genome, makes their prioritization challenging. Here, we demonstrate the power of RNA sequencing to provide a confirmed genetic diagnosis for 10% (5 of 48) of undiagnosed mitochondrial disease patients and identify strong candidate genes for patients remaining without diagnosis. We found a median of 1 aberrantly expressed gene, 5 aberrant splicing events, and 6 mono-allelically expressed rare variants in patient-derived fibroblasts and established disease-causing roles for each kind. Private exons often arose from sites that are weakly spliced in other individuals, providing an important clue for future variant prioritization. One such intronic exon-creating variant was found in three unrelated families in the complex I assembly factor TIMMDC1, which we consequently established as a novel disease-associated gene. In conclusion, our study expands the diagnostic tools for detecting non-exonic variants of Mendelian disorders and provides examples of intronic loss-of-function variants with pathological relevance.

In summary, both projects not only showed the conceptual benefit of a joint DNA-RNA analysis, but also provided statistical models and bioinformatic tools that can be used to drive future studies.

Publications

Negative feedback buffers effect of regulatory variants

Ref [1]

Daniel M Bader, Stefan Wilkening, Gen Lin, Manu M Tekkedil, Kim Dietrich, Lars M Steinmetz, Julien Gagneur
(2015) Molecular Systems Biology, DOI:10.15252/msb.20145844

Author contribution JG conceived and designed the experiments. SW, MT, and KD performed the experiments. DMB, LG, and JG analyzed the data. DMB, JG, and LMS wrote the paper.

Biallelic Mutations in NBAS Cause Recurrent Acute Liver Failure with Onset in Infancy

Ref [2]

Tobias B Haack*, Christian Staufner*, Marlies G Köpke, Beate K Straub, Stefan Kölker, Christian Thiel, Peter Freisinger, Ivo Baric, Patrick J McKiernan, Nicola Dikow, Inga Harting, Flemming Beisse, Peter Burgard, Urania Kotzaeridou, Joachim Kühr, Urban Himbert, Robert W Taylor, Felix Distelmaier, Jerry Vockley, Lina Ghaloul-Gonzalez, Johannes Zschocke, Laura S Kremer, Elisabeth Graf, Thomas Schwarzmayr, **Daniel M Bader**, Julien Gagneur, Thomas Wieland, Caterina Terrile, Tim M Strom, Thomas Meitinger, Georg F Hoffmann**, Holger Prokisch**
* joint first authors, ** joint senior authors
(2015) American Journal of Human Genetics, DOI:10.1016/j.ajhg.2015.05.009

Genetic diagnosis of Mendelian disorders via RNA sequencing

Ref [3]

Laura S Kremer*, **Daniel M Bader***, Christian Mertes, Robert Kopajtich, Garwin Pichler, Arcangela Iuso, Tobias B Haack, Elisabeth Graf, Thomas Schwarzmayr, Caterina Terrile, Eliska Konafikova, Birgit Repp, Gabi Kastenmüller, Jerzy Adamski, Peter Lichtner, Christoph Leonhardt, Benoit Funalot, Alice Donati, Valeria Tiranti, Anne Lombes, Claude Jardel, Dieter Gläser, Robert W Taylor, Daniele Ghezzi, Johannes A Mayr, Agnes Rötig, Peter Freisinger, Felix Distelmaier, Tim M Strom, Thomas Meitinger, Julien Gagneur**, Holger Prokisch**
* joint first authors, ** joint senior authors
(2017) bioRxiv, DOI:10.1101/066738

License: The copyright holder for this preprint is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license. Article is now published at Nature communications [4].

Author contribution Authors are ordered in each category by their contribution to the complete paper. Project planning: TM, JG, HP. Experimental design: HP. Review of phenotypes, sample collection and biochemical analysis: CL, BF, AD, VT, AL, DG, RT, DG, JAM, AR, PF, FD, and TM. Investigation LSK, DMB, and CM. Data curation and analysis: LSK, DMB, CM, TMS, and HP. Cell biology experiments: LSK, RK, AI, CT, EK, and BR. Exome, genome, and RNA sequencing; LSK, RK, EG, TS, PL and TMS. Exome analysis: LSK, RK, TBH, and HP. Quantitative proteomics: LSK and GP. Metabolomic studies: LSK, GK, and JA. Manuscript writing: LSK, DMB, CM, JG, and HP. Visualization LSK, DMB, and CM. Critical revision of the manuscript: all authors.

Contents

Chapter 1

Introduction

1.1 Biological background

Life is organized in cells that separate living matter from non-living matter. Cells store their building plan in a molecule called Deoxyribonucleic acid or short DNA. In order to reproduce, the mother cell duplicates its DNA and passes it to the daughter cell. The information encoded in the DNA gets activated via transcription of certain regions (genes) into Ribonucleic acids (RNAs), also called gene expression. These RNAs can perform already many essential functions in a cell, but most work is done by proteins that are created by translating some of the RNAs. This cascade of activating DNA information by transcription into RNA and consequently translation into proteins is called the central dogma of biology [5, 6].

During the lifetime of an organism spontaneous mutations can occur in the DNA and change its sequence and thereby its stored information about the cells building plan. These mutations can be passed to the next generation and can later be identified as inborn variants to the reference genome of the corresponding organism. This is common for both uni-cellular organisms like yeast and complex multi-cellular organisms like humans.

Most inborn variants have no effect on the organism as whole, although they are present in every cell. This buffering of variant effects is achieved in part by regulation of the different steps from transcription of genes, translation of RNA into proteins, and their corresponding degradation. Consequently, variants that are not too deleterious can accumulate, which might change the regulation of gene expression at some point enough to form a new species. However, the time scale for new species to form is within millions of years. To this day, it is not fully understood which gene regulatory mechanisms buffer the consequences of DNA variation.

Inborn variants in essential genes can be lethal for the affected organism. Other inherited variants can lead to severe diseases. These two types of variants are the most extreme with respect to organismal phenotype and correspondingly rare. So is the prevalence for a disease that can be caused by these variants. For example the European Union defines a disease as rare, if it affects less than 1 in 2,000 individuals[1] (see also this recent article [7]). More than 8,000 rare diseases are de-

[1] Regulation (EC) No 141/2000 of the European Parliament and of the Council of 16 December 1999 on orphan medicinal products (celex:32000R0141).

scribed in Orphanet[2], an online database with 40 countries contributing. Many of these rare diseases have genetic origin and are inherited according to Mendel's laws – Mendelian diseases. Such rare Mendelian diseases are hard to diagnose and treat, because there are few individuals with the same disease, their clinical symptoms are often severe and the treating clinician is by chance seldom an expert for the corresponding disease. The consequence for patients is often an odyssey of various clinical tests that try to diagnose the cause. Therefore a genetic approach that can improve the diagnosis of rare Mendelian diseases would help many clinicians and patients.

1.1.1 History of early genome sequencing

In 1965 Holley and Sanger together with their colleagues were the first to determine not only the composition of nucleic acids of a RNA molecule, but also its sequence [8, 9]. RNA was partially digested by ribonuclease and fractionated across two dimensions according to its length and base composition. The beginning of nucleic acid sequencing also started a new research field driven by technological advances. In contrast to ribonuclease digestion, Wu and colleagues used DNA polymerase to incorporate radioactive nucleotides, one at a time to determine their sequence at overhanging 5' ends in a phage [10, 11]. With these techniques available the first complete RNA sequence of the coat protein and shortly after the complete RNA genome of the bacteriophage MS2 was sequenced [12, 13]. The next step forward came with the separation of nucleic acid fragments via polyacrylamide gels [14, 15], but the leap towards the sequencing era came with the chain-termination technique by Sanger [16]. Here, mixing of radiolabelled dideoxynucleotides together with normal deoxynucleotides for DNA extension produces fragments of all possible lengths in a single reaction (Fig 1.1). The introduction of dideoxynucleotides improved not only protocol duration but also complexity to allow sequencing of more complex organisms. Shortly after, Sanger and colleagues sequenced the first complete DNA genome of bacteriophage ϕX174 [17].

The sequencing of the human genome [19, 20] around the millennium was still based on modifications of the *Sanger sequencing* technique [16]. Complete biological understanding of humans was the goal of the sequencing project, i.e. understanding disease, diversity and aging. Yet, the first findings were almost disappointing: 30,000 - 40,000 genes, only twice as many as in worm or fly [19]. With these two landmark studies the foundation was laid for the next generation of sequencing technology. The history of sequencing is also nicely reviewed in [18].

1.1.2 Rapid developments in high throughput sequencing

The completion of the human genome together with model organisms allowed to develop a new generation of sequencing building upon these reference genomes [21, 22]. The first of the next generation sequencers was developed by 454 Life sciences corporation and relies on pyro-sequencing in picoliter-sized wells [23]. However, they were outperformed by Illumina, which is now the most common sequencing

[2] Orphanet: an online rare disease and orphan drug data base. ©INSERM 1997. Available on `http://www.orpha.net`. Accessed 10 April 2017.

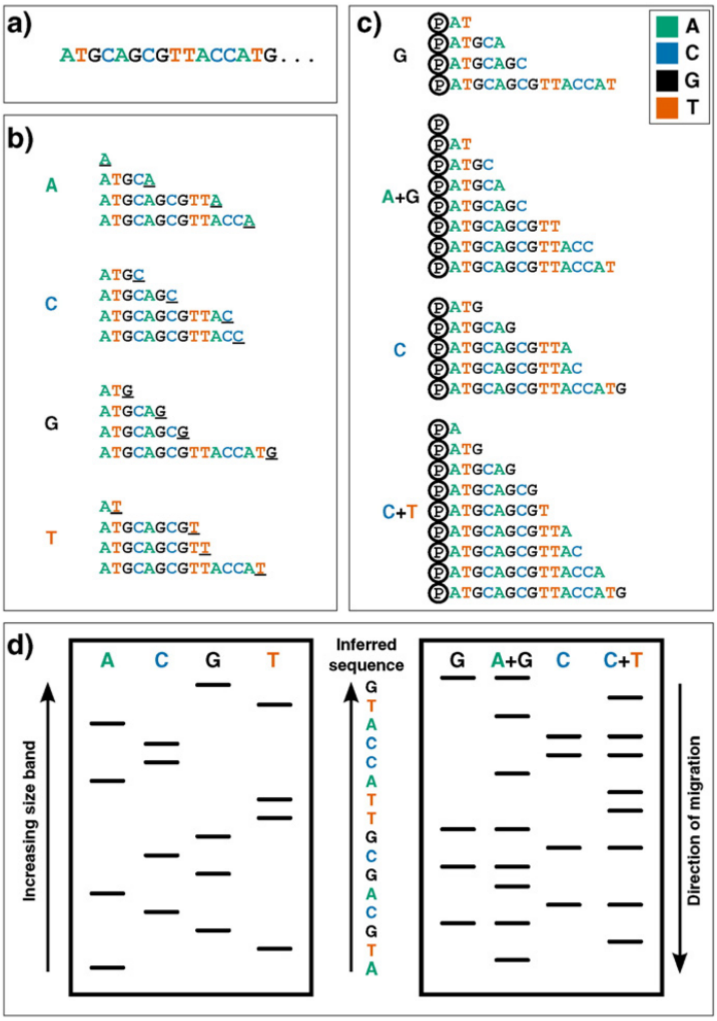


Figure 1.1: **First-generation DNA sequencing technologies.** Taken from Ref [18](Fig 1). Example DNA to be sequenced **(a)** is illustrated undergoing either Sanger (b) or Maxam–Gilbert (c) sequencing. **(b)** Sanger's 'chain-termination' sequencing. Labeled ddNTP nucleotides of a given type are included in DNA polymerization reactions at low concentrations. Therefore in each of the four reactions, sequence fragments are generated with 3' truncations as a ddNTP is randomly incorporated at a particular instance of that base (underlined 3' terminal characters). **(c)** Maxam and Gilbert's 'chemical sequencing' method. DNA must first be labeled, typically by inclusion of radioactive P^{32} in its 5' phosphate moiety (℗). Hydrazine removes bases from pyrimidines (cytosine and thymine). Acid can be used to remove the bases from purines (adenine and guanine). Piperidine is used to cleave the phophodiester backbone at the abasic site, yielding fragments of variable length. **(d)** Fragments can be visualized via electrophoresis on a high-resolution polyacrylamide gel: sequences are then inferred by reading 'up' the gel, as the shorter DNA fragments migrate faster. In Sanger sequencing (left) the sequence is inferred by finding the lane in which the band is present for a given site, as the 3' terminating labelled ddNTP corresponds to the base at that position. Maxam–Gilbert sequencing (right) requires a small additional logical step: Ts and As can be directly inferred from a band in the pyrimidine or purine lanes respectively, while G and C are indicated by the presence of dual bands.

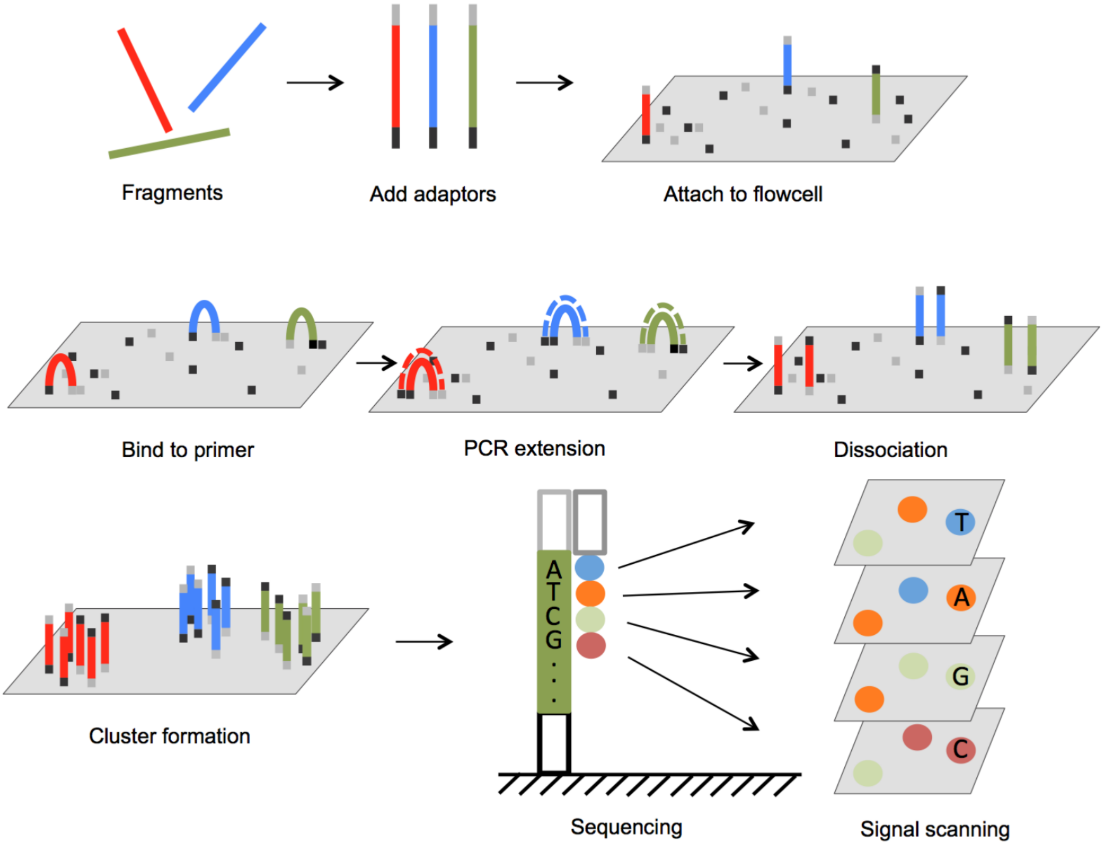


Figure 1.2: **Outline of Illumina genome analyzer sequencing process.** Taken from Ref [26, Fig 1]. (1) Adaptors are annealed to the ends of sequence fragments. (2) Fragments bind to primer-loaded flow cell and bridge PCR reactions amplify each bound fragment to produce clusters of fragments. (3) During each sequencing cycle, one fluorophore attached nucleotide is added to the growing strands. Laser excites the fluorophores in all the fragments that are being sequenced and an optic scanner col- lects the signals from each fragment cluster. Then the sequencing terminator is removed and the next sequencing cycle starts.

platform world-wide [24]. The Illumina protocol is structured in three main steps (Fig 1.2) [25]:

First, the DNA is denatured and fragmented into smaller pieces. Adapters are added to allow targeted post processing. In the clustering step, each of these modified fragments gets amplified on the flow cell. A flow cell is a glass plate with multiple separated reaction lanes. Each lane has two kinds of oligonucleotides attached to its surface complementary to the two adapters of the fragment. Consequently, the fragments hybridize via base-pairing and a polymerase synthesizes the complementary strand to the template fragment. The original template strand gets washed away and only the surface-attached strands remain.

Second, the newly synthesized strands that are attached to the surface on one end are now amplified by bridging to the other type of oligonucleotides. Here, the strand bends and hybridizes to a neighboring oligonucleotide with its free end. A polymerase synthesizes the complementary strand forming a double stranded bridge. Denaturation leads to two single strands attached to surface of flow cell. This bridging process is repeated and parallelized over the flow cell to amplify template strands for sequencing.

Third, after amplification, sequencing starts at the primer encoded in the adapter sequence. All four nucleotides carry a base-specific fluorophore and compete for binding at this step. After incorporation of the correct base the identity of the bound nucleotide is determined by laser-induced excitation and imaging. The fluorophore gets removed, leaving the growing strand ready for further elongation. This process is called sequencing by synthesis through reversible terminator chemistry.

1.1.3 RNA sequencing

The DNA contains functional elements called genes that are transcribed into RNAs. These RNAs can take over important enzymatic, regulatory, or messenger functions, such as those RNAs acting in the ribosome. The ribosome for example translates messenger RNAs into proteins. Those RNAs are also called *coding*, since they encode proteins. Analogously, there are non-coding RNAs with mostly regulatory functions, and also RNAs that are named after their function, if any. The collection of all RNAs in a cell is called the transcriptome.

The quantification standard for the transcriptome shifted from microarrays [27, 28] to RNA sequencing (RNA-seq) driven by the developments in DNA sequencing. Through reverse transcription of RNA into complementary DNA (cDNA) the same DNA sequencing technologies can be used to quantify the transcriptome [29]. In 2008 many pilot studies determined the transcriptome of model organisms, e.g. yeast [30] and human [31]. RNA-seq has the following advantages over microarray technologies [29]: RNA-seq is not limited to existing reference genomes, has lower background signal, has a higher dynamic range with no upper limit, and it is also highly accurate with respect to quantitative polymerase chain reaction and spike-in RNA controls of known concentration.

One difficulty of RNA-seq is to keep the polarity information of the transcript, i.e. from which of the two DNA strands it originates. Especially in prokaryotes and lower eukaryotes the genome is very compact, e.g. for fast replication. Here, compact means not only most of the DNA is encoding transcripts, but also that different transcripts overlap on opposite strands. Consequently, preserving the polarity or strand information of a transcript is crucial to distinguish RNAs. During reverse transcriptase of single stranded RNA into double stranded DNA the polarity information gets lost. However, there are different strategies to overcome this drawback [32]. For example Parkhomchuk et al. [33] incorporate deoxyuridine triphosphate (dUTP) instead of deoxythymidine triphosphate (dTTP) during the synthesis of the second strand. This allows targeted degradation of the second strand by Uracil-N-Glycosylase (Fig 1.3).

1.1.4 Modeling RNA sequencing data

The latest sequencing systems are creating a constant flood of DNA and RNA data, together with the need for sophisticated analysis pipelines. With the recent Illumina's NovaSeq sequencing system[3] it is (soon) possible to produce 10 billion reads or 3 terabytes of data per run. For this amount of available data it is crucial

[3] `https://www.illumina.com/systems/sequencing-platforms/novaseq.html`

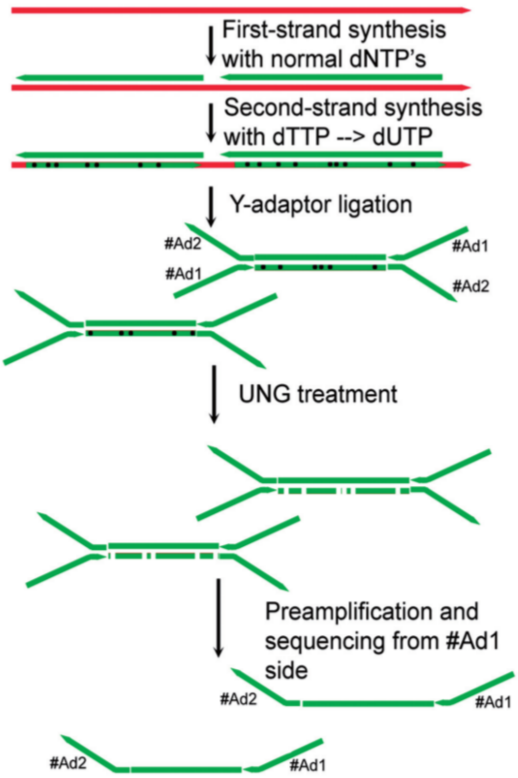

First-strand synthesis with normal dNTP's
Second-strand synthesis with dTTP --> dUTP
Y-adaptor ligation
#Ad2
#Ad1
#Ad1
#Ad2
UNG treatment
Preamplification and sequencing from #Ad1 side
#Ad2
#Ad1
#Ad2
#Ad1

Figure 1.3: **Strand-specific RNA sequencing.** Figure is taken from [33, Fig 1A]. Flowchart of the ssRNA-Seq procedure. RNA is shown in red, DNA in green. Arrows are in the 5' to 3' direction. UNG, Uracil-N-Glycosylase; dNTP, deoxynucleoside triphosphate.

to model gene expression measured via RNA-seq most stringently to extract a high signal to noise ratio from the data and not be fooled by spurious events.

To quantify RNA-seq reads for the estimation of gene expression each read is aligned to its originating genome and assigned to the corresponding target transcript. This alignment information can be summarized as read counts per gene which functions as approximation of the corresponding gene product abundance in the cell. One of the most basic questions asked for in gene expression analysis is the difference in transcript abundance between two or more biological conditions, i.e. treated versus untreated with respect to a certain stimulus. To assess significant differences in read counts for a given gene, an appropriate statistical test should be applied. In one of the first comparisons between microarray and RNA-seq expression [31], a Poisson distribution was suggested to model the distribution of read counts, since read counts represent discrete events of a transcript being present. Furthermore, Marioni et al. [31] showed the potential to overcome the limitations of microarray-based expression studies. The Poisson distribution has a single parameter, its mean, and the other parameters are derived from it, i.e. the variance is equal to the mean. This variance model was shown to be too restrictive for RNA-seq read count data, producing too many false positives (type-I error) [34].

The two software packages *edgeR* and *DESeq* were the first to model read count data via a negative binomial distribution [35, 36] allowing for variance larger than the mean, i.e. over-dispersion. In *DESeq* the read count $K_{i,j}$ for gene i in sample j (with $j \in [1, m]$) is described with a generalized linear model of the negative binomial with a logarithmic link function:

$$
\begin{aligned}
K_{i,j} &\sim \mathrm{NB}(mean = \mu_{i,j}, dispersion = \alpha_i) && (1.1) \\
\mu_{i,j} &= s_j q_{i,j} && (1.2) \\
\log_2 q_{i,j} &= \sum_r x_{j,r} \beta_{i,r} && (1.3)
\end{aligned}
$$

where NB is the negative binomial distribution, α_i is a gene-specific dispersion parameter; r corresponds to the number of conditions that are modeled; x is the design matrix assigning samples j to conditions r; βs are the coefficients that represent the effect of a condition r on a specific gene i; s_j is the size factor of sample j. Size factors allow to normalize for sample-specific effects, e.g. amount of material loaded into the sequencing machine. They are estimated with the median-of-ratios method:

$$
s_j = \underset{i}{\mathrm{median}} \frac{K_{i,j}}{\left(\prod_{j=1}^{m} K_{i,j}\right)^{1/m}} \tag{1.4}
$$

Throughout this thesis the updated *DESeq2* [37] is used to model RNA-seq read count data. It has the advantage of empirical Bayes based shrinkage for dispersion and fold change estimation. Gene-wise dispersion is estimated via a maximum likelihood approach which relies only on the data for each gene separately. A smooth curve is fitted for expected dispersion by expression strength. The final dispersion estimates are obtained by shrinking the gene-wise estimates towards the predicted values. Fold changes are shrunken towards zero also in an empirical Bayes based way stronger for low counts, high dispersion, or few degrees of freedom [37].

Consequently, with this software we are not only able to identify differential expression from RNA-seq data, but also robustly rank differentially expressed genes based on their fold changes [37].

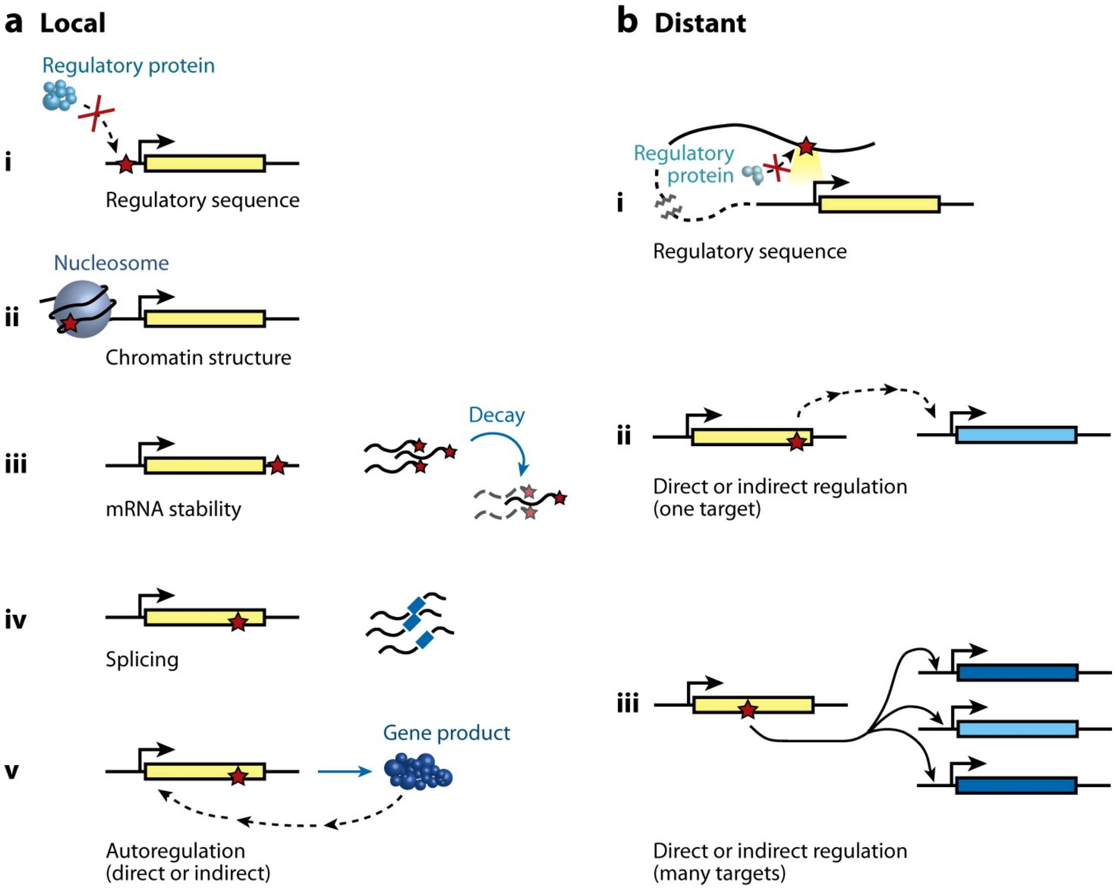
1.2 Studying consequences of DNA variation

1.2.1 DNA variation drives evolution through gene expression changes

DNA can affect the phenotype through changes not only in regions encoding proteins, but also in regions regulating gene expression. Already in 1975, King and Wilson [38] concluded that the differences in protein sequence and biochemical properties between humans and chimpanzees are too small to explain their phenotypic diversity. The crucial differences do not lie in protein-coding, but rather in regulatory variations that alter gene expression. One of the first genome-wide studies investigating intraspecies variation for yeast strains was conducted by Brem et al. [39]. They used microarrays to measure the expression of more than 6,000 genes in two yeast strains and their haploid segregants and systematically identified local and distant regulation by linkage. Linkage means that the expression of a gene is linked to its local or some distant genotype. In a parallel study [40], expression differences between alleles in mice hybrids were proposed to identify cis-regulation, since the two alleles in a hybrid are subject to the same trans effects. Trans effects can be caused by regulatory elements that are not inherited together with the gene. The approaches of Brem et al. [39] and Cowles et al. [40] were combined not only to achieve a systematic distinction of positional linkage of an expression Quantitative Trait Loci (eQTL) into local and distant, but also for mechanistic classification into cis and trans [41]. Ronald et al. [41] investigated self-linkage of expression in 112 segregants of two laboratory yeast strains in 5,727 transcripts. Looking more detailed into local regulation they selected 77 genes with strong self-linkage to be tested for allelic differential expression (ADE) in a diploid hybrid. Of these, 78% showed ADE besides self-linkage classifying them as cis-regulated. The remaining self-linkage could be explained by a nearby linked gene or feedback on the gene itself, as demonstrated for AMN1.

The findings and terms of gene expression regulation are nicely reviewed later [42–44] (Fig 1.4)[4]. Briefly, if an eQTL is allele-specific, it is regulated through a cis mechanism; if an eQTL affects expression of both alleles of a diploid organism it acts through trans. Moreover, local and distant eQTLs are distinguished by their genomic position with respect to the gene they influence [43].

[4] Permissions: “Material may be republished in a thesis / dissertation without obtaining additional permission from Annual Reviews.”



Skelly DA, et al. 2009.
Annu. Rev. Genomics Hum. Genet. 10:313–32

Figure 1.4: **Molecular mechanisms of local and distant regulatory variation.** Figure taken from [43]. **(a)** Local regulatory variation acts from a position near the gene of interest. This type of variation can impact gene expression levels by affecting (i) the binding of regulatory proteins to regulatory sequences, (ii) nucleosome binding or chromatin remodeling to influence chromatin structure, (iii) sequences that contribute to transcript-specific decay rates to determine mRNA stability, (iv) transcript structure as determined by the fidelity of intron splicing, and (v) regulation of the gene by its own product or the product of a gene downstream in the transcriptional regulatory network. **(b)** Distant regulatory variation acts from a position far from the gene of interest. This type of variation can impact gene expression levels by affecting (i) the binding of regulatory proteins to distant regulatory sequences or (ii and iii) regulation of one or more genes directly or at some point downstream in the transcriptional regulatory network.

1.2.2 Quantification of cis and trans regulation in hybrid organisms

After these conceptual studies on single genes many genome-wide approaches set out to quantify the different regulatory mechanisms in different species. Tirosh et al. [45] conducted the first genome-wide study in a diploid hybrid of two laboratory yeast strains still using microarrays to measure gene expression. They found that cis-regulation drives interspecies differences, whereas trans-regulation is condition specific for sensory signals. Cis expression differences were confirmed to shape adaptive expression divergence between species using a RNA-seq approach again of two yeast strains and their hybrid [46]. After preliminary work by Wittkopp et al. [47, 48], McManus et al. [49] conducted the first genome-wide quantification of cis and trans regulation in Drosophila again with the help of RNA-seq. Mammals followed two years after with another RNA-seq approach, here on two laboratory mice strains and their F1 hybrid [50].

The analysis of gene regulation across species highly profited from the fast progress in sequencing technologies. Their common findings report cis-regulatory changes to contribute more to interspecies differences and the presence of antagonistic cis-trans regulation at various levels across species. Yet, conclusion across biological kingdoms are hard to draw since the statistical methods evolved together with the measurement techniques. Summarizing, it is still an open question which mechanisms buffer cis-regulatory differences, although this buffering is crucial for further evolution.

1.2.3 Possible buffering mechanisms for the effects of regulatory variants

The introduction presented in this section is part of the manuscript "Negative feedback buffers effect of regulatory variants" from Bader et al. 2015 [1].

In 2014, two studies have assessed the role of translation in buffering variations in RNA expression [51, 52]. In both studies, ADE was compared to allelic differential translation efficiency estimated from allele-specific ribosome occupancies in a cross of the yeast species *S. cerevisiae* and *S. paradoxus*. ADE indicates effects of cis variants, i.e. regulatory variants that act on one but not on both alleles of a gene [40, 53]. Focusing on genes with both a significant ADE and significant allele-specific translation efficiency differences, these studies reported an excess of translation efficiency differences opposing to the allelic differential expression. In contrast, Muzzey et al. [54] reported a genome-wide trend for reinforcing ADE during translation in the yeast *C.albicans*. As these studies used distinct statistical procedures and species, it is hard to compare them and conclude about the generality of these findings. It is appealing to conceive translation as a check point to counter allelic expression imbalance (Fig 2.1A). However, a general mechanism that could sense mRNA allelic imbalance and regulate translation accordingly is hard to imagine. Instead, the most likely explanation for translational buffering is the selection for compensatory mutations [51, 52]. Hence, variation in translation efficiency might contribute to buffering but does not appear as an intrinsic mechanism that yields robustness against newly arisen regulatory variants.

Alternatively, Denby et al. [55] have proposed that negative feedback controlling the level of RNA expression could be a common mechanism to buffer effects of regulatory variants (Fig 2.1A). Negative feedback would buffer expression differences by exerting a stronger repression on alleles with higher expression levels and a weaker repression on alleles with lower expression levels. Screening for auto-regulated transcription factors in yeast, Denby et al. [55] found *ROX1* to be under strong negative feedback. Mutant experiments showed that this negative feedback confers robustness to the expression of *ROX1* in the face of naturally occurring allelic variants present in a set of divergent yeast strains. This study demonstrated for a single gene that negative feedback could act as a buffering mechanism for regulatory variants. However, data about the extent of feedback mechanisms genome-wide and its importance for buffering regulatory variants is still lacking.

1.2.4 DNA variation in human disease

One way to study the consequences of DNA variation, is to monitor all or as many DNA differences as possible and associate them with your trait of interest – also known as a genome-wide association study (GWAS) [56–58]. In a GWAS approximately 500,000 to almost three million markers [57, 59] are genotyped to compare case and control subjects, e.g. people suffering from a likely inherited disease and healthy people with otherwise matching characteristics like ethnicity, age, and sex. Sequence and copy-number variations can be identified in this hypothesis-free approach. The first GWASs investigating human disease found strong risk loci for age-related macular degradation [60] and Parkinson disease [61]. However, GWASs also come with major limitations: i) the identified loci are spanning more than thousand bases, they are not as detailed as individual genes or single nucleotide variants, ii) only common alleles can be detected (minor allele frequency about 5%), and iii) large sample sizes are required for both identification and replication [57]. One example of the large sample size requirements is the association of 18 new loci with body mass index in an analysis of 249,796 individuals [59].

The limitations of GWAS do not allow to explain the missing heritability of complex traits observed for their associated loci [62]. GWAS studies are blind towards rare variants that might also contribute to common traits. Furthermore, rare variants can not only contribute to common traits, but a single variant can also cause a severe disease (Fig 1.5). Since these diseases can be inherited through the variants according to Mendel's laws [63] they are called Mendelian disorders. Their genetic diagnosis can be achieved, e.g. by targeted sequencing of candidate genes or molecular assays that can capture the effects of variants [64].

1.2.5 Exome sequencing in genetic diagnosis

Parallel to studies on gene regulation, genetic diagnosis got a boost up through the new sequencing technologies built upon reference genomes. Especially the field of Mendelian disorders profited from the development of whole exome sequencing (WES) [66, 67]. WES is agnostic to the disease, consequently among all the exonic variants that are routinely detected it is possible to find also new disease-causing variants in contrast to the pre-defined marker set used for GWAS. In the

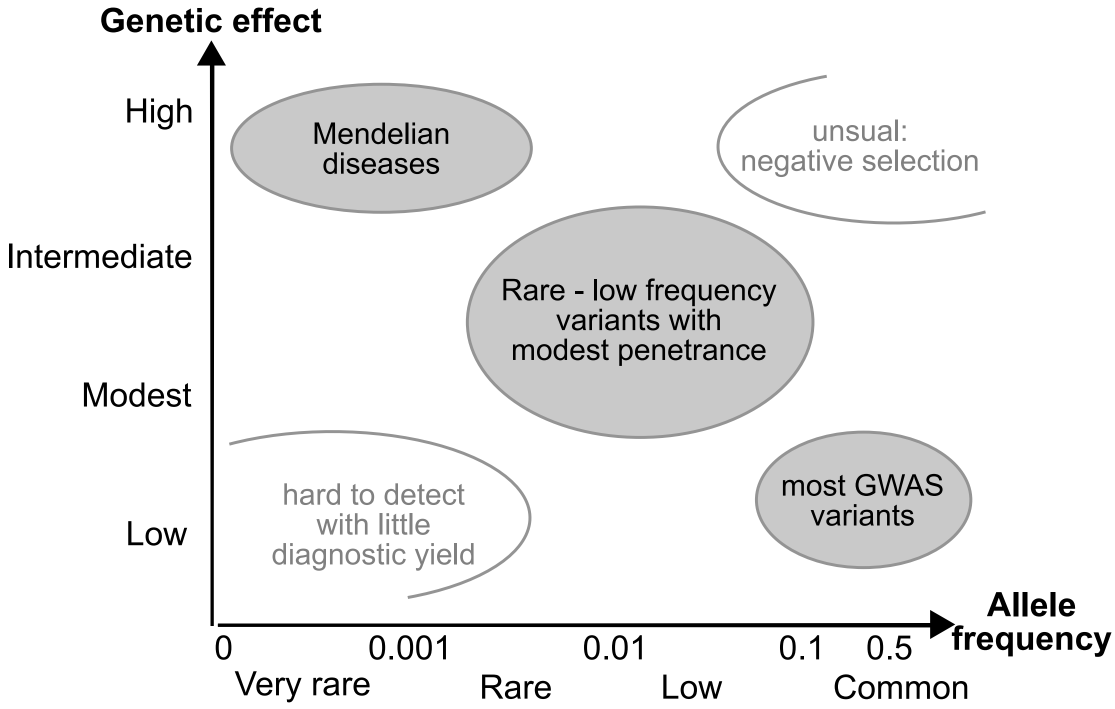


Figure 1.5: **Variant frequency versus genetic effect.** Figure adopted from [62, 65].

first diagnosis study WES of four patients with the rare dominantly inherited Freeman–Sheldon syndrome was performed and the previously known causal variants were recovered [67]. This landmark study paved the way for many success stories of new disease-causing genes being discovered, e.g. for mitochondrial complex I deficiency [68], recessive Miller syndrome [69] and dominant Schinzel–Giedion syndrome [70]. With the growing number of genetic diagnosis studies the need for standards arose and was discussed by an expert group in 2012 and later published [71]. One of their key remarks was the need for global sharing of data to build a resource that can serve as a reference; for example sequencing data for minor allele frequencies of variants or variant information together with their evidence for genetic diagnosis. Both goals were achieved shortly after with a database for disease-causing variants ClinVar [72, 73] and a catalog of WES data of more than 60,000 human individuals [74, 75].

1.2.6 Diagnosis rate of Mendelian diseases

WES has become a cornerstone for genetic diagnosis. This change in routine is reflected in the number of Mendelian disease genes discovered by year and technology (see Fig 1 in [76]). Since 2013 this number grew almost thrice as high for sequencing based discoveries compared to non-sequencing techniques (see Fig 4 in [77]). When diagnosing patients with a high suspicion of a Mendelian disorder, WES achieves an overall diagnosis rate of about 30% [77–79]. Within these unselected patient cohorts, the highest diagnosis rate of about 47% was achieved by diseases affecting vision [78, 79]. On complex Mendelian phenotypes such as mitochondrial disorders with more than 250 known disease-causing genes, diagnosis rates with WES vary [80, 81]. Yet, in a set of 53 patients with biochemical evidence of respiratory chain defects the diagnosis rate of WES reached 60% [82].

1.2.7 Limitations of genome sequencing

WES has proven very useful in genetic diagnosis, however it is limited by design to the coding part of the DNA, i.e. regions specified by the exon capture kit. Mendelian disorders are not exclusively caused by these kind of variants. The straight forward solution to this limitation is to sequence the whole genome to get all variants from both coding and non-coding regions. This DNA sequence information is difficult to interpret without a reference database or additional molecular measurements that corroborate variant frequency or gene function, correspondingly. A pilot study of the WGS500[5] program analyzed 156 independent cases with a broad range of disorders that were suspected to have a strong genetic component [83]. Overall, a pathogenic variant could be identified in 21% (33/156 cases), whereas 15% (5/33 variants) of these causal variants would likely be missed by standard WES.

The WGS500 program leaves open questions on the improvements of whole genome sequencing over WES. Despite the wealth of genomic variation data, the interpretation of the non-coding genome is challenging. With many ongoing sequencing projects like the UK10K project [84], the *All of Us* project[6] (discussed here [85]), or the goal to sequence 100 million chinese people (discussed here [86]) the distribution of rare variants will be better understood, especially with respect to different populations. However, functional interpretation of these variants with purely genetic information remains another challenge.

1.2.8 Personalized transcriptomics

One way to overcome the limitation of identifying only vague loci in GWAS is to measure both genotype and gene expression. With this additional layer of information, it was possible not only to identify which genomic regions differ between case and controls, but also which genes differ in expression. If the expression changes were limited to one gene for an associated loci, follow-ups to detect the variant within the associated loci could be more targeted. Analog to studies in model organisms (section 1.2.2), eQTL studies in mammals started with microarrays to measure the expression for a fraction of transcripts [87, 88] up to almost transcriptome scale in humans [89]. Emilsson et al. [89] were the first to study expression association with human disease phenotypes in primary tissues, i.e. blood and subcutaneous fat.

Again, the advances in sequencing technology brought a better quantification of transcribed RNA: that is a larger dynamic scale, more and new transcripts [90, 91]. In both studies RNA-seq was performed on lymphoblastoid cell lines of already genotyped individuals from the HapMap project [92, 93]. They highlighted the potential of RNA-seq combined with genetic information to investigate new regulatory mechanisms and haplotypes together with transcript abundance and

[5] http://www.well.ox.ac.uk/wgs500

[6] https://www.nih.gov/research-training/allofus-research-program

structure. Thus, personalized transcriptomics [94] represents another way around the limitations of WES and interpretability of whole genome sequencing in genetic diagnosis. Personalized transcriptomics offers not only a second layer of information beyond the DNA but also possible functional consequences of regulatory variants.

1.2.9 RNA features with diagnostic potential

The introduction presented in this section is part of the manuscript "Genetic diagnosis of Mendelian disorders via RNA sequencing" from Kremer, Bader et al. 2016 [3].

With RNA sequencing (RNA-seq), limitations of the sole genetic information can be complemented by directly probing variations in RNA abundance and in RNA sequence, including allele-specific expression and splice isoforms. At least three extreme situations can be directly interpreted to prioritize candidate disease-causing genes for a rare disorder. First, the expression level of a gene can lie outside its physiological range. Genes with expression outside their physical range can be identified as expression outliers, often using a stringent cutoff on expression variations, for instance using the Z-score [95] or statistics at the level of whole gene sets[96, 97]. The genetic causes of such aberrant expression includes rare variants in the promoter [98] and enhancer but also in coding or intronic regions [95]. Second, RNA-seq can reveal extreme cases of allele-specific expression (mono-allelic expression), whereby one allele is silenced, leaving only the other allele expressed. When assuming a recessive mode of inheritance, genes with a single heterozygous rare coding variant identified by WES or WGS analysis are not prioritized. However, mono-allelic expression of such variants fits the recessive mode of inheritance assumption. Detection of mono-allelic expression can thus help re-prioritizing heterozygous rare variants. Reasons for mono-allelic expression can be genetic. A pilot study validated compound heterozygous variants within one gene as cause of TAR syndrome, where one allele is deleted and the other harbors a non-coding variant that reduces expression [99]. Mono-allelic expression can also have epigenetic causes such as X-chromosome inactivation or imprinting on autosomal genes, possibly by random choice [100, 101]. Third, splicing of a gene can be affected. Aberrant splicing has long been recognized as a major cause of Mendelian disorders (reviewed in ref. [102–104]). However, the prediction of splicing defects from genetic sequence is difficult because splicing involves a complex set of cis-regulatory elements that are not yet fully understood. Some of them can be deeply located in intronic sequences [105] and are thus not covered by WES. Hence, direct probing of splice isoforms by RNA-seq is important, and has led to the discovery of multiple splicing defects based on single gene studies: skipping of multiple exons (exon 45-55) [106] and creation of a new exon by a deep intronic variant in DMD [107], intron retention in LMNA caused by a 5' splice site variant [108], and skipping of exon 7 in SMN1 caused by a variant in a splicing factor binding site [109]. Altogether, RNA-seq promises to be an important complementary tool to facilitate molecular diagnosis of rare genetic disorders. However, no systematic study to date has been conducted to assess its power.

1.3 Aims and scope of this thesis

In my PhD thesis I studied the consequences of DNA variation by the means of RNA sequencing data in two ways.

First, I investigate gene expression in two closely related yeast lab strains to detect and quantify regulatory mechanisms that buffer the consequences of DNA variation. I will build upon the success in quantifying local and distant as well as cis and trans regulation. Notably, local trans regulation was never measured genome-wide. This thesis will fill the gap on gene expression regulation and answer the following questions:

- What is the amount and direction of local trans regulation?
- How can cis regulatory differences be buffered in general?
- What are the mechanisms of this buffering?

In the second part of this thesis I will combine DNA variation quantified by WES with RNA-seq data to set new standards for genetic diagnosis of Mendelian disorders. The following questions will be answered:

- Which RNA features can be used for genetic diagnosis in addition to exome variants?
- Can RNA features prioritize genes missed by WES analysis?
- Can we build a general pipeline to use RNA-seq for genetic diagnosis?

Chapter 2

Negative feedback buffers effect of regulatory variants

The results presented in this section are part of the manuscript "Negative feedback buffers effect of regulatory variants" from Bader et al. 2015 [1].

Here, we sought to quantify the extent of buffering by feedback against naturally occurring regulatory variants genome-wide. To this end, we devised a novel experimental design in which ADE in a hybrid of two yeast strains is compared against ADE in a pool of spores of the same cross (Fig 2.1B). We distinguish three types of regulatory variants [42]. First, cis-regulatory variants affect by definition only the allele of the same chromosome and induce ADE in both the hybrid and the pool of spores (Fig 2.1C, left column). Instances of cis regulatory elements include transcription factor binding sites and regulatory elements in the UTR. Second, local trans mechanisms, which act in trans and are inherited together with the gene they affect, induce ADE in the pool of spores. However, as any trans effect [40, 53], local trans mechanisms act in the hybrid unspecifically on both alleles and thus do not induce ADE in the hybrid (Fig 2.1C, middle column). Local trans regulation can be due to the product of the gene itself (feedback) or to another gene in linkage disequilibrium such as a nearby encoded transcription factor [41]. Local trans regulation can reduce the ADE in the spores compared to the hybrid, if it counteracts the cis regulation (Fig 2.1B). Third, distant trans mechanisms, which are encoded on another chromosome or at a distant, unlinked locus of the same chromosome, are inherited independently of their target genes in the spores. Hence, effects of distant trans mechanisms are averaged out across the population of spores and thus do not contribute to ADE (Fig 2.1C, right column). Altogether, comparison of ADE in the hybrid against the pool of spores thus enables the dissection of local regulation into cis and local trans (including feedback) effects.

We find that buffering through local trans regulation is widespread, typically compensating for 15% of cis-regulatory effects on individual genes. It is stronger for genes with essential function and with low to middle level of expression. In contrast, re-analysis of published ribosome profiling data [51] did not support buffering at the translational level. Altogether, our results indicate that negative feedback plays an important role in buffering regulatory consequences of genetic variants.

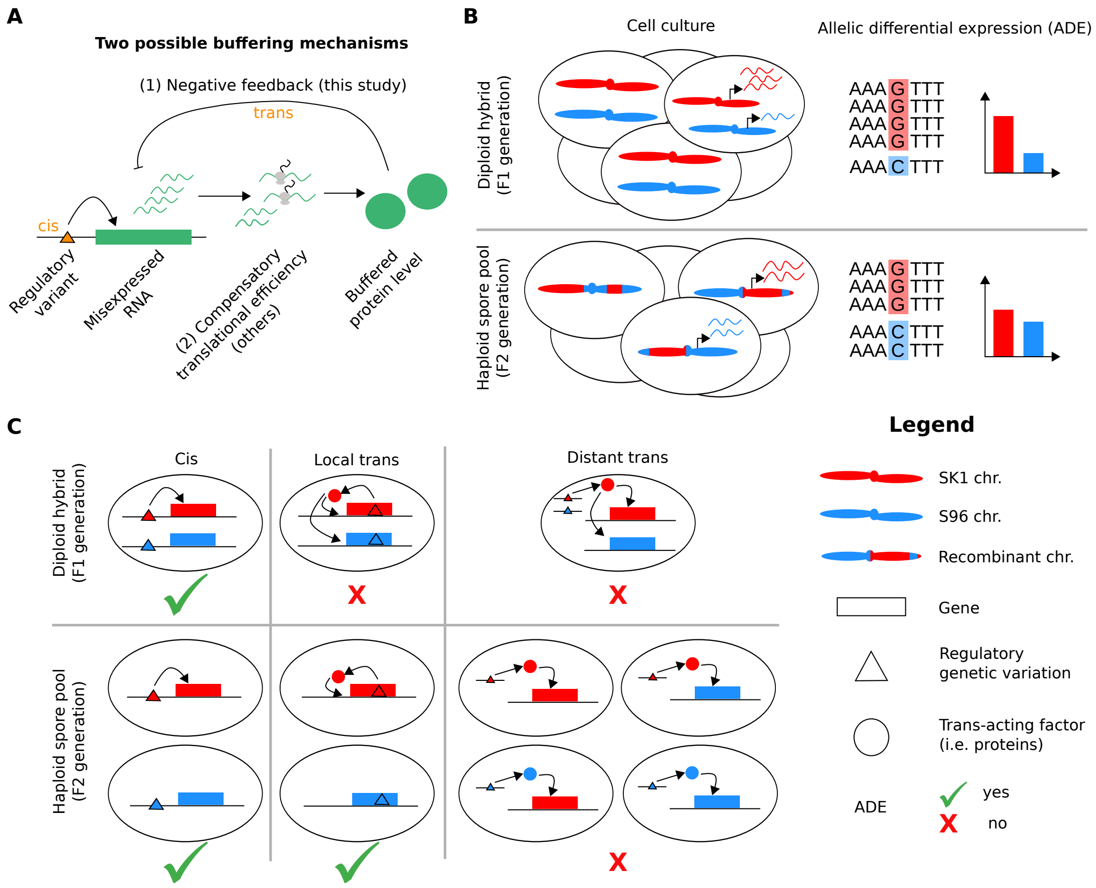

A
Two possible buffering mechanisms
(1) Negative feedback (this study)
trans
cis
Regulatory variant
Misexpressed RNA
(2) Compensatory translational efficiency (others)
Buffered protein level
B
Cell culture
Allelic differential expression (ADE)
Diploid hybrid (F1 generation)
AAA G TTT
AAA G TTT
AAA G TTT
AAA G TTT
AAA C TTT
Haploid spore pool (F2 generation)
AAA G TTT
AAA G TTT
AAA G TTT
AAA C TTT
AAA C TTT
C
Cis
Local trans
Distant trans
Diploid hybrid (F1 generation)
Haploid spore pool (F2 generation)
Legend
SK1 chr.
S96 chr.
Recombinant chr.
Gene
Regulatory genetic variation
Trans-acting factor (i.e. proteins)
ADE
yes
no

Figure 2.1: **Tested hypothesis and experimental design (A)** Effects of RNA misexpression due to cis-acting regulatory variants (orange triangle) could be buffered through (1) negative feedback of a gene product onto its RNA expression level as investigated here or (2) through compensatory translation efficiency effects as recently proposed [51, 52]. **(B)** Allelic differential expression (ADE) was estimated from allele-specific read counts in RNA-sequencing (right column) from a cross (F1 generation, top row) of the yeast strains SK1 (red) and S96 (blue) and compared against ADE from its pool of spores (F2 generation, bottom row). **(C)** Cis effects yield to ADE in both the hybrid and the pool of spores (left column). In contrast, local trans effects including feedback only yield to ADE in the pool of spores (center column). Distant trans effects do not yield to ADE neither in the hybrid nor in the pool of spores (being averaged out).

2.1 Dissecting cis and local trans regulatory effects

The reference lab strain S96 [110, 111] was crossed with the wild isolate SK1 [112, 113]. Sporulation, germination, and overnight growth of the pool of spores led to enrichment of alleles due to natural selection as well technical selection for a single mating type [114–116]. To control for this bias, allele frequencies were robustly estimated from DNA sequence data of the pools (Methods). S96 and SK1 are genetically distant strains (0.7% divergence, [113]), allowing investigation of a large set of regulatory polymorphisms and alleles. We identified 7,231 genes of a comprehensive S96 transcriptome annotation [117] that are common to both backgrounds by reciprocal best alignments with at least an identity of 95% (Methods). Out of these, the 6,934 (96%) genes that showed expression for both alleles and carried at least one polymorphism were amenable to allele-specific expression profiling by RNA-sequencing (Fig 2.1B, Methods).

RNA-sequencing showed high reproducibility between biological replicates, though higher between hybrids than between pools of spores (Supplementary Fig 2.2, Spearman correlation 0.98 and median coefficient of variation of expression level of 14% in hybrids versus 0.96 and 24% in spores, respectively). Deep sequencing led to 6,691 genes (93%, 5,078 coding and 1,613 non-coding) with more than 10 allele-specific reads on average per sample (median 1,044), for which we considered to have enough data to investigate their allele-specific regulation quantitatively. Cis and local trans effects were estimated using a generalized linear model of allele-specific RNA-sequencing read counts (using the software DESeq2 by [36], Methods). In contrast to standard methods that estimate allelic differential expression from RNA-sequencing data [46, 49, 118], our approach (i) jointly modelled all replicates, avoiding summarizations of per-replicate results that do not take between-replicate variance into account, (ii) modelled over-dispersion of RNA-sequencing read counts, limiting false positive results in comparison to Poisson or binomial models [36], and (iii) flexibly allowed controlling for covariates with known (genomic allele frequency) or with unknown (replicate, ploidy) effects. Lack of correlation of cis effect estimates with genomic allele frequency (Supplementary Fig 2.3) and L-shaped distribution of P-values (Supplementary Fig 2.4 center) indicated the validity of the method.

Overall, 984 (15%) genes showed strong and significant cis effects (cis genes, effect $>$ 1.5-fold and FDR $<$ 0.2, Benjamini-Hochberg correction here and in the following) and 54 (1%) genes showed strong and significant local trans effects (effect $>$ 1.5-fold and FDR $<$ 0.2, Supplementary Fig 2.4, Methods). When not filtering by effect size, the prevalence of cis effects in this cross (23%, 1,552) was in line with former reports in yeast (∼33%, 1400 of 4140 genes in [45]; 19% cis, 830 of 4282 genes in [46]), fly (18% cis, 1,359 of 7,631 in [119]), and mice (31% cis, 3149 of 10,090 genes in [50]). Local trans genes were enriched for genes encoding proteins that localize in the extracellular region (Gene Ontology enrichment [120], Fisher test, FDR= 0.02), in agreement with trans effects acting often due to variations in sensory processes [45]. Most of the local trans genes do not encode transcription factors (Methods) in line with the lack of enrichment of transcription factors among trans-acting regulatory loci [121] and thus were missed in the previous transcription factor screen [55]. On the other hand, *ROX1* showed no

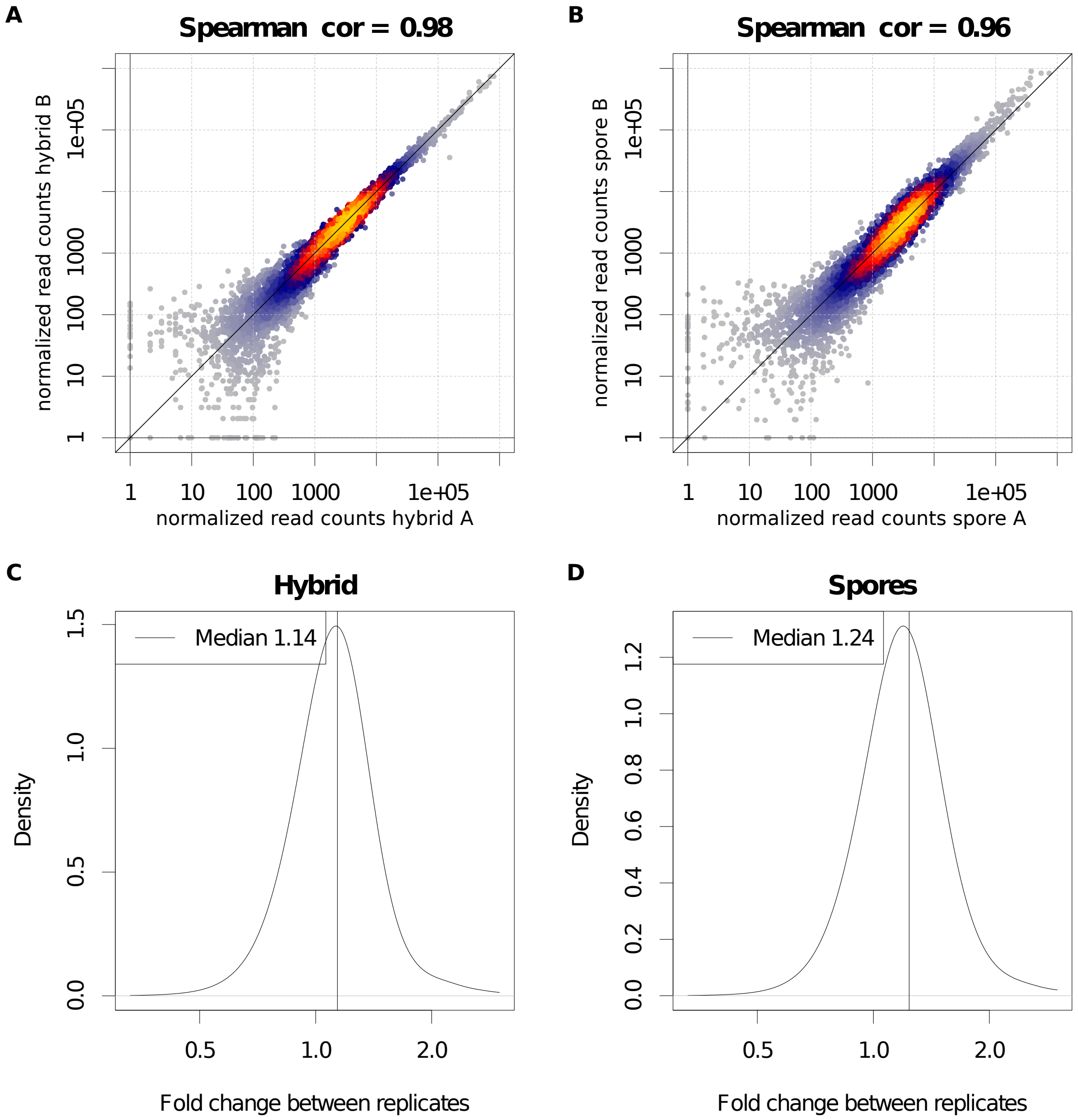


Figure 2.2: **Biological replicate variation. (A-B)** Scatter plot of gene-level allelic read counts corrected for sequencing depth and genomic allele frequency (Methods) for hybrids (A) and pools of spores (B). **(C-D)** Distribution of the gene-level fold change between the biological replicates for hybrids (C) and pools of spores (D).

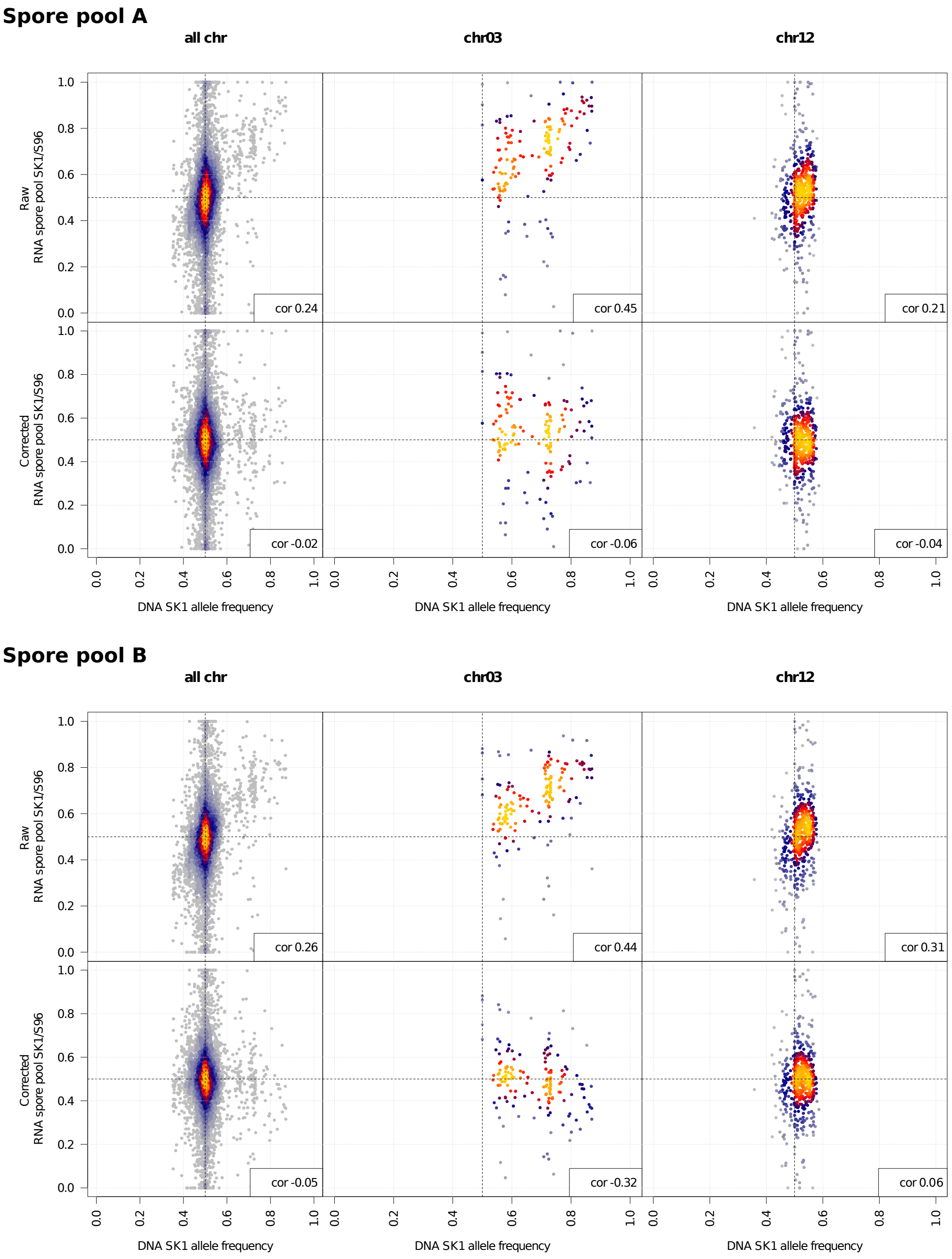


Figure 2.3: **Correction for genomic allele frequency.** For the spore pool A (top) and for the spore pool B (bottom): RNA count ratios (y-axis, top row) and RNA count ratios corrected for genomic allele frequency (y-axis, bottom row, Methods) versus genomic allele frequency (x-axis) and respective Spearman correlation (lower right corner). Artificial selection (MAT Locus on chromosome III) and natural selection (presumably for $HAP1$ on chromosome XII) leading to genomic allele frequency imbalance in the pool of spores.

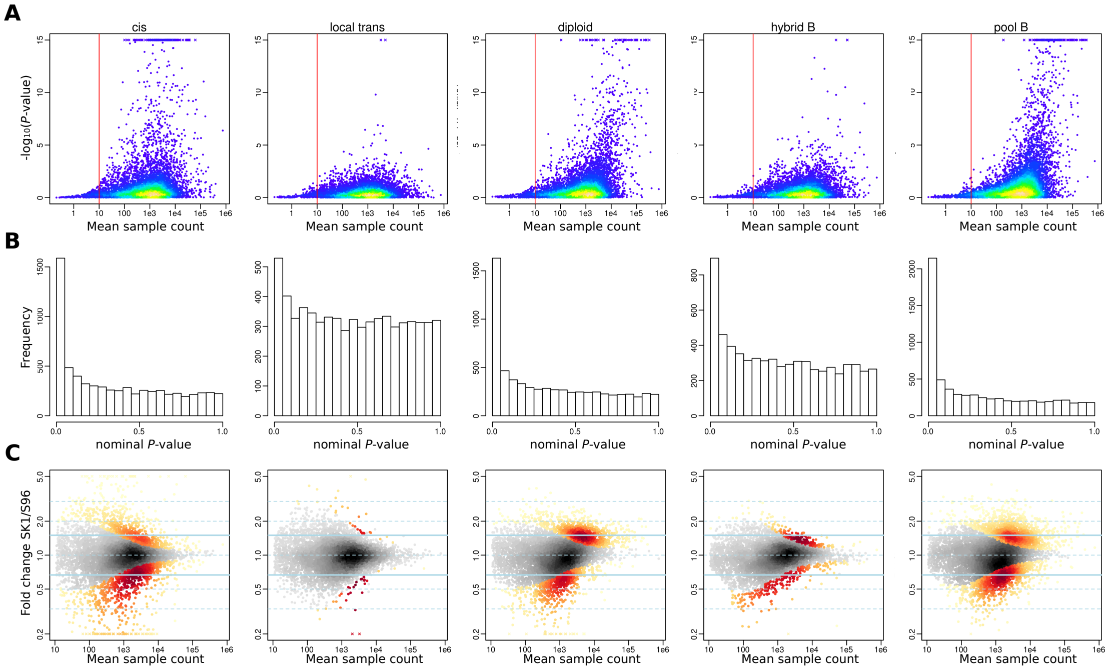


Figure 2.4: **DESeq2 statistics**. **(A)** Effect of minimum read coverage filter. No small P-value s (y-axis) at a mean sample count (x-axis) smaller than ten (red vertical line) are reported due to poor statistical power. These genes are filtered out for further analysis. Outlier with $P < 10^{-15}$ are indicated with a cross. **(B)** Histograms of nominal P-value s. All effects show the expected L-shaped (and not a J-shaped nor U-shaped) distribution of P-value s indicating that P-value s are not overestimated. **(C)** Scatter plot of fold change (y-axis) versus mean sample count (x-axis). Genes with an FDR < 0.2 are highlighted (red to yellow). Highlighted genes with a fold change greater than 1.5 (solid light blue line) were considered significant for the corresponding effect (Methods).

evidence for local trans regulation in our study, most likely because its feedback works under hypoxic conditions [55]. The much smaller amount of genes with significant local trans effects in comparison to the amount of genes with significant cis effects does not prove that local trans effects are less prevalent. Instead, this difference is likely a consequence of the limited statistical power for calling local trans effects, which relies on determining a difference between spore ADE and hybrid ADE. In comparison, there is much higher power to detect cis effects which mainly relies on determining hybrid ADE. Nonetheless, genes under documented feedback regulation including *PHO84* [122] and *AMN1* [41, 121, 123] were identified (Supplementary Fig 2.5 top). This shows that genuine strong local trans effects could be detected. Moreover, 14 out of the 54 genes showed complete buffering of cis effects through local trans regulation, i.e. they exhibited a strong ADE in the hybrid and essentially equal allelic expression in the pool of spores (hybrid count ratios larger than 1.5 and spore count ratios smaller than 1.5, examples in Supplementary Fig 2.5 bottom). Together, these findings indicate that buffering through local trans regulation might be frequent.

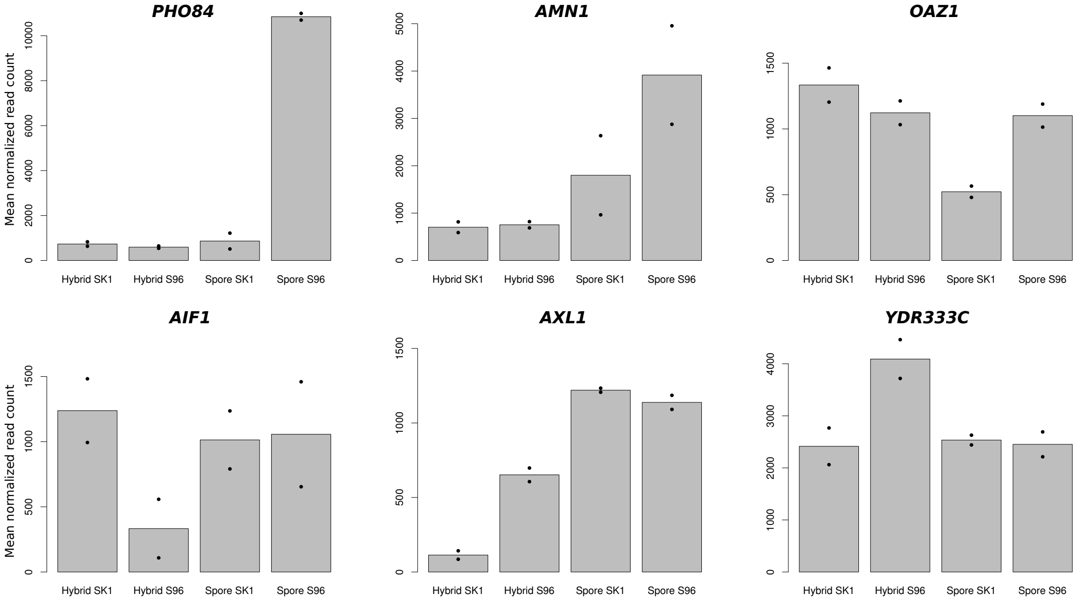


Figure 2.5: **Read counts corrected for sequencing depth and genomic allele frequency for six local trans genes.** Top row shows three genes with at least 1.5 fold difference between the count ratio (SK1/S96) of the spores, but not for the hybrid. *PHO84* is a reported case of positive feedback [122, 124], which leads to ADE in the pool of spores but not in the hybrid. *AMN1* is known to regulate itself through a negative feedback loop and to carry a coding mutation in the reference lab strain that impairs this feedback [41]. In the case of a mutation affecting the negative feedback loop itself, negative feedback is exerted only in the half of the spore population that inherited the functional feedback. Thus allelic differential expression is specific or at least stronger in the pool of spores than in the hybrid. Consistently, *AMN1* showed only allelic differential expression in the spores. Bottom row shows three genes with at least 1.5 fold difference between the count ratio of the hybrid strains, but not for the spores. Individual replicate measures are indicated by black dots.

2.2 Local trans effects buffer cis effects genome-wide

As statistical power on individual genes is limited, we also analyzed local trans regulation genome-wide. In this experimental setup, buffering can only be assessed for genes showing a cis effect in the first place. For the 984 cis genes, allelic expression imbalances typically agreed in direction, but were weaker for the pool of spores compared to the hybrid (Fig 2.6A, mass of the data subdiagonal). To quantify the amount of buffering of cis effects, we defined the buffering coefficient C as one minus the log-ratio of allele-specific expression in the spores versus the hybrid (See Methods for definition and unbiased estimation). The buffering coefficient has a value of 0 in the absence of buffering (equal ADE in the pool of spore and hybrid), 1 for complete compensation (ADE in the hybrid but no ADE in the pool of spores). The buffering coefficient is greater than 1 in case of over-compensation and is negative if local trans effects enhance cis effects. More than half of the genes with cis effects showed at least partial buffering (60% with C above 0). Local trans buffering appeared to affect all classes of genes, since no gene ontology category was significantly enriched (Fisher test, FDR < 0.1). Moreover, no significant association was found between buffering coefficient and gene features that have been associated with gene expression variability (TATA box) or dosage compensation in fly (gene length) (Supplementary Fig 2.7). The trend for buffering was robust to the definition of cis genes as it was still detectable across all genes (Supplementary Fig 2.8A). Hence, genome-wide cis effects tend to be partially buffered by local trans-regulatory mechanisms. These local trans mechanisms buffer typically 15% (Fig 2.6C, median $C = 0.148$; $P = 6.5 \times 10^{-15}$, one-sided Wilcoxon test) of allelic expression log-ratios caused by cis-regulatory variants (Fig 2.1A).

To compare the amount of buffering by local trans mechanisms against buffering by translation efficiency, we re-analyzed one ribosome-profiling dataset [51] following the same statistical procedure as above. Here, the ribosome profiles of the hybrid substitute for the transcription profiles in the pool of spores (Methods). A total of 592 genes were identified as having cis differences on RNA expression (effect > 1.5-fold and FDR < 0.2). For these genes, allelic differential levels of ribosome-bound RNAs had typically the same extent as allelic differential levels of expression of the RNAs in the hybrid (Fig 2.6B, mass of the data along the diagonal; Fig 2.6C, median buffering coefficient -0.058, 54% with $C < 0$). This observation was robust with respect to the definition of cis genes, since no support for translation efficiency buffering was detectable across all genes, too (Supplementary Fig 2.8B). We did not find an enrichment for translation efficiency opposite to ADE either when we focused on genes with both a significant ADE and significant allele-specific translation efficiency differences as the original study did (164, 54%, genes out of 303 genes with FDR < 0.2 for both effects had opposing ADE and translation efficiency, $P = 0.17$ two-sided Binomial test). Both previous publications [51, 52] could have been misled by the fact that translation efficiency estimates were technically anti-correlated with RNA levels estimates [126] and by the fact that the measurement variance was larger than assumed (Supplementary information).

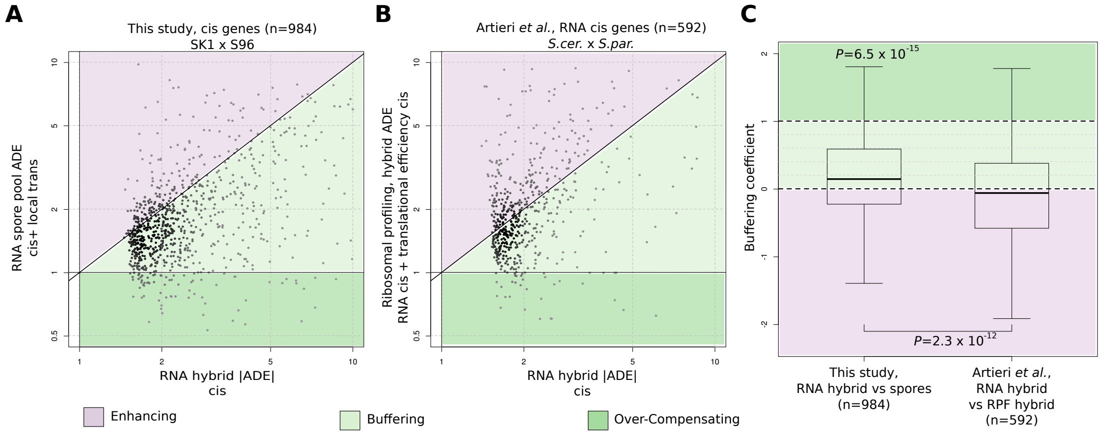


Figure 2.6: **Local trans effects, but not translation, buffer ADE (A)** Scatter plot of allele-specific expression ratios in the pool of spores (y-axis) against hybrid (x-axis) for the genes with cis effect (984 cis genes). For both axes and on a gene basis, the allele with the lower expression level in the hybrid is taken as reference (denominator). ADE in the hybrid measures cis-regulatory effects (x-axis). Three categories of genes are distinguished depending on the resulting ADE in the pool of spores (y-axis, due to cis and local trans regulation): compensated (dark green background) with canceled or opposite ADE (over compensation), buffered (light green) with reduced ADE, and enhanced ADE (purple). Most of the genes are buffered. **(B)** Analogous to (A) but for the 592 RNA cis genes of the Artieri et al. [51] dataset. Ribosomal profiling ratios (y-axis) of a cross between *S. cerevisiae* and *S. paradoxus* are compared against RNA ratios (x-axis) of the same hybrid. The mass of the data lies at the diagonal indicating that RNA cis effects in the hybrid are not buffered translationally. **(C)** Quartiles (boxes) and 1.5 times the interquartile range (whiskers) of the buffering coefficient for the gene sets from (A), left and (B), right. The buffering coefficients at RNA level are significantly greater than zero (left, median = 0.147, $P < 6.5 \times 10^{-15}$, one-sided Wilcoxon test), whereas they are not at translational level (right).

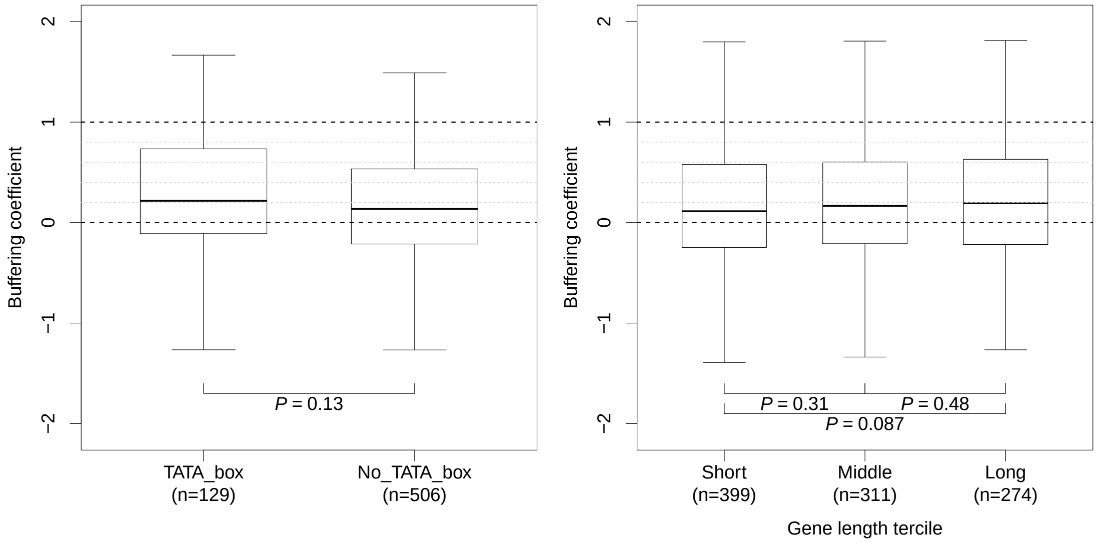


Figure 2.7: **Buffering compared to gene features**. No correlation between buffering coefficient and TATA box presence as well as gene length [125] defined as mean of SK1 and S96 length among cis genes. P-values were computed using a two-sided Wilcoxon test.

2.3 Local trans buffering is stronger for essential genes

If local trans regulation confers robustness against regulatory variants, then one would expect it to be stronger at genes important for fitness. We tested this hypothesis by classifying genes into three categories with increasing fitness relevance: 1,613 non-coding genes (24%, ncRNA), 4,004 non-essential protein-coding genes (60%, non-essential), and 1,074 essential protein-coding genes (16%, essential). The proportion of cis genes in each category was inversely related to fitness relevance (Fig 2.9A), whereby ncRNAs were enriched for cis genes (20%, $P = 1.6 \times 10^{-10}$, Fisher test) and essential genes were depleted for cis genes (11%, $P = 9.2 \times 10^{-5}$, Fisher test). This result also held when controlling for expression level and considering the combination of two FDR thresholds (0.1 and 0.2), with and without fold change cutoff (Supplementary Fig 6.1). The association of cis effects with gene categories is in line with former reports limited to protein-coding genes [45, 46] and consistent with the idea that selection on regulatory elements is more important for coding than non-coding genes and for essential than non-essential genes. Surprisingly, the buffering coefficient and fitness relevance did not correlate (Fig 2.9B). However, stratifying genes into three equally large groups with low, middle and high average expression levels revealed that highly expressed genes showed lower buffering coefficients compared to the two other groups (Fig 2.9C, median buffering coefficient= -0.036 versus 0.284 and 0.202 with $P = 3.6 \times 10^{-7}$ and $P = 6.0 \times 10^{-7}$ for low and middle levels, respectively. Wilcoxon test, Methods, Supplementary Fig 6.2 top). This result held when considering combinations of FDR and fold change cutoffs as above (Supplementary Fig 6.3). A plausible explanation for this observation is that buffering is less needed for highly expressed genes because RNAs are produced in excess and

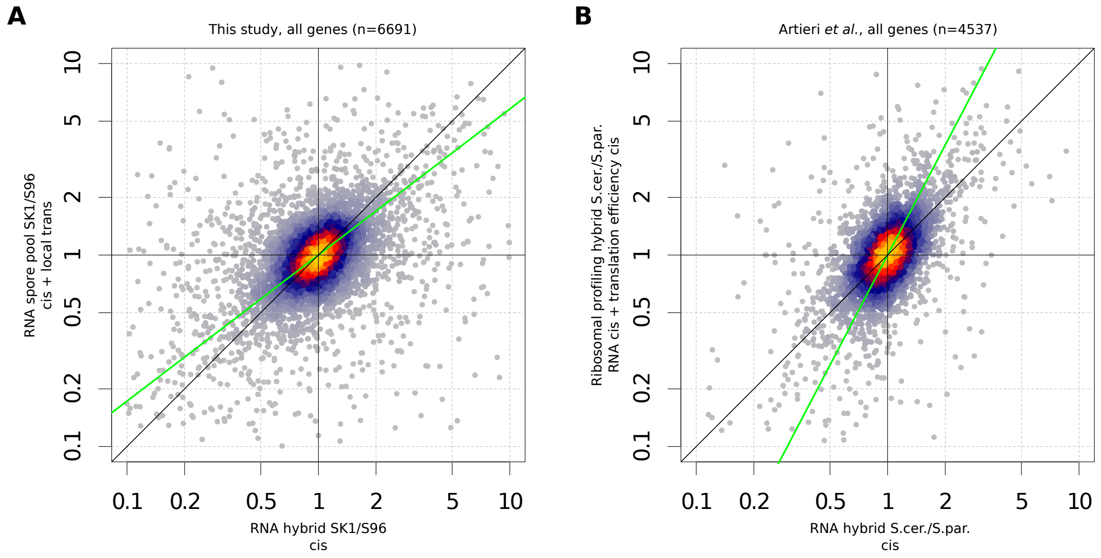


Figure 2.8: **Analysis of buffering trend across all genes (A)** Scatterplot of allelic ratio corrected for sequencing depth and genomic allele frequency in pool of spores (y-axis) against the hybrid (x-axis) for all genes. We used principal component analysis to estimate buffering across all genes because the buffering coefficient is ill-defined for non-cis genes (Methods). The trendline (green) is the direction of the first principal component. Its slope (0.75) is lower than 1 indicating genome-wide trend for buffering of cis effects by local trans effects. Note that this analysis is conservative since larger replicate variance for the pool of spores (y-axis) than for the hybrid (x-axis) leads to overestimation of the first pincipal component slope. **(B)** Scatterplot of allelic ratio in ribosome profiling data (y-axis) against allelic ratio in ribosome profiling data (x-axis) in the *S. cerevisiae* x *S. paradoxus* hybrid for all genes (data from [51]). The first principal component is above diagonal (slope=1.90, green line), thus does not provide evidence for buffering genome-wide at the translational level. Here, the slope overestimation of the principal component analysis might confound this result. Moreover, because variance between biological replicates is larger across RNAseq for pools of spores than for ribosomal profiling across hybrids, a buffering effect as large as the one seen in the spores would have been detected, if it were present at the translational level.

thus variation in their expression level has less phenotypic impact. Consistent with this hypothesis, the buffering coefficient was found to be positively associated with fitness relevance when restricted to genes with low and middle levels of expression (Fig 2.9D, Supplementary Fig 6.2 bottom). These results provide clear evidence for two regulatory strategies conferring robustness against regulatory variants: Excess amount of RNA on the one hand, and buffering through local trans regulation for low to middle levels of expression on the other hand.

2.4 Local trans buffering is primarily due to negative feedback

Buffering by local trans regulation can be caused by the gene itself (negative feedback) or by any other gene in linkage disequilibrium with it. Although negative feedback provides a simpler explanation for our data since the buffering is accomplished without the need for compensatory mutations, both mechanisms could be at play. To understand which of these two mechanisms is the major contributor to buffering, we revisited data of a previous study in which protein levels of 730 genes in diploid strains with one gene copy deleted were compared to wildtype levels [127]. In this experiment, compensatory mutations had no time to occur since the deletion was introduced artificially. Consequently, only the effect of feedback was measured. Springer and colleagues' screen was technically limited to non-essential genes and to genes with high levels of expression (63% in the highly expressed tercile, Fig 2.10A), i.e. for two gene categories for which we detected lower amounts of buffering than genome-wide. Nonetheless, we found evidence for buffering in this dataset (Fig 2.10B; median $C = 0.055$, $P = 2.1 \times 10^{-15}$ for [127], one-sided Wilcoxon-test). Moreover, buffering in these data was comparable to the amount of local trans buffering we observed for genes with matched properties (Fig 2.10B, median $C = 0.058$, Methods and Supplementary information). Hence, these deletion experiments indicate that negative feedback is the primary mechanism for local trans buffering. A further feature distinguishing negative feedback from compensatory mutation is that negative feedback also confers robustness to environmental variations. Consistently, the buffering coefficient of the cis genes negatively associated with expression response to more than 1,500 environmental perturbations [45] (median buffering coefficient=0.22 for the low versus 0.07 for the high tercile of environmental response, P-value = 0.031, one-sided Wilcoxon test, Fig 2.10C, Supplementary Fig 6.4). Altogether, these results indicate that local trans buffering is primarily due to negative feedback rather than due to compensatory mutations.

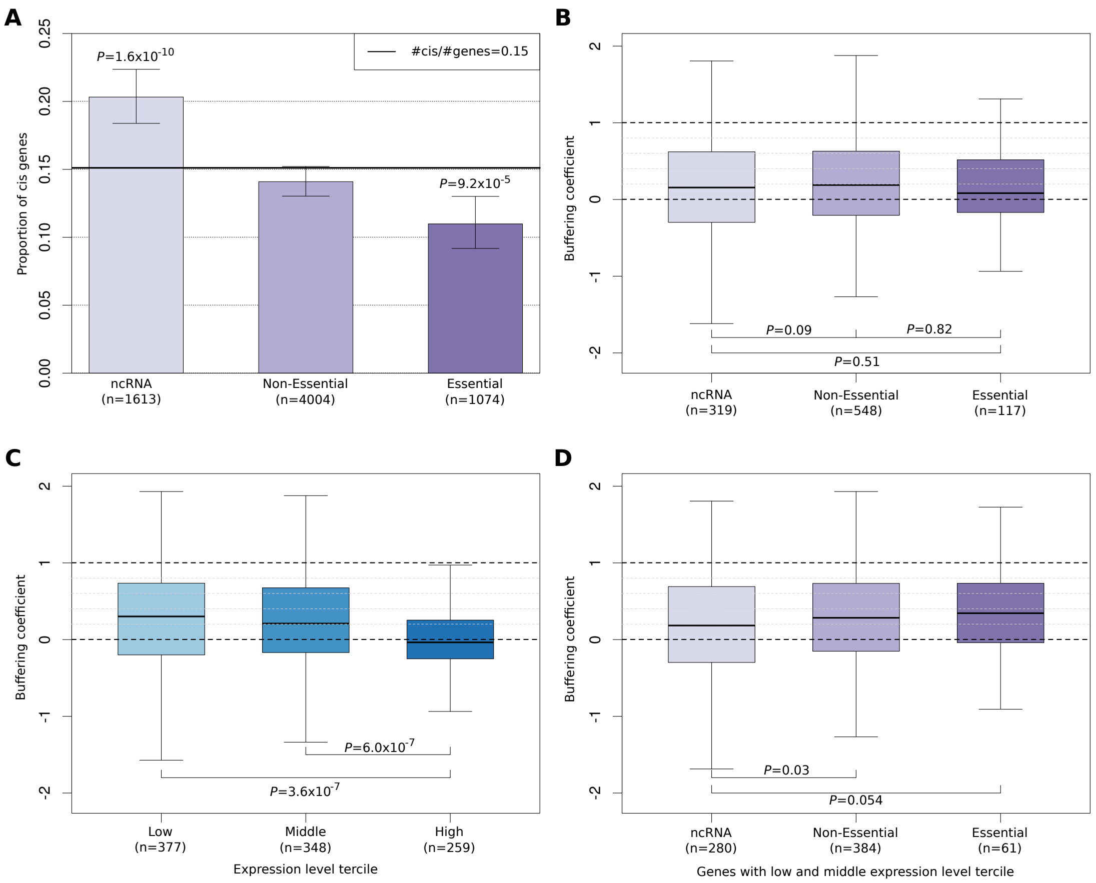


Figure 2.9: **Local trans buffering is stronger for genes important for fitness and with low to middle levels of expression. (A)** Proportion of cis genes by gene category. Essential genes show a lower cis gene proportion than genome-wide (horizontal line), whereas non-coding RNAs are enriched for cis genes (P-value from two-sided Fisher test, Error bars indicate 95% confidence intervals for binomial proportions). **(B)** Distribution of buffering coefficient for cis genes grouped by gene category. No significant differences detectable. P-value s are computed with an one-sided Wilcoxon test with the alternative hypothesis that essential genes are more buffered than ncRNA, analogously for non-essential. **(C)** Distribution of buffering coefficient for cis genes grouped by expression level tercile. Highly expressed genes are less buffered than genes with low and middle expression levels. P-value s are computed with a two-sided Wilcoxon test. **(D)** Same as (B) but for cis genes only at low and middle expression levels. At these levels of expression, buffering positively associates with fitness relevance category.

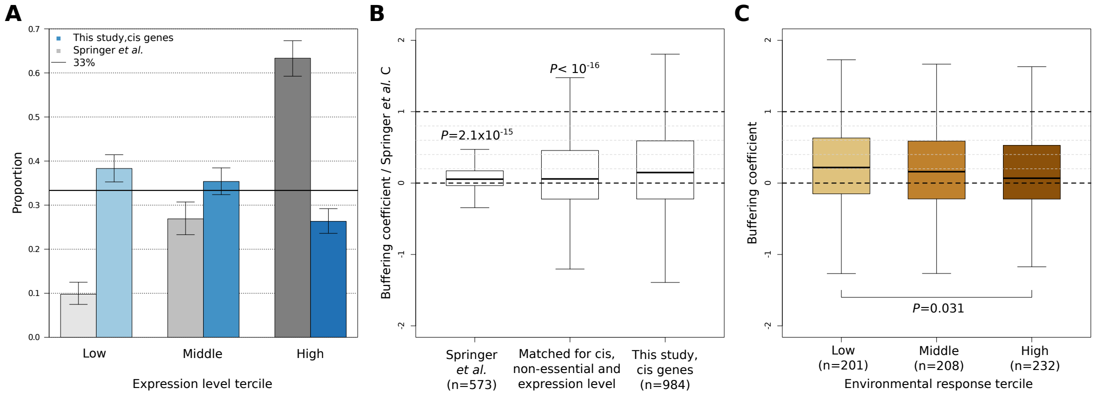


Figure 2.10: **Local trans buffering is primarily due to negative feedback.** **(A)** Proportion of expression levels in [127] dataset (gray) and from cis genes in this study (blue). Due to technical limitations, Springer and colleagues' dataset is enriched for genes with high levels of expression. Error bars indicate 95% confidence intervals for binomial proportions. **(B)** Quartiles (boxes) and 1.5 times the interquartile range (whiskers) of Springer and colleagues' C coefficient (left), of the buffering coefficient estimated in this study for cis genes with expression level distribution and gene category matching Springer and colleagues dataset (Methods, center), and of the buffering coefficient estimated in this study for all cis genes (right). Springer and colleagues' C mathematically corresponds to the here defined buffering coefficient under simple assumptions (Supplementary information). Significant buffering is found in Springer's gene ($P = 2.1 \times 10^{-15}$, one-sided Wilcoxon test). The significantly lower amount of buffering (left, median=0.055) compared to the genome-wide amount of buffering reported here (right, median=0.148) is explained by the bias for non-essential and highly expressed genes in Springer and colleagues experimental setup (median=0.058 for matched distribution, center). **(C)** Distribution of buffering coefficient for cis genes (y-axis) by tercile of median absolute value of gene expression $log2$-ratio in response to more than 1,500 environmental changes ([45], x-axis). Environmental expression data were available for coding genes only.

Chapter 3

Genetic diagnosis of Mendelian disorders via RNA sequencing

The results presented in this section are part of or adapted from the manuscript "Genetic diagnosis of Mendelian disorders via RNA sequencing" from Kremer, Bader et al. 2016 [3]. Supplementary Data are provided in the same order as in the paper and are located on our webserver[1].

Here, we established an analysis pipeline to systematically detect instances of i) aberrant expression, ii) aberrant splicing, and iii) mono-allelic expression of the alternative allele to complement whole exome sequencing based genetic diagnosis. We considered applying our approach on patients diagnosed with a mitochondrial disorder for three reasons. First, mitochondrial diseases collectively represent one of the most frequent inborn errors of metabolism affecting 2 in 10,000 individuals [128]. Second, the broad range of unspecific clinical symptoms and the genetic diversity in mitochondrial diseases makes molecular diagnosis difficult and WES often results in variants of unknown significance. As a consequence of the bi-genomic control of the energy-generating oxidative phosphorylation (OXPHOS) system, mitochondrial diseases may result from pathogenic mutations of the mitochondrial DNA (mtDNA) or nuclear genome. More than 1,500 different nuclear genes encode mitochondrial proteins [129] and causal defects have been identified in approximately 300 genes and presumably more additional disease-associated genes still awaiting identification [80]. Third, since the diagnosis often relies on biochemical testing of a tissue sample, fibroblast cell lines are usually available from those patients. Moreover, for many patients, the disease mechanisms can be assayed in epidermal fibroblast cell lines even though the disease may manifest in different tissues [130]. This allows rapid demonstration of the necessary and sufficient role of candidate variants by perturbation and complementation assays [68]. This also indicates that disease-causing expression defects, if any, should be detectable in these cell lines.

We performed RNA-seq on 105 fibroblast cell lines from patients with a suspected mitochondrial disease including 48 patients for which whole exome sequencing based variant prioritization did not yield a genetic diagnosis (Fig 3.1, Table 3.1, Methods 4.2.2, Supplementary Data 1). After discarding lowly expressed genes,

[1] `https://i12g-gagneurweb.informatik.tu-muenchen.de/public/paper/mitoMultiOmics/bioRxiv_2016_12_28/paper_supplement_data/`

RNA-seq identified 12,680 transcribed genes (at least 10 reads in 5% of all samples, Methods 4.2.4). We systematically prioritized genes with the following three strategies: i) genes with aberrant expression level [96–98], ii) genes with aberrant splicing [107, 131], and iii) mono-allelic expression of rare variants [99] to estimate their disease association (Fig 3.1). All strategies are based on the comparison of one patient against the rest. We assumed the causal defects to differ between patients, which is reasonable for mitochondrial disorders with a diversity of 300 known disease-causing genes (Supplementary Data 2). Therefore, the patients serve as good controls for each other.

Table 3.1: Sample numbers. Number of samples measured with the specified quantification method binned by their diagnosis status.

Method	Diagnosed	Not diagnosed	Total
RNA	57	48	105
RNA & WES	40	48	88
RNA & proteomics	11	20	31

3.1 Aberrant expression

Before we could assess aberrant expression, we normalized for technical biases, sex, and biopsy site as follows: Hierarchical clustering revealed three main clusters that could not be linked to biological or technical properties of the samples (Fig 3.2A). These clusters were considered as groups of unknown technical biases. We could improve the correlation between replicated samples significantly via correction for the technical biases (Two-sided Wilcoxon-test P-value 0.02, Fig 3.2B). Furthermore, 5 HOX genes were among the 150 genes with most variable expression (3.3% w.r.t. 0.2% HOX genes in all 12,680 genes, Fig 3.2C). HOX genes are important regulators of the body plan during development of the anterior-posterior axis [132]. Since the fibroblast cell lines are taken from different body parts depending on the clinic, the first diagnosis and other factors, we hypothesized that the expression pattern of the HOX genes is a good proxy for the body parts of the biopsy. To identify likely biopsy site groups we performed hierarchical clustering of the RNA expression across all samples based only on the HOX genes (identified as genes with names starting with "HOX"), which revealed four major sample clusters (Fig 3.2D). After normalization for technical biases, sex, and biopsy site hierarchical clustering did not reveal further biases (Fig 3.2E, Methods 4.2.5).

The samples typically presented few aberrantly expressed genes (median of 1, Fig 3.3, Supplementary Table 1) with a large effect ($|\mathrm{Z-score}| > 3$) and significant differential expression (Hochberg adjusted P-value < 0.05). Among the most aberrantly expressed genes across all samples, we found 2 genes encoding mitochondrial proteins, MGST1 (one case) and TIMMDC1 (two cases) to be significantly down-regulated (Fig 3.4). For both genes, WES did not identify any variants in the respective patients, no variant is reported to be disease-associated, and no case of potential bi-allelic rare variant is listed in our in-house database comprising more than 1,200 whole-exomes from mitochondrial patients and 15,000 WES dataset available to us from different ongoing research projects.

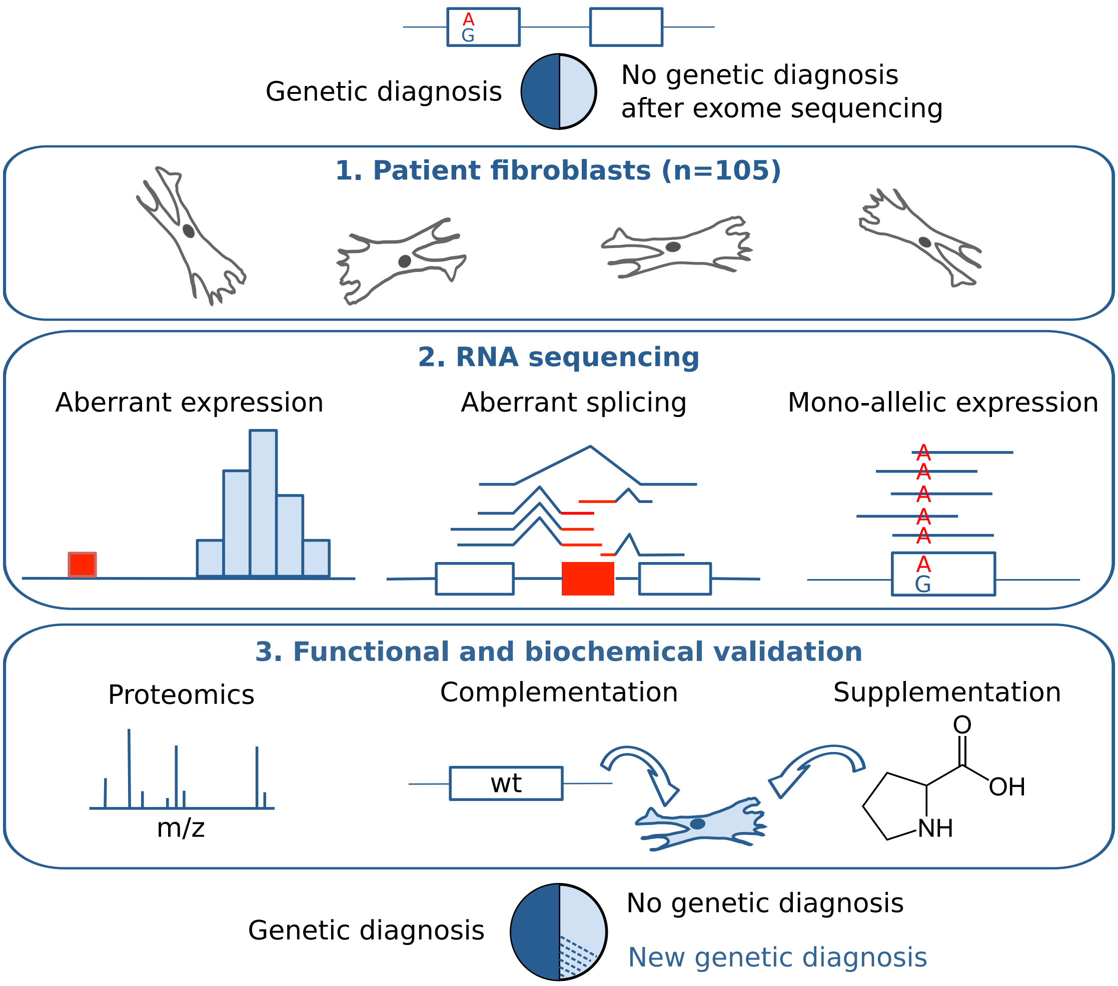


Figure 3.1: **Strategy for genetic diagnosis using RNA-seq.** The approach we followed started with RNA sequencing of fibroblasts from unsolved WES patients. Three strategies to facilitate diagnosis were pursued: Detection of aberrant expression (e.g. depletion), aberrant splicing (e.g. exon creation) and mono-allelic expression of the alternative allele (i.e. A as alternative allele). Candidates were validated by proteomic measurements, lentiviral transduction of the wildtype (wt) allele or, in particular cases, by specific metabolic supplementation.

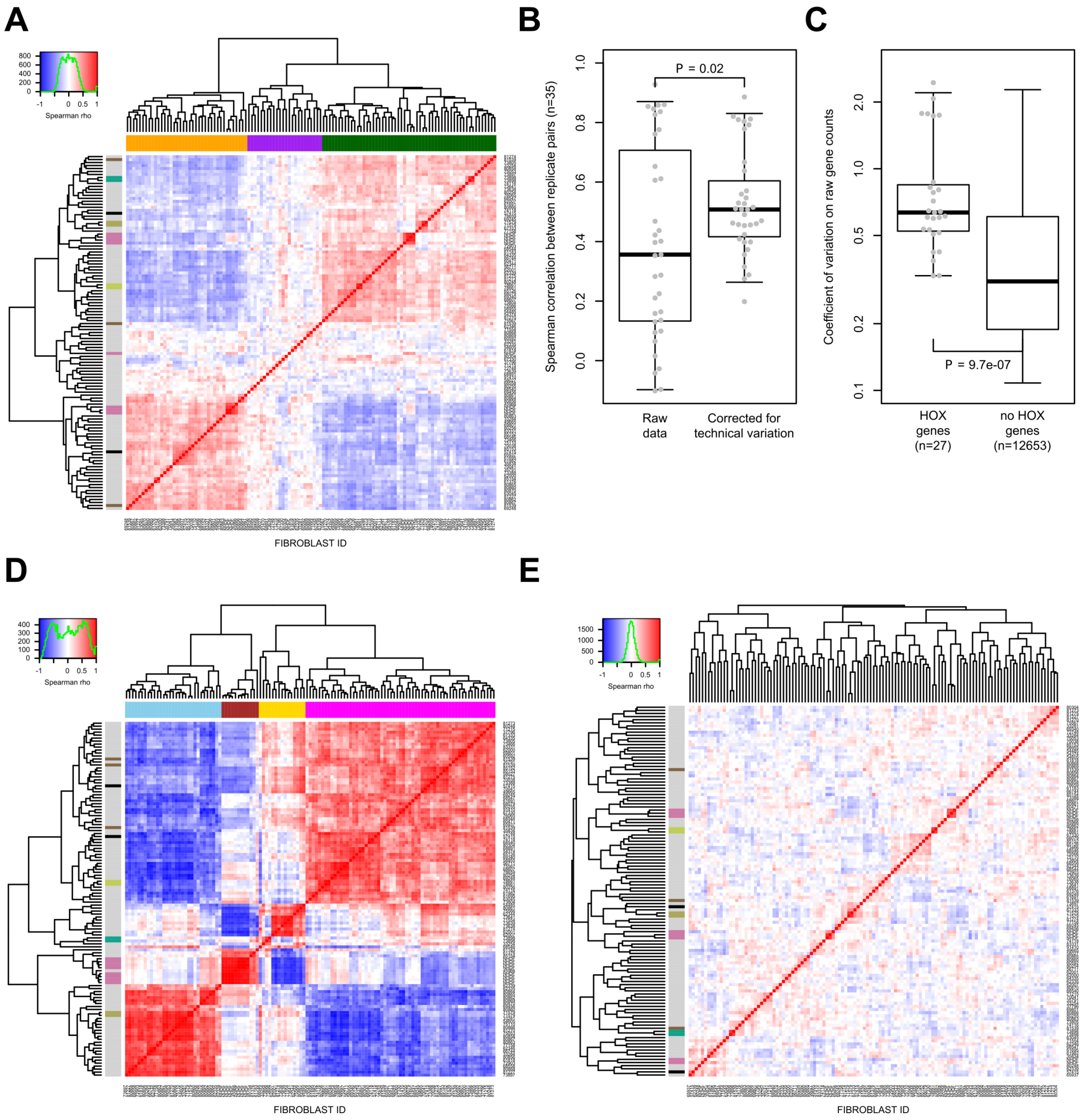


Figure 3.2: **RNA normalization.** **(A)** Spearman-correlation heat map of size-factor normalized gene expression between all fibroblasts (n=119) including biological replicates (left side color code). The dendrogram represents the sample-wise hierarchical clustering. The color code on the top depicts the top three clusters. The color key of the spearman rho value (top left) includes a histogram based on the values (green line). **(B)** Boxplot of the spearman correlation between all replicate pairs (n=35) before and after normalizing for technical variation. Equi-tailed 95% interval (whiskers), 25th, 75th percentile (boxes) and median (bold horizontal line) are indicated. The *P*-value is based on a two-sided Wilcoxon test. **(C)** Boxplot as in (B) of coefficients of variation (standard deviation / mean) for the 30 HOX genes and the remaining genes based on raw gene counts. The *P*-value is based on a two-sided Wilcoxon test. **(D)** Same as (A), but correlation is computed only on the HOX genes among all samples. The top four clusters are highlighted (color code top). **(E)** Same as (A) after normalization for the technical biases, sex variation and four HOX gene groups.

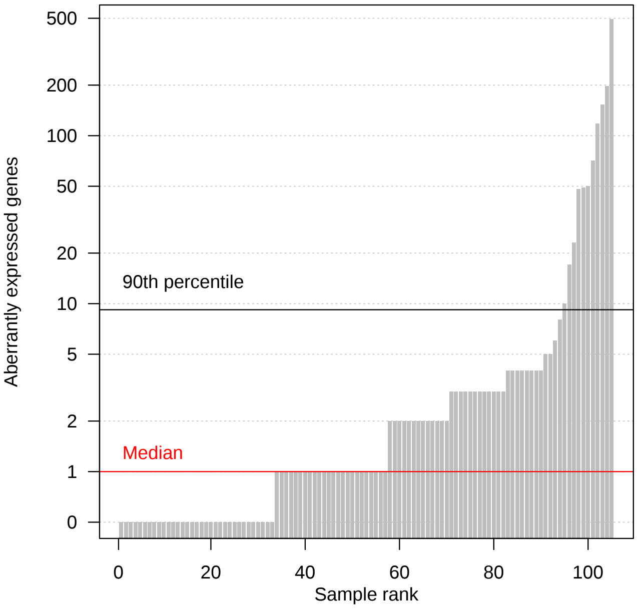


Figure 3.3: **Overview aberrant expression.** Aberrantly expressed genes (Hochberg corrected P-value < 0.05 and $|\mathrm{Z} - \mathrm{score}| > 3$) for each patient fibroblast cell line.

To evaluate the consequences of diminished RNA expression at the protein level, we performed quantitative proteomics in a total of 31 fibroblast cell lines (including these three patients, Table 3.1, Methods 4.2.9, Supplementary Data 3) from a second aliquot of cells taken at the same time as the RNA-seq aliquot. Normalized RNA and protein expression levels showed a median rank correlation of 0.59, comparable to what has been previously reported [95, 133] (App. Fig 6.7). Patient #73804 showed 2% of control MGST1 level whilst the lack of detection of TIMMDC1 in both patients (#35791 and #66744) confirmed an even stronger effect on protein expression, indicating loss of function (Fig 3.5).

MGST1, a microsomal glutathione S-transferase, is involved in the oxidative stress defense [134]. Accordingly, MGST1 depletion results in significantly increased reactive oxygen species (ROS) levels compared to a healthy individual (Fig 3.6A, paper methods on ROS). Magnetic resonance images showed also a progressive brain atrophy for this patient (Fig 3.6B), who suffers from an infantile-onset neurodegenerative disorder similar to a recently published case with another defect in the ROS defense system (App. 6.3.2) [135]. Consequently, the loss of expression of MGST1 is not only a likely cause of the disease of this patient, but also suggests a treatment with antioxidants.

Both TIMMDC1 patients presented with muscular hypotonia, developmental delay, and neurological deterioration, which led to death in the first 3 years of life (App. 6.3.2). Consistent with the described function of TIMMDC1 as a respiratory chain complex I assembly factor [136, 137], we found isolated complex I deficiency in muscle (Fig 3.7A,B), and globally decreased levels of complex I subunits in fibroblasts by quantitative proteomics (Fig 3.5) and western blot (Fig 3.7C). Re-expression of TIMMDC1 in these cells increased complex I subunit levels (Fig 3.7C). These results not only validate TIMMDC1-deficiency as

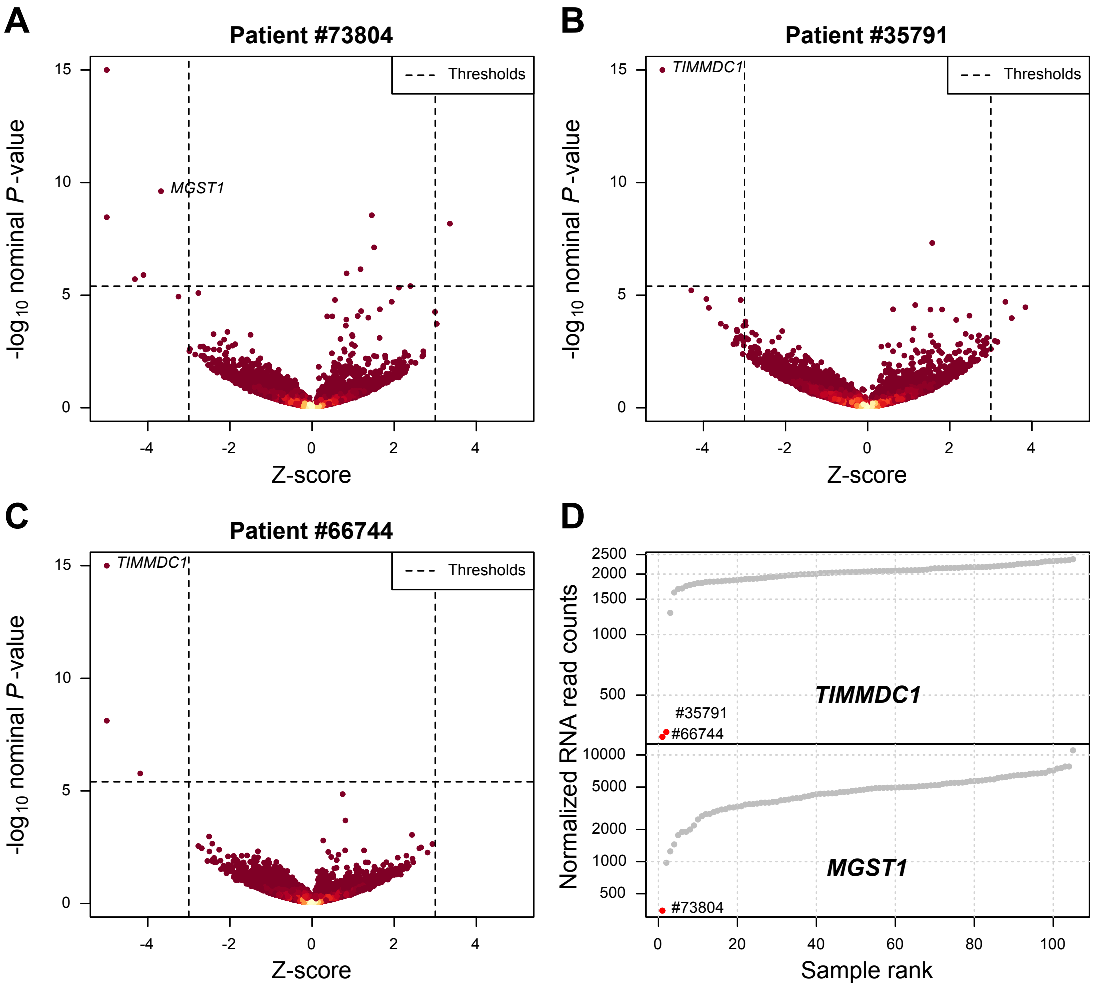


Figure 3.4: **Examples aberrant expression.** **(A)** Gene-wise RNA expression volcano plot of nominal P-value s (- log10 P-value) against Z-scores of the patient #73804 compared against all other fibroblasts. Absolute Z-scores greater than 5 are plotted at ±5, respectively. The y-axis is limited to 15, more extreme values are shown at 15. **(B)** Same as (a) for patient #35791. **(C)** Same as (A) for patient #66744. **(D)** Sample-wise RNA expression is ranked for the genes TIMMDC1 (top) and MGST1 (bottom). Samples with aberrant expression for the corresponding gene are highlighted in red (#35791, #66744, and #73804).

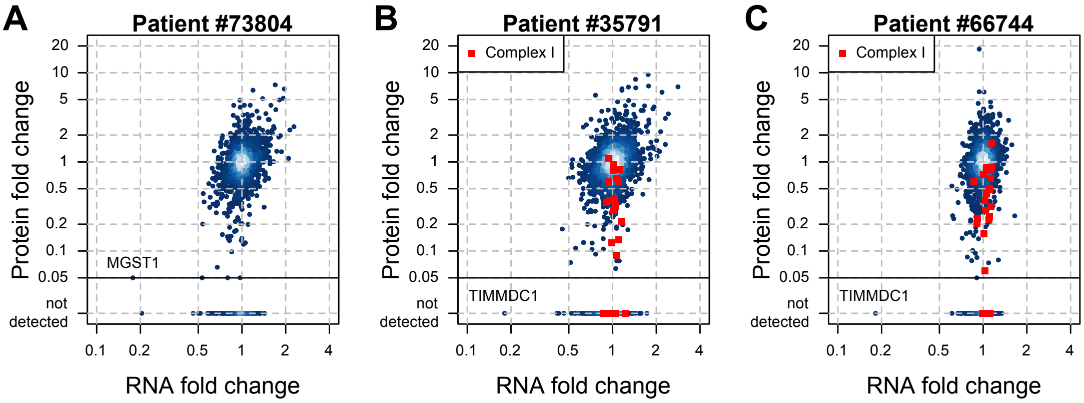


Figure 3.5: **Comparison of RNA and protein changes. (A)** Gene-wise comparison of RNA and protein fold changes of patient #73804 against all other fibroblast cell lines. Protein fold changes lower than 0.05 are plotted on the horizontal line. Reliably detected proteins that were not detected in this sample are shown separately with their corresponding RNA fold changes (points below solid horizontal line). **(B)** Same as (A) for patient #35791. Subunits of the mitochondrial respiratory chain complex I are highlighted (red squares). **(C)** Same as (B) for patient #66744.

disease causing but also provide compelling evidence for an important function of TIMMDC1 in complex I assembly.

3.2 Aberrant splicing

We identified aberrant splicing events by testing for differential splicing in each patient against the others, using an annotation-free algorithm able to detect also novel splice sites (median of 5 abnormal events per sample, Fig 3.8A, Methods 4.2.6). Among the 175 aberrant spliced genes detected in the undiagnosed patients, the most abundant events were, apart from differential expression of isoforms, exon skipping followed by the creation of new exons (Fig 3.8B). Two genes encoding mitochondrial proteins, TIMMDC1 and CLPP, which were among the 20 most significant genes, caught our attention (Supplementary Table 3).

Out of 136 exon-junction reads overlapping the acceptor site of CLPP exon 6 for patient #58955, 82 (percent spliced in [138] $\Psi = 60\%$) skipped exon 5, and 14 ($\Psi = 10\%$) showed a 3'-truncated exon 5, in striking contrast to other samples (Fig 3.9A). The likely genetic cause of these two splice defects is a rare homozygous variant in exon 5 of CLPP affecting the last nucleotide of exon 5 (c.661G>A, p.Glu221Lys 1.2×10^{-5} frequency in the ExAC database [74], Fig 3.9B). Both detected splice defects result in truncated CLPP and western blots corroborated the complete loss of full-length CLPP (Fig 3.9C). Our WES variant filtering reported this variant as a variant of unknown significance (VUS) and classified CLPP as one among 30 other potentially bi-allelic affected candidate genes (see Supplementary Table 1 of [3]). Since the variant was of unknown significance, the patient remained without genetic diagnosis. The loss of function found by RNA-seq and confirmed by Western blotting now highlights clinical relevance of the variant within CLPP. CLPP encodes a mitochondrial ATP-dependent endopeptidase [139] and CLPP-

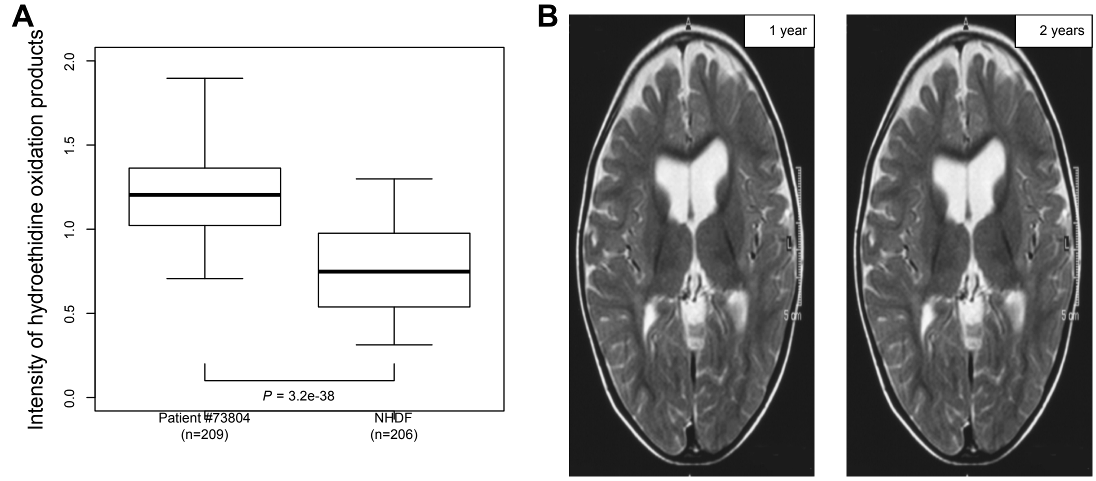


Figure 3.6: **Details for patient #73804. (A)** Quantification of cellular reactive oxygen species production. Hydroethidine oxidation production was measured using epifluorescence microscopy. Equi-tailed 95% interval (whiskers), 25th, 75th percentile (boxes) and median (bold horizontal line) are indicated. The P-value is based on a two-sided Wilcoxon test. **(B)** Magnetic resonance imaging of the brain of patient #73804 at the age of one and two years (left and right panel, respectively).

deficiency causes Perrault syndrome [140, 141] (OMIM #601119) which is overlapping with the clinical presentation of the patient investigated here including microcephaly, deafness, and severe psychomotor retardation (App. 6.3.2). Moreover, a study recently showed that Clpp-/- mice are deficient for complex IV expression [142], in line with complex IV deficiency of this patient (Fig 3.9D).

Split read distribution indicated that both TIMMDC1-deficient patients expressed almost exclusively a TIMMDC1 isoform with a new exon in intron 5 (Fig 3.10A). This new exon introduces a frameshift yielding a premature stop codon (p.Gly199_Thr200ins5*, Fig 3.10B). Moreover, this new exon contained a rare variant (c.596+2146A>G) not listed in the 1,000 Genomes Project [143, 144]. Whole genome sequencing demonstrated that this variant is homozygous in both patients (Fig 3.10B,C), the only rare variant in this intron, and close to the splice site (+6 of the new exon). We could not identify any rare variant in the promoter region or in any intron-exon boundary of TIMMDC1. Additionally, when testing six prediction tools for splicing events, this deep intronic rare variant is predicted by SpliceAid2 [145] to create multiple binding sites for splice enhancers. Together with the correctly predicted new acceptor and donor sites by SplicePort [146] (Feature generation algorithm score 0.112 and 1.308, respectively) this emphasizes the influence of this variant in the creation of the new exon. Besides, the four other tools predicted no significant change in splicing [147–150].

We further discovered an additional family in our in-house WGS database (consisting of 36 patients with a suspected mitochondrial disorder and 232 further patients with unrelated diseases) carrying the same homozygous intronic variant. In this family three affected siblings presented with similar clinical symptoms although without a diagnosis of a mitochondrial disorder (Fig 3.10C). Two siblings died before the age of 10. A younger brother (#96687), now 6 years of age,

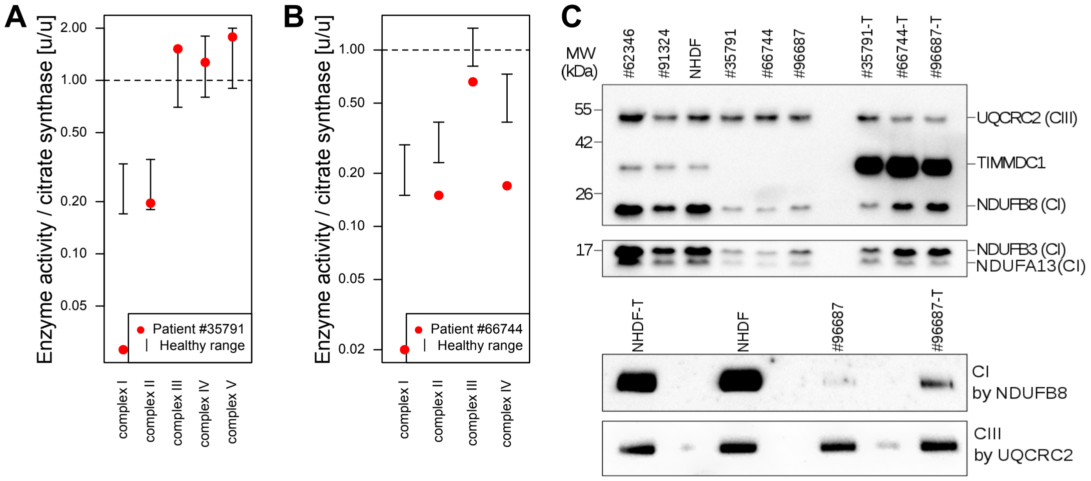


Figure 3.7: **TIMMDC1 validation. (A)** Enzyme activities of respiratory chain complexes I-V of #35791. Activities were measured in a muscle biopsy and normalized to citrate synthase. **(B)** Enzyme activities of respiratory chain complexes I-IV of #66744. Analogous to (A). **(C-TOP)** Western blot of TIMMDC1, NDUFA13, NDUFB3, and NDUFB8 protein in three fibroblast cell lines without (#62346, #91324, NHDF) and three with a variant in TIMMDC1 (#35791, #66744, #96687), and fibroblasts re-expressing TIMMDC1 ("-T") (#35791-T, #66744-T, #96687-T). UQCRC2 was used as loading control. MW, molecular weight; CI, complex I subunit; CIII, complex III subunit. **(C-BOTTOM)** Blue native PAGE blot of the control fibroblasts re-expressing TIMMDC1 (NHDF-T), the control fibroblasts (NHDF), patient fibroblasts (#96687), and patient fibroblast re-expressing TIMMDC1 (#96687-T). Immunodecoration for complex I and complex III was performed using NDUFB8 and UQCRC2 antibodies, respectively. CI, complex I subunit; CIII, complex III subunit.

presented with muscle hypotonia, failure to thrive and neurological impairment (App. 6.3.2), similar to the patients described above. Western blot analysis confirmed TIMMDC1-deficiency (Fig 3.7C) and impaired complex I assembly, which was restored after re-expression of TIMMDC1 (Fig 3.7C). The discovery of the same intronic TIMMDC1 variant in three unrelated families from three different ethnicities provides convincing evidence on the causality of this variant for the TIMMDC1 loss-of-function.

In almost all non-TIMMDC1-deficiency samples, we noticed a few split reads supporting inclusion of the new exon (Fig 3.10A), consistent with an earlier report that many cryptic splice sites are not entirely repressed but active at low levels [151]. We set out to quantify this phenomenon and to interrogate the frequency of private exons originating from weakly spliced exons, independent of their possible association with disease. Consequently, we modeled the distribution of Ψ for the 1,603,042 splicing events detected genome-wide in 105 samples as a mixture of three components (Methods 4.2.6). The model classified splicing frequencies per splice site as strong (20%, with $\Psi > 5.3\%$), weak (16%, with $0.16\% < \Psi < 5.3\%$), or background (64%, with $\Psi < 0.16\%$, Fig 3.11, App. Fig 6.8). Strikingly, the majority (70%, 4.4-fold more than by chance) of the 17 discovered private exons originated from weak splice sites (Fig 3.11 bottom). These data confirm that

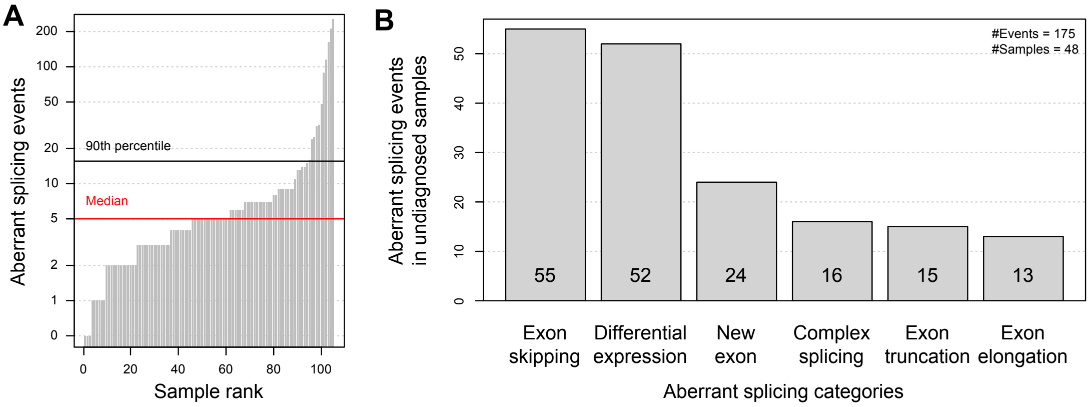


Figure 3.8: **Aberrant splicing detection. (A)** Aberrant splicing events (Hochberg corrected *P*-value < 0.05) for all fibroblasts. **(B)** Aberrant splicing events (n=175) grouped by their splicing category in undiagnosed patients (n=48) after manual inspection.

weakly spliced cryptic exons are loci more susceptible to turn into strongly spliced sites than other intronic regions. These weak splicing events are usually dismissed as 'noise' since they are only supported by few reads in a given sample. Our analysis shows that they can be detected as accumulation points across multiple individuals. Hence, these results suggest that the prioritization of deep intronic variants of unknown significances gained through whole genome sequencing could be improved by annotating weak splice sites and their resulting cryptic exons.

3.3 Mono-allelic expression

As a third approach, we searched for mono-allelic expression (MAE) of rare variants. In median per sample, 35,521 heterozygous single nucleotide variant (SNV)s were detected by WES, of which 7,622 were sufficiently covered by RNA-seq to call mono-allelic expression (more than 10 reads), 20 showed MAE (Hochberg adjusted *P*-value $<$ 0.05, allele frequency $\geq$ 0.8), of which 6 were rare variants (minor allele frequency $<$ 0.001, Methods 4.2.7, Fig 3.12).

Amongst the 18 rare mono-allelic expressed variants in patient #80256 was a VUS in ALDH18A1 (c.1864C>T, p.Arg622Trp, Fig 3.13A-C), encoding an enzyme involved in mitochondrial proline metabolism [152]. This VUS had been seen in WES as compound heterozygous with a nonsense variant (c.1988C>A, p.Ser663*, Fig 3.13A-C). Variants in ALDH18A1 had been reported to be associated with cutis laxa III (OMIM #138250) [153, 154], yet the patient did not present cutis laxa. Because of this inconsistent phenotype and the unknown significance of the non-synonymous variant, the variants in ALDH18A1 were not regarded as disease causing. However, RNA-seq-based aberrant expression (Fig 3.13D,E) and mono-allelic expression analysis prioritized ALDH18A1 again. Our systematically performed validation by quantitative proteomics revealed severe reduction down to 2% ALDH18A1 (Fig 3.14A), indicating that the rare MAE variant affects translation or protein stability. Metabolomics profile of blood plasma was in accordance with a defect in proline metabolism (Fig 3.14B) and the following changes in urea cycle. Patient fibroblasts showed a growth defect that was rescued by supplementa-

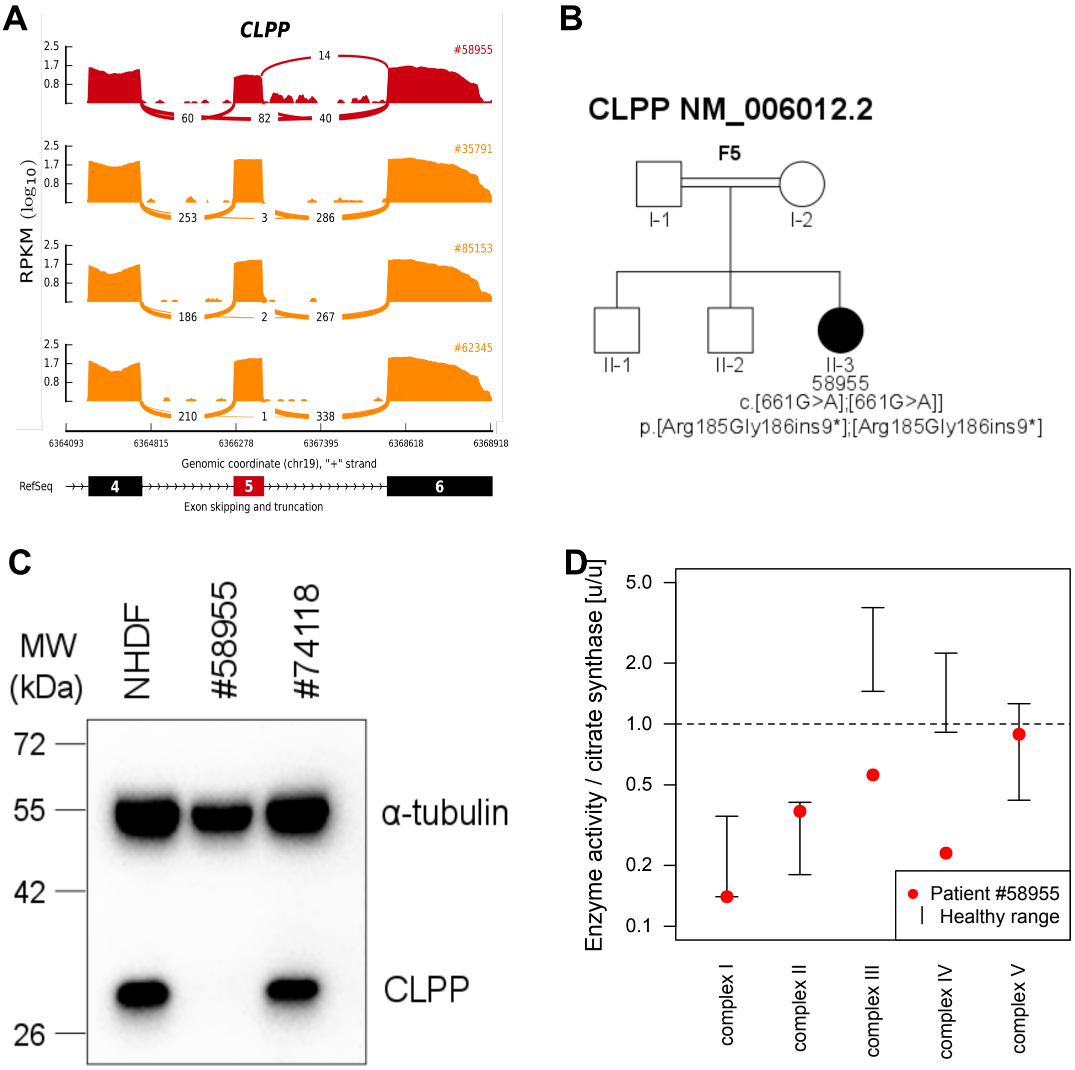


Figure 3.9: **CLPP aberrant splicing.** **(A)** CLPP Sashimi plot of exon skipping and truncation events in affected and unaffected fibroblasts (red and orange, respectively). The RNA coverage is given as the log10 RPKM-value and the number of split reads spanning the given intron is indicated on the exon-connecting lines. At the bottom the gene model of the RefSeq annotation is depicted. The aberrantly spliced exon is colored in red. **(B)** Pedigree of the family with mutations in CLPP showing the mutation status. **(C)** Western Blot showing the amount of CLPP for the NHDF cell line, the patient carrying variants in CLPP (#58955), and a patient not carrying variants in CLPP (#74118). α-tubulin was used as loading control. **(D)** Enzyme activities of respiratory chain complexes I-V of #58955. Activities were measured in a muscle biopsy and normalized to citrate synthase.

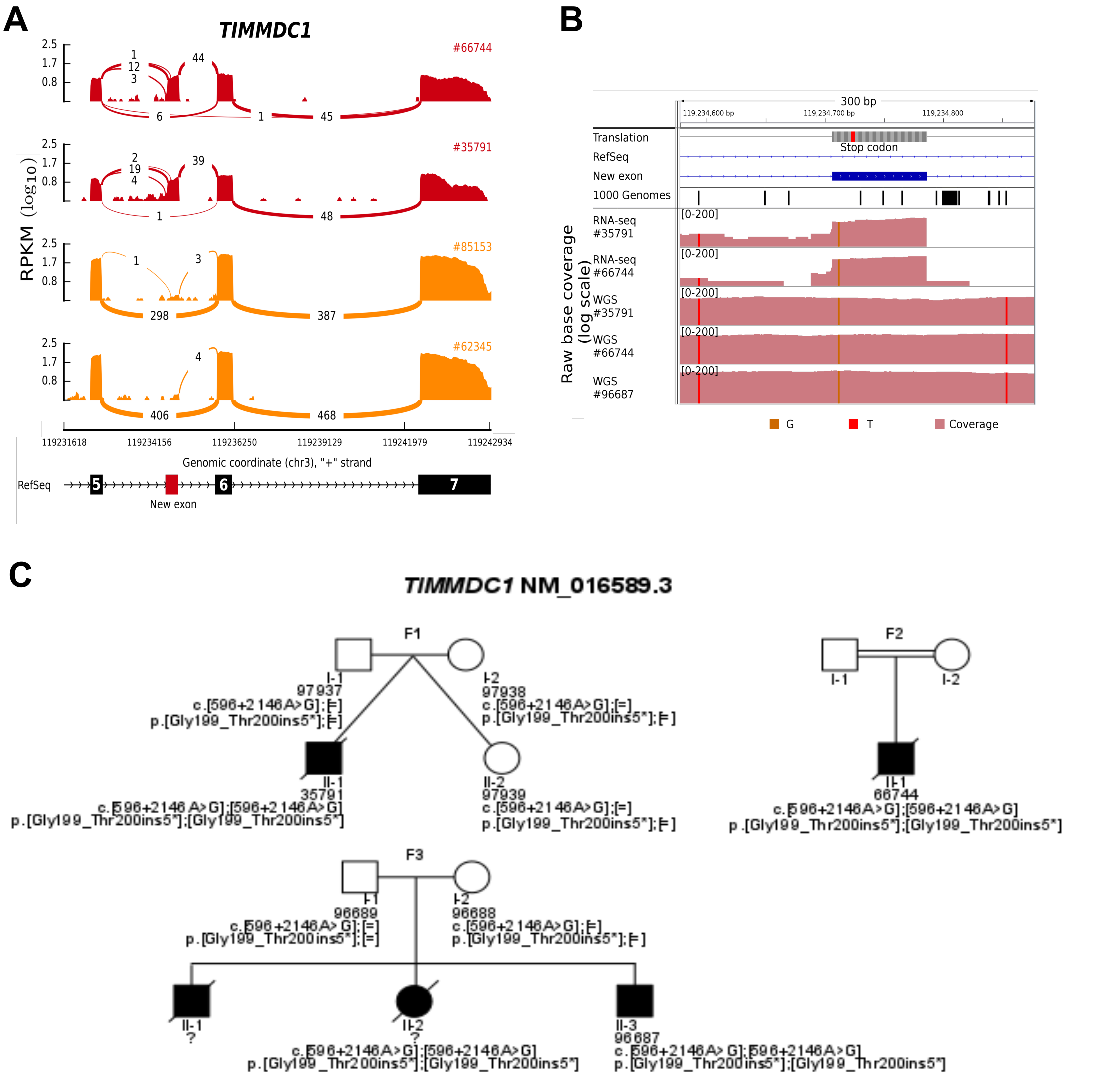


Figure 3.10: **TIMMDC1 aberrant splicing.** **(A)** TIMMDC1 Sashimi plot of exon creation events in affected and unaffected fibroblasts (red and orange, respectively). The RNA coverage is given as the log10 RPKM-value and the number of split reads spanning the given intron is indicated on the exon-connecting lines. At the bottom the newly created exon is depicted in red within the RefSeq annotation track. **(B)** Coverage tracks (light red) for patients #35791, #66744, and #91324 based on RNA and whole genome sequencing. For patient #91324 only WGS is available. The homozygous SNV c.596+2146A>G is present in all coverage tracks (vertical orange bar). The top tracks show the genomic annotation: genomic position on chromosome 3, DNA sequence, amino acid translation (grey, stop codon in red), the RefSeq gene model (blue line), the predominant additional exon of TIMMDC1 (blue rectangle), and the SNV annotation of the 1000 Genomes Project (each black bar represents one variant). **(C)** Pedigrees of the families with mutations in TIMMDC1 showing the mutation status. Mutations are confirmed by Sanger sequencing.

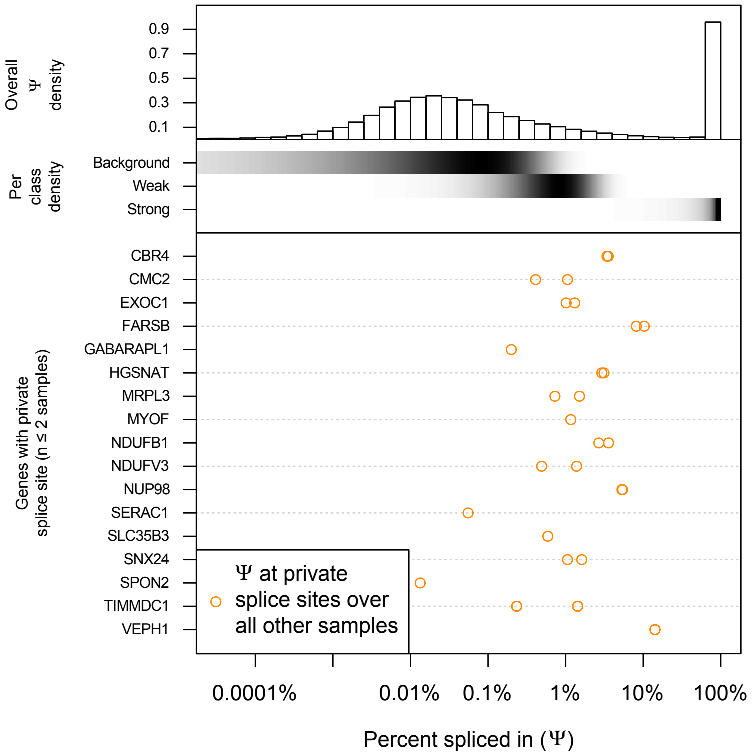


Figure 3.11: **Weak splicing.** Percent spliced in (Ψ) distribution for different splicing classes and genes. Top: Histogram giving the genome-wide distribution of the 3' and 5' Ψ-values based on all reads over all samples. Middle: The shaded horizontal bars represent the densities (black for high density) of the background, weak and strong splicing class, respectively (Methods 4.2.6). Bottom: Ψ-values of the predominant donor and acceptor splice sites of genes with private splice sites (i.e. found dominant in at most two samples) computed over all other samples.

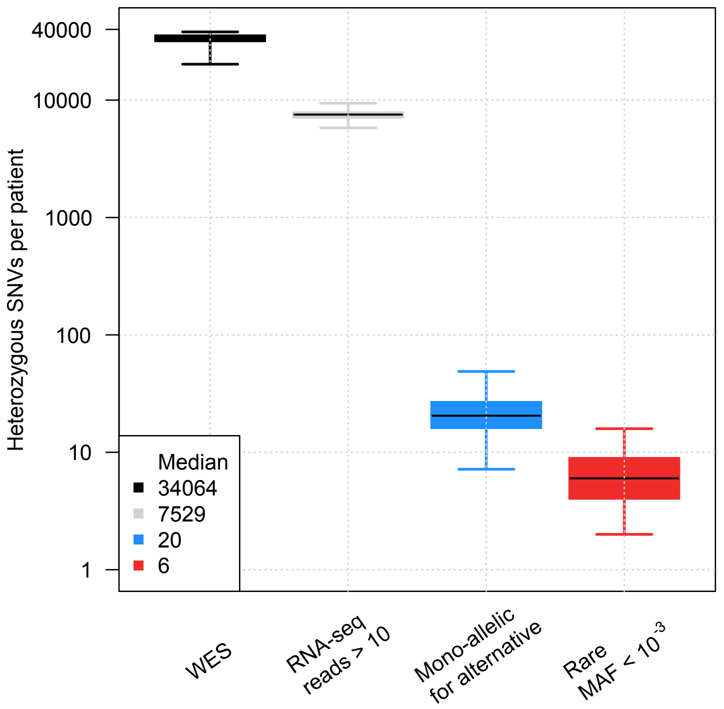


Figure 3.12: **Distribution of heterozygous single nucleotide variants (SNVs) across samples for different consecutive filtering steps.** Heterozygous SNVs detected by exome sequencing (black), SNVs with RNA-seq coverage of at least 10 reads (gray), SNVs where the alternative allele is mono-allelically expressed (alternative allele frequency $\geq$ 0.8 and Benjamini-Hochberg corrected P-value $<$ 0.05, blue), and the rare subset of those (ExAC minor allele frequency $<$ 0.001, red).

tion of proline, validating impaired proline metabolism as the underlying molecular cause (Fig 3.14C). Our experimental evidence strongly suggests that the two observed variants are causal. Moreover, a recent report [155] on ALDH18A1 patients extended the phenotypic spectrum to spastic paraplegia (OMIM #138250), which resembles the symptoms of our patient (App. 6.3.2).

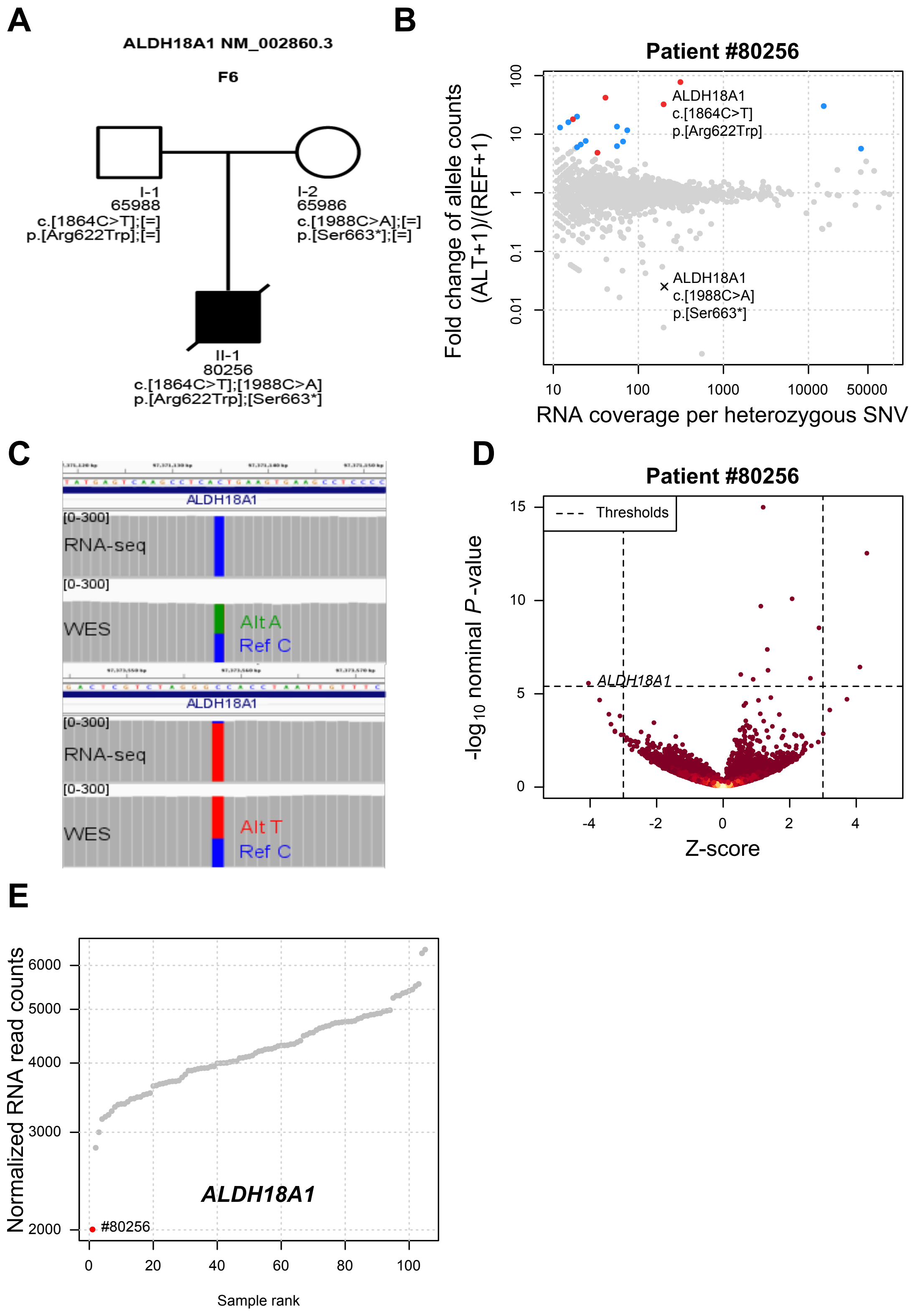

A
ALDH18A1 NM_002860.3
F6
I-1
65988
c.[1864C>T];[=]
p.[Arg622Trp];[=]
I-2
65986
c.[1988C>A];[=]
p.[Ser663*];[=]
II-1
80256
c.[1864C>T];[1988C>A]
p.[Arg622Trp];[Ser663*]
B
Patient #80256
Fold change of allele counts
(ALT+1)/(REF+1)
100
10
1
0.1
0.01
ALDH18A1
c.[1864C>T]
p.[Arg622Trp]
ALDH18A1
c.[1988C>A]
p.[Ser663*]
10
100
1000
10000
50000
RNA coverage per heterozygous SNV
C
ALDH18A1
[0-300]
RNA-seq
[0-300]
WES
Alt A
Ref C
ALDH18A1
[0-300]
RNA-seq
[0-300]
WES
Alt T
Ref C
D
Patient #80256
Thresholds
ALDH18A1
15
10
5
0
-log10 nominal P-value
-4
-2
0
2
4
Z-score
E
Normalized RNA read counts
6000
5000
4000
3000
2000
ALDH18A1
#80256
0
20
40
60
80
100
Sample rank

Figure 3.13: **Mono-allelic expression of ALDH18A1. (A)** Pedigree of the family with mutations in ALDH18A1 showing the mutation status. Mutations are confirmed by Sanger sequencing. **(B)** Fold change between alternative (ALT+1) and reference (REF+1) allele read counts for the patient #80256 compared to total read counts per SNV within the sample. Points are colored according to the groups defined in Fig 3.12. **(C)** Exome and RNA sequencing read coverage tracks (gray) of the two SNVs indicated in (B) for ALDH18A1 (antisense strand). Alternative (Alt) and reference (Ref) nucleotides are indicated by their corresponding color (A green, G brown, T red). **(D)** Gene-wise RNA expression volcano plot of nominal P-value s (-log10 P-value) against Z-scores of the patient #80256 compared against all other fibroblasts. Absolute Z-scores greater than 5 are plotted at ±5, respectively. The y-axis is limited to 15, more extreme values are shown at 15. **(E)** Sample-wise RNA expression is ranked for ALDH18A1. Samples with aberrant expression for the corresponding gene are highlighted in red (#80256).

In another patient (#62346) we found borderline non-significant low expression of MCOLN1 (Fig 3.15A,B) with 10 of 11 reads expressing an intronic VUS (c.681-19A>C, Fig 3.15C). This intronic variant was detected as part of a retained intron, which introduced a nonsense codon (p. Lys227_Leu228ins16*, Fig 3.15D). When looking at the WES data we could additionally identify a heterozygous nonsense variant (c.832C>T, p.Gln278*, Fig 3.15E). The allele with the exonic nonsense mutation was not expressed, most likely due to nonsense-mediated decay. Mutations in MCOLN1 are associated with mucolipidosis (OMIM #605248). The symptoms of the patient were initially suggestive for mucolipidosis, but none of the enzymatic tests available for mucolipidosis type 1, 2, and 3 revealed an enzyme deficiency in blood leukocytes (App. 6.3.2). Moreover, MCOLN1 was missed by our WES variant filter since the intronic variant was not prioritized. Hence, the WES data could not be conclusive about MCOLN1. In contrast, the RNA-seq data demonstrated two loss-of-function alleles in MCOLN1 and therefore established the genetic diagnosis.

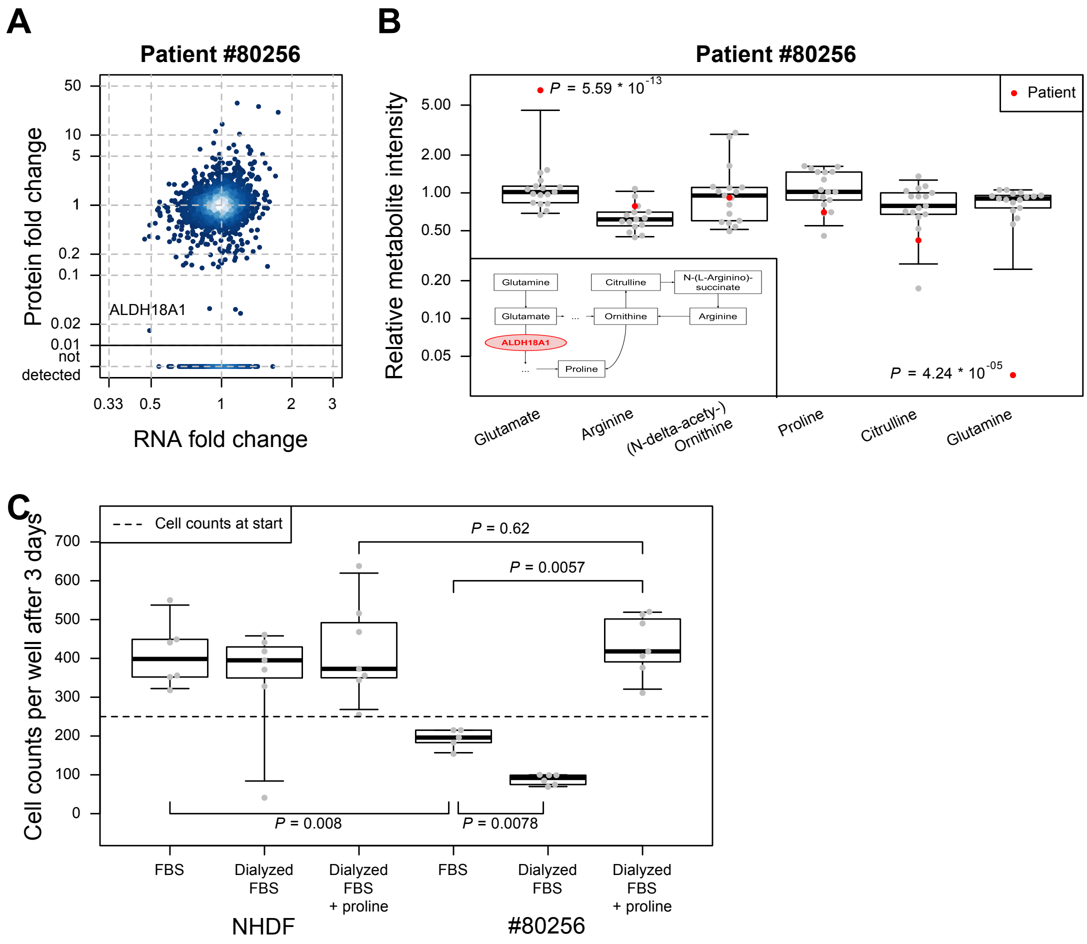


Figure 3.14: **Proline rescue for patient #80256. (A)** Gene-wise comparison of RNA and protein fold changes of the patient #80256 against all other patients' fibroblasts. The position of the gene ALDH18A1 is highlighted. Reliably detected proteins that were not detected in this sample are shown separately with their corresponding RNA fold changes (points below solid horizontal line). **(B)** Relative intensity for metabolites of the proline biosynthesis pathway (inlet) for the patient #80256 and 16 healthy controls of matching age. Equi-tailed 95% interval (whiskers), 25th, 75th percentile (boxes) and median (bold horizontal line) are indicated. Data points belonging to the patient are highlighted (red circles, *P*-value s are computed with Student's t-test). **(C)** Cell counts under different growth conditions for the normal human dermal fibroblast (NHDF) and patient #80256. Both fibroblasts were grown in fetal bovine serum (FBS), dialyzed FBS (without proline) and dialyzed FBS with proline added. Boxplot as in (a). *P*-value s are based on a two-sided Wilcoxon test.

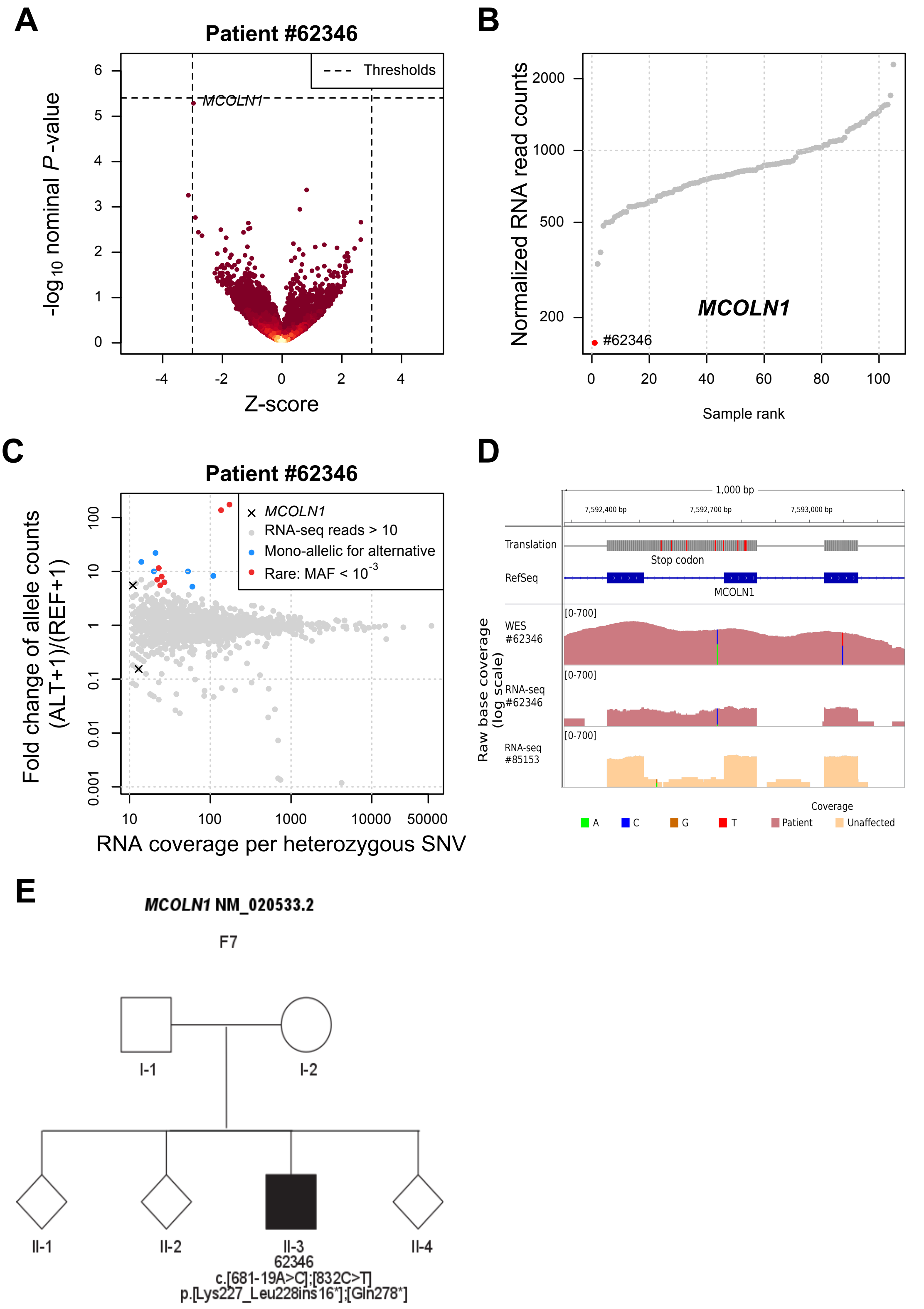

A
Patient #62346
Thresholds
MCOLN1
-log10 nominal P-value
6
5
4
3
2
1
0
-4
-2
0
2
4
Z-score
B
Normalized RNA read counts
2000
1000
500
200
MCOLN1
#62346
0
20
40
60
80
100
Sample rank
C
Patient #62346
× MCOLN1
RNA-seq reads > 10
Mono-allelic for alternative
Rare: MAF < 10^-3
Fold change of allele counts
(ALT+1)/(REF+1)
100
10
1
0.1
0.01
0.001
10
100
1000
10000
50000
RNA coverage per heterozygous SNV
D
1,000 bp
7,592,400 bp
7,592,700 bp
7,593,000 bp
Translation
Stop codon
RefSeq
MCOLN1
Raw base coverage
(log scale)
[0-700]
WES
#62346
[0-700]
RNA-seq
#62346
[0-700]
RNA-seq
#85153
Coverage
A
C
G
T
Patient
Unaffected
E
MCOLN1 NM_020533.2
F7
I-1
I-2
II-1
II-2
II-3
62346
c.[681-19A>C];[832C>T]
p.[Lys227_Leu228ins16*];[Gln278*]
II-4

Figure 3.15: **Mono-allelic expression of MCOLN1.** **(A)** Gene-wise RNA expression volcano plot of nominal *P*-value s (-log10 *P*-value) against Z-scores of the patient #62346 compared against all other fibroblasts. Absolute Z-scores greater than 5 are plotted at ±5, respectively. **(B)** Sample-wise RNA expression is ranked for MCOLN1. Samples with aberrant expression for the corresponding gene are highlighted in red (#62346). **(C)** Fold change between alternative (ALT+1) and reference (REF+1) allele read counts for the patient #62346 compared to total read counts per SNV within the sample. Points are colored according to the steps of the mono-allelic expression variant filter cascade. **(D)** Intron retention for MCOLN1 in patient #62346. Tracks from top to bottom: genomic position on chromosome 19, amino acid translation (red for stop codons), RefSeq gene model, coverage of whole exome sequencing of patient #62346, RNA-seq based coverage for patients #62346 and #85153 (red and orange shading, respectively). SNVs are indicated by non-reference colored bars with respect to the corresponding reference and alternative nucleotide. **(E)** Pedigree of the family with mutations in MCOLN1 showing the mutation status.

Chapter 4

Methods

4.1 Methods for the chapter on negative feedback

This section explains the methods used to generate the results presented in chapter 2. The following methods are additionally described in the paper [1]: Data availability, Yeast strains, DNA sequencing, Transcriptome profiling.

The methods presented in this section are part of the manuscript "Negative feedback buffers effect of regulatory variants" from Bader et al. 2015 [1].

4.1.1 Genotyping and allele frequencies

S96 is isogenic to S288c besides the mating type and therefore we could use the reference genome of the *S. cerevisiae* database [111]. We used the allele frequencies computed earlier by [116]. The coverage of the spore pool B DNA sample was lower than for the other three samples (see DNA sequencing section), hence we have allele frequencies for about 60,000 and 10,000 SNPs, respectively. To adjust the SNP coordinates we lifted them from S288c version R63 to R64. We smoothed the allele frequencies over a window of 28,000bp ($\sim$ 10 Centimorgan) using local binomial likelihood estimation (R CRAN package locfit). We observed a mapping bias towards the S288c genome (median S288c allele frequency 0.52), most likely due to the better annotated reference genome. This artificial bias was used to correct the spore frequency estimations. Those mapping-bias-corrected spore allele frequencies were used to correct the read counts for the statistical model. A similar mapping issue was not observed for the hybrid RNA counts.

4.1.2 Gene annotation

To include also recent non-coding RNAs we used the gene annotation of Xu and colleagues [117] for gene coordinates in the S96 strain (isogenic to S288c). The SK1 gene annotation was generated via bidirectional best hits: Using the coordinates from Xu and colleagues we extracted the S96 gene sequences from the S288c genome version R64 of the Saccharomyces Genome Database [111]. These sequences were searched in the SK1 genome using BLAST [156] with default parameters. The best hit of this first search became query of the second search in the

S96 genome. If this second search resulted in the query of the first, we considered the gene pair as ortholog candidates. Every pair with an alignment identity of more than 95% was considered orthologous. This includes also longer indels and does not restrict our analysis to single nucleotide variants.
Additionally, expression levels for each gene are defined as the average read counts divided by the mean gene length over both strains. These levels were sorted and categorized into three equally sized groups: *Low*, *Middle* and *High* using *cut2* (R package Hmisc). Transcription factor annotation was taken from [157].

4.1.3 Mapping and read counts

RNA-seq reads were mapped to the both genomes of S96 and SK1 jointly. GSNAP [158] was used allowing for four mismatches with novel splice site detection enabled, apart from that we used default parameters. We classified mapped read pairs into three categories: common, only SK1, only S96. Common reads matched equally well to both genomes and therefore are not apt to measure ADE. Only the strain specific and proper-paired alignments can led to ADE and were filtered by their SAM flags (i.e. 83/163 and 99/147) for our statistical model. Additionally, if one read had one proper pair and one mate aligned to the same chromosome on the other allele, it was considered as specific, too. All other reads were discarded together with the common reads.
The filtered alignments were processed with *htseq-count* [159] using *intersection-strict* as overlap mode to generate read counts per gene. *Strict* means that a read or read pair has to align completely inside the annotated gene region to be counted. As gene annotation we used our expressed orthologs with start and end extended by 50bp to increase sensitivity.

4.1.4 Statistical modeling of cis and local trans effects

The raw counts of reads (integer values) per annotated gene are prone to systematic biases that need to be corrected. During the growth of the spores artificial (one mating type) and natural selection takes place [114, 115]. To deal with this bias, we used the genomic allele frequencies of the spores for correction (Supplementary Fig 2.3, see genotyping and allele frequencies section). Additionally, we corrected for length differences between the strains gene-wise as well as the standard sample size factors by DESeq2 [36]. Furthermore, we modelled additional confounding factors for diploid cells, and the biological replicate of each hybrid and spore pool (design matrix, Table 4.1). Hence, allele-specific read counts $K_{i,j}$ were modelled according to the following generalized linear model:

$$K_{i,j} \sim \mathrm{NB}(\mu_{i,j}, \alpha_i) \tag{4.1}$$

$$\mu_{i,j} = s_j \times f_{i,j} \times q_{i,j} \times l_i \tag{4.2}$$

$$\log_2(q_{i,j}) = \boldsymbol{\beta}_i^0 + \boldsymbol{\beta}_i^{cis}\mathbf{x}_{i,j}^{cis} + \boldsymbol{\beta}_i^{localtrans}\mathbf{x}_{i,j}^{localtrans} + \boldsymbol{\beta}_i^{nuis^T}\mathbf{x}_{i,j}^{nuis} \tag{4.3}$$

where NB is the negative binomial distribution, α_i is a gene-specific dispersion parameter; s_j is the size factor of sample j; $f_{i,j}$ is the allele frequency of gene i in sample j; l_i is the length difference for the orthologous gene pair i. This value is

Table 4.1: DESeq design matrix. A cell denotes whether we can observe an effect of the modelled factor (column) in the specified sample (row). Samples split by strain and biological replicate.

SAMPLE \ FACTOR	cis	local trans	diploid	hybrid B	spore B
hybrid A only SK1	1	1	1	0	0
hybrid A only S96	0	1	1	0	0
hybrid B only SK1	1	1	1	1	0
hybrid B only S96	0	1	1	1	0
spore A only SK1	1	1	0	0	0
spore A only S96	0	0	0	0	0
spore B only SK1	1	1	0	0	1
spore B only S96	0	0	0	0	1

0.5 in the hybrid sample and is robustly estimated from genomic DNA sequencing in the pool. $\mathbf{x}_{i,j}^{cis}$ is 1 for allele K and 0 otherwise. $\mathbf{x}_{i,j}^{localtrans}$ is 1 in the pool for allele K and 0 otherwise. $\mathbf{x}_{i,j}^{nuis}$ represents all nuisance parameters to control for: *diploid*, *hybrid B*, *pool B* (Table 4.1). The model was implemented with the R/Bioconductor package DESeq2 [36], which provides robust estimation of the size factors and of the dispersion parameters.

After the correction and fitting process we removed genes from further analysis that had less than ten reads average count over all samples, in order to increase our detection power at the same type I error (Supplementary Fig 2.4 top row, [36, 160]). This minimal expression filtering resulted in 6,691 genes. Accordingly, we corrected the P-values for multiple testing using false discovery rate [161]. Supplementary table 1 provides normalized counts together with fitted coefficients and further gene annotation.

4.1.5 Analysis of ribosome profiling data

We re-analyzed read count data kindly provided by Carlo Artieri and Hunter Fraser (personal communication, supplementary table 2), adopting our model to the hybrid data from RNA-seq and ribosomal profiling. The ribosome-bound fraction was assumed to be the product of the expression level and the binding affinity to RNA, a proxy for translation efficiency [162]. Accordingly, allele specific read counts $K_{i,j}$ were modelled according to the following generalized linear model:

$$K_{i,j} \sim \mathrm{NB}(\mu_{i,j}, \alpha_i) \tag{4.4}$$

$$\mu_{i,j} = s_j \times q_{i,j} \tag{4.5}$$

$$\log_2(q_{i,j}) = \boldsymbol{\beta}_i^0 + \boldsymbol{\beta}_i^{cisRNA}\mathbf{x}_{i,j}^{cisRNA} + \boldsymbol{\beta}_i^{cisTE}\mathbf{x}_{i,j}^{cisTE} + \boldsymbol{\beta}_i^{nuis^T}\mathbf{x}_{i,j}^{nuis} \tag{4.6}$$

where NB is the negative binomial distribution, α_i is a gene-specific dispersion parameter; s_j is the size factor of sample j; $\mathbf{x}_{i,j}^{cisRNA}$ is 1 for the *S. paradoxus* allele and 0 otherwise. $\mathbf{x}_{i,j}^{cisTE}$ is 1 in the ribosome-bound fraction for the *S. paradoxus* allele and 0 otherwise. $\mathbf{x}_{i,j}^{nuis}$ represents nuisance parameters that were controlled for: baseline translation efficiency and overall replicate effect (Table 4.2). The model was implemented with the R/Bioconductor package DESeq2 [36]. Supplementary table 3 provides normalized counts together with fitted coefficients and further gene annotation.

Table 4.2: DESeq design matrix for ribosome profiling data. Value of covariates by sample for the Equation 4.4.

SAMPLE	RNA cis	TE cis	RNA bias	hybrid rep2
hybrid RNA 1 SCER	1	0	1	0
hybrid RNA 2 SCER	1	0	1	1
hybrid RNA 1 SPAR	0	0	1	0
hybrid RNA 2 SPAR	0	0	1	1
hybrid RIBO 1 SCER	1	1	0	0
hybrid RIBO 2 SCER	1	1	0	1
hybrid RIBO 1 SPAR	0	0	0	0
hybrid RIBO 2 SPAR	0	0	0	1

4.1.6 Improvements on the original analyses of ribosomal profiling data

We improved the assessment of translational buffering compared to the original studies [51, 52] in the following three aspects:

1. **Modeling the biological variance.** In the two original studies, tests for allelic differential expression were performed for each biological replicate separately. One of these tests is a binomial test [49] and the other one is a more conservative test controlling for differences in mappability and nucleotide content [118]. To call significant effects over the two biological replicates, the largest P-value of the two samples had to be smaller than a threshold and allelic expression imbalance had to agree in direction. Hence, both of these approaches assess the within-sample significance but do not assess the significance of allelic expression ratios compared to the variability of expression levels between biological replicates. We found that allelic expression ratios for genes called significant according to these procedures often had low fold-changes in comparison to the median biological standard deviation (20.2% less than 1.96 times the median standard deviation at a nominal P-value of 0.05 for [51], Supplementary Fig 6.5), indicating that the extent of many reported effects did not significantly replicate between biological replicates. As comparison, assuming known variance, Gaussian distribution, and same sample size ($n = 2$), a nominal P-value of 0.05 is reached for differences of about 1.96 or more standard deviations. With our test, which models both the so-called shot noise (Poisson noise dominating low counts) and the biological noise (dominating the high counts), only 3.1% of the called genes at a nominal P-value of 0.05 show less than 1.96 times the median standard deviation (Supplementary Fig 6.5). Consequently, P-value s were underestimated with the original statistical tests leading to an abundant fraction of rejected null hypotheses. The same issue affected the significance assessment of translation efficencies.

2. **Independent estimates.** Both studies estimated translational efficiencies as the ratio of RNA levels in the ribosome-bound fraction divided by the RNA expression level. Hence, estimates for translational efficiencies and for expression levels were not independent. Specifically, noise in RNA expres-

sion level measurements induce anticorrelation between translation efficiency estimates and expression level estimates. A scatterplot of allelic log-ratios of translation efficiencies versus allelic log-ratios of RNA levels gave the misleading impression that the two quantities are biologically anticorrelated (Fig 3B in [52] and Fig 2A in [51]). In contrast, scatterplot of the untransformed data does not indicate a trend for translational buffering (mass of the data above diagonal, Fig 2.8B). Because the original statistical tests did not assess the between-replicate variability, most of the effects that were called significant for allelic differences in expression and in translation efficiency were likely due to random variations. Estimated allelic ratio of expression and translation efficiencies of these genes therefore tended to suffer from the anti-correlation and thus to spuriously show opposite effects. Re-analysing the data of [51] with our test and with filtering criteria matching those of the original analysis (FDR=0.05 and no cut-off on fold change), we found much fewer instances (99) significant for both translation efficiency and cis effects. Among these 99 genes, only 55 (56%) show opposing effects which is not statistically significant ($P = 0.31$, two-sided binomial test).

3. **Considering noise in explanatory variable.** In one of the two original studies, genome-wide trend for compensation at the translational level was estimated by a regression of allelic ratios in the ribosome-bound fraction over the allelic expression ratios [51]. An important assumption of linear regression is that there is no noise in the explanatory variable. This was not the case here because the RNA levels are noisy estimates. Linear regression in case of noise in the explanatory variable is known to underestimate the slope (regression to the mean effect), which had led to underestimation of the trend. Compare Fig 2B in (Artieri2014) with supplementary figure 2.8, here we are instead using principal component analysis.

The two latter points were also noticed by [126].

4.1.7 Buffering coefficient

Here we define a measure to quantify the amount of buffering on gene expression. We show that under some assumptions our measure is the same than the compensation metric C of [127].
We write a gene expression level y as :

$$y = \alpha^{1-C}\beta^{C} \tag{4.7}$$

where C is the coefficient of compensation, α is the expression level that the gene would reach in the absence of compensation (i.e. if $C = 0$), β is the expression level that would be reached under full compensation ($C = 1$).

Estimation of C in this study

We assume the unlogged expression level of an allele to be the product of cis and trans effects: $y = \text{cis} \times \text{trans}$. Moreover, we assume the cis effect to be independent of the compensation C. Thus the allele expression ratio in the hybrid is independent of C and is the same as in absence of compensation:

$$\frac{y_B^{\text{HYBRID}}}{y_A^{\text{HYBRID}}} = \frac{\text{cis}_B}{\text{cis}_A} = \frac{\alpha_B}{\alpha_A} \tag{4.8}$$

In the spores carrying allele A or allele B, respectively, the level of expressions are:

$$y_A^{\text{SPORE}} = \alpha_A^{1-C}\beta^C \tag{4.9}$$

$$y_B^{\text{SPORE}} = \alpha_B^{1-C}\beta^C \tag{4.10}$$

Hence the allelic expression ratio in the pool of spores is:

$$\frac{y_B^{\text{SPORE}}}{y_A^{\text{SPORE}}} = \left(\frac{\alpha_B}{\alpha_A}\right)^{1-C} = \left(\frac{y_B^{\text{HYBRID}}}{y_A^{\text{HYBRID}}}\right)^{1-C} \tag{4.11}$$

$$\tag{4.12}$$

We therefore use as working definition of the coefficient of compensation C in this study:

$$C = 1 - \frac{\log_2(y_B^{\text{SPORE}}/y_A^{\text{SPORE}})}{\log_2(y_B^{\text{HYBRID}}/y_A^{\text{HYBRID}})} \tag{4.13}$$

$$\tag{4.14}$$

Definition

To quantitatively estimate how much cis effects are buffered by local trans effects, we defined the buffering coefficient C as:

$$C = 1 - \frac{\log(y_{\text{spore, SK1}}/y_{\text{spore, S96}})}{\log(y_{\text{hybrid, SK1}}/y_{\text{hybrid, S96}})} \tag{4.15}$$

where y denotes the RNA expression level.
In order to estimate buffering at the transcriptional level, we also defined buffering coefficient when comparing ribosome profiling data (RP) and RNA levels in the *S. par.* x *S. cer.* cross.

$$C_{translation} = 1 - \frac{\log(y_{\text{RP, }\textit{S. par.}}/y_{\text{RP, }\textit{S. cer.}})}{\log(y_{\text{RNA, }\textit{S. par.}}/y_{\text{RNA, }\textit{S. cer.}})} \tag{4.16}$$

where y_{RNA} denotes the RNA expression level, and y_{RP} the ribosome occupancy.
Note that both for the local trans regulation case and for the translation efficiency case, C is ill-defined for hybrid RNA ratios close to zero. This is equivalent to say that buffering can only be assessed if there is a cis effect in the first place. We therefore restricted the analysis of buffering for genes with significant and sufficiently large cis effects.

Calibration

We defined as raw buffering coefficient the quantity:

$$C_{raw} = 1 - \frac{\log(\#\text{reads}_{\text{spore, SK1}}/\#\text{reads}_{\text{spore, S96}})}{\log(\#\text{reads}_{\text{hybrid, SK1}}/\#\text{reads}_{\text{hybrid, S96}})} \tag{4.17}$$

C_{raw} is a biased estimator of the buffering coefficient C defined by Equation 4.15. We empirically derived an unbiased estimator of C by inferring the relationship between C_{raw} and C from simulations for all values of C in $[0, 0.5]$ with a 0.005 spacing. For each simulated value of C, read counts for every gene i were simulated by random draws according to Equations (1-3), keeping all the parameters fixed to their estimated values on the primary dataset, except for substituting $\beta_i^{localtrans}$ with $-C\beta_i^{cis}$. On these simulated genome-wide read counts, the exact same analysis as for the primary dataset was performed (i.e. including filter for minium read counts, DESeq2 normalization and fits, and filter for large and significant cis effects) and the median C_{raw} across cis genes was computed. To obtain an unbiased estimator of translational buffering for the ribosome dataset, the same procedure was applied substituting β^{cisTE} with $-C\beta^{cisRNA}$. For both datasets, we observed a linear relationship between the simulated true C and the median C_{raw} (Supplementary Fig 6.6A-B, Pearson correlation ¿0.99) and used the linear regression fit as calibration function. This linear transformation of C_{raw} was then used for all further analysis as buffering coefficient C.

Significance

To assess the significance of the median buffering coefficient, data were simulated under the null hypothesis of independence between cis effects and local trans effects in a semi-parametric fashion. A total of $B = 1000$ bootstrap genome-wide datasets were generated by permuting the estimated local trans effects $\beta_i^{localtrans}$ between genes while keeping all remaining parameters fixed to their estimated values on the primary dataset and drawing counts according to Equations (1-3). On these simulated genome-wide read counts, the exact same analysis as for the primary dataset was performed (i.e. including filter for minium read counts, DESeq2 normalization and fits, and filter for large and significant cis effects) and the median buffering coefficient across cis genes was computed.
One-sided P-value was then estimated by [163]:

$$P = \frac{1 + \#\{\bar{C}_i^* \geq \bar{C}\}}{B + 1} \tag{4.18}$$

where $\bar{C}$ is the median buffering coefficient in the observed dataset and $\bar{C}_i^*, i = 1 \dots B$ are the bootstrap values of the median buffering coefficient (Supplementary Fig 6.6C). The same procedure was applied to the ribosome dataset whereby the estimated translation efficiency estimates β_i^{cisTE} were permuted across genes (Supplementary Fig 6.6D).

4.1.8 Comparison with Springer's C

Comparison with Springer et al. [127] data was done for the same growth medium as the one used in this study (rich growth medium YPD). Distribution of our buffering coefficient under matching distribution of gene category and expression levels (Fig 2.10A, central box) was obtained by (i) restricting to non-essential genes and (ii) randomly sampling 1,000 times with replacement the same number of genes in each tercile of expression as in Springer and colleagues' dataset.

4.1.9 Equivalence with Springer's C

Springer and colleagues [127] assess buffering of a protein's expression using diploid strains in which one of the two alleles is marked by GFP. In the so-called wild type strain, the unmarked allele is kept intact whereas in the heterozygote strain the unmarked allele is deleted. Springer's compensation metric is defined as the $\log_2$-ratio of the GFP expression in the heterozygote over the wild type strain.
We index with A the unmarked allele's constants and with B the GFP-tagged allele's constants. In this experiment, the expression levels of the two alleles are assumed to be the same in absence of compensation (i.e. $\alpha_A = \alpha_B := \alpha/2$) and in presence of compensation ($y_A = y_B$). The feedback acts on the overall expression level $y := y_A + y_B = 2y_B$.
According to Equation 4.7, we expect the GFP expression level y_B^{WT} to be:

$$y_B^{\mathrm{WT}} = \frac{1}{2}y = \frac{1}{2}\alpha^{1-C}\beta^C \tag{4.19}$$

The GFP expression level in the heterozygote strain, for which $y_A = 0$, is:

$$y_B^{\mathrm{HET}} = y = (\alpha/2)^{1-C}\beta^C \tag{4.20}$$

Hence we get:

$$\log_2\left(\frac{y_B^{\mathrm{HET}}}{y_B^{\mathrm{WT}}}\right) = \log_2\left(\frac{(1/2)^{1-C}}{1/2}\right) = C \tag{4.21}$$

4.2 Methods for the chapter on genetic diagnosis

This section explains the methods used to generate the results presented in chapter 3. Where the Author contributions information of the paper [3] is not detailed enough, I gave further explanations at the beginning of the corresponding method's section. The following methods are additionally described in the paper [3]: Sanger sequencing, exome sequencing, whole genome sequencing, cell culture, mass spectrometric sample preparation, mass spectrometric data acquisition, Immunoblotting, Blue native PAGE, metabolomics, Proline supplementation growth assay, Cellular reactive oxygen species production.

The methods presented in this section are part of the manuscript "Genetic diagnosis of Mendelian disorders via RNA sequencing" from Kremer, Bader et al. 2016 [3].

4.2.1 Variant of unknown significance

"A variation in a genetic sequence whose association with disease risk is unknown. Also called unclassified variant, variant of uncertain significance, and VUS." (see `https://www.cancer.gov/publications/dictionaries/genetics-dictionary?cdrid=556493`)

4.2.2 Exome alignment and variant prioritization

Read alignment to the human reference genome (UCSC Genome Browser build hg19) was done using Burrows-Wheeler Aligner (v.0.7.5a)[164]. Single-nucleotide variants and small insertions and deletions (indels) were detected with SAMtools (version 0.1.19)[165, 166]. Variants with a quality score below 90, a genotype quality below 90, a mapping quality below 30, and a read coverage below 10 were discarded. The reported variants and small indels were annotated with the most severe entry by the Variant Effector Predictor [167] based on The Sequence Ontology term ranking [168]. The candidate variants for one patient are filtered to be rare, affect the protein sequence and potentially both alleles. Variants are rare with a minor allele frequency $<$ 0.001 within the ExAC database [74] and a frequency $<$ 0.05 among our samples. Variants affect the protein, if they are a coding structural variant or their mutation type is one of ablation, deletion, frameshift, incomplete, start lost, insertion, missense, splice, stop gain, stop retain, unstart, unstop. A potential biallelic effect can be caused by either a homozygous or at least two heterozygous variants in the same gene, whereas in latter case we assume that the heterozygous variants are on different alleles (Appendix Fig 6.9). This filter is designed for a recessive type disease model and does not account for a single heterozygous variant that could be disease-causing in a dominant way.

4.2.3 RNA sequencing

Non-strand specific, polyA-enriched RNA sequencing was performed as described earlier [130]. Briefly, RNA was isolated from whole-cell lysates using the AllPrep RNA Kit (Qiagen) and RNA integrity number (RIN) was determined with the Agilent 2100 BioAnalyzer (RNA 6000 Nano Kit, Agilent). For library preparation, 1 µg of RNA was poly(A) selected, fragmented, and reverse transcribed with the Elute, Prime, Fragment Mix (Illumina). End repair, A-tailing, adaptor ligation, and library enrichment were performed as described in the Low Throughput protocol of the TruSeq RNA Sample Prep Guide (Illumina). RNA libraries were assessed for quality and quantity with the Agilent 2100 BioAnalyzer and the Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies). RNA libraries were sequenced as 100 bp paired-end runs on an Illumina HiSeq2500 platform.

4.2.4 Processing of RNA sequencing files

RNA-seq reads were demultiplexed and mapped with STAR (version 2.4.2a)[169] to the hg19 genome assembly (UCSC Genome Browser build). In addition to the default parameters we detected gene fusions and increased sensitivity for novel splice junctions (chimSegmentMin=20, twopassMode="Basic"). Analysis was restricted to the 27,682 UCSC Known Genes (genome annotation version hg19)[170] of chromosomes 1 to 22, M, X, or Y. Per gene, reads that are paired with mates from opposite strands and that overlapped completely within the gene on either strand orientation were counted using the summarizeOverlaps function of the R/Bioconductor GenomicAlignments [171] package (parameters: mode=intersectionStrict, singleEnd=FALSE, ignore.strand=TRUE, fragments=FALSE). If the 95th percentile of the coverage across all samples was below 10 reads the gene was considered “not expressed” and discarded from later analysis.

4.2.5 Computing RNA fold changes and differential expression

Before testing for differential expression between one patient of interest and all others, we controlled for technical batch effect, sex, and biopsy site as inferred from the expression of hox genes (Supplementary Data 4). We modeled the RNA-seq read counts K$_{i,j}$ of gene i in sample j with a generalized linear model using the R/Bioconductor DESeq2 package [36, 37]:

$$K_{i,j} \sim NB(s_j \times q_{i,j}, \alpha_i) \tag{4.22}$$

$$\log_2(q_{i,j}) = \beta_i^0 + \beta_i^{condition}\mathbf{x}_{i,j}^{condition} + \beta_i^{batch}\mathbf{x}_{i,j}^{batch} + \beta_i^{sex}\mathbf{x}_{i,j}^{sex} + \beta_i^{hox}\mathbf{x}_{i,j}^{hox} \tag{4.23}$$

Where NB is the negative binomial distribution; α_i is a gene specific dispersion parameter; s_j is the size factor of sample j; β_i^0 is the intercept parameter for gene i. The value of $\mathbf{x}_{i,j}^{condition}$ is 1 for all RNA samples j of the patient of interest, thereby allowing for biological replicates, and 0 otherwise. The resulting vector $\beta_i^{condition}$ represents the log$_2$-fold changes for one patient against all others. Z-scores were computed by dividing the fold changes by the standard deviation of the normalized expression level of the respective gene. The P-value s corresponding to the $\beta_i^{condition}$ were corrected for multiple testing using the Hochberg family-wise error rate method [172].

4.2.6 Detection of aberrant splicing

The methods presented in this section were developed by Christian Mertes, co-author of [3].

The LeafCutter [173] software was utilized to detect aberrant splicing. Each patient was tested against all others. To adjust LeafCutter to the rare disease setting, we modified the parameters to detect rare clusters, capture local gene fusion events and to detect junctions unique to a patient (minclureads=30; maxintronlen=500,000; mincluratio=1e-5, Supplementary Data 4). Furthermore, one sample was tested against all other samples (min_samples_per_group=1; min_samples_per_intron=1). The resulting P-value s were corrected for multiple testing using a family-wise error rate approach [172].
The significant splice events (Hochberg adjusted P-value < 0.05) detected in the undiagnosed patients were visually classified as exon skipping, exon truncation, exon elongation, new exon, complex splicing (any other splicing event or a combination of the aforementioned ones) and false positives (n=73, Fig 3.8). However, due to LeafCutter's restriction to split reads it is difficult to detect intron retention events, since in a perfect intron retention scenario no split-reads are present.
For further analysis, only reads spanning a splice junction, so called split reads, were extracted with a mapping quality of greater than 10 to reduce the false positive rate due to mapping issues. Each splice site was annotated as belonging to the start or end of a known exon or to be entirely new. For the reference exon annotation the GENCODE release 24 based on GRCh37 was used [174]. The percent spliced in (Ψ) values for the 3' and 5' sites were calculated as described earlier [138]:

$$\Psi_5(D,A) = \frac{n(D,A)}{\sum_{A'} n(D,A')} \quad \textit{and} \quad \Psi_3(D,A) = \frac{n(D',A)}{\sum_{D'} n(D',A)} \tag{4.24}$$

Where D is a donor site and A is an acceptor site. $n(D,A)$ denotes the number of reads spanning the given junction. D' and A' represent all possible donor and acceptor sites, respectively.
Classification of splice sites into background, weak and strong was done by modeling the distribution of the Ψ_5 and Ψ_3-values with three components. Identifiability of the three components was facilitated by considering three groups of junctions depending on previous annotation of splice sites: 'no side is annotated', 'one side is annotated' and 'both sides are annotated'. Specifically, the number of split reads $n(D,A)$ of a junction conditioned on the total number of reads $N(D,A) = \sum_{A'} n(D,A')$, for Ψ_5, and $N(D,A) = \sum_{D'} n(D',A)$, for Ψ_3, was modeled as:

$$P\left(n(D,A)|N(D,A)\right) = \sum_{c\in\{bg,wk,st\}} \sum_{s\in\{0,1,2\}} \pi_{s,c} BetaBin(n(D,A)|N(D,A), \alpha_c, \beta_c) \tag{4.25}$$

where c is the component index, s the number of annotated sites and BetaBin the beta-binomial distribution. Hence, the components were modeled to have the same parameters α_c, β_c in all three groups but their mixing proportions $\pi_{s,c}$ to be group-specific. Fitting was performed using the expectation-maximization algorithm. For the initial step, the data points were classified as background ($\Psi < 0.001$), weak spliced ($\Psi < 0.1$) and canonical ($\Psi \geq 0.1$). After convergence of the clustering the obtained parameters were used to estimate the probability for each junction side to belong to a given class.

4.2.7 Detection of mono-allelic expression

The methods presented in this section were developed jointly by Christian Mertes, co-author of [3] and Daniel M Bader.

For mono-allelic expression analysis only heterozygous single nucleotide variants with only one alternative allele detected from exome sequencing data were used. The same quality filters were used as mentioned in the exome sequencing part, but no frequency filter was applied. To get allele counts from RNA sequencing for the remaining variants the function *pileLettersAt* from the R/Bioconductor package *GenomicAlignments* [171] was used. The data was further filtered for variants with coverage of at least 10 reads on the transcriptome.

The DESeq2 package [36, 37] was applied on the final variant set to estimate the significance of the allele-specific expression. Allele-specific expression was estimated on each heterozygous variant independently of others (i.e. without phasing the variants). For each sample, a generalized linear model was fitted with the contrast of the coverage of one allele against the coverage of the other alleles (*condition*). Specifically, we modeled $K_{i,j}$ the number of reads of variant i in sample j as:

$$K_{i,j} \sim NB(s_j \times q_{i,j}, \alpha) \tag{4.26}$$

$$\log_2(q_{i,j}) = \beta_i^0 + \beta_i^{condition} \mathbf{x}_{i,j}^{condition} \tag{4.27}$$

Where NB is the negative binomial distribution; the dispersion parameter α was fixed for all variants to $\alpha = 0.05$, which is approximately the average dispersion over all samples based on the gene-wise analysis; s_j is the size factor of each condition; β_i^0 is the intercept parameter for variant i. The value of $\mathbf{x}_{i,j}^c ondition$ is 1 for the alternative alleles and 0 otherwise. The resulting $\beta_i^{condition}$ represents the $\log_2$-fold changes for the alternative allele against the reference allele. The independent filtering by DESeq2 was disabled ($independentFiltering = FALSE$) to keep the coverage outliers among the results. To classify a variant as mono-allelically expressed a cutoff of $|\beta_i^{condition}| \geq 2$ was used, which corresponds to an allele frequency ≥ 0.8, and we filtered Hochberg adjusted P-value s to be smaller than 0.05.

4.2.8 Processing of proteome intensities

Label-free protein quantification was done using the MaxLFQ algorithm [175] integrated into MaxQuant (for detailed parameters see [3]).

The LFQ intensities and gene names were extracted for 6,566 protein groups from the MaxQuant output file *proteinGroups.txt*. For protein groups with more than one member, the first member was chosen to represent the group as single protein with a distinct gene name (similar to earlier studies [176]). MaxLFQ intensities of

0 actually represent non-quantified peaks and were therefore replaced with missing values (NA). The 10 samples that had a frequency of missing values higher than 50% were considered bad quality and were discarded. Furthermore, proteins were discarded because they had no gene name assigned (n=198), were not the most abundant among their duplicates (n=295), were not expressed in any sample (n=93), because their 95th percentile was not detected (n=549), which was also considered as not expressed, analogously to RNA filtering. Finally, 5,431 proteins and 31 samples were considered for further analysis (Supplementary Data 3).

4.2.9 Computing protein fold changes and differential expression

Since the mass spectrometric measurements of all samples were done in a single run, no technical artifacts could be found with a hierarchical clustering. Protein differential expression for each patient compared to the others was tested using moderated T-test approach as implemented in the R/Bioconductor limma package [177]. The transcriptome covariates for sex and HOX effects were used in the linear model for normalization.

4.2.10 Transduction and Transfection

Overexpression of TIMMDC1 in fibroblast cell lines was performed by lentivirus-mediated expression of the full-length TIMMDC1 cDNA (DNASU Plasmid Repository) using the ViraPower HiPerform Lentiviral TOPO Expression Kit (Thermo Fisher Scientific)[178]. TIMMDC1 cDNA was cloned into the pLenti6.3/V5-TOPO expression vector and cotransfected into 293FT cells with the packaging plasmid mix using Lipofectamine 2000. After 24 h, the transfection mix was replaced with high glucose DMEM supplemented with 10% FBS. After further 72 h, the viral particle containing supernatant was collected and used for transduction of the fibroblast cell lines. Selection of stably expressing cells was performed using 5 μg/mL Blasticidin (Thermo Fisher Scientific) for 2 weeks.

Chapter 5

Discussion

The past decade marks the rise of DNA and RNA sequencing at ever decreasing costs. Consequently, it has become feasible to investigate multiple genome-wide layers of information in one project organized by one lab. These advances allowed me to investigate the consequences of DNA variation on RNA expression using sequencing data in two projects: i) identify and quantify regulatory mechanisms that buffer the effect of DNA variation and ii) use RNA-seq information to improve genetic diagnosis of Mendelian disorders beyond genome-information-based approaches.

5.1 Discussion of negative feedback results

The discussion presented in this section is part of the manuscript "Negative feedback buffers effect of regulatory variants" from Bader et al. 2015 [1].

We found that compensatory local trans-regulatory mechanisms buffer typically 15% of RNA level log-ratios caused by naturally occurring cis-regulatory variants in *S. cerevisiae*. Local trans mechanisms involve the gene itself (feedback) or trans-acting variants in its genetic vicinity. Analysis of expression data of heterozygous deletions indicates that this buffering is primarily due to negative feedback regulation and not due to compensatory mutations. In addition, we did not find evidence for translational buffering to be common when reanalyzing ribosome profiling data of a cross between two yeast species, even though translational buffering occurs for specific instances. The intensity of buffering through local trans regulation was lower for highly expressed genes, suggesting that the sheer amount of transcripts available for these genes confer robustness against cis-regulatory variants. In low to middle range of expression, buffering was increasing across the three categories, non-coding, non-essential coding, and essential coding genes, correlating with presumed functional importance.

We dissected local regulation into its cis and trans components using a novel experimental design in which ADE in a yeast hybrid strain was compared against ADE in a pool of its spores. In contrast, former dissection of local regulation was performed in two steps [41]. First, polymorphisms in the vicinity of genes that significantly associated with their expression across a population of spores were identified (eQTL mapping). Second, the estimated effect of these local eQTLs was compared to allelic differential expression in a hybrid strain. The advantage of our experimental design is first economic, because the spores are pooled whereas eQTL mapping requires typically dozens of individual spores to be transcription profiled. Second, our design suffers less from confounders such as batch effects that can give false associations in eQTL mapping. Third, ADE in the hybrid is more comparable to ADE in the pool of spores than to eQTL effects because in the former case the same experimental protocol and the same analysis are applied. One should note that amplification and sequencing biases could favor one allele thereby leading to overestimated ADE. However, the same bias applies similarly to the pool and to the hybrid and thus does not affect our observation that ADE is lower in the pool than in the hybrid. Our experimental design could be applied to study other levels of gene regulation where local trans mechanisms, and in particular regulatory feedback, could play a significant role, including synthesis and decay of RNA, translation, and protein levels [179].

Our findings have implications for the understanding of dosage compensation, i.e. the buffering of expression level in case of gene copy number variation. Unlike for sex chromosomes, the prevalence and the mechanisms for dosage compensation on autosomes are poorly understood. Buffering in the 10% to 20% range was reported for a set of seven autosomal single copy deletions in fruit fly [125]. In contrast, Springer et al. [127] reported a general lack of dosage compensation in yeast. Our study shows that these observations are more in agreement to each other than they seem to be. We found that buffering against cis-acting regulatory variants in yeast is typically of 15% genome-wide, and that Springer and colleagues' heterozygous deletion screen was biased for genes with little buffering (about 5%). Hence, the extent of buffering appears to be conserved from yeast to fly. Moreover, we found that buffering is primarily due to negative feedback which confers robustness against single nucleotide polymorphisms and short indels as well, as supported by the fact that we assessed genes with more than 95% identity between the two parental strains. Together, these results suggest that dosage compensation of autosomal genes in higher eukaryotes might be explained to a large extent by negative feedback, i.e. by a mechanism that generally buffers regulatory variants rather than by a copy number surveillance pathway.

In 1942, Waddington hypothesized the existence of buffering mechanisms against genetic variants that would explain the remarkable stability of developmental processes among individuals [180, 181]. It is still unclear to date, which buffering mechanisms act across the stages of phenotypic expression, from DNA to RNA, protein and cellular phenotypes, and what their respective contribution is. Robustness against coding variations can be explained by redundancy, such as diploidy, copy number variation, and functional duplication [182, 183]. Our data show that already at the level of RNA expression, buffering is widespread. We could estimate its effect and identified negative feedback as the predominant mechanism. Protein abundance of orthologous genes has been shown to be more conserved than mRNA abundance across all domains of life ranging from bacteria to fungi and primates [184–186]. Thus, further mechanisms buffering regulatory variants downstream of RNA expression remain to be identified [187, 188]. One possibility is that negative feedback is also common for controlling protein levels.

Buffering plays an important role in evolution because it confers robustness to mutations on the one hand and allows the accumulation of cryptic genetic variants in the population that might give selective advantage under new environmental conditions on the other hand. In this context, a capacitor is a switch capable of releasing previously cryptic heritable variation [189]. Since feedback loops themselves can be impaired, through mutations as in the case of *ROX1* or environmental changes, we suggest that negative feedback loops could function as capacitors.

5.1.1 Conclusion

In the negative feedback project discussed above, gene expression from a hybrid of two yeast strains and its spores was measured. We developed a new statistical model that allowed us to quantify local trans regulatory effects with our new experimental design using an established statistical software package. Local trans regulation was found to buffer the effect of cis-regulatory elements, in contrast to translational efficiency. Furthermore, this buffering is stronger for lower expressed genes and genes that are essential. We could also clarify that local trans buffering acts primarily through negative feedback loops. In general, negative feedback loops are likely to confer robustness at multiple stages of gene regulation not only in yeast, but also in higher organisms.

5.1.2 Outlook

Our advanced statistical model can be applied to all existing hybrid gene expression and ribosome profiling studies (section 1.2.2). With such a reanalysis cis and trans regulatory mechanisms can be finally compared across multiple species in terms of affected genes, direction of effect, and effect size. Two recent studies of RNA-seq in intraspecies hybrids of higher organisms do neither resolve existing conflicts (section 1.2.3) nor establish analysis standards: A study in drosophila hybrids reported cis-trans compensation using RNA-seq data [190]; In the other study RNA-seq and ribosome profiling data were used to show translational buffering in mice hybrids [191]. Further mechanistic insights can also be expected from allele-specific protein expression data from hybrid organisms [179]. To combine all these ideas, RNA-seq, ribosome profiling, and allele-specific protein expression could be investigated in mouse hybrids viable with one allele of one chromosome deleted. Depending on the haploid region's size in the mice hybrids, local trans effects can be measured genome-wide in mammals for RNA, translation, and protein regulatory mechanisms.

5.2 Discussion of genetic diagnosis results

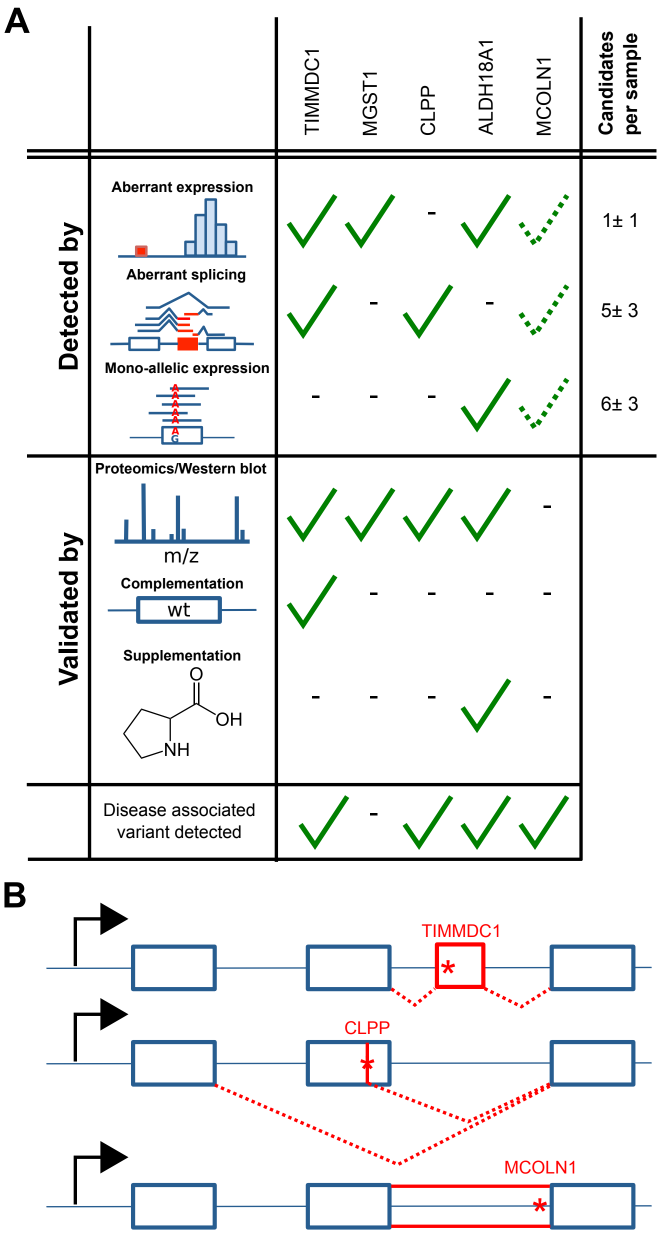
The discussion presented in this section is part of or adapted from the manuscript "Genetic diagnosis of Mendelian disorders via RNA sequencing" from Kremer, Bader et al. 2016 [3].

Altogether, our study demonstrates the utility of RNA sequencing in combination with bioinformatics filtering criteria for genetic diagnosis by i) discovering a new disease-associated gene, ii) providing a diagnosis for 10% (5 of 48) of undiagnosed cases, and iii) identifying a limited number of strong candidates. We established a pipeline for the detection of aberrant expression, aberrant splicing and mono-allelic expression of rare variants, that is able to detect significant outliers, i.e. a median of 1, 5, and 6, respectively. Overall, for 36 patients our pipeline provides a strong candidate gene, i.e. a known disease-causing or mitochondrial protein-encoding gene, like MGST1 (Fig 5.1A, Supplementary Table 1). This manageable amount, similar to the median number of 16 genes with rare potentially bi-allelic variants detected by WES, allows manual inspection and validation by disease experts. While filtering by frequency is highly efficient when focusing on the coding region, frequency filtering is not as effective for intronic or intergenic variants identified by whole genome sequencing. The loss-of-function character observed on RNA level thus improved interpretation of VUS identified by genotyping.

We focused our analysis on one sample preparation pipeline, which has several advantages. Based on our experience, expression outliers can only reliably be detected after extensive normalization process. This needs information about all technical details starting from the biopsy, growth of the cells, to the RNA extraction and library preparation. Usually not all this information is available in published data sets. For detecting aberrant splicing such as new exons, we would recommend not to mix different tissues because splicing can be tissue-specific. Mono-allelic expression is the most robust of all criteria in this respect because it only relies on read counts within a sample. Overall, we recommend not relying on a single sample being compared to public RNA-seq datasets. Instead, RNA-seq should be included in the pipeline of diagnostic centers in order to generate matching controls over time. The situation is similar for whole exome and whole genome sequencing, where the control for platform-specific biases is important.

Here, we included genetically diagnosed patients in our RNA-seq analysis pipeline to increase the power for the detection of aberrant expression and aberrant splicing in fibroblast cell lines. However, when applied to the 40 diagnosed cases with WES and RNA-seq available, aberrant splicing detected 6 out of 8 cases with a causal splicing variant, mono-allelic expression recovered 3 out of 6 patients with heterozygous missense variants compound with a stop or frameshift variant, and aberrant expression recovered 3 out of 9 stop variants. Counterintuitively, only one of the 9 frame-shift variants did lead to a detectable RNA defect, i.e. mono-allelic expression of a near splice site intronic variant within a retained intron. The partial recovery of stop and frameshift variants may reflect incomplete non-sense mediated decay. For none of the 14 genes where missense variants were disease causing, a RNA defect could be detected with our pipeline. This is expected, since missense variants more likely affect protein function rather than RNA expression (Supplementary Table 4).

To our surprise, many newly diagnosed cases were caused by a defective splicing event, which caused loss of function (Fig 5.1B), confirming the increasing role of splicing defects in both Mendelian [192, 193] and common disorders [131]. In the case of TIMMDC1, the causal variant was intronic, and thus not covered by WES. Even when detected by WGS, such deep intronic variants are difficult to prioritize from the sequence information alone. Here, we showed that RNA-seq of large cohorts can provide important information about intronic positions that are particularly susceptible to affect splicing when mutated. We showed that private exons often arise from loci with weak splicing of about 1%. This suggests that rare variants affecting such cryptic splice sites are more likely to affect splicing and that these can be detected as positions with low yet consistent splicing. We reason that analysis of a RNA-seq compendium of healthy donors across multiple



		TIMMDC1	MGST1	CLPP	ALDH18A1	MCOLN1	Candidates per sample
Detected by	Aberrant expression	✓	✓	-	✓	✓	1± 1
	Aberrant splicing	✓	-	✓	-	✓	5± 3
	Mono-allelic expression	-	-	-	✓	✓	6± 3
Validated by	Proteomics/Western blot	✓	✓	✓	✓	-	
	Complementation	✓	-	-	-	-	
	Supplementation	-	-	-	✓	-	
	Disease associated variant detected	✓	-	✓	✓	✓	

Figure 5.1: **Validation summary. (A)** Discovery and validation of genes with RNA defects in newly diagnosed patients, i.e. TIMMDC1 (n=2 patients), CLPP, ALDH18A1, and MCOLN1, and patients with strong candidates, i.e. MGST1. The median number (± median absolute deviation) of candidate genes is given per detection strategies. Dotted check, manual inspection not statistically significant. **(B)** Schematic representation of variant causing splicing defects for TIMMDC1 (top, new exon red box), CLPP (middle, exon skipping and truncation), and MCOLN1 (bottom, intron retention). Variants are depicted by a red star.

tissues such as GTEx [194] could provide tissue-specific maps of cryptic splice sites useful for prioritizing intronic variants.

Genetic disorders typically show specificity to some tissues, some of which might not be easily accessible for RNA-sequencing. It is therefore natural to question whether transcriptome sequencing of an unaffected tissue can help diagnosis. Here, we performed RNA-seq on patient derived dermal fibroblast cell lines. The fibroblast cell lines are the byproducts of muscle biopsies routinely undertaken in the clinic to biochemically diagnose mitochondrial disorders with enzymatic assays. By using fibroblast cell lines we overcome the limited accessibility of the affected tissues, which in the case of mitochondrial disorders are often high energy demanding tissues like brain, heart, skeletal muscle or liver. It turns out that many genes with a mitochondrial function are expressed in most tissues [195], including fibroblasts. Hence, extreme regulatory defects such as loss of expression or aberrant splicing of genes encoding mitochondrial proteins can be detected in fibroblasts, even though the physiological consequence on fibroblasts might be negligible. This property might be true for other diseases: the tissue-specific physiological consequence of a variant does not necessarily stem from tissue-specific expression of the gene harboring the variant. In many cases, tissue-specificity might be due to environmental or cellular context, or from tissue-specific expression of further genes. Hence, tissue-specificity does not preclude RNA-seq of unaffected tissues from revealing the causative defect for a large number of patients. Moreover, non-affected tissues have the advantage that the regulatory consequences on other genes are limited and therefore the causal defects are more likely to stand out as outliers [196].

Parallel to our effort, another study systematically investigated the usage of RNA-seq for molecular diagnosis with a similar sample size, using muscle biopsies from rare neuromuscular disease patients [193]. Analogously to our approach, not only exome sequencing-based VUS candidates were validated, but also new disease-causing mechanisms identified using RNA-seq data. Despite a few differences in the approach (expression outliers were not looked for, only samples of the affected tissues were considered and using samples of healthy donors as controls), the results are in line with ours whereby aberrant splicing also turns out to be a frequent disease-causing event. Moreover, the success rate was even higher (35%) confirming the relevance of using RNA-seq for diagnosis of Mendelian disorders.

In conclusion, we predict that RNA sequencing will become an essential companion of genome sequencing to address undiagnosed cases of genetic disease.

5.2.1 Conclusion

In the genetic diagnosis project discussed above, RNA-seq was performed on cell lines from patients that were likely to have mitochondrial disorder. We implemented three strategies to systematically prioritize genes for RNA defects: i) aberrant expression, ii) aberrant splicing, and iii) mono-allelic expression. Our approach complements the common diagnosis pipeline based on DNA variants and therefore helped to diagnose the disease-causing gene in 7 unrelated families. The diagnosis pipeline presented here is generally applicable to other Mendelian diseases and can become routine in addition to WES or whole genome sequencing.

5.2.2 Outlook

A systematic large-scale application of proteomics for diagnosis of Mendelian disorders has not been conducted, yet. Thus, including proteomics into the diagnosis process would complete the view on the central dogma's products. With proteomics not only aberrant protein expression can be detected, but also variants can be prioritized that have a direct effect on proteins. Yet, protein information needs to be interpreted with great care, since observed protein defects can indicate both cause and consequences of a disease. It is necessary to gather more experience on diagnosis standards with transcriptomics and proteomics analysis for further diseases. As a vision for future disease diagnosis, clinicians ask for and interpret a patient's omics-profile the same way they use blood panels today.

Chapter 6

Appendix

6.1 Acronyms

ADE allelic differential expression

eQTL expression Quantitative Trait Loci

GWAS genome-wide association study

MAE mono-allelic expression

NHDF normal human dermal fibroblasts

RNA-seq RNA sequencing

ROS reactive oxygen species

SNV single nucleotide variant

VUS variant of unknown significance

WES whole exome sequencing

6.2 Appendix for the chapter on negative feedback

This section corresponds to the results presented in chapter 2
The supplementary information presented in this section is part of the manuscript "Negative feedback buffers effect of regulatory variants" from Bader et al. 2015 [1].

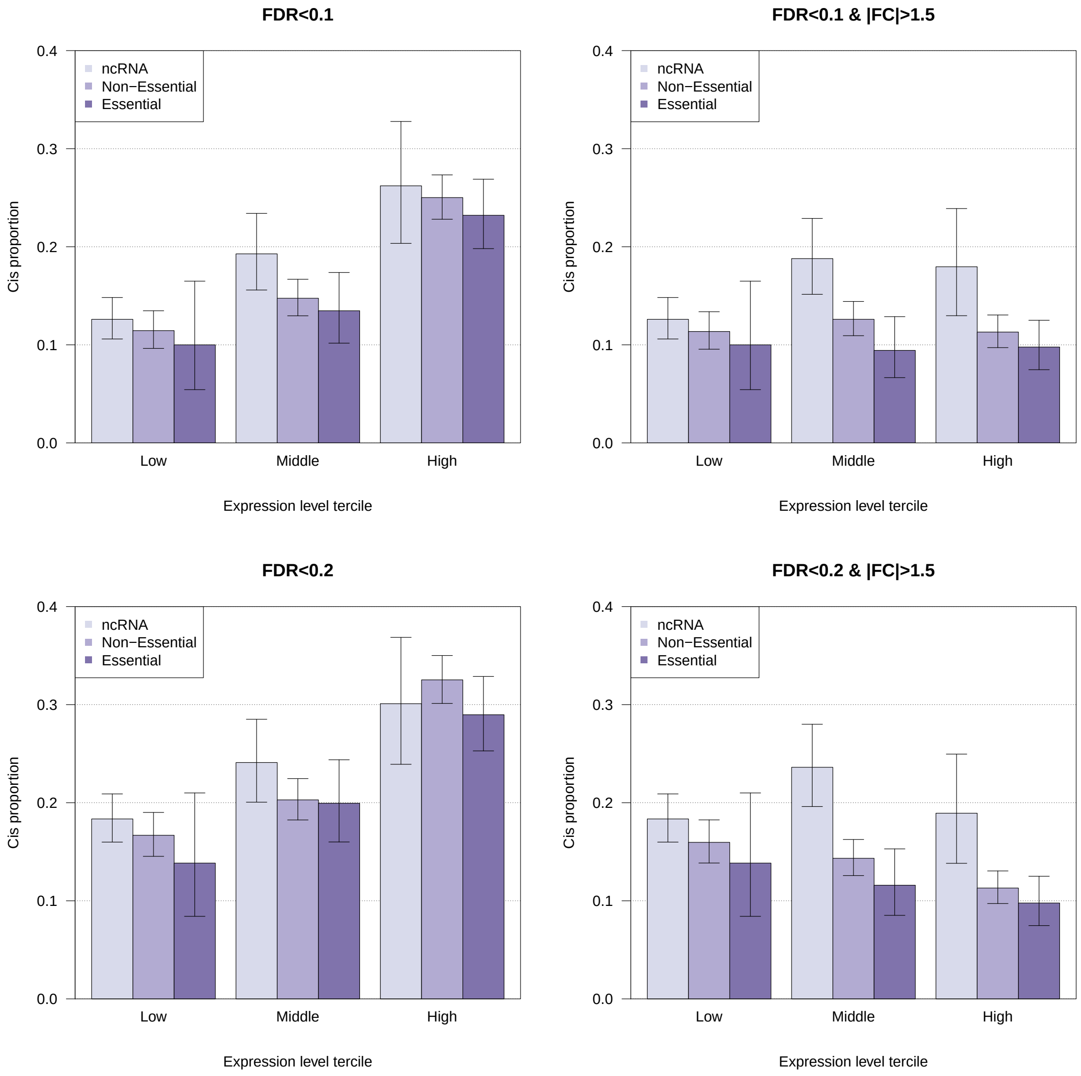


Figure 6.1: **Correlation between cis detection and expression.** Analog to figure 2.9A: proportion of cis genes for gene categories (purple shadings) and expression level terciles (grouped bars). The applied thresholds for false discovery rate (FDR) and absolute fold change ($|FC|$) are indicated for each plot (title).

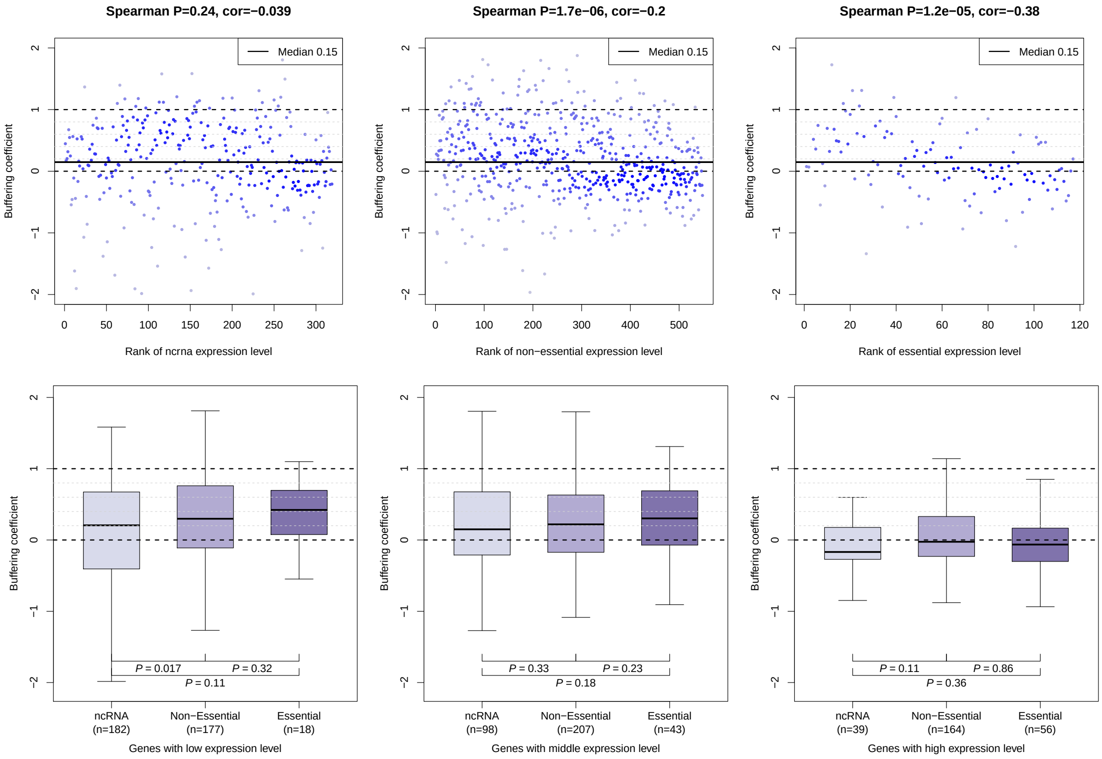


Figure 6.2: **Detailed figure 2.9C and 2.9D**. Buffering coefficient compared against the ranks of expression level for all gene categories (top). A significant one-sided Spearman correlation test (caption) confirmed the trend in Fig 2.9C as well as for Fig 2.9D (one-sided Wilcoxon test, bottom)

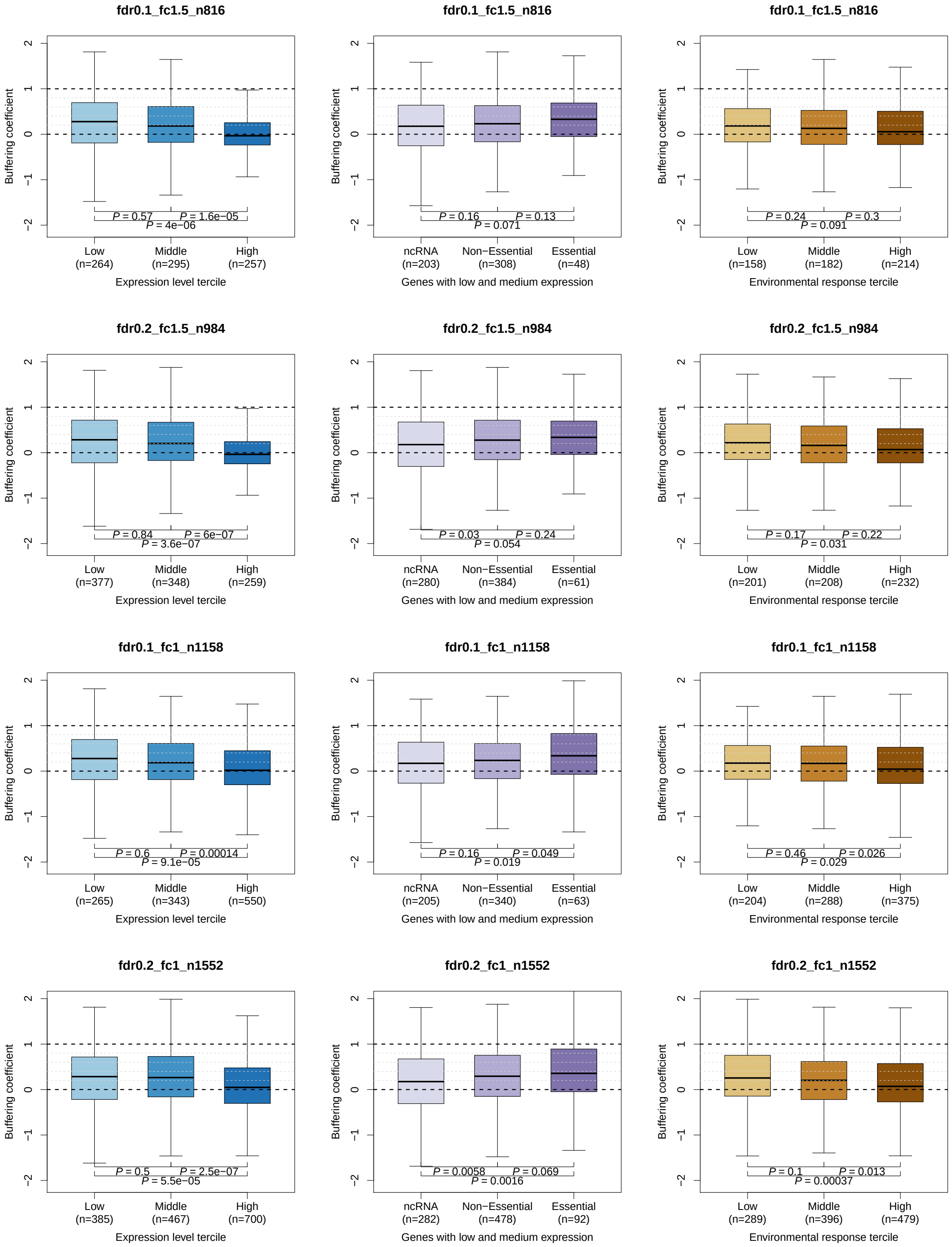


Figure 6.3: **Evaluation of different thresholds.** The figures 2.9C, 2.9D and 2.10C (left to right) are generated for different sets of cis genes. These cis gene sets vary in the false discovery rate (fdr) and absolute fold change (fc) filter that was applied (title) and therefore also in size (n, increasing top to bottom).

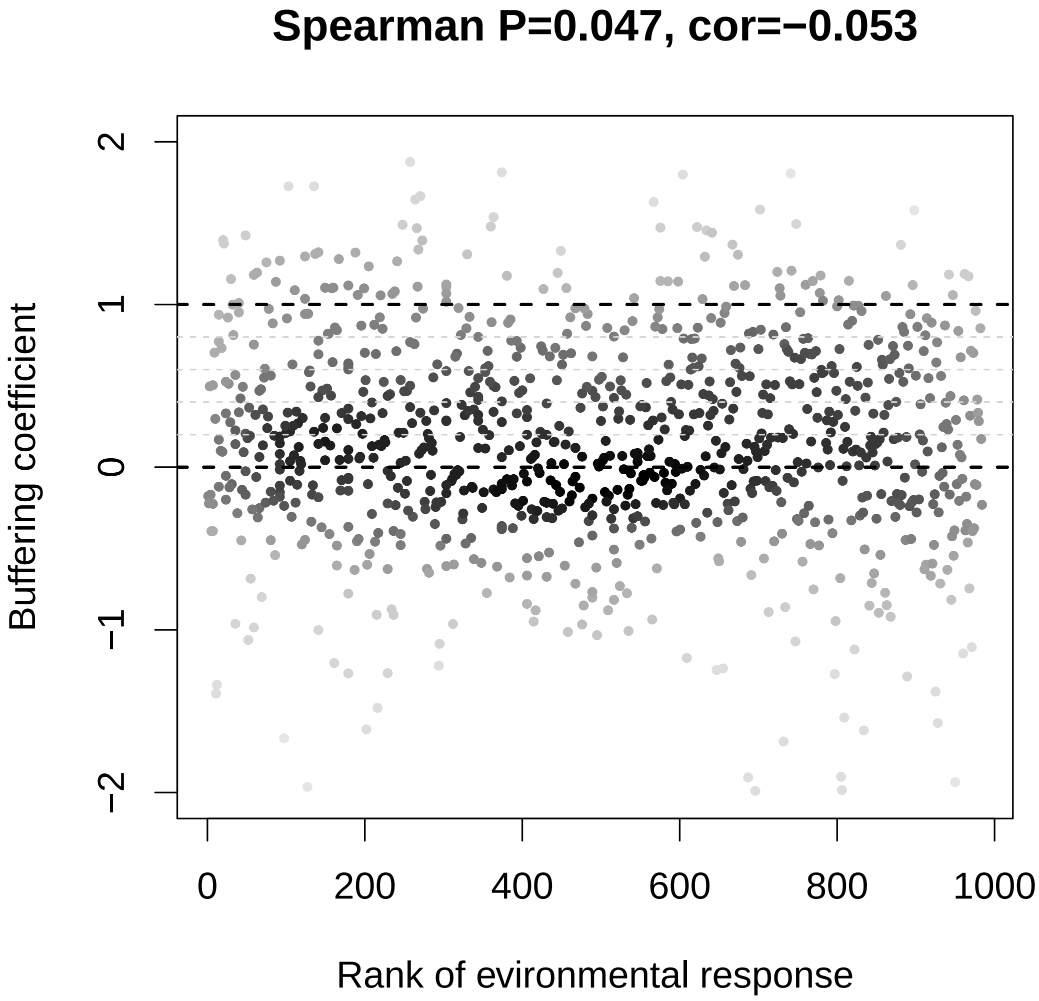


Figure 6.4: **Detailed buffering environmental response.** Buffering coefficient compared against the ranks of environmental response [45]. A one-sided Spearman correlation test (title) confirmed the trend in Fig 2.10C.

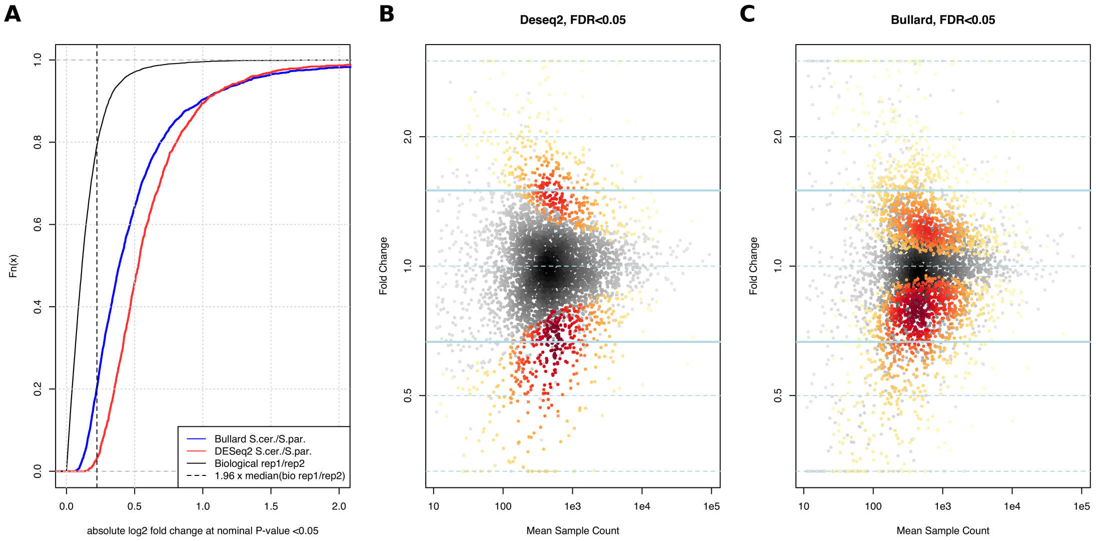


Figure 6.5: **Comparison of DESeq2-based test for ADE and the test described by [118].** **(A)** Empirical cumulative distribution of log$_2$-fold change of RNA expression level between the two biological replicates in the *S. cer* x *S. par* hybrid (black), 1.96 times the median standard deviation of log$_2$-expression levels across biological replicates in the same hybrid (vertical dashed line), log$_2$-fold change of allelic expression ratio among genes with a significant allelic differential expression at a nominal P-value of 0.05 according to the originally used statistical test based on [118] (blue) and according to the approach developed here based on DeSeq2 (red). **(B)** Same as Supplementary Fig 2.4C, but the FDR threshold of 0.05 (red) is used instead of 0.2 to match the cutoff used by [51] based on Bullard test. **(C)** Same as (B) using the originally used statistical test based on [118].

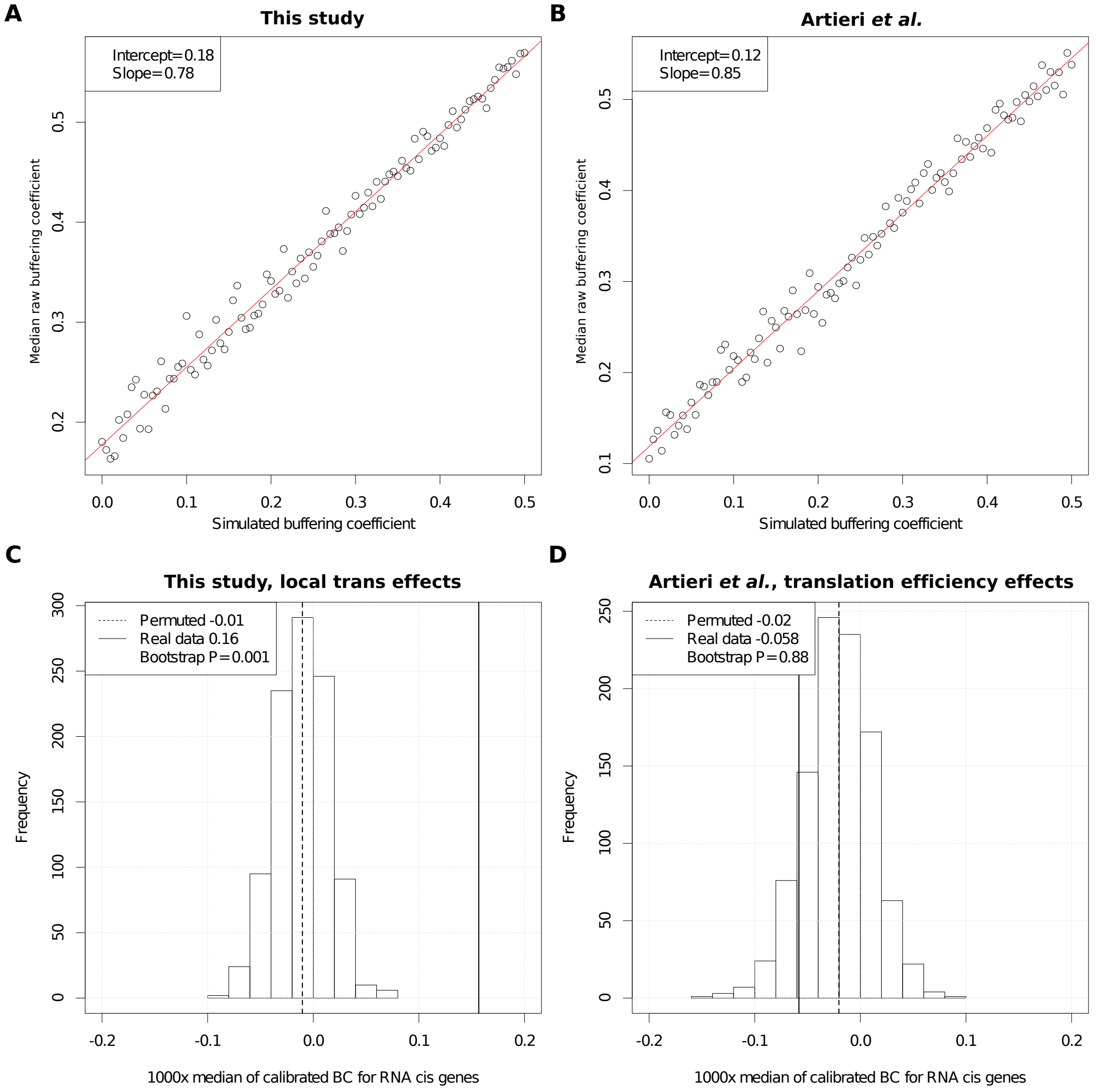


Figure 6.6: **Calibration of buffering coefficient and testing.** **(A)** Simulated genome-wide buffering coefficient (x-axis, Methods) versus median observed raw buffering coefficient (y-axis). Linear regression (red line) is used to calibrate observed raw buffering coefficients. **(B)** Analogous to (A), but for [51] data (Methods). Here too, a linear model gives a good calibration. **(C)** Distribution of median buffering coefficient across genes under independence of cis and local trans effects (1,000 permutations and dataset simulations, methods). The distribution is centered at zero (dashed line, median=-0.01) confirming with a distinct simulation scheme the correctness of the calibration. The observed median buffering coefficient (solid black line) is larger than on any dataset simulated under independence assumption (Bootstrap P-value $= 0.001$). **(D)** Analogous to (C), but for [51] data (Methods). The distribution is centered at zero (dashed line, median=-0.02) confirming with a distinct simulation scheme the correctness of the calibration. However, the observed median buffering coefficient (solid black line) is not significantly large (Bootstrap P-value $= 0.88$).

6.3 Appendix for the chapter on genetic diagnosis

This section corresponds to the results presented in chapter 3.
The supplementary information presented in this section is part of the manuscript "Genetic diagnosis of Mendelian disorders via RNA sequencing" from Kremer, Bader et al. 2016 [3].

6.3.1 Appendix figures for the chapter on genetic diagnosis

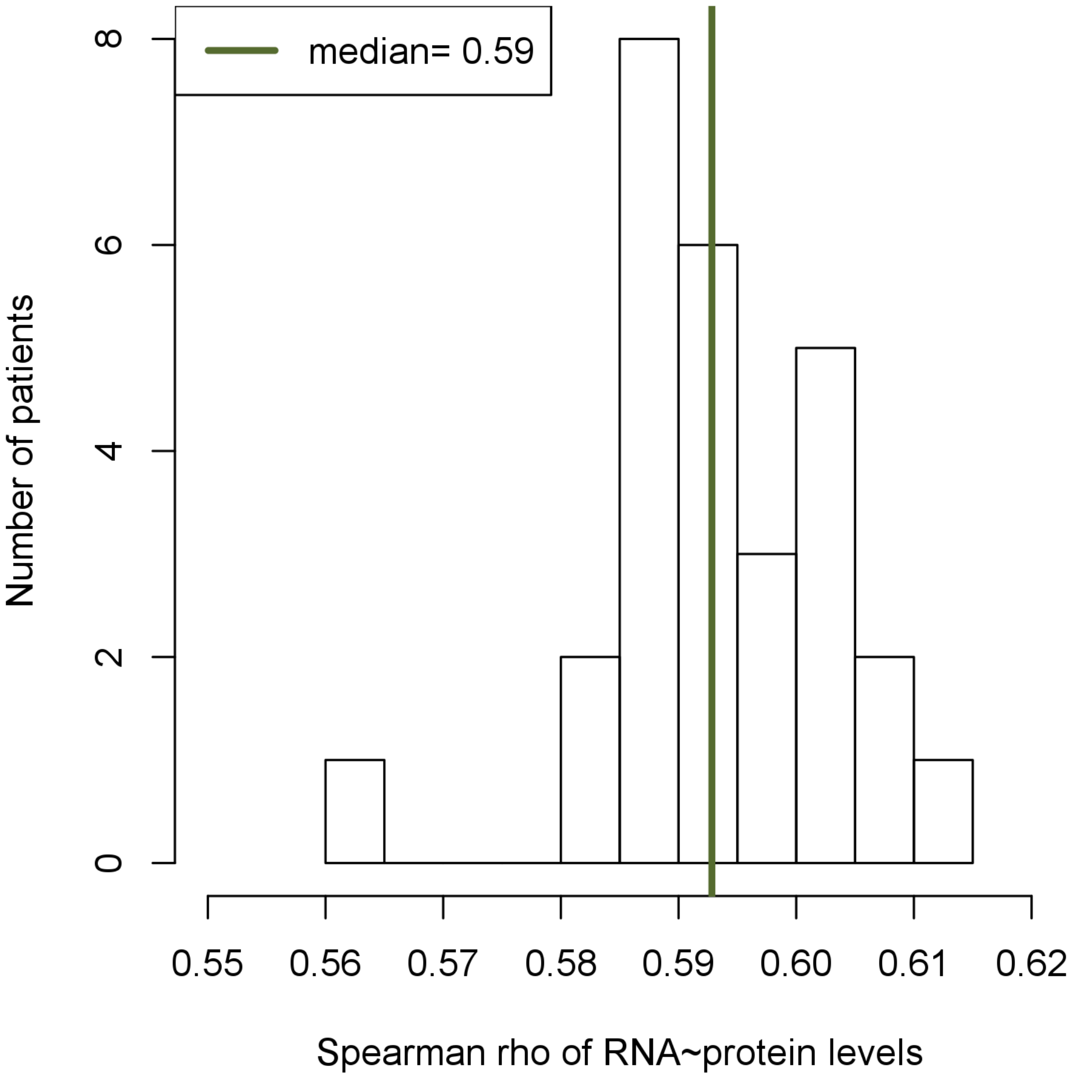


Figure 6.7: **RNA protein rank correlation.** Histogram of spearman correlation of RNA and protein levels for each patient with proteomics data available (n=31). Median is highlighted (green line).

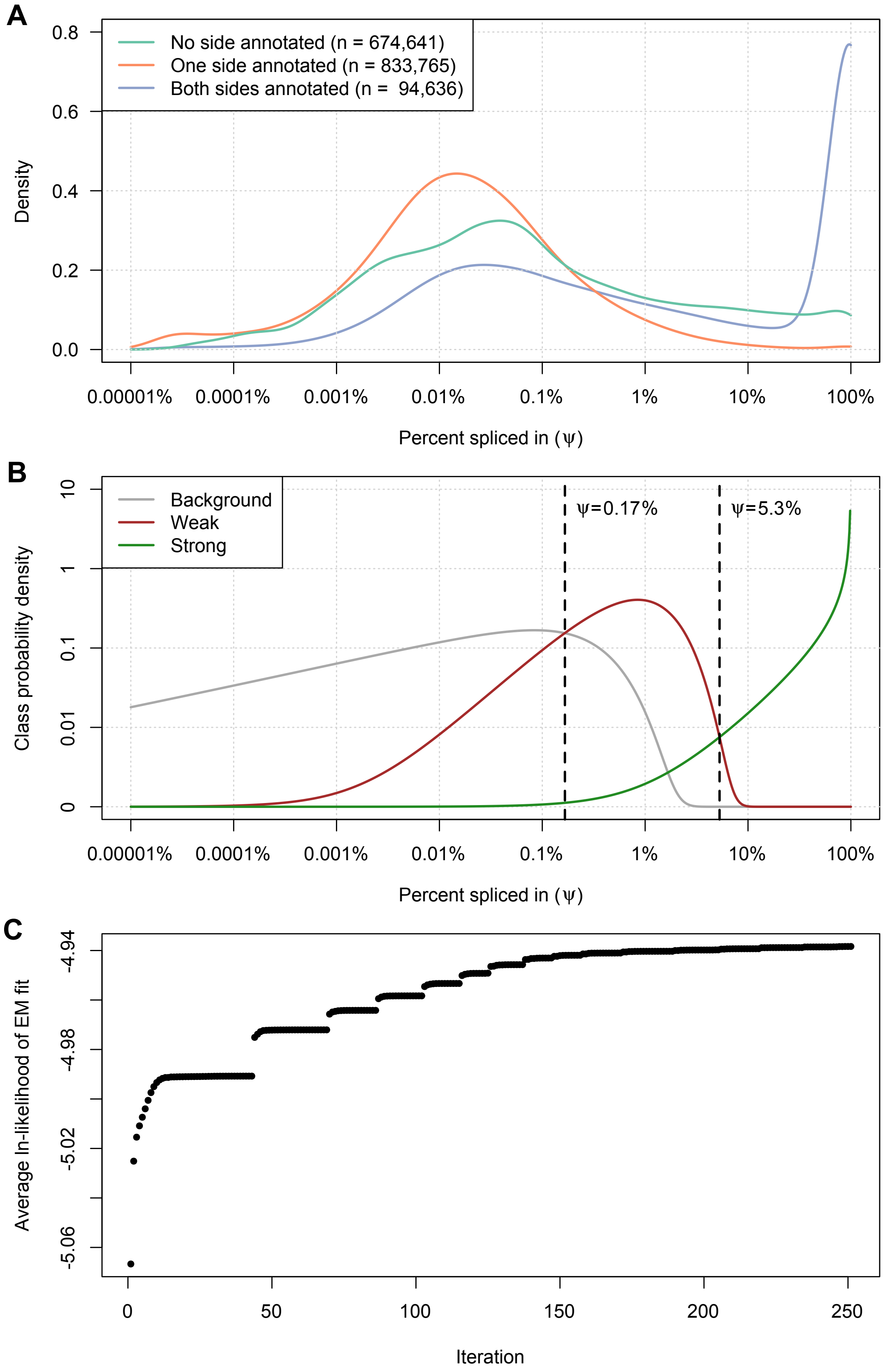

A
No side annotated (n = 674,641)
One side annotated (n = 833,765)
Both sides annotated (n = 94,636)
Density
0.0
0.2
0.4
0.6
0.8
0.00001%
0.0001%
0.001%
0.01%
0.1%
1%
10%
100%
Percent spliced in (ψ)
B
Background
Weak
Strong
ψ=0.17%
ψ=5.3%
Class probability density
0
0.01
0.1
1
10
0.00001%
0.0001%
0.001%
0.01%
0.1%
1%
10%
100%
Percent spliced in (ψ)
C
Average ln-likelihood of EM fit
-5.06
-5.02
-4.98
-4.94
0
50
100
150
200
250
Iteration

Figure 6.8: **Percent spliced in distributions. (A)** The densities of genome-wide percent spliced in (Ψ) 5' and 3' values grouped by their GENCODE annotation status: Both sides of junction are annotated (green), only one side of the junction is annotated (orange), and no side of the junction is annotated (blue). **(B)** The expectation maximization (EM) fitted splice class model based on the GENCODE annotation status. Each line represents the probability density belonging to a splice class given a Ψ-value. **(C)** The convergence of the EM algorithm. Each point represents the average ln-likelihood of the EM-fit after a specific iteration cycle (n=250).

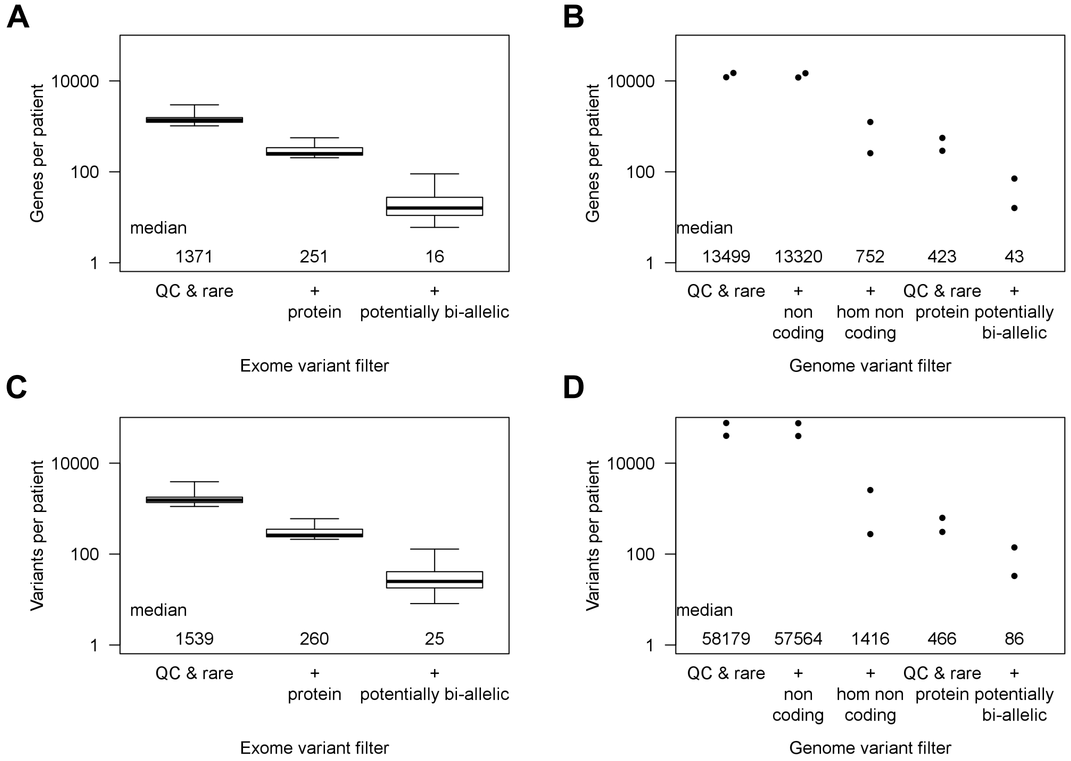


Figure 6.9: **Variant filter for exome and genome sequencing.** The different filtering steps are explained in detail in the Methods section. **(A)** Whole exome sequencing data filtered for candidate genes per patient that match the filter criteria. **(B)** Analog to (A), but for whole genome sequencing data. **(C)** Analog to (A), but for variants that match the filter criteria. **(D)** Analog to (B), but for variants that match the filter criteria. The whole genome sequencing data is based on the two TIMMDC1 deficient patients #35791 and #66744.

6.3.2 Appendix case reports for the chapter on genetic diagnosis

Informed consent was obtained from all affected individuals or their guardians in case of minor study participants. The study was approved by the ethical committee of the Technische Universität München.

Patient #35791 TIMMDC1

Variant: TIMMDC1 c.[596+2146A>G]; [596+2146A>G], variants not listed in any database.
This boy was born at term to non-consanguineous Greek parents after uneventful twin pregnancy (dizygotic twins) via cesarean delivery (weight 2450 g, length 48 cm, head circumference 33 cm). He did not show obvious dysmorphia. Shortly after birth, he was noted to have muscular hypotonia and poor feeding behavior. During his first year of life developmental delay and failure to thrive became evident. Diagnostic work-up revealed sensorineural deafness and brain MRI showed enlarged ventricles and megacisterna magna. MR-spectroscopy as well as metabolic work-up in body fluids failed to detect specific abnormalities. Biochemical analysis of fibroblast and muscle tissue demonstrated a severe isolated

complex I deficiency (16% of lowest control). He developed muscle wasting and a dyskinetic movement disorder, never achieving his developmental milestones and suffered from recurrent respiratory infections, finally leading to his death at the age of 30 months.

Patient #66744 TIMMDC1

Variant: TIMMDC1 c.[596+2146A>G]; [596+2146A>G], variants not listed in any database.
This boy was the only child of consanguineous parents from Northern Africa. Pregnancy, delivery and birth parameters were normal. Symptoms were first noted at the age of 6 months when he presented with muscular hypotonia, delayed acquisition of motor milestones, and nystagmus with altered electroretinogram and evoked visual potentials. An acute episode with abnormal eye movements, myoclonus, and loss of consciousness, followed by cerebellar syndrome, led to the diagnosis of Leigh syndrome, confirmed on a CT scan showing hypersignal in basal ganglia and mildly elevated lactate levels in blood. Subsequent NMR brain imaging was however normal. Biochemical analysis of muscle tissue showed a predominant complex I defect (15% of lowest control). Large-scale deletions and depletion of muscle mitochondrial DNA and common mtDNA mutations, including those involving MT-ATP6, were excluded. At one year of age, he presented with profound hypotonia, cerebellar syndrome with severe dysmetria, delayed mental development, and peripheral neuropathy. His lactate levels in blood and urine were repeatedly normal despite severe clinical condition. He died at 20 months of age.

Patient #96687 TIMMDC1

Variant: TIMMDC1 c.[596+2146A>G]; [596+2146A>G], variants not listed in any database.
This boy was born after uneventful pregnancy via spontaneous delivery to healthy, non-consanguineous parents from Germany (weight 4180 g, length 57 cm, head circumference 36 cm). Starting from the age of 3 months poor feeding behaviour, muscular hypotonia, and failure to thrive were noted. In the following, developmental delay and muscle wasting became evident. He showed severe cognitive/language impairment, and never achieved ambulation. Starting from the age of four years, the patient developed severe therapy-resistant epilepsy. Brain MRI studies as well as metabolic work-up did not reveal any specific abnormalities. Of note, two older siblings died due to unexplained neurodegenerative disorders with severe epilepsy.

Patient #73804 MGST1

This boy was born after uneventful pregnancy via spontaneous delivery to healthy, non-consanguineous German parents (weight 4050 g, length 56 cm, head circumference 37 cm). Soon after birth, his parents noted that he was less active and unusually quiet and did not fix or follow objects. Eye examinations suggested cortical or central blindness. In addition, hearing testing was repeatedly abnormal. During the first year of life severe developmental delay became evident and he developed epilepsy. Brain MRIs demonstrated rapidly progressive brain atrophy. Biochemical analysis of muscle tissue revealed a combined deficiency of complex III

(75% of lowest control) and complex V (67% of lowest control). Metabolic work-up was otherwise normal. He never showed developmental progress and at the current age of 17 years he has severe intellectual disability and is wheelchair-bound. Of note, his clinical course as well as biochemical and neuroimaging findings, showed similarities to a patient suffering from thioredoxin 2 deficiency, a recently described disorder of mitochondrial oxidative stress regulation [135].

Patient #58955 CLPP

Variants: CLPP, c.[661G>A];[661G>A], allele frequency 1.2×10^{-5} in ExAC.
This girl was the third child of healthy consanguineous parents (first cousins) of Turkish origin. She was born at term by spontaneous delivery after an uneventful pregnancy (weight 2755 g, length 49 cm, head circumference 37 cm). One brother, aged 17 years is healthy; one brother, aged 13 years, suffers from sensorineural deafness diagnosed at age 2 years.

Her parents noted muscular weakness from the first week of life. When she presented at age 2 months, she was microcephalic (1 cm < 3rd percentile) and showed generalized muscular hypotonia. Echocardiography demonstrated mild hypertrophic cardiomyopathy. Metabolic analysis disclosed repeatedly metabolic acidosis with elevated lactate levels (2.8 to 9.0 mmol/l, normal < 2.3 mmol/l). Fumaric acid, 2-oxo-glutaric acid, and methylmalonic acid were mildly elevated in urine. Brain MRI at the age of 3 months showed no significant abnormality. Biochemical analysis of fresh muscle tissue showed decreased complex IV activity (26% of lowest control) and the coenzyme Q10 content was decreased (30% of lowest control). Muscle immunohistochemistry revealed near absent of immunoreactivity for complex IV subunits. Subsequently, progressive developmental delay, persistent microcephaly, and deafness became evident. A percutaneous gastroenterostomy (PEG) was inserted for chronic feeding difficulties. She developed epilepsy (West-syndrome) and was treated with Topiramat, Levetiracetam and Lamotrigen. Left ventricular hypertrophy initially progressed but stabilized under treatment with propranolol. She is currently 5 years of age and has microcephaly, deafness, severe psychomotor retardation, and moderate left ventricular hypertrophy.

Patient #80256 ALDH18A1

Varaints: ALDH18A1, c.[1864C>T];[1988C>A], allele frequency 0.82×10^{-5}; not listed in ExAC.
This boy was the first child of healthy non-consanguineous German parents. Prenatal ultrasound revealed intrauterine growth retardation. He was born without complications after 39 weeks of gestation (weight 2640 g, length 49 cm, head circumference 33.5 cm). Following birth, bilateral congenital cataract and multiple small haemangiomas were noted. In the first months of life, he developed muscular hypotonia, developmental delay, severe failure to thrive, and microcephaly. Brain MRI revealed hypoplasia of the corpus callosum, lack of insular opercularization, and reduced myelination. Moreover, the child developed drug-resistant epilepsy. A metabolic work-up showed no significant abnormalities. Muscle biopsy at age 2.5 years showed no significant morphological abnormality, whilst substrate oxidation and enzyme activities of mitochondrial complexes I-V were within normal limits. His onward clinical course was characterized by severe, global developmental delay,

dyskinetic movement disorder, and epilepsy. He died at the age of 4 years from pneumonia and respiratory failure.

Patient #62346 MCOLN1

Varaints: MCOLN1, c.[681-19A>C];[832C>T], allele frequency 0.83×10^{-5}; not listed in ExAC.
This boy was the third child of healthy, non-consanguineous French parents. Pregnancy and delivery were uneventful, whilst birth parameters and early psychomotor development of the child were normal. However, speech development was delayed, the patient acquiring language at the age of 4 years. At the age of 11 years, he began to experience psychomotor regression and progressive visual loss due to degenerative retinopathy. He developed cerebellar ataxia, hyperreflexia, external ophthalmoparesis, bilateral corneal clouding, and abnormal behavior. The association of corneal clouding with a degenerative retinopathy and psychomotor regression was suggestive of mucolipidosis, but none of the enzymatic tests available for mucolipidosis type 1, 2, and 3 revealed an enzyme deficiency in blood leukocytes. Muscle biopsy showed moderate subsarcolemmal accumulation of mitochondria. At the current age of 47 years he has severe walking difficulties due to ataxia and blindness. On examination, he has cerebellar ataxia, hyperreflexia, external ophthalmoparesis predominating in vertical gaze, bilateral corneal clouding, and abnormal behavior (easily frightened, sometimes aggressive). Spontaneous speech is markedly reduced.

List of Figures

List of Tables

Bibliography

[1] Bader, D. M. *et al.* Negative feedback buffers effects of regulatory variants. *Molecular Systems Biology* **11**, 785–785 (2015). URL `http://msb.embopress.org/cgi/doi/10.15252/msb.20145844`.

[2] Haack, T. B. *et al.* Biallelic Mutations in NBAS Cause Recurrent Acute Liver Failure with Onset in Infancy. *American journal of human genetics* **97**, 163–9 (2015). URL `http://www.ncbi.nlm.nih.gov/pubmed/26073778`.

[3] Kremer, L. S. *et al.* Genetic diagnosis of Mendelian disorders via RNA sequencing. *bioRxiv* (2016). URL `http://biorxiv.org/lookup/doi/10.1101/066738`.

[4] Kremer, L. S. *et al.* Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nature Communications* **8**, 15824 (2017). URL `http://www.nature.com/doifinder/10.1038/ncomms15824`.

[5] Crick, F. H. C. On protein synthesis. *Symposia of the Society for Experimental Biology* **12**, 138–63 (1958). URL `http://www.ncbi.nlm.nih.gov/pubmed/13580867`.

[6] Crick, F. H. C. Central dogma of molecular biology. *Nature* **227**, 561–3 (1970). URL `http://www.ncbi.nlm.nih.gov/pubmed/4913914`. `arXiv:1011.1669v3`.

[7] Baldovino, S., Moliner, A. M., Taruscio, D., Daina, E. & Roccatello, D. Rare Diseases in Europe: from a Wide to a Local Perspective. *The Israel Medical Association journal : IMAJ* **18**, 359–63 (2016). URL `http://www.ncbi.nlm.nih.gov/pubmed/27468531`.

[8] Holley, R. W. *et al.* Structure of a ribonucleic acid. *Science (New York, N.Y.)* **147**, 1462–5 (1965). URL `http://www.ncbi.nlm.nih.gov/pubmed/14263761`.

[9] Sanger, F., Brownlee, G. G. & Barrell, B. G. A two-dimensional fractionation procedure for radioactive nucleotides. *Journal of molecular biology* **13**, 373–98 (1965). URL `http://www.ncbi.nlm.nih.gov/pubmed/5325727`.

[10] Wu, R. & Kaiser, A. D. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *Journal of molecular biology* **35**, 523–37 (1968). URL `http://www.ncbi.nlm.nih.gov/pubmed/4299833`.

[11] Wu, R. & Taylor, E. Nucleotide sequence analysis of DNA. *Journal of Molecular Biology* **57**, 491–511 (1971). URL `http://www.sciencedirect.com/science/article/pii/0022283671901057`.

[12] Min Jou, W., Haegeman, G., Ysebaert, M. & Fiers, W. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* **237**, 82–8 (1972). URL `http://www.ncbi.nlm.nih.gov/pubmed/4555447`.

[13] Fiers, W. *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**, 500–7 (1976). URL `http://www.ncbi.nlm.nih.gov/pubmed/1264203`. `arXiv:1011.1669v3`.

[14] Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology* **94**, 441–8 (1975). URL `http://www.ncbi.nlm.nih.gov/pubmed/1100841`.

[15] Maxam, a. M. & Gilbert, W. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences* **74**, 560–564 (1977). URL `http://www.pnas.org/cgi/doi/10.1073/pnas.74.2.560`.

[16] Sanger, F., Nicklen, S. & Coulson, a. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5463–7 (1977). URL `http://www.ncbi.nlm.nih.gov/pubmed/271968`. `0402594v3`.

[17] Sanger, F. *et al.* The nucleotide sequence of bacteriophage ϕX174. *Journal of Molecular Biology* **125**, 225–246 (1978). URL `http://linkinghub.elsevier.com/retrieve/pii/0022283678903467`.

[18] Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8 (2016). URL `http://dx.doi.org/10.1016/j.ygeno.2015.11.003`.

[19] Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001). URL `http://www.nature.com/doifinder/10.1038/35057062`. `11237011`.

[20] Venter, J. C. *et al.* The sequence of the human genome. *Science (New York, N.Y.)* **291**, 1304–51 (2001). URL `http://www.sciencemag.org/cgi/doi/10.1126/science.1058040`.

[21] Goffeau, A. *et al.* Life with 6000 genes. *Science (New York, N.Y.)* **274**, 546, 563–7 (1996). URL `http://www.ncbi.nlm.nih.gov/pubmed/8849441`.

[22] Human Genome Sequencing Consortium, I. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004). URL `http://www.nature.com/doifinder/10.1038/nature03001`. `NIHMS150003`.

[23] Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–80 (2005). URL `http://www.ncbi.nlm.nih.gov/pubmed/16056220`.

[24] Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature reviews. Genetics* **17**, 333–51 (2016). URL `https://www.ncbi.nlm.nih.gov/pubmed/27184599`.

[25] Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008). URL `http://www.nature.com/doifinder/10.1038/nature07517`. `NIHMS150003`.

[26] Lu, Y., Shen, Y., Warren, W. & Walter, R. Next Generation Sequencing in Aquatic Models. In *Next Generation Sequencing - Advances, Applications and Challenges*, Chapter 2 (InTech, 2016). URL `http://dx.doi.org/10.5772/61657`.

[27] David, L. *et al.* A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 5320–5325 (2006). URL `http://www.pnas.org/cgi/doi/10.1073/pnas.0601091103`.

[28] Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science (New York, N.Y.)* **306**, 2242–6 (2004). URL `http://www.sciencemag.org/cgi/doi/10.1126/science.1103388`. `arXiv:1011.1669v3`.

[29] Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* **10**, 57–63 (2009). URL `http://www.ncbi.nlm.nih.gov/pubmed/19015660`. `NIHMS150003`.

[30] Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)* **320**, 1344–9 (2008). URL `http://www.ncbi.nlm.nih.gov/pubmed/18451266`. `1006.1266v2`.

[31] Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research* **18**, 1509–17 (2008). URL `http://www.ncbi.nlm.nih.gov/pubmed/18550803`.

[32] Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nature reviews. Genetics* **12**, 87–98 (2011). URL `http://www.nature.com/doifinder/10.1038/nrg2934`. `NIHMS150003`.

[33] Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic acids research* **37**, e123 (2009). URL `http://www.ncbi.nlm.nih.gov/pubmed/19620212`.

[34] Robinson, M. D. & Smyth, G. K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics (Oxford, England)* **23**, 2881–7 (2007). URL `http://www.ncbi.nlm.nih.gov/pubmed/17881408`.

[35] Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* **26**, 139–40 (2010). URL `http://www.ncbi.nlm.nih.gov/pubmed/19910308`.

[36] Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11**, R106 (2010). URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-10-r106.

[37] Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014). URL http://genomebiology.com/2014/15/12/550.

[38] King, M. & Wilson, A. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975). URL http://www.sciencemag.org/cgi/doi/10.1126/science.1090005. arXiv:1011.1669v3.

[39] Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science (New York, N.Y.)* **296**, 752–5 (2002). URL http://www.ncbi.nlm.nih.gov/pubmed/11923494.

[40] Cowles, C. R., Hirschhorn, J. N., Altshuler, D. & Lander, E. S. Detection of regulatory variation in mouse genes. *Nature genetics* **32**, 432–7 (2002). URL http://www.ncbi.nlm.nih.gov/pubmed/12410233.

[41] Ronald, J., Brem, R. B., Whittle, J. & Kruglyak, L. Local Regulatory Variation in Saccharomyces cerevisiae. *PLoS Genetics* **1**, e25 (2005). URL http://dx.plos.org/10.1371/journal.pgen.0010025.

[42] Rockman, M. V. & Kruglyak, L. Genetics of global gene expression. *Nature reviews. Genetics* **7**, 862–72 (2006). URL http://www.ncbi.nlm.nih.gov/pubmed/17047685.

[43] Skelly, D. a., Ronald, J. & Akey, J. M. Inherited variation in gene expression. *Annual review of genomics and human genetics* **10**, 313–32 (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19630563.

[44] Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nature reviews. Genetics* **16**, 197–212 (2015). URL http://www.nature.com/doifinder/10.1038/nrg3891.

[45] Tirosh, I., Reikhav, S., Levy, A. a. & Barkai, N. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science (New York, N.Y.)* **324**, 659–62 (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19407207.

[46] Emerson, J. J. *et al.* Natural selection on cis and trans regulation in yeasts. *Genome research* **20**, 826–836 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20445163.

[47] Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Evolutionary changes in cis and trans gene regulation. *Nature* **430**, 85–8 (2004). URL http://www.ncbi.nlm.nih.gov/pubmed/15229602.

[48] Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Regulatory changes underlying expression differences within and between Drosophila species. *Nature genetics* **40**, 346–50 (2008). URL http://dx.doi.org/10.1038/ng.77.

[49] McManus, C. J. *et al.* Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome research* **20**, 816–25 (2010). URL `http://genome.cshlp.org/content/20/6/816.abstract`.

[50] Goncalves, A. *et al.* Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome research* **22**, 2376–84 (2012). URL `http://genome.cshlp.org/content/22/12/2376`.

[51] Artieri, C. C. G. & Fraser, H. B. H. Evolution at two levels of gene expression in yeast. *Genome research* **24**, 411–21 (2014). URL `http://www.ncbi.nlm.nih.gov/pubmed/24318729. 1311.7140`.

[52] McManus, C. J., May, G. E., Spealman, P., Shteyman, A. & McManus, J. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome research* **24**, 422–30 (2014). URL `http://www.ncbi.nlm.nih.gov/pubmed/24318730`.

[53] Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B. & Kinzler, K. W. Allelic variation in human gene expression. *Science (New York, N.Y.)* **297**, 1143 (2002). URL `http://www.sciencemag.org/cgi/doi/10.1126/science.1072545`.

[54] Muzzey, D., Sherlock, G. & Weissman, J. S. Extensive and coordinated control of allele-specific expression by both transcription and translation in Candida albicans. *Genome Research* **24**, 963–973 (2014). URL `http://genome.cshlp.org/cgi/doi/10.1101/gr.166322.113`.

[55] Denby, C. M., Im, J. H., Yu, R. C., Pesce, C. G. & Brem, R. B. Negative feedback confers mutational robustness in yeast transcription factor regulation. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 3874–8 (2012). URL `http://www.pnas.org/content/109/10/3874.long`.

[56] Wang, W. Y. S., Barratt, B. J., Clayton, D. G. & Todd, J. A. Genome-wide association studies: theoretical and practical concerns. *Nature reviews. Genetics* **6**, 109–18 (2005). URL `http://www.ncbi.nlm.nih.gov/pubmed/15716907`.

[57] Hardy, J. & Singleton, A. Genomewide association studies and human disease. *The New England journal of medicine* **360**, 1759–68 (2009). URL `http://www.nejm.org/doi/abs/10.1056/NEJMra0808700`.

[58] Manolio, T. A. Genomewide Association Studies and Assessment of Risk of Disease. *New England Journal of Medicine* **363**, 2076–2077 (2010). URL `http://www.nejm.org/doi/abs/10.1056/NEJMc1010310`.

[59] Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* **42**, 937–948 (2010). URL `http://www.nature.com/doifinder/10.1038/ng.686`.

[60] Klein, R. J. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science* **308**, 385–389 (2005). URL `http://www.sciencemag.org/cgi/doi/10.1126/science.1109557`.

[61] Maraganore, D. M. *et al.* High-resolution whole-genome association study of Parkinson disease. *American journal of human genetics* **77**, 685–93 (2005). URL `http://www.ncbi.nlm.nih.gov/pubmed/16252231`.

[62] Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009). URL `http://www.nature.com/doifinder/10.1038/nature08494`.

[63] Mendel, G. Versuche ueber Pflanzen-Hybriden. *Verhandlungen des Naturfoschenden Vereines in Bruenn* **4**, 3–47 (1866). URL `http://www.biodiversitylibrary.org/item/124139`.

[64] Shashi, V. *et al.* The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. *Genetics in Medicine* **16**, 176–182 (2014). URL `http://www.nature.com/doifinder/10.1038/gim.2013.99`.

[65] McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics* **9**, 356–69 (2008). URL `http://www.nature.com/doifinder/10.1038/nrg2344`.

[66] Hodges, E. *et al.* Genome-wide in situ exon capture for selective resequencing. *Nature genetics* **39**, 1522–7 (2007). URL `http://www.ncbi.nlm.nih.gov/pubmed/17982454`.

[67] Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–6 (2009). URL `http://dx.doi.org/10.1038/nature08250`.

[68] Haack, T. B. *et al.* Exome sequencing identifies ACAD9 mutations as a cause of complex I deficiency. *Nature genetics* **42**, 1131–4 (2010). URL `http://www.ncbi.nlm.nih.gov/pubmed/21057504`.

[69] Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics* **42**, 30–5 (2010). URL `http://dx.doi.org/10.1038/ng.499`. `15334406`.

[70] Hoischen, A. *et al.* De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nature genetics* **42**, 483–5 (2010). URL `http://www.ncbi.nlm.nih.gov/pubmed/20436468`.

[71] MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014). URL `http://www.nature.com/doifinder/10.1038/nature13127`. `NIHMS150003`.

[72] Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research* **42**, D980–5 (2014). URL `http://www.ncbi.nlm.nih.gov/pubmed/24234437`.

[73] Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic acids research* **44**, D862–8 (2016). URL `http://www.ncbi.nlm.nih.gov/pubmed/26582918`.

[74] Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016). URL `http://www.nature.com/doifinder/10.1038/nature19057`. `030338`.

[75] Ruderfer, D. M. *et al.* Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nature Genetics* **48**, 1107–1111 (2016). URL `http://www.nature.com/doifinder/10.1038/ng.3638`.

[76] Boycott, K. M., Vanstone, M. R., Bulman, D. E. & MacKenzie, A. E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature reviews. Genetics* **14**, 681–91 (2013). URL `http://www.nature.com/doifinder/10.1038/nrg3555`.

[77] Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *American journal of human genetics* **97**, 199–215 (2015). URL `http://linkinghub.elsevier.com/retrieve/pii/S0002929715002451`.

[78] Lee, H. *et al.* Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA* **312**, 1880–7 (2014). URL `http://www.ncbi.nlm.nih.gov/pubmed/25326637`. `NIHMS150003`.

[79] Retterer, K. *et al.* Clinical application of whole-exome sequencing across clinical indications. *Genetics in Medicine* **18**, 696–704 (2016). URL `http://www.nature.com/doifinder/10.1038/gim.2015.148`.

[80] Mayr, J. A. *et al.* Spectrum of combined respiratory chain defects. *Journal of Inherited Metabolic Disease* **38**, 629–640 (2015). URL `http://link.springer.com/10.1007/s10545-015-9831-y`.

[81] Lightowlers, R. N., Taylor, R. W. & Turnbull, D. M. Mutations causing mitochondrial disease: What is new and what challenges remain? *Science* **349**, 1494–1499 (2015). URL `http://www.sciencemag.org/cgi/doi/10.1126/science.aac7516`.

[82] Taylor, R. W. *et al.* Use of Whole-Exome Sequencing to Determine the Genetic Basis of Multiple Mitochondrial Respiratory Chain Complex Deficiencies. *JAMA* **312**, 68 (2014). URL `http://www.ncbi.nlm.nih.gov/pubmed/25058219`.

[83] Taylor, J. C. *et al.* Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nature genetics* **47**, 717–26 (2015). URL `http://www.nature.com/doifinder/10.1038/ng.3304`.

[84] UK10K Consortium *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015). URL `http://www.nature.com/doifinder/10.1038/nature14962`.

[85] Reardon, S. Giant study poses DNA data-sharing dilemma. *Nature* **525**, 16–17 (2015). URL `http://www.nature.com/doifinder/10.1038/525016a`.

[86] Cyranoski, D. China embraces precision medicine on a massive scale. *Nature* **529**, 9–10 (2016). URL `http://www.nature.com/doifinder/10.1038/529009a`.

[87] Schadt, E. E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003). URL `http://www.nature.com/doifinder/10.1038/nature01434`.

[88] Schadt, E. E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics* **37**, 710–717 (2005). URL `http://www.nature.com/doifinder/10.1038/ng1589`.

[89] Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423–428 (2008). URL `http://www.nature.com/doifinder/10.1038/nature06758`.

[90] Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–72 (2010). URL `http://www.nature.com/doifinder/10.1038/nature08872`.

[91] Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–7 (2010). URL `http://www.ncbi.nlm.nih.gov/pubmed/20220756`.

[92] Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007). URL `http://dx.doi.org/10.1038/nature06258`.

[93] Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010). URL `http://www.nature.com/doifinder/10.1038/nature09298`.

[94] Montgomery, S. B. & Dermitzakis, E. T. From expression QTLs to personalized transcriptomics. *Nature reviews. Genetics* **12**, 277–82 (2011). URL `http://www.ncbi.nlm.nih.gov/pubmed/21386863`.

[95] Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535–50 (2016). URL `http://dx.doi.org/10.1016/j.cell.2016.03.014`.

[96] Zeng, Y. *et al.* Aberrant Gene Expression in Humans. *PLOS Genetics* **11**, e1004942 (2015). URL `http://dx.plos.org/10.1371/journal.pgen.1004942`.

[97] Guan, J. *et al.* Exploiting aberrant mRNA expression in autism for gene discovery and diagnosis. *Human Genetics* **135**, 797–811 (2016). URL `http://www.ncbi.nlm.nih.gov/pubmed/27131873`.

[98] Zhao, J. *et al.* A Burden of Rare Variants Associated with Extremes of Gene Expression in Human Peripheral Blood. *American journal of human genetics* **98**, 299–309 (2016). URL `http://dx.doi.org/10.1016/j.ajhg.2015.12.023`.

[99] Albers, C. A. *et al.* Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. *Nature genetics* **44**, 435–9, S1–2 (2012). URL http://www.nature.com/doifinder/10.1038/ng.1083.

[100] Reinius, B. & Sandberg, R. Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. *Nature Reviews. Genetics* **16**, 653–664 (2015). URL http://www.nature.com/doifinder/10.1038/nrg3888.

[101] Eckersley-Maslin, M. A. & Spector, D. L. Random monoallelic expression: regulating gene expression one allele at a time. *Trends in genetics : TIG* **30**, 237–44 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24780084. NIHMS150003.

[102] Tazi, J., Bakkour, N. & Stamm, S. Alternative splicing and disease. (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/18992329.

[103] Singh, R. K. & Cooper, T. A. Pre-mRNA splicing in disease and therapeutics. *Trends in Molecular Medicine* **18**, 472–482 (2012). URL http://dx.doi.org/10.1016/j.molmed.2012.06.006.

[104] Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nature Reviews. Genetics* **17**, 19–32 (2015). URL http://www.nature.com/doifinder/10.1038/nrg.2015.3.

[105] Xiong, H. Y. *et al.* RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science (New York, N.Y.)* **347**, 1254806 (2015). URL http://www.sciencemag.org/cgi/doi/10.1126/science.1254806. 9605103.

[106] Muntoni, F., Torelli, S. & Ferlini, A. Dystrophin and mutations: one gene, several proteins, multiple phenotypes. *The Lancet. Neurology* **2**, 731–740 (2003). URL http://www.ncbi.nlm.nih.gov/pubmed/14636778.

[107] Gonorazky, H. *et al.* RNAseq analysis for the diagnosis of muscular dystrophy. *Annals of clinical and translational neurology* **3**, 55–60 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/26783550.

[108] Morel, C. F. *et al.* A LMNA Splicing Mutation in Two Sisters with Severe Dunnigan-Type Familial Partial Lipodystrophy Type 2. *The Journal of Clinical Endocrinology {&} Metabolism* **91**, 2689–2695 (2006). URL http://press.endocrine.org/doi/10.1210/jc.2005-2746.

[109] Qu, Y.-j. *et al.* A rare variant (c.863G{>}T) in exon 7 of SMN1 disrupts mRNA splicing and is responsible for spinal muscular atrophy. *European Journal of Human Genetics* **24**, 864–870 (2016). URL http://www.nature.com/doifinder/10.1038/ejhg.2015.213.

[110] Mortimer, R. K. & Johnston, J. R. Genealogy of principal strains of the yeast genetic stock center. *Genetics* **113**, 35–43 (1986). URL http://www.ncbi.nlm.nih.gov/pubmed/3519363.

[111] Cherry, J. M. *et al.* Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic acids research* **40**, D700–5 (2012). URL `http://nar.oxfordjournals.org/content/40/D1/D700`.

[112] Kane, S. M. & Roth, R. Carbohydrate metabolism during ascospore development in yeast. *Journal of bacteriology* **118**, 8–14 (1974). URL `http://www.ncbi.nlm.nih.gov/pubmed/4595206`.

[113] Nishant, K. T. *et al.* The baker's yeast diploid genome is remarkably stable in vegetative growth and meiosis. *PLoS genetics* **6**, e1001109 (2010). URL `http://dx.plos.org/10.1371/journal.pgen.1001109`.

[114] Ehrenreich, I. M. *et al.* Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature* **464**, 1039–42 (2010). URL `http://www.ncbi.nlm.nih.gov/pubmed/20393561`.

[115] Parts, L. *et al.* Revealing the genetic structure of a trait by sequencing a population under selection. *Genome Research* **21**, 1131–1138 (2011). URL `http://genome.cshlp.org/cgi/doi/10.1101/gr.116731.110`.

[116] Wilkening, S. *et al.* An Evaluation of High-Throughput Approaches to QTL Mapping in Saccharomyces cerevisiae. *Genetics* **196**, 853–865 (2014). URL `http://www.genetics.org/cgi/doi/10.1534/genetics.113.160291`.

[117] Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**, 1033–7 (2009). URL `http://www.nature.com/doifinder/10.1038/nature07728`.

[118] Bullard, J. H., Mostovoy, Y., Dudoit, S. & Brem, R. B. Polygenic and directional regulatory evolution across pathways in Saccharomyces. *Proceedings of the National Academy of Sciences* **107**, 5058–5063 (2010). URL `http://www.pnas.org/cgi/doi/10.1073/pnas.0912959107`.

[119] Suvorov, A. *et al.* Intra-specific regulatory variation in Drosophila pseudoobscura. *PloS one* **8**, e83547 (2013). URL `http://dx.plos.org/10.1371/journal.pone.0083547`.

[120] Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2000). URL `http://www.nature.com/doifinder/10.1038/75556. 10614036`.

[121] Yvert, G. *et al.* Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nature Genetics* **35**, 57–64 (2003). URL `http://www.nature.com/doifinder/10.1038/ng1222`.

[122] Wykoff, D. D., Rizvi, A. H., Raser, J. M., Margolin, B. & O'Shea, E. K. Positive feedback regulates switching of phosphate transporters in S. cerevisiae. *Molecular cell* **27**, 1005–13 (2007). URL `http://www.ncbi.nlm.nih.gov/pubmed/17889672`.

[123] Wang, Y., Shirogane, T., Liu, D., Harper, J. W. & Elledge, S. J. Exit from exit: resetting the cell cycle through Amn1 inhibition of G protein signaling.

Cell **112**, 697–709 (2003). URL http://www.ncbi.nlm.nih.gov/pubmed/12628189.

[124] Gagneur, J. *et al.* Genome-wide allele- and strand-specific expression profiling. *Molecular systems biology* **5**, 274 (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19536197.

[125] Lundberg, L. E., Figueiredo, M. L. a., Stenberg, P. & Larsson, J. Buffering and proteolysis are induced by segmental monosomy in Drosophila melanogaster. *Nucleic acids research* **40**, 5926–37 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22434883.

[126] Albert, F. W., Muzzey, D., Weissman, J. S. & Kruglyak, L. Genetic Influences on Translation in Yeast. *PLoS genetics* **10**, e1004692 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/25340754.

[127] Springer, M., Weissman, J. S. & Kirschner, M. W. A general lack of compensation for gene dosage in yeast. *Molecular Systems Biology* **6**, 368 (2010). URL http://onlinelibrary.wiley.com/doi/10.1038/msb.2010.19/full.

[128] Gorman, G. S. *et al.* Mitochondrial diseases. *Nature Reviews Disease Primers* **2**, 16080 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27775730.

[129] Elstner, M. *et al.* MitoP2: an integrative tool for the analysis of the mitochondrial proteome. *Molecular biotechnology* **40**, 306–15 (2008). URL http://link.springer.com/10.1007/s12033-008-9100-5.

[130] Haack, T. B. *et al.* ELAC2 mutations cause a mitochondrial RNA processing defect associated with hypertrophic cardiomyopathy. *American Journal of Human Genetics* **93**, 211–223 (2013). URL http://dx.doi.org/10.1016/j.ajhg.2013.06.006.

[131] Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science (New York, N.Y.)* **352**, 600–4 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27126046.

[132] Lewis, E. B. A gene complex controlling segmentation in Drosophila. *Nature* **276**, 565–570 (1978). URL http://www.ncbi.nlm.nih.gov/pubmed/103000. 1011.1669.

[133] Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382–387 (2014). URL http://www.nature.com/doifinder/10.1038/nature13438.

[134] Lee, K. K., Shimoji, M., Hossain, Q. S., Sunakawa, H. & Aniya, Y. Novel function of glutathione transferase in rat liver mitochondrial membrane: role for cytochrome c release from mitochondria. *Toxicology and applied pharmacology* **232**, 109–118 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/18634816.

[135] Holzerova, E. *et al.* Human thioredoxin 2 deficiency impairs mitochondrial redox homeostasis and causes early-onset neurodegeneration. *Brain : a journal of neurology* **139**, 346–354 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/26626369.

[136] Guarani, V. *et al.* TIMMDC1/C3orf1 functions as a membrane-embedded mitochondrial complex I assembly factor through association with the MCIA complex. *Molecular and cellular biology* **34**, 847–61 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24344204.

[137] Andrews, B., Carroll, J., Ding, S., Fearnley, I. M. & Walker, J. E. Assembly factors for the membrane arm of human complex I. *Proceedings of the National Academy of Sciences* **110**, 18934–18939 (2013). URL http://www.pnas.org/cgi/doi/10.1073/pnas.1319247110.

[138] Pervouchine, D. D., Knowles, D. G. & Guigó, R. Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics (Oxford, England)* **29**, 273–4 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/23172860.

[139] Halperin, T., Zheng, B., Itzhaki, H., Clarke, A. K. & Adam, Z. Plant mitochondria contain proteolytic and regulatory subunits of the ATP-dependent Clp protease. *Plant molecular biology* **45**, 461–8 (2001). URL http://www.ncbi.nlm.nih.gov/pubmed/11352464.

[140] Jenkinson, E. M. *et al.* Perrault syndrome: further evidence for genetic heterogeneity. *Journal of Neurology* **259**, 974–976 (2012). URL http://link.springer.com/10.1007/s00415-011-6285-5.

[141] Jenkinson, E. M. *et al.* Perrault Syndrome Is Caused by Recessive Mutations in CLPP, Encoding a Mitochondrial ATP-Dependent Chambered Protease. *The American Journal of Human Genetics* **92**, 605–613 (2013). URL http://dx.doi.org/10.1016/j.ajhg.2013.02.013.

[142] Szczepanowska, K. *et al.* CLPP coordinates mitoribosomal assembly through the regulation of ERAL1 levels. *The EMBO Journal* **35**, 2566–2583 (2016). URL http://emboj.embopress.org/lookup/doi/10.15252/embj.201694253.

[143] 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). URL http://www.nature.com/doifinder/10.1038/nature15393. 15334406.

[144] Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015). URL http://www.nature.com/doifinder/10.1038/nature15394. NIHMS150003.

[145] Piva, F., Giulietti, M., Burini, A. B. & Principato, G. SpliceAid 2: a database of human splicing factors expression data and RNA target motifs. *Human mutation* **33**, 81–85 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/21922594.

[146] Dogan, R. I., Getoor, L., Wilbur, W. J. & Mount, S. M. SplicePort–An interactive splice-site analysis tool. *Nucleic Acids Research* **35**, W285–W291 (2007). URL `http://www.ncbi.nlm.nih.gov/pubmed/17576680`.

[147] Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *Journal of molecular biology* **268**, 78–94 (1997). URL `http://www.ncbi.nlm.nih.gov/pubmed/9149143`.

[148] Yeo, G., Hoon, S., Venkatesh, B. & Burge, C. B. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 15700–15705 (2004). URL `http://www.pnas.org/cgi/doi/10.1073/pnas.0404901101`.

[149] Desmet, F.-O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic acids research* **37**, e67 (2009). URL `http://www.ncbi.nlm.nih.gov/pubmed/19339519`.

[150] Hebsgaard, S. M. *et al.* Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. *Nucleic acids research* **24**, 3439–52 (1996). URL `http://www.ncbi.nlm.nih.gov/pubmed/8811101`.

[151] Kapustin, Y. *et al.* Cryptic splice sites and split genes. *Nucleic Acids Research* **39**, 5837–5844 (2011). URL `http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkr203`.

[152] Adams, E. & Frank, L. Metabolism of Proline and the Hydroxyprolines. *Annual review of biochemistry* **49**, 1005–1061 (1980). URL `http://www.annualreviews.org/doi/pdf/10.1146/annurev.bi.49.070180.005041`.

[153] Baumgartner, M. R. Hyperammonemia with reduced ornithine, citrulline, arginine and proline: a new inborn error caused by a mutation in the gene encoding Delta1-pyrroline-5-carboxylate synthase. *Human Molecular Genetics* **9**, 2853–2858 (2000). URL `https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/9.19.2853`.

[154] Fischer-Zirnsak, B. *et al.* Recurrent De Novo Mutations Affecting Residue Arg138 of Pyrroline-5-Carboxylate Synthase Cause a Progeroid Form of Autosomal-Dominant Cutis Laxa. *American journal of human genetics* **97**, 483–492 (2015). URL `http://www.ncbi.nlm.nih.gov/pubmed/26320891`.

[155] Coutelier, M. *et al.* Alteration of ornithine metabolism leads to dominant and recessive hereditary spastic paraplegia. *Brain : a journal of neurology* **138**, 2191–2205 (2015). URL `http://www.ncbi.nlm.nih.gov/pubmed/26026163`.

[156] Altschul, S., Gish, W. & Miller, W. Basic local alignment search tool. *Journal of molecular Biology* **251**, 403–410 (1990). URL `http://www.sciencedirect.com/science/article/pii/S0022283605803602`.

[157] MacIsaac, K. D. *et al.* An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC bioinformatics* **7**, 113 (2006). URL `http://www.ncbi.nlm.nih.gov/pubmed/16522208`.

[158] Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics (Oxford, England)* **26**, 873–881 (2010). URL `http://www.ncbi.nlm.nih.gov/pubmed/20147302`.

[159] Anders, S., Pyl, P. T. & Huber, W. HTSeq - A Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)* **31**, 166–9 (2014). URL `http://www.ncbi.nlm.nih.gov/pubmed/25260700`.

[160] Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 9546–51 (2010). URL `http://www.pnas.org/content/107/21/9546.short`.

[161] 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* **57**, 289–300 (1995). URL `http://www.jstor.org/stable/2346101. 95/57289`.

[162] Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science (New York, N.Y.)* **324**, 218–23 (2009). URL `http://www.ncbi.nlm.nih.gov/pubmed/19213877`.

[163] Davison, A. C. & Hinkley, D. V. Bootstrap Methods and Their Application. *Engineering* **42**, 216 (1997). URL `http://www.jstor.org/stable/1271471?origin=crossref`.

[164] Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754–60 (2009). URL `http://www.ncbi.nlm.nih.gov/pubmed/19451168`.

[165] Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)* **27**, 2987–93 (2011). URL `http://www.ncbi.nlm.nih.gov/pubmed/21903627. 1203.6372`.

[166] Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**, 2078–9 (2009). URL `http://www.ncbi.nlm.nih.gov/pubmed/19505943. 1006.1266v2`.

[167] McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics (Oxford, England)* **26**, 2069–70 (2010). URL `http://www.ncbi.nlm.nih.gov/pubmed/20562413`.

[168] Eilbeck, K. *et al.* The Sequence Ontology: a tool for the unification of genome annotations. *Genome biology* **6**, R44 (2005). URL `http://genomebiology.biomedcentral.com/articles/10.1186/gb-2005-6-5-r44`.

[169] Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15–21 (2013). URL `http://www.ncbi.nlm.nih.gov/pubmed/23104886`.

[170] Hsu, F. *et al.* The UCSC Known Genes. *Bioinformatics (Oxford, England)* **22**, 1036–46 (2006). URL `http://www.ncbi.nlm.nih.gov/pubmed/16500937`.

[171] Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology* **9**, e1003118 (2013). URL `http://dx.plos.org/10.1371/journal.pcbi.1003118`.

[172] Hochberg, Y. A sharper bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–802 (1988). URL `http://biomet.oxfordjournals.org/cgi/doi/10.1093/biomet/75.4.800`.

[173] Li, Y. I., Knowles, D. A. & Pritchard, J. K. LeafCutter: Annotation-free quantification of RNA splicing. *bioRxiv* 044107 (2016). URL `http://biorxiv.org/lookup/doi/10.1101/044107`.

[174] Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research* **22**, 1760–1774 (2012). URL `http://genome.cshlp.org/cgi/doi/10.1101/gr.135350.111`.

[175] Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & cellular proteomics : MCP* **13**, 2513–26 (2014). URL `http://www.ncbi.nlm.nih.gov/pubmed/24942700`.

[176] Cheng, Z. *et al.* Differential dynamics of the mammalian mRNA and protein expression response to misfolding stress. *Molecular Systems Biology* **12**, 855–855 (2016). URL `http://msb.embopress.org/cgi/doi/10.15252/msb.20156423`.

[177] Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* **43**, e47 (2015). URL `http://www.ncbi.nlm.nih.gov/pubmed/25605792`.

[178] Van Haute, L. *et al.* Deficient methylation and formylation of mt-tRNA(Met) wobble cytosine in a patient carrying mutations in NSUN3. *Nature communications* **7**, 12039 (2016). URL `http://www.nature.com/doifinder/10.1038/ncomms12039`.

[179] Khan, Z. *et al.* Quantitative measurement of allele-specific protein expression in a diploid yeast hybrid by LC-MS. *Molecular systems biology* **8**, 602 (2012). URL `http://www.ncbi.nlm.nih.gov/pubmed/22893000`.

[180] Waddington, C. H. Canalization of Development and the Inheritance of Acquired Characters. *Nature* **150**, 563–565 (1942). URL `http://www.nature.com/doifinder/10.1038/150563a0`.

[181] Flatt, T. The evolutionary genetics of canalization. *The Quarterly review of biology* **80**, 287–316 (2005). URL http://www.ncbi.nlm.nih.gov/pubmed/16250465.

[182] Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47–52 (1999). URL http://www.ncbi.nlm.nih.gov/pubmed/10591225.

[183] Hartman, J. L., Garvik, B. & Hartwell, L. Principles for the buffering of genetic variation. *Science (New York, N.Y.)* **291**, 1001–4 (2001). URL http://www.ncbi.nlm.nih.gov/pubmed/11232561.

[184] Schrimpf, S. P. *et al.* Comparative Functional Analysis of the Caenorhabditis elegans and Drosophila melanogaster Proteomes. *PLoS Biology* **7**, e1000048 (2009). URL http://dx.plos.org/10.1371/journal.pbio.1000048.

[185] Laurent, J. M. *et al.* Protein abundances are more conserved than mRNA abundances across diverse taxa. *Proteomics* **10**, 4209–12 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/21089048.

[186] Khan, Z. *et al.* Primate Transcript and Protein Expression Levels Evolve Under Compensatory Selection Pressures. *Science* **342**, 1100–1104 (2013). URL http://www.sciencemag.org/cgi/doi/10.1126/science.1242379. NIHMS150003.

[187] Dahan, O., Gingold, H. & Pilpel, Y. Regulatory mechanisms and networks couple the different phases of gene expression. *Trends in genetics : TIG* **27**, 316–22 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21763027.

[188] Vogel, C. Evolution. Protein expression under pressure. *Science (New York, N.Y.)* **342**, 1052–3 (2013). URL http://www.sciencemag.org/cgi/doi/10.1126/science.1247833.

[189] Masel, J. & Siegal, M. L. Robustness: mechanisms and consequences. *Trends in Genetics* **25**, 395–403 (2009). URL http://linkinghub.elsevier.com/retrieve/pii/S0168952509001462.

[190] Fear, J. M. *et al.* Buffering of Genetic Regulatory Networks in Drosophila melanogaster. *Genetics* **203**, 1177–90 (2016). URL http://www.genetics.org/cgi/doi/10.1534/genetics.116.188797.

[191] Hou, J. *et al.* Extensive allele-specific translational regulation in hybrid mice. *Molecular Systems Biology* **11**, 825–825 (2015). URL http://msb.embopress.org/cgi/doi/10.15252/msb.156240.

[192] Sibley, C. R., Blazquez, L. & Ule, J. Lessons from non-canonical splicing. *Nature reviews. Genetics* **17**, 407–421 (2016). URL http://www.nature.com/doifinder/10.1038/nrg.2016.46.

[193] Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Science translational medicine* **9**, eaal5209 (2017). URL http://stm.sciencemag.org/lookup/doi/10.1126/scitranslmed.aal5209.

[194] Gibson, G. Human genetics. GTEx detects genetic effects. *Science (New York, N.Y.)* **348**, 640–641 (2015). URL `http://www.ncbi.nlm.nih.gov/pubmed/25953996`.

[195] Vafai, S. B. & Mootha, V. K. Mitochondrial disorders as windows into an ancient organelle. *Nature* **491**, 374–383 (2012). URL `http://www.nature.com/doifinder/10.1038/nature11707`.

[196] Gagneur, J. *et al.* Genotype-environment interactions reveal causal pathways that mediate genetic effects on phenotype. *PLoS genetics* **9**, e1003803 (2013). URL `http://dx.plos.org/10.1371/journal.pgen.1003803`.