

# Combinatorial Biological Complexity

## A study of amino acid side chains and alternative splicing

Dissertation

zur Erlangung des akademischen Grades

“doctor rerum naturalium” (Dr. rer. nat.)



---

seit 1558

vorgelegt dem Rat der Biologisch-Pharmazeutischen Fakultät  
der Friedrich-Schiller-Universität Jena

**von Konrad Grützmann**

**geboren am 22.07.1981 in Halle (Saale)**

---

Gutachter:

1. Prof. Dr. Stefan Schuster - Friedrich-Schiller Universität Jena
2. PD Dr. Kerstin Voigt - Hans-Knöll-Institut Jena, Friedrich-Schiller Universität Jena
3. Prof. Dr. Ivo Große - Martin-Luther-Universität Halle-Wittenberg

Datum der öffentlichen Verteidigung

1. April 2015

Die vorliegende Arbeit wurde am Lehrstuhl für Bioinformatik der biologisch-pharmazeutischen Fakultät der Friedrich Schiller Universität unter Leitung von Prof. Dr. Stefan Schuster angefertigt.

## Abstract

Both, laymen and experts have always been intrigued by nature's vast complexity and variety. Often, these phenomena arise from combination of parts, be they species in ecosystems, cell types of the human body, or the many proteins of a cell. In this thesis, I apply bioinformatic means to investigate combinatorial complexity in an exemplary manner in three different aspects: combinations of aliphatic amino acid side chains, alternative mRNA splicing in fungi, and mutually exclusively spliced exons in human and mouse.

The first part addresses the question of how many theoretically possible aliphatic amino acid side chains there are, and how many of them are realized in nature. Structural combinations yield a vast potential for amino acids, yet we find that only a fraction of them are realized. Reasons for this phenomenon are discussed, especially in the light of restrictions posed by the genetic code. Moreover, strategies for the need for increased diversity are examined.

In the second part, the extent and prevalence of alternative splicing (AS) in the fungal kingdom is investigated. To this end, a genome-wide and comparative multi-species study is conducted. The emphasis lies on normalization of the rates of AS to allow for a comparability of the species. AS is found to be a common process in fungi, but with lower frequency compared to plants and animals. Increasing AS rates are found for more complex fungi. Also, fungal AS is involved in the intricate process of virulence. The results hint at the contribution of AS to multi-cellular complexity in fungi.

In the third part, mutually exclusive exon (MXE) splicing is addressed. MXE is an interdependent type of AS, which has not been so well researched in transcriptome-wide respect so far. Here, MXEs of mouse and human are detected and characterized. Rather unexpected patterns are found: the majority of MXEs originate from non-adjacent exons and frequently appear in clusters. These properties make most of the found regulatory mechanisms of MXEs of other species unsuitable. New mechanisms have to be sought for mammals.

Summarizing, several instances in which combinations contribute to biological complexity are investigated in this thesis. It is hypothesized that complexity from combinations constitutes a rather universal principle in biology. However, there seems to be a need to restrict the combinatorial potential. This is highlighted, for example, in the interdependence of MXEs and the low number of realized amino acids in the genetic code. Combinatorial complexity and its restriction are discussed with respect to other biological examples to further substantiate the hypothesis.

## Zusammenfassung

Laien und Experten sind seit jeher beeindruckt von der Komplexität und Vielfalt der Natur. Diese Phänomene entstehen oft durch Kombination von “Teilen”, seien es Spezies in Ökosystemen, Zelltypen des menschlichen Körpers oder die zahlreichen Proteine einer Zelle. In dieser Dissertation wird kombinatorische Komplexität mittels bioinformatischer Methoden exemplarisch hinsichtlich dreier Aspekten untersucht: Kombination aliphatischer Aminosäureseitenketten, alternatives Spleißen in Pilzen und sich gegenseitig ausschließend gespleißte Exons in Mensch und Maus.

Im ersten Teil wird der Frage nachgegangen, wie viele theoretisch mögliche aliphatische Aminosäureseitenketten es gibt und wie viele davon in der Natur realisiert werden. Kombination in der Seitenkettenstruktur eröffnet ein großes Potential für Aminosäuren. Trotzdem findet nur ein Bruchteil der möglichen Aminosäuren in der Natur Verwendung. Gründe für dieses Phänomen werden diskutiert, besonders hinsichtlich Einschränkungen durch den genetischen Code. Darüber hinaus werden Strategien der Natur diskutiert, um trotzdem einen erhöhten Bedarf an Vielfalt zu decken.

Im zweiten Teil wird das Ausmaß des alternativen Spleißens (AS) im Pilzreich untersucht. Dazu wurde eine genomweite, vergleichende Studie mit mehreren Spezies durchgeführt. Ein Schwerpunkt war die Normalisierung der Raten des AS, so dass Spezies vergleichbar sind. Es wurde herausgefunden, dass AS ein im Pilzreich üblicher Prozess ist, der jedoch im Vergleich zu Pflanzen und Tieren seltener auftritt. Komplexere Pilze weisen erhöhte Raten von AS auf. Außerdem ist AS im komplexen Prozess der Virulenz mancher Pilze involviert. Die Ergebnisse deuten auf einen Beitrag von AS zur multizellulären Komplexität in Pilzen hin.

Im dritten Teil werden sich gegenseitig ausschließend gespleißte Exons (MXEs, englisch *mutually exclusive exons*) untersucht. MXEs sind voneinander abhängige Spleißereignisse, die bisher wenig transkriptomweit untersucht wurden. Im Rahmen dieser Arbeit wurden MXEs in Mensch und Maus detektiert und charakterisiert. Dabei wurden ungewöhnliche Muster entdeckt: die meisten MXEs stammen von nicht-benachbarten Exons und treten häufig in Gruppen auf. Diese Eigenschaften stellen

die Eignung bisher für andere Tiere bekannter Regulationsmechanismen von MXEs in Frage. Neue Mechanismen müssen für Säugetiere gefunden werden.

In dieser Arbeit werden mehrere Aspekte untersucht, bei denen Kombination zu biologischer Komplexität beiträgt. Es wird die Hypothese aufgestellt, dass Komplexität durch Kombination ein eher universelles Prinzip der Natur darstellt. Dennoch scheint die Beschränkung des kombinatorischen Potentials nötig zu sein. Das wird anhand der gegenseitigen Abhängigkeit von MXEs und der kleinen Zahl tatsächlich verwendeter Aminosäuren im genetischen Code herausgestellt. Kombinatorische Komplexität und deren Begrenzung wird an weiteren biologischen Beispielen diskutiert, um die Hypothese zu bekräftigen.

## Danksagung

Mein besonderer Dank gilt meinem Betreuer Stefan Schuster, der mir ermöglichte, an seinem Lehrstuhl in so einer produktiven und angenehmen Gruppe zu arbeiten. Ihm verdanke ich spannende Forschungsthemen und Nebenthemen, das Ankurbeln wertvoller Kooperationen und das Einrühren sportlicher Aktivitäten voller Diskussion und Erholung.

Ich danke sehr Kerstin Voigt für den Kontakt zum Reich der Pilze, für anregende Diskussionen und Hilfe in Forschungs- und Publikationsfragen.

Großer Dank geht an Matthias Platzer, der mit wertvollen Diskussionen zum Gelingen meiner Arbeit beitrug.

Ich danke sehr Karol Szafranski, der mit Vehemenz an das Gelingen meiner wohl schwersten Publikation geglaubt hat und mir mit unermüdlichem Einsatz und wissenschaftlichem Gespür zum Ziel verhalf.

Weiterer Dank geht an Günter Theißen, der enorm zum Florieren der oben genannten sportlichen Aktivitäten beitrug und durch dessen Ideen ich auf kombinatorische Komplexität als Thema für meine Dissertation kam.

Ich danke sehr Ivo Große für die Einladung zum damaligen Vortrag mit anschließender, sehr anregender Diskussion und für das Begutachten meiner Dissertation.

Ganz besonderer Dank gilt meiner Frau Julia und meinem Sohn Luca. Ohne Euch wäre die Sache wohl nie so gut gelungen. Danke für die erholsame Ablenkung, die Gespräche und nicht zuletzt die Freiräume, die manchmal für die Forschung nötig sind.

Ich danke auch meinen Eltern, meinem Bruder und all meinen Verwandten für die Unterstützung und die schöne Zeit.

Nicht zuletzt möchte ich natürlich auch meinen ehemaligen Mitarbeitern des Lehrstuhls danken, mit denen die Zusammenarbeit immer eine große Freude war.

---



# Contents

<b>Abstract</b>	<b>iii</b>
<b>Zusammenfassung</b>	<b>v</b>
<b>Danksagung</b>	<b>vii</b>
<b>Glossary</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Organismic Complexity . . . . .	1
1.2 Amino Acids . . . . .	4
1.3 The Genetic Code . . . . .	5
1.4 Splicing . . . . .	6
1.5 Alternative Splicing . . . . .	6
1.6 The Fungal Kingdom . . . . .	9
<b>2 Publications</b>	<b>13</b>
2.1 Combinatorics of aliphatic amino acids . . . . .	14
2.2 The alternative messages of fungal genomes . . . . .	23
2.3 Fungal alternative splicing is associated with multicellular complexity and virulence – A genome-wide multi-species study . . . . .	28
2.4 Mutually exclusive spliced exons show non-adjacent and grouped patterns . . . . .	70
2.5 Alternative splicing of mutually exclusive exons – A review . . . . .	77
<b>3 General Discussion</b>	<b>87</b>
3.1 Amino acid side chain combinations restrained to a fraction of their potential . . . . .	87
3.2 Alternative splicing in fungi . . . . .	91
3.3 Mutually exclusive exons - A interdependent type of combinatorial splicing . . . . .	97
3.4 Combinatorial complexity - final discussion . . . . .	101

## CONTENTS

---

3.4.1	Complexity through combination - a universal principle . . . . .	101
3.4.2	Combinatorial complexity beyond the genetic code . . . . .	102
3.4.3	Combinatorial complexity in alternative splicing . . . . .	104
3.4.4	Costs of the combinatorial principle . . . . .	106
3.4.5	Restriction of combinations . . . . .	107
<b>References</b>		<b>109</b>
<b>Beitrag der Autoren</b>		<b>117</b>
<b>Lebenslauf</b>		<b>118</b>
<b>Erklärung</b>		<b>121</b>

# Glossary

<b>A3'SS</b>	alternative 3' intron splice site	<b>MXE</b>	mutually exclusive exon
<b>A5'SS</b>	alternative 5' intron splice site	<b>NGS</b>	next generation sequencing
<b>AA</b>	amino acid	<b>NMD</b>	nonsense-mediated mRNA decay
<b>ABU</b>	$\alpha$ -aminobutyric acid	<b>PTC</b>	premature termination codon
<b>AS</b>	alternative splicing	<b>PTM</b>	post-translational modification
<b>CE</b>	cassette exon	<b>Pyl</b>	pyrrolysine
<b>EST</b>	expressed sequence tag	<b>R</b>	purine
<b>GMX</b>	glucuronoxylomannan	<b>RI</b>	retained intron
<b>hnRNP</b>	heterogeneous nuclear ribonucleoprotein	<b>SE</b>	skipped exon
<b>IR</b>	intron retention	<b>Sel</b>	selenocysteine
<b>mRNA</b>	messenger ribonucleic acid	<b>snRNP</b>	small nuclear ribonucleoprotein
		<b>SR protein</b>	serine-arginine rich protein
		<b>SS</b>	splice site
		<b>SUMO protein</b>	Small Ubiquitin-like Modifier protein
		<b>TF</b>	transcription factor
		<b>tRNA</b>	transfer ribonucleic acid
		<b>Y</b>	pyrimidine

## GLOSSARY

---

# 1

## Introduction

### 1.1 Organismic Complexity

The unraveling of the human genome in around the year 2001 was a surprise for the biosciences<sup>1</sup>. While early estimations stated that there are about 100,000 human genes, this number was constantly refined to lower values during the successive sequencing and analysis of the human genome. Current estimates claim around 21,000 protein coding genes<sup>2</sup>. How can such a complex system as the human be built from only 21,000 genes? And not less intriguing, why do seemingly less complex species as the worm *Caenorhabditis elegans* (21,733 genes<sup>3</sup>) or wheat (> 94,000 genes<sup>4</sup>) harbor a similar number of genes or even more? Neither the haploid DNA mass (C value paradox<sup>5</sup>) nor the number of genes<sup>6</sup> correlate with what is perceived as biologic complexity. There are many concepts of what complexity constitutes and how it is achieved. The main theme of this thesis is how biological complexity is attained, and which role combinations play. The central dogma of molecular biology is quite simple, and thus, appealing in the quest for understanding biology: information is transferred from linear DNA to RNA to proteins and does not flow back<sup>7</sup>. The dogma deals with the basic steps of gene expression: DNA transcription into messenger RNA (mRNA), and mRNA translation into proteins. With a more refined view, information does not only flow in one direction during gene expression. It rather interferes and can also be transferred back. RNA and proteins determine which DNA is transcribed and which mRNA is translated. Further, RNA can be recoded to DNA and incorporated into the genome sequence by reverse transcription<sup>8</sup>. Thus, even when neglecting all intricate intermediate steps of gene expression, life is a highly dynamic system. On the one hand, organisms can proceed from one developmental state to another, accompanied by great changes in morphology and cell composition. On the other hand, once being in such a state, healthy organisms reside in an equilibrium (e.g., maintaining cell identity, metabolite concentrations, gene expression levels),

## 1. INTRODUCTION

---

which necessitates a dynamical adaptation to their changing environment<sup>9,10</sup> and the ability to regenerate<sup>11</sup>.

In its pure form, the central dogma stated a one to one relation, a mapping, of genetic information. Again, in detail it is more complicated. There are bifurcations on many stages of gene expression. Alternative transcription start sites yield different primary transcripts from one gene<sup>12</sup>. Alternative splicing and RNA editing give rise to multiple mature isoforms of eukaryotic mRNA. Proteins are diversified by post-translational modifications<sup>8</sup>. Hence, counting the raw number of a species' genes is not even a rough guide to assess its complexity<sup>13</sup>. Yet another loosening of the central dogma is the discovery that a huge part of the transcribed RNA of eukaryotes is not translated into proteins<sup>2</sup>. A cellular function has been shown for a remarkable part of these non-protein-coding RNAs, and thus, puts the theory of the junk-harboring genome into perspective<sup>14</sup>. Furthermore, from a more sophisticated view, genetic information is not only stored in a linear series of symbols. Genomic DNA is organized in a ordered secondary structure, which taps potential to store additional information<sup>15</sup>. Additionally, there is modification of histones, the "packaging material" of chromosomes. Variable acetylation and methylation of histone amino acids are involved in silencing and activation of gene expression<sup>16</sup>. In this respect, the notion of epigenetic code is introduced by which further genetic information can be stored and altered on a short-term basis, as opposed to the slowly evolving sequences of genomic DNA<sup>17</sup>.

In order to grasp complex phenomena as species and living matter, a multi-disciplinary approach with different perspectives and methodologies is beneficial. From a computer science perspective, species can be viewed as information processing entities. Thus, a species' complexity can be measured by assessing the information it bears. There are two well-known measures of complexity in computer science: Shannon entropy and Kolmogorov complexity. The first one is understood as the average uncertainty about a sequence of letters, and only involves probabilities of the letters<sup>18</sup>. Thus, Shannon entropy disregards the sequential order of letters or the structure of a text. However, Shannon theory was often successfully applied in bioinformatics, for example, for quantification of methylation differences<sup>19</sup>, and for automatic detection of a certain type of cardiac arrhythmia<sup>20</sup>. Another measure derived from Shannon entropy is mutual information, which quantifies the information that two objects, as DNA sequences, share<sup>21</sup>. As examples, it was used to analyze nucleosome positioning motifs<sup>22</sup>, and to predict coding regions in DNA independent of organism-specific features<sup>23</sup>. The Kolmogorov complexity or algorithmic complexity from algorithmic information theory can be interpreted as the length of the shortest program to describe a sequence of letters<sup>24</sup>. With respect to Kolmogorov complexity, many repetitive and low-complexity parts<sup>25</sup> make the human genome very compressible. This means, it could store more information as it actually does. The third measure of complexity I want

to mention, Trifonov's linguistic complexity, allows for a closer appreciation of the genome's informational structure. It measures the fraction of different substrings of a sequence to the number of possible substrings<sup>26</sup>. Using it, one can detect that several codes are superimposed on the genome, e.g. the genetic code and information for nucleosome positioning<sup>27</sup>. Calculations of complexity have helped to understand many biological aspects, e.g., to identify disease genes in genetic association studies<sup>28</sup>, to quantify splicing disorder in cancer<sup>29</sup>, and to quantify nematodes' morphological aging based on image texture entropy<sup>30</sup>.

To recapitulate the observation from the beginning, organismic complexity is not a mere question of how many and which genes an organism has. It rather comprises the complex interplay and the spatio-temporal combinatorial expression of subsets of the whole gene inventory. In a similar manner, the beauty and worth of a building can not simply be assessed from the shape and number of its bricks. A better way to numerically grasp complexity is to account for the number of macro-level occurrences assembled from the micro-level parts. Thus, many scientists rather ascribe importance to the number of proteins, tissue types, and developmental stages a species can produce, than to its gene number<sup>13,31</sup>. One common observation in nature is the use of combinations to achieve variety. The central principle leveraged by nature is that a vast number of different compositions can be produced from only a small set of components. This multiplication is also known in subjects besides biology and called combinatorial explosion<sup>32</sup>.

In this thesis, biological complexity is investigated from the view point of combinatorial diversity. As opposed to the approach in systems biology, I rather do not investigate complexity from interrelations in biological systems. In the core of this thesis, complexity by combinatorial diversity is studied on two levels. First, the combinatorics of aliphatic amino acid side chains is investigated. The theoretical potential of an amino acid inventory is contrasted by the actual realization in nature. Theories for this discrepancy are discussed, as well as limitations and consequences for the genetic code. In the second part, two aspects of AS are investigated. The scope and extent of AS in the fungal kingdom is studied, and potential involvement in virulence and multi-cellular complexity are investigated and discussed. Finally, human and mouse mutually exclusive splicing events are mined and characterized using bioinformatics methods. Consequences for earlier proposed regulatory mechanisms are discussed in light of the unexpected findings.

Throughout this thesis I will use the word "we" when referring to the work involved in producing the publications embedded in the thesis, as these arose by collaborative work. I will use "I" when referring to content that solely belongs to this thesis.

## 1. INTRODUCTION

---

### 1.2 Amino Acids

Amino acids (AAs) are organic compounds of living cells and constitute the building blocks of proteins. They are connected in linear fashion by peptide bonds. In aqueous solution as the cytosol, an AA chain folds to form a protein with spatial structure and physico-chemical properties. These allow proteins to execute their manifold functions<sup>8</sup>. The first AAs asparagine and cysteine were discovered in the beginning of the 19th century<sup>33,34</sup>. Still, it took another century until proteins were proposed to be composed of AAs<sup>35</sup>. An AA consists of a central carbon atom with the following four binding partners: a hydrogen, a carboxyl group (COOH or COO<sup>-</sup>), an amino group (NH<sub>2</sub> or NH<sub>3</sub><sup>+</sup>) and a variable side chain. The side chains have a high potential variability. Mostly, they consist of branched, unbranched or cyclic carbon atom structures, often times supplemented with nitrogen, oxygen or sulfur, and saturated with hydrogens<sup>36</sup>. The proteinogenic AAs can be partitioned according to their physico-chemical properties into polar, non-polar, and charged<sup>37</sup>. Charged and polar AAs are hydrophilic, non-polar AAs are hydrophobic<sup>8</sup>. Charged amino acids can further be subdivided into electrical positively (alkaline) or negatively (acidic) charged<sup>8</sup>. Hydrophobic AAs are subdivided into aliphatic and aromatic<sup>8</sup>. In my first publication, aliphatic amino acids are investigated. These are hydrophobic, contain carbon and hydrogen and have no aromatic rings<sup>36</sup>. Non-aromatic rings and multiple bonds are excluded for simplification of the calculation, though both would not conflict the definition of being aliphatic. Furthermore, sulphur, nitrogen and oxygen are allowed in a less strict definition of aliphatic AAs, as long as they retain a certain hydrophobicity.

During the formation of the covalent peptide bond, also called condensation reaction, the amino group of one AA is bound to the carboxyl group of another AA, thereby releasing water. Once a polypeptide chain of a certain length has formed it can fold to a spatial structure<sup>36</sup>. During the thought experiment of Levinthal's paradox, a cell sequentially samples all possible combinations of the peptide bond angels to find the correct protein structure. Even if these combinations could be sampled at a time-scale of 10<sup>-13</sup> seconds, protein folding would take longer than the existence of the universe because of the inconceivable number of angel combinations<sup>36,38</sup>. In reality, folding is guided by the properties of the contained AAs, which cause attracting or repelling interactions. In aqueous solutions as the cytosol, hydrophobic AAs strive to be positioned in the center of a protein, whereas hydrophilic AAs frequently occur on a protein's surface<sup>36</sup>. These forces make sampling of the whole possible fold space unnecessary. Also, polypeptide chains are thought to form small modules first, which are partially correctly folded transition states. These substructures then funnel the global protein structure formation, so that folding takes only a few minutes<sup>36</sup>.

Once proteins attain their natural conformation, they can carry out their cellular function. They



dynamically bind other molecules to form cell structures or catalyze biochemical reactions<sup>8</sup>.

There are 20 canonical AAs found in proteins of biological organisms. The set was extended by the finding of selenocysteine, the 21st<sup>39,40</sup>, and pyrrolysine, the 22nd<sup>41</sup>, non-canonical AAs. They are not directly represented in the genetic code, but are incorporated by the repurpose of stop codons in the presence of certain RNA structures<sup>42</sup>. Besides these proteinogenic AAs, there are many others that are not incorporated into proteins. They are, for example, secondary metabolites or precursors of these, e.g., 2-amino butanoic acid for chemical communication in *Globodera rostochiensis*<sup>43</sup>, or L-norvaline that is an anti-inflammatory arginase inhibitor in human<sup>44</sup>. Others are metabolic intermediates, as e.g. citrulline and ornithine from the urea cycle<sup>36</sup>. There is a vast diversity of non-proteinogenic AAs and new ones are discovered frequently, as, e.g., 2-amino-9,13-dimethyl-heptadecanoic acid secondary metabolite from *Streptomyces* sp. 1010<sup>45</sup>.

### 1.3 The Genetic Code

During translation of mRNA into proteins, a ribosome sequentially reads the codons (nucleotide triplets) of the messenger transcript with the help of transfer RNA (tRNA). The codon sequence is deciphered into an AA sequence, which will form a new protein. The actual code for translation is implemented in the tRNA where the anti-codon that matches a specific mRNA codon is linked to the corresponding AA. These linkage reactions are conducted by specific aminoacyl tRNA synthetases<sup>8</sup>. During the 1960's, codons were first shown to consist of three nucleotides<sup>46</sup>, and subsequently, the code assignments were deciphered by several scientists<sup>47</sup>. There are 61 codons coding for 20 AAs. The code is called degenerate since for some AAs more than one codon exists. However, it is not ambiguous, meaning that no codon can be translated into different AAs. The degeneracy makes the code more stable towards mutations. Consider, e.g., a so-called fourfold degenerate codon site, where four codons that differ only in one position code for the same AA. If such a codon mutates at the said position, the encoded AA stays the same. This is called a synonymous mutation. The methionine codon AUG is also interpreted as translational start when not positioned within a coding sequence. Furthermore, there are three codons (UAA, UAG, UGA) interpreted as translational stop<sup>8</sup>.

There is a multitude of theoretically possible assignments of AAs to codons. However, the canonical genetic code can be found in all species with only small variations<sup>48</sup>. An example are vertebrate mitochondria, where e.g. AGA and AGG are additional stop codons<sup>49</sup>. The high similarity of genetic codes hint at a common origin. At the same time, it shows that such a central part of the gene expression machinery is evolving very slowly, if at all<sup>36</sup>. An obvious reason is that a small deviation in the code would have a huge effect on the expression

## 1. INTRODUCTION

---

of all proteins of an organism. This can hardly be tolerated. It will be discussed later that the expansion of the AA repertoire is impeded for similar reasons.

### 1.4 Splicing

Splicing is a process during expression of eukaryotic genes taking place before translation. The pre-mRNA, resulting from transcription of DNA by RNA-polymerase, contains sequence parts called exons interspersed with so called introns. The introns are removed during splicing and the remaining exons are ligated together to form the mature mRNA. Splicing is carried out by a large complex of ribonucleoproteins in the cell nucleus called the spliceosome. The five small nuclear ribonucleoproteins (snRNPs) U1, U2, U4, U5, and U6 assemble in step-wise fashion onto the intron by recognizing exonic and intronic sequence motifs, and catalyze the splice reactions. The assembly of the spliceosome needs several other factors, such as U2AF and SF1<sup>25</sup>. In sum, several hundred molecules take part in coordinating one splicing reaction<sup>50</sup>, and it is regarded one of “the most complex macromolecular machine in the cell”<sup>51</sup>.

The recognition motifs in introns are the 5' end GU (donor) and the 3' end AG (acceptor) dinucleotides, the polypyrimidine upstream of the 3' end, and the branch site containing an adenine<sup>25</sup>. These signals are supposed to contribute around 50% of the information needed for proper intron recognition<sup>52</sup>. These motifs have surrounding sequences that harbor further information, some reach into exonic regions and are conserved. Additionally, there are other *cis*-acting (i.e. on the mRNA) splice motifs. These are called intronic and exonic splice enhancers and silencers, depending on their position and kind of splice regulation<sup>25</sup>. Increasing evidence shows that splicing takes part co-transcriptionally at the nascent mRNA parts<sup>53</sup>. Also, there is evidence that several spliceosomes can assemble onto one mRNA at the same time<sup>54</sup>.

Two sequential transesterification reactions take place during splicing. In the first one, the 5' end G is attached to the branch site A nucleotide. In the second reaction, the 3' intron end is excised and the exons are joined. The branched, loop-shaped intron structure that is formed is called lariat and degraded afterwards<sup>25</sup>.

Besides the above described splicing pathway conducted by the so called major spliceosome there is another very similar pathway. The minor spliceosome excises introns that are flanked by other consensus dinucleotides. It shares the U5 subunit, but differs in the other ones<sup>55</sup>.

### 1.5 Alternative Splicing

Alternative splicing (AS) was first discovered in adenoviruses in 1977<sup>56</sup>. Several years later it was found in cancer cell lines of a self-replicating organism<sup>57</sup>. AS was considered rather an exception

for long time. With the rise of new technologies as whole genome sequencing, Sanger sequencing of ESTs, and microarrays in the 1990s, and even more with the high-throughput sequencing, the predicted phylogenetic scope and individual extent of AS seems to be ever increasing<sup>58,59,60</sup>. During AS, different acceptor and/or donor splice sites (SSs) can be chosen by the spliceosome, and alternative mRNA isoforms are produced. This may lead to alternative protein isoforms, degradation of spurious transcripts<sup>52</sup>, or regulatory effects of alternative untranslated regions<sup>61</sup>. There are four basic types of AS: cassette exons (CE, or skipped exons), intron retentions (IRs), alternative 5' intron ends (A5'SSs) or donor sites, and alternative 3' intron ends (A3'SSs) or acceptor sites<sup>52</sup>. Especially in species with many introns, these basic events can affect a gene's transcript in combined fashion and SS choice may be dependent. For example, two exons may be skipped together<sup>62</sup>. Furthermore, there are more complex types of AS, for example mutually exclusive exons (MXEs), in which in its basic form exactly one exon out of two appears in mature mRNA<sup>63</sup>.

There are several factors involved in AS. Most important, a SS can be less conserved, meaning that the sequence surrounding the dinucleotide consensus deviates from the one found in constitutively spliced introns. Thus, the spliceosome can bind less strongly to it. Also, there has to be a similar strong nearby SS to which the spliceosome can bind alternatively<sup>64</sup>. Other factors are exonic and intronic splice enhancers and silencers present in the mRNA<sup>25</sup>. These *cis*-acting factors are bound by *trans*-acting factors, proteins and ribonucleoproteins, which influence spliceosome assembly in a concerted manner. One class of *trans*-factors are serine-arginine rich (SR) proteins that are generally assumed to enhance splicing<sup>65</sup>. Another class are heterogeneous nuclear ribonucleoproteins (hnRNPs) that are rather splice repressors<sup>65</sup>. Interestingly, a splicing factor can have enhancing or repressing effects depending on its distance to the SS and the context<sup>65,66</sup>.

The co-transcriptional occurrence of splicing<sup>53</sup> opens the potential of regulatory coupling of both processes. Several ways of coupling have been found. For example, splicing factors can be recruited to the transcription site with help of the RNA polymerase II<sup>67</sup>. Also, transcription factor presence can determine alternative SS usage<sup>68</sup>. For example, different steroid hormones make hormone receptors recruit different promoter coregulators. Evidence was found that these not only affect transcription, but also influence the production of AS variants of the target gene<sup>69</sup>.

Furthermore, it was demonstrated that transcription speed can influence AS. A spliceosome has more time to assemble on an upstream SS when elongation is slow, and the downstream SS will be chosen less preferentially<sup>70</sup>. Nucleosomes impede RNA-polymerase II elongation. Thus, in turn, differential remodelling of chromatin, that is tightening and loosening of nucleosomes, can influence AS outcomes. Post-translational modifications can set histone marks in chromatin,

## 1. INTRODUCTION

---

which help recruit splicing factors and spliceosome components, and thus, enhance splicing<sup>68</sup>. There were correlations found between absence and presence of specific histone marks and certain AS outcomes<sup>71</sup>. Intriguingly, the so-called Hu proteins, which also affect chromatin structure, affect splicing in yet another way. They compete for an intronic splice enhancer that is normally bound by TIA-1/TIAR proteins, and thus, cause skipping of an exon<sup>72</sup>. In sum, AS is a intricate process and tightly embedded in other eukaryotic cell activities.

AS takes place in eukaryotes. As a trend, more genes are associated with AS in more highly developed species, and vertebrates show more AS events than invertebrates<sup>73</sup>. Mammals are considered to have the highest AS rates, and human presumably reach the top with 74%<sup>74</sup> to 94%<sup>75,76</sup> of affected genes. It is generally assumed, that plants show less AS events than animals. For example, in the monocot rice and dicot *Arabidopsis thaliana* about 16% and 18% of the TIGR annotated genes are AS associated, respectively<sup>77</sup>. The eukaryote kingdom fungi, is less well investigated with respect to AS. While yeasts show nearly no AS events<sup>78,79</sup>, other filamentous fungi with more complex lifestyles have higher rates, for example, *Aspergillus oryzae* 8.6%<sup>80</sup>, *Cryptococcus neoformans* 4.2%<sup>81</sup>, and *Coccidioides posadasii* with around 1000 AS events<sup>79</sup>.

AS can have different consequences for the encoded proteins. It can influence intracellular localization, affinity of binding sites and catalytic activity of enzymes and others<sup>65</sup>. AS takes effect on the development and function of whole tissues and organs. For instance, there are many studies on the role of AS during neuronal development. AS may provide the diversity necessary for differentiation into such an intricate structure as the mammalian brain<sup>82</sup>. Another example is the sex determination of *Drosophila melanogaster*, which involves a cascade of splicing factors (Sxl, Tra). They lead to splicing of the transcription factor Dsx into sex specific isoforms, which ultimately determine sex on molecular level via repression and activation of sex specific genes<sup>83</sup>. The third example of AS affecting physiology is about the tropomyosin family that is comprised of four genes that are alternatively spliced into 15-20 isoforms with specific properties. As example, exon 6A and 6B are mutually exclusive and are expressed specifically in either smooth or skeletal muscle cells<sup>84</sup>. Interestingly, it is not always the presence of one or the other splice variant that determines the cellular effect, but sometimes the exact ratio of the variants, as can be seen for isoforms of the murine membrane protein Prominin-1<sup>85</sup>.

There is an ongoing debate what fraction of the known AS events are functionally relevant for cells. Most of the A5'SSs and A3'SSs are only three nucleotides long, and thus, are called tandem acceptors and donors<sup>86</sup>. They were proposed to be 'noise' of the unreliably binding spliceosome<sup>87</sup>. However, other scientists speculate that these events may 'fine-tune' the proteome<sup>88</sup>. A more recent study shows, that tandem acceptors are regulated in tissue-specific way, and thus, likely have a function<sup>89</sup>.

Only 66% of alternatively skipped exons conserved between human and mouse keep the reading

frame, as their length is a multiple of three, the codon length<sup>52</sup>. Frame-shifting AS events unlikely produces sensible coding sequences, because it is very unlikely to have two overlapping protein-coding reading frames. On top, when considering frame-shifted sequences as random sequences, the chance of stop codons is very high (three out of 64 possible codons). Hence, the resulting proteins are truncated because they contain premature termination codons (PTCs). In order to prevent these spurious transcripts to cause harm to the cell, many species have the nonsense-mediated mRNA decay (NMD) pathway to degrade them. Research shows that NMD is a widely spread mechanism in the three eukaryotic crown groups<sup>90,91,92</sup>. While it is generally accepted that NMD is for removal of aberrant transcripts<sup>93</sup>, evidence is accumulating that NMD and AS may act in conjunction to control gene expression<sup>94</sup>. Simply put, gene expression can be switched off by splicing the mRNA into an alternative isoform that has a PTC, and thus, gets degraded via NMD. Hence, splice factors can control the fate of mRNA. As splice factor expression and recruitment are regulated by the cell, intricate gene expression patterns can be produced via this pathway. Thus, coupling of AS and NMD can be understood as another layer of gene expression control, taking place between transcription and translation<sup>52</sup>, allowing for additional regulatory complexity. However, it is currently not known to which extent frame-shifting AS events are actually used for gene expression regulation via this coupling.

AS is part of the explanation for the vast amount of more than one million different human proteins that are expressed from only about 21,000 genes. The potential of AS lies in joining the existing gene material to new compositions, thereby enhancing combinatorial complexity. The potential is corroborated by the fact that 94% of the human genes are multi-exonic, and the average human gene has 7.8 introns<sup>95</sup>. Indeed, it was proposed that multiexon genes are affected by at least seven AS events on average<sup>76</sup>.

## 1.6 The Fungal Kingdom

Fungi are one of the five kingdoms<sup>1</sup> of the eukaryotes, next to Protozoa, Animalia, Plantae and Chromista<sup>100</sup>. Fungi form a phylogenetic sister group to the multicellular animals, from which they have diverged around one billion years ago<sup>101</sup>. Fungi occupy many ecological niches and spread across diverse habitats around the globe. Their appearance reaches from being single cellular micro-organisms of a few micrometers in diameter (e.g. the yeast *Saccharomyces cerevisiae*) to mushrooms having filaments that reach an extent of hundreds of hectares<sup>102,103</sup>. It is estimated that around 5.1 million fungal species exist on earth based on high-throughput sequencing experiments<sup>104</sup>. Fungi take a crucial role in the global carbon cycle. They decompose

---

<sup>1</sup>Note, there are other classifications that divide eukaryotes into, e.g., four<sup>96</sup> or six<sup>97,98</sup> groups. The correct division is under debate and no single consensus has been found<sup>99</sup>.

## 1. INTRODUCTION

---

extant material of bacteria, plants and animals. The resulting compounds are taken into the cycle again<sup>102</sup>.

Fungi live on wood, in soil, water, on plant and animal debris, on dead organisms or make diverse forms of symbioses with other species. Many fungi spread their spores via air. They are not capable of doing photosynthesis, and hence, are heterotrophic. Fungi do not ingest food as animals do, but get nutrients only by absorption. They are usually saprotrophs, i.e., they digest organic material extracellularly and import the resulting compounds. Their diverse appearances and versatility have brought about some bizarre behavior, as e.g., a fungus that traps roundworms *Arthrobotrys oligospora*<sup>102</sup>.

The fungal kingdom can be subclassified into six groups (*phyla*): *Chytridiomycota*, *Neocallimastigomycota*, *Blastocladiomycota*, *Glomeromycota*, *Ascomycota*, and *Basidiomycota*<sup>105</sup>. The *Microsporidia*, which Hibbett et al.<sup>105</sup> still classify as fungi, are now considered a sister group of fungi<sup>106</sup>. *Ascomycota* and *Basidiomycota* form the subkingdom *Dikarya*, and are often called higher fungi. Further, there are five subphyla *incertae sedis* with uncertain taxonomic placement<sup>105,107</sup>. *Rhizopus oryzae* belongs to one such taxon, the *Mucoromycotina*<sup>105</sup>.

Fungi mostly possess hyphae (filaments) or are closely related to hyphal species. Their cell walls harbor chitin but no cellulose during most part of their life<sup>108</sup>. Most fungi are known to have a sexual phase. During this, they produce spores, which are useful for a fungus's classification because they have distinctive shapes. *Ascomycota* (more informally Ascomycetes) and *Basidiomycota* (informally Basidiomycetes) have cross-walls (septa) that divide the hyphae into compartments. Basidiomycetes account for many of the well known mushrooms, toadstools, rusts and smuts. Some higher fungi lack a sexual phase but build asexual spores. They are called mitosporic (from mitoses) fungi<sup>102</sup>.

Besides the predominant hyphal fungi there are yeasts, which are usually unicellular. These are, for example, the baker's or brewer's yeast *Saccharomyces cerevisiae* and the fission yeast *Schizosaccharomyces pombe*. Some of the normally multicellular growing fungi are capable to switch to unicellular growth and *vice versa*. This is called dimorphism or dimorphic switching and is influenced by the environmental conditions of a fungus<sup>102</sup>. The dimorphic switch allows pathogenic fungi to adapt to the micro environment during invasion of the host tissue. Thus, pathogens can make use of changing nutrients supply and cope with the host defense<sup>109</sup>.

One wide-spread form of fungal symbiosis is mutualism where both species benefit from their alliance. Intriguingly, more than 90% of plants have fungi associated with their roots (mycorrhizae). They help the plants to take up nutrients from the soil. Another interesting form of mutualism are lichens, symbioses of fungi with algae or cyanobacteria. They live in humid areas afar from human civilization and cover 8% of the earth's surface<sup>102</sup>.

Fungal pathogens are responsible for parasitic symbioses causing great negative effects on mankind.

On the one hand, there are plant pathogens causing billions of crop loss in the USA alone<sup>110</sup>. On the other hand, there are human pathogens responsible for diverse, wide-spread diseases, some of which result in high mortality rates<sup>111,112</sup>. There are two types of human pathogenic fungi. The true pathogens invade healthy humans, e.g. *Coccidioides immitis*<sup>102</sup>. Then there are opportunistic pathogens, which almost exclusively infect immunocompromised individuals. The immune system is hampered and the already present opportunist has a chance to attack the body<sup>113</sup>. Infections have several stages of severity reaching from superficial to systemic. In the latter case, multiple body parts are infected, treatment is nearly impossible, and the mortality rate is extremely high<sup>111,112</sup>. There are several virulence factors necessary for an infection. Hyphae growth facilitates penetration of host tissue. Typically, an arsenal of secreted lytic enzymes enables the pathogen to advance by destroying host cells. Another factor is the ability to grow in the unusual host environment. Fungal pathogens must cope with the high body temperature and a different pH. The presence of minerals, as iron and copper, in different organs are often times virulence factors too, making the pathogens more aggressive as they use these minerals. 'Successful' pathogens have developed different strategies to ward off attacks of the immune system<sup>114,115</sup>. Other fungi make parasitic symbioses with other animals, as insects, Amphibia and fish. Plant diseases as rust, smut and mildew are caused by plant pathogens<sup>102</sup>. Besides the mentioned threads, there is a wide range of benefits from fungi. For centuries, fungi have been used for human nutrition in form of edible mushrooms, yeasts for bakery and brewing, and fungi for diverse soft cheeses and drinks. Other fungi are used to produce antibiotics (e.g. penicillin), immunosuppressive drugs, steroids<sup>102</sup>, citric acids and numerous commercial enzymes<sup>116</sup>. Next to bacteria, yeasts are used to produce so-called 'biofuels' from lignocellulosic biomass for alternative energy supply<sup>117</sup>. For decades, fungi as *S. cerevisiae* and *N. crassa* have been scientific study objects, which enabled advances in genetics and molecular biology<sup>102</sup>. The thousands not yet investigated fungi carry a vast potential for discovery of natural products as secondary metabolites. There are many scientists and whole institutes, as the Jena Hans-Knöll Institute, dedicated to the quest for new drugs.

The choice of fungi as object for AS analysis had three reasons. First, the understanding of fungi's lifestyles and cellular functions is medically highly relevant. Second, one of the main research fields of the Leibniz Institute for Natural Product Research and Infection Biology - Hans-Knöll-Institute (HKI) Jena - is the infection biology of fungi. And third, the Jena Microbial Resource Collection (JMRC) localized at the HKI has great expertise in researching fungi. These circumstances led to the fact that Jena offered a great network of scientific collaboration partners to study fungi.

## 1. INTRODUCTION

---



**2**

## **Publications**

## 2. PUBLICATIONS

---

### 2.1 Combinatorics of aliphatic amino acids

In the publication of Grützmann et al. 2011<sup>118</sup> we calculate the number of theoretically possible aliphatic amino acids as a function of the number of carbon atoms in the side chain. We discuss which of the theoretically possible structures actually occur in living organisms. Our results reflect the general phenomenon in biology that usually a small number of molecules are used as building blocks to assemble a huge number of different macro-molecules, thereby giving rise to biological complexity.

## Combinatorics of aliphatic amino acids

Konrad Grützmann · Sebastian Böcker ·  
Stefan Schuster

Received: 20 August 2010 / Revised: 12 November 2010 / Accepted: 13 November 2010 / Published online: 1 December 2010  
© Springer-Verlag 2010

**Abstract** This study combines biology and mathematics, showing that a relatively simple question from molecular biology can lead to complicated mathematics. The question is how to calculate the number of theoretically possible aliphatic amino acids as a function of the number of carbon atoms in the side chain. The presented calculation is based on earlier results from theoretical chemistry concerning alkyl compounds. Mathematical properties of this number series are highlighted. We discuss which of the theoretically possible structures really occur in living organisms, such as leucine and isoleucine with a chain length of four. This is done both for a strict definition of aliphatic amino acids only involving carbon and hydrogen atoms in their side chain and for a less strict definition allowing sulphur, nitrogen and oxygen atoms. While the main focus is on proteinogenic amino acids, we also give several examples of non-proteinogenic aliphatic amino acids, playing a role, for instance, in signalling. The results are in agreement with a general phenomenon found in biology: Usually, only a small number of molecules are chosen as building blocks to assemble an inconceivable number of different macromolecules as proteins. Thus, natural biological complexity arises from the multifarious combination of building blocks.

**Keywords** Aliphatic amino acids · Aliphatic side chain · Amino acid signalling · Enumeration of isomers · Pólya's enumeration theorem · Ternary tree graphs

### Introduction

Obviously, life on earth forms complex structures. On the other hand, on many scales of living systems, only relatively few building blocks are used. For example, out of more than 100 chemical elements, only six are mainly used: carbon, hydrogen, oxygen, nitrogen, sulphur and phosphorus (Lodish et al. 2000). Only four nucleobases appear in DNA: guanine, adenine, cytosine and thymine. Proteins are built from a limited set of amino acids (Alberts et al. 2007). Currently, 22 different amino acids with specific codons in the genetic code are known, with selenocysteine (Lee et al. 1989; Leinfelder et al. 1989) and pyrrolysine (Srinivasan et al. 2002) being the 21st and 22nd. In addition, there are a number of amino acids arising from post-translational modification, such as hydroxyproline (Colley and Baenziger 1987) and hydroxylysine (Miller and Robertson 1973; Kannicht 2002). Moreover, there are a number of non-proteinogenic amino acids such as citrulline and ornithine, which are metabolites in the urea cycle (Berg et al. 2002). A nice overview of proteinogenic and non-proteinogenic amino acids (about 50 in total) was given by Karas (1954). He provided a systematisation in terms of their side chain length and functional groups.

The number of encoded amino acids is relatively small in view of the enormous number of possible chemical structures of side chains. Analysing the complete combinatorics under consideration of all the five chemical elements mentioned above, single and double bonds, branched and ring-shaped structures etc. is next to impossible. In contrast,

K. Grützmann (✉) · S. Böcker · S. Schuster  
Jena Centre for Bioinformatics, Friedrich Schiller University Jena,  
Ernst-Abbe-Platz 2,  
07743, Jena, Germany  
e-mail: konrad.g@uni-jena.de

S. Böcker  
e-mail: sebastian.boecker@uni-jena.de

S. Schuster  
e-mail: stefan.schu@uni-jena.de

if we restrict the analysis to a certain class of amino acids such as the aliphatic amino acids, combinatorial analysis is feasible and still provides insightful and non-trivial results. Therefore, we use, in this manuscript, a number of restrictions specified below.

In the “**Mathematical approach**” section, we show how to calculate the number of theoretically possible aliphatic amino acids, whose side chains are composed of  $n=1, 2, 3, \dots$  carbon atoms. Similar combinatorics studies have been performed for alcohols (Henze and Blair 1931) and hydrocarbons (Balaban et al. 1988), but not yet for amino acids. More specifically, we consider  $\alpha$ -amino acids. Their central branching point is the  $C_\alpha$  atom, to which an H atom, the amino group and the carboxyl group are bound (while in  $\beta$ -amino acids, the amino group is bound to the  $C_\beta$  atom). The fourth binding partner is the side chain. In aliphatic amino acids, the side chain is a hydrocarbon.

Moreover, we use the restrictions that each carbon in the side chain is saturated with H atoms and that the side chain only contains single bonds and no rings. While the term ‘aliphatic’ would not exclude saturated rings, this would render the calculation much more complicated. Chirality is disregarded, that is, different stereoisomers are not counted separately.

In the “**Biochemical implications**” section, we examine how many of the theoretically possible aliphatic amino acids actually occur in proteins or with other biological functions. Finally, in the “**Discussion**” section, we consider some evolutionary aspects.

### Mathematical approach

The side chains defined above can be considered mathematically as a graph, where carbon atoms (except  $C_\alpha$  and C in the carboxyl group) are nodes and the bonds between them are edges of the graph. More precisely, it is a tree graph, since we exclude cycles. The  $C_\beta$  atom, which is the first C atom of the side chain, can be viewed as the root of the tree. Every C atom can bind a maximum of four other C atoms. When considering this tree as a directed graph, each node can have at most three child nodes because one bond is the edge to the father node. That is, the maximum out-degree is three. So the task is to find the number of all trees with  $n$  nodes and out-degree less than, or equal to, three. For the case of hydrocarbons, this task has been approached already in the end of the nineteenth century by Cayley (1874), up to  $n=6$  carbon atoms. A recursion formula has been found by Henze and Blair in 1931, and the numbers were given up to  $n=20$ .

Now, we recapitulate how to calculate the number of trees with the above-mentioned properties and apply them to the theoretically possible aliphatic amino acids. In each case, we refer to the ‘biological realisations’.

Let  $x_n$  be the number in question. We will advance in a recursive manner for increasing  $n$ . We start with  $x_0=1$ , which is realised in the amino acid glycine. Assume we know all  $x_i$  for  $i=0$  to  $k$ . To determine  $x_{k+1}$ , we start from the  $C_\beta$  atom and attach exactly three side chains to it, each of which can use a certain number of carbon atoms such that the sum of these numbers equals  $k$  (Fig. 1).

Because the number of possible structures results from multiplication of the numbers of possibilities for the subgraphs, each  $x_{k+1}$  can be calculated by multiplying terms of the series with smaller indices. In the following calculation, it is convenient to sort the side chains by the number of atoms in increasing manner.

$$x_1 = x_0 x_0 x_0 = 1 (\text{alanine}) \quad (1)$$

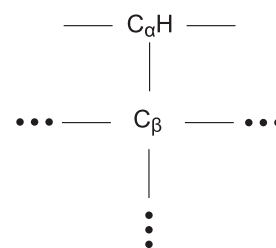
$$\begin{aligned} x_2 &= x_1 x_0 x_0 \\ &= 1 \text{ (not realized by a proteinogenic amino acid)} \end{aligned} \quad (2)$$

$$\begin{aligned} x_3 &= x_2 x_0 x_0 + x_1 x_1 x_0 \\ &= 2 \text{ (one of which is realized by valine)} \end{aligned} \quad (3)$$

In the following, the tuple  $(i, j, l)$  represents three attached subgraphs with  $i, j$  and  $l$  C atoms, respectively. For  $n=4$ , the combinations  $(3,0,0)$ ,  $(2,1,0)$  and  $(1,1,1)$  yield 2, 1 and 1 possible subgraphs, respectively. Their sum is  $x_4 = 4$ . The biological realisations will be discussed in the following section. For  $n=5$ , the combinations  $(4,0,0)$ ,  $(3,1,0)$ ,  $(2,2,0)$  and  $(2,1,1)$  yield 4, 2, 1 and 1 possible subgraphs, respectively. Their sum is  $x_5 = 8$ . Analogously, we obtain  $x_6 = 17$ . Notably, Cayley (1874) wrongly calculated this number to be 13.

In the calculation of  $x_7$ , the combination  $(3,3,0)$  occurs. For this, one has to account for symmetry, since attaching a branched chain with  $k=3$  to one side and an unbranched chain with  $k=3$  to the other side amounts to the same as doing this the other way round (Fig. 2). We can think of pulling two balls from an urn containing two balls (corresponding to branched and unbranched chains), with placing balls back. Thus, the selected items are not necessarily distinct. In mathematical

**Fig. 1** Schematic picture of the structure of aliphatic amino acids. Three dots stand for a tree subgraph, with  $i, j$  and  $l$  nodes, where  $i+j+l=k$ . If one of these indices equals zero, there is only an H atom attached



terms, this is a combination with repetition. The number is given by  $\binom{2+2-1}{2} = 3$  (rather than  $2 \cdot 2 = 4$ ). Together with the other combinations, we obtain  $x_7 = 39$ . The next term is  $x_8 = 89$ , where we again took into account that the combination (3,3,1) yields  $\binom{3}{2} = 3$ .

The following recurrence formula with initialisation  $x_0=1$  can be used (Pólya 1937).

$$x_n = \frac{1}{6} \left( \sum_{i+j+k=n-1} x_i x_j x_k + 3 \sum_{i+2j=n-1} x_i x_j + 2 \sum_{3i=n-1} x_i \right) \quad (4)$$

$x_0=1.$

where the sums are taken over all  $i, j, k \geq 0$ , and we set

An illustration of this formula is given in Fig. 2. The sketches of two alternative proofs are given in the Appendix. No closed formula is known for the exact calculation of the series.

The sequence  $x_1, x_2, \dots$  increases very rapidly: 1, 1, 1, 2, 4, 8, 17, 39, 89, 211, 507, ... In fact, it increases more rapidly than  $2^{n-2}$ , as can be shown as follows. From  $x_5$  on, the recursion is

$$x_n = x_{n-1}x_0x_0 + x_{n-2}x_1x_0 + x_{n-3}x_2x_0 + x_{n-3}x_1x_1 + \dots \quad (5)$$

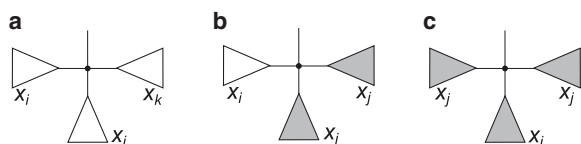
Thus,  $x_n \geq x_{n-1} + x_{n-2} + 2x_{n-3}$ . By induction, we conclude

$$\begin{aligned} x_n &\geq 2^{n-3} + 2^{n-4} + 2 \cdot 2^{n-5} = 2^{n-3} + 2^{n-4} + 2^{n-4} \\ &= 2^{n-3} + 2^{n-3} = 2^{n-2} \end{aligned} \quad (6)$$

Below, we will give the exact basis for the asymptotic exponential growth. It is easy to see already here that the series grows more slowly than  $4^n$ : At each carbon atom, there are four possibilities of continuation, attaching 0, 1, 2 or 3 carbon atoms. Thus, the basis lies between 2 and 4. In 1937, Pólya showed that the exponential base of growth is in the interval (2.77, 2.86). In 1948, Otter calculated the exact asymptotic behaviour of  $x_n$  to be

$$x_n \sim 0.5178760 \cdot 2.81546^n \cdot n^{-3/2} \quad (7)$$

This result builds upon a generating function  $T(z) = x_0z^0 + x_1z^1 + x_2z^2 + \dots$  (see Appendix) and on an analysis



**Fig. 2** Scheme illustrating the recursive calculation in Eq. 4. Sub-figures a, b and c essentially correspond to the three terms in Eq. 4. Grey (white) triangles correspond to subtrees that are (not) identical. Note the role of symmetry, in that the exchange of grey subtrees does not lead to a different structure

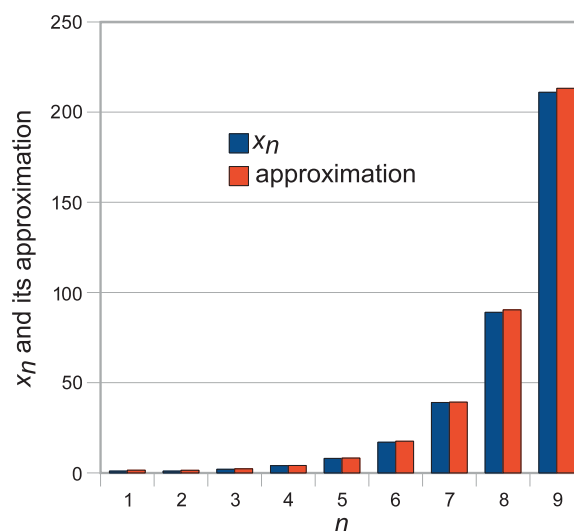
of its asymptotic behaviour. We omit the sophisticated mathematical details. For  $n=1$ , the approximation gives a value of 1.458 with the correct value being 1. For very low  $n$  values, the approximation formula is not needed, though, because the exact values can be calculated easily. It is remarkable that from  $n=3$  on, estimates from Eq. 7 are very reliable (Fig. 3). For  $n=3$  and  $n=19$ , for example, the approximation yields 2.22 and 2,175,376.00, respectively. Rounding the former value to an integer gives the correct value. The relative deviation of the latter value to the exact number of 2,156,010 is about 1%.

### Biochemical implications

As briefly mentioned in the “Mathematical approach” section, some of the possible aliphatic amino acids do occur in proteins while others do not. Moreover, some non-proteinogenic amino acids fulfil other biological functions such as molecular signalling. Based on a comprehensive literature search, Table 1 summarises the role of aliphatic amino acids in living organisms.




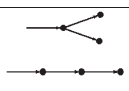
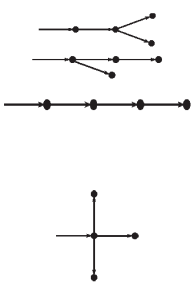
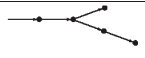
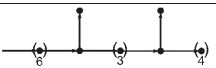
Strictly speaking, side chains of length three in proteinogenic amino acids are only realised by valine. In the broader sense, proline also belongs to that group since it has three C atoms in its side chain. These C atoms form a five-ring with the  $C_\alpha$  atom and the N atom of the imino group.

It is interesting to plot the series  $x_n/(n+2)$  (Fig. 4). This gives the number of possible variants of aliphatic amino acids per carbon ‘invested’ by the organism. The ‘invest-



**Fig. 3** Diagram showing a comparison of the series  $x_n$  (blue/dark grey bars) and its approximation (red/light grey bars) given in Eq. 7 for  $n=1, 2, 3, \dots, 9$ . Note that for  $n=0$  the approximation is not applicable since in this case the power term  $n^{-3/2}$  is not defined

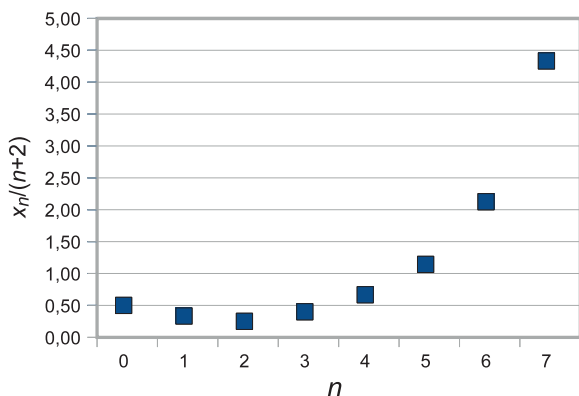
**Table 1** Biological functions of aliphatic amino acids (AAs)

Length $n$ of side chain	No. of possible structures	No. of known proteinogenic AAs	Amino acids	Chemical formula	Structure of side chain	Occurrence	Proteinogenic AAs if definition is less strict
0	1	1	Gly	$\text{CO}_2^- \text{CH}_2 \text{-NH}_3^+$		Proteins	
1	1	1	Ala	$\text{CO}_2^- \text{C}_2\text{H}_4 \text{-NH}_3^+$		Proteins	
2	1	0	2-amino butanoic acid	$\text{CO}_2^- \text{C}_3\text{H}_6 \text{-NH}_3^+$		Chemical communication in <i>Globodera rostochiensis</i> (Riga et al. 1997)	Cys, Sec, Ser
3	2	1	Val L-norvaline	$\text{CO}_2^- \text{C}_4\text{H}_8 \text{-NH}_3^+$		Proteins Anti-inflammatory arginase inhibitor (human) (Ming et al. 2009)	Pro, Thr
4	4	2	Leu Ile L-norleucine  3-methyl valine	$\text{CO}_2^- \text{C}_5\text{H}_{10} \text{-NH}_3^+$		Proteins Proteins Metabolite in orchidaceous plants (Kikuchi et al. 1981)  Not yet found	Met, Cystine
5	8	0	4-methyl L-norleucine	$\text{CO}_2^- \text{C}_6\text{H}_{12} \text{-NH}_3^+$		Seed of <i>Aesculus californica</i> (horse chestnut) (Fowden and Smith 1968)	Lys
6	17	0	-	$\text{CO}_2^- \text{C}_7\text{H}_{14} \text{-NH}_3^+$			
7	39	0	-	$\text{CO}_2^- \text{C}_8\text{H}_{16} \text{-NH}_3^+$			Arg
...							
17	321,198	0	2-amino-9,13-dimethyl-heptadecanoic acid	$\text{CO}_2^- \text{C}_{18}\text{H}_{36} \text{-NH}_3^+$		secondary metabolite from <i>Streptomyces</i> sp. 1010 (Ivanova et al. 2001)	

Proteinogenic AAs are given in the three-letter code (*Sec* selenocysteine). Other atoms than C and H are allowed for the less strict definition of AAs (last column)

ment' is represented by  $n+2$  because the  $\text{C}_\alpha$  atom and the carbon in the carboxyl group must also be taken into account. It can be seen that the series has a minimum at  $n=2$ . Thus, it may be speculated that  $x_2$  is not realised because it provides only one possible side chain, as in the case of  $x_0$  and  $x_1$ , but implies higher carbon 'costs'. A major reason for the observation that side chains longer than  $n=4$  are not used in proteins is probably the fact that their solubility in aqueous medium decreases with increasing length (see also the "Discussion" section).

There are no unbranched side chains of two, three or four C atoms in proteinogenic amino acids. However, when sulphur and selenium atoms are allowed besides carbons, side chains of length two and three can be found in cysteine, selenocysteine and methionine (Table 1 and Fig. 5). Indeed, methionine, which has a lower dipole moment than cysteine, is often classified as an aliphatic amino acid. When two cysteines form as disulphide bond, the resulting chain (sometimes called cystine) can be considered as aliphatic with  $n=4$ . The two dipole moments

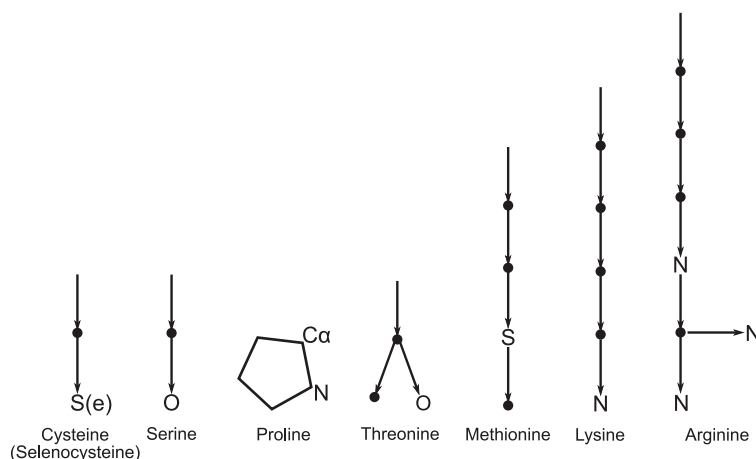


**Fig. 4** Diagram showing quantity  $x_n/(n+2)$  as a function of the number of carbon atoms in the side chain,  $n$ . That quantity can be regarded as the number of possible variants of aliphatic amino acids per ‘invested’ carbon by the organism. The  $n+2$  atoms include  $n$  carbons for the side chain, the  $C_\alpha$  atom and the carbon in the carboxyl group. Note the minimum at  $n=2$  (interpretation see text)

of the cysteines cancel each other, so that cystine is hydrophobic (Taylor 1986).

In the broader sense, lysine could be considered as an aliphatic amino acid with five nodes in its side chain. The last node consists of nitrogen. Although lysine is regarded as a polar amino acid, it is quite hydrophobic in the central region due to the methylene groups. With such an extension to heteroatoms, also serine and threonine (involving oxygen atoms in the side chains) can be included (Fig. 5). If we relax the definition even more, we could include arginine, having three nitrogen atoms and one double bond in the side chain. It corresponds to a graph with seven nodes. In addition to the proteinogenic amino acids given in the column of Table 1, there are non-proteinogenic amino acids with a less strict definition of aliphatic properties. Examples are homocysteine ( $n=3$ ) and ornithine ( $n=4$ ).

**Fig. 5** Graph-theoretical representation of the side chains of ‘aliphatic’ amino acids with heteroatoms. Note that, in proline, the nitrogen from the backbone amino group and the  $C_\alpha$  atom are not counted as part of the side chain. In arginine, the double bond to the N branching from the main side chain is represented by a single edge



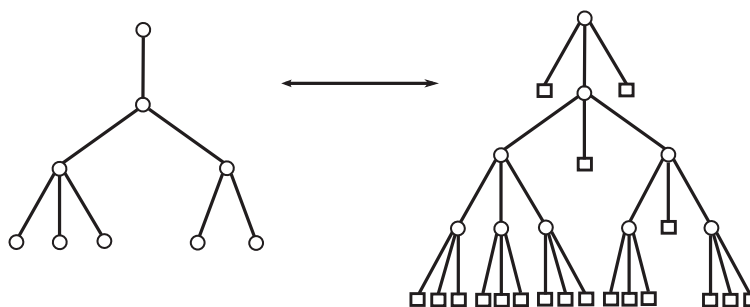
Both lysine and arginine can be methylated at their side chain nitrogen atoms when incorporated in proteins, for example, histones (Kim et al. 2009). Other post-translational histone modifications include acetylation, ubiquitination and sumoylation (with SUMO proteins, Small Ubiquitin-like Modifier) of lysines, and phosphorylation of serine and threonine (Kim et al. 2009). These mechanisms increase the number of possibilities realised in nature even more.

## Discussion

Here, we have analysed the combinatorics of aliphatic amino acids. Although the rules of molecular structure of these amino acids are relatively simple, it is a non-trivial task to compute the number of different isomers as a function of the number of carbon atoms. Under several restrictions such as exclusion of rings and double bonds, that function can be expressed as a series of numbers, which shows asymptotic exponential growth. This series has several other applications in theoretical chemistry, for example the number of isomeric primary alcohols ignoring stereoisomers (Henze and Blair 1931), the number of rooted ternary trees (Otter 1948; Balaban et al. 1988), that is, rooted trees with exactly three children at every node.

For  $n=0, 1$  or  $2$ , there is one possible structure each. The former two cases really occur in proteins: glycine and alanine (Table 1). The amino acid with  $n=2$ , 2-amino butanoic acid, is not proteinogenic but plays a role in signalling. With increasing  $n$ , a smaller and smaller fraction of the theoretical possibilities are realised. Thus, in the course of evolution, only a small number of all possible amino acids have been chosen as building blocks to assemble an inconceivable number of different proteins. So, natural biological complexity is made up of a small

**Fig. 6** Illustration of the replacement of rooted trees with  $n$  nodes and at most three children at every node (a) by rooted trees with exactly three children at every node, and  $n$  internal (non-leaf) vertices (b). The nodes of the original tree (circles) become internal nodes of the new tree by adding new leaves (squares)



tractable alphabet. Actual multiplicity arises from combination. This principle can also be observed for other biological macromolecules such as DNA and RNA (Alberts et al. 2007). The use of a controllable amount of units for the building of proteins favours a controllable genetic code for the transmission of information. Multiplicity by combination can also be observed in the case of glycans and glycolipids (Gabius 2009).

For proteins, this observation may have several practical reasons. Aliphatic amino acids with long side chains are practically insoluble in the aqueous environment of the cytosol. This would lead to the precipitation of either the free amino acid upon its synthesis or the peptides or proteins containing it. Secondly, large amino acids would require longer and, thus, slower biosynthesis pathways and their synthesis would be more costly (Seligmann 2003). Moreover, the atoms of longer side chains with a high degree of branching would sterically hinder each other. At the level of protein structures, bulky side chains are cumbersome for folding and functioning of macromolecules. Binding pockets and catalytic sites might be blocked and conformation changes could be impossible. These are crucial evolutionary drawbacks of unflexible proteins.

Beside the argumentation for structural and functional optimisation, another reason for nature using exactly those amino acids mentioned above is illustrated by the notion of ‘frozen accident’ (Crick 1968). According to this hypothesis, the set of present amino acids is relatively ancient. The set could not be optimised further because it is used by higher-level mechanisms such as the transcription–translation machinery and the genetic code that strongly rely on them and cannot tolerate any change in, or addition of, building blocks. However, there are exceptions to this rule. Selenocysteine and pyrrolysine were probably recruited later, and former stop codons are now used alternatively for these amino acids (Leinfelder et al. 1989; Srinivasan et al. 2002).

Miller and Urey (1959) showed in their famous experiment that in a simulated atmosphere of pre-biotic earth only consisting of water, methane ( $\text{CH}_4$ ), ammonia ( $\text{NH}_3$ ) and hydrogen ( $\text{H}_2$ ), several amino acids can be produced

spontaneously. These include the aliphatic amino acids glycine,  $\beta$ -alanine (in which the amino group is bound to the  $\text{C}_\beta$  atom), alanine, sarcosine (glycine with an additional methyl group bound to the amino group),  $n$ -methylalanine (alanine methylated at the amino group),  $\alpha$ -amino- $n$ -butyric acid and  $\alpha$ -aminoisobutyric acid. The latter does not belong to the class of amino acids considered here because it involves two methyl groups linked to the  $\text{C}_\alpha$  atom. In contrast,  $\alpha$ -amino- $n$ -butyric acid represents the  $x_2=1$  aliphatic amino acid with two carbons in the side chain. Thus, it is not disfavoured by chemical reasons; there must have been additional evolutionary pressure against its usage in proteins. We have here suggested that the absence of isomers of the side chain and, thus, the lack of variability in spite of the relatively high investment of four carbons may be a reason for not using it in proteins.

The present analysis shows that biochemistry and mathematics can nicely be combined. Starting from a relatively simple question, rather complex mathematics can be done. This combination may help students to put a classification into the set of amino acids. This may also facilitate to learn their structures, which is rather uninspiring otherwise.

Throughout, we have neglected chirality, that is, stereoisomers have been considered equivalent. Of course, chirality is important for biological activity. Among the proteinogenic amino acids, isoleucine and threonine are chiral in their side chains. The stereoisomers, alloisoleucine and allothreonine, are practically not used in proteins. When this effect is taken into account, the series of numbers of possible structures increases even more rapidly starting from  $n=4$  on, with  $x_4=5$  and  $x_5=11$  rather than 4 and 8, respectively.

It is worth mentioning that nature has brought about further structures by replacing the hydrogen atom that is bound to the  $\text{C}_\alpha$  atom by an alkyl chain. For example, in 2-ethyl-norvaline, an unbranched propyl group and an ethyl group are bound to the  $\text{C}_\alpha$  atom. This amino acid is part of a cytotoxic peptide and antibiotic of the fungus *Tohyopocladium geodes* (Tsantrizos et al. 1996). Further amino acids



with deviating basic structures such as sarcosine (*N*-methylglycine) were listed by Karas (1954).

The huge diversity of proteins arises in at least two ways: in addition to the versatile combination of the building blocks, their modification after incorporation into the protein plays an important role, as is realised, for example, by post-translational modification of amino acids. From an evolutionary point of view, one may argue that in early evolution, a smaller set of elementary units was sufficient. Later, with increasing complexity of the organisms, evolution has not only combined these elements but also ‘tinkered’ with them by multitudinous modifications.

**Acknowledgements** We kindly thank Gunnar Brinkmann from the University of Gent, Belgium, for suggestions about the very first ideas for this manuscript. Further acknowledgements go to Dr Ina Weiß and Heike Göbel for literature search. We also thank Christian Bodenstein who inspired us to the idea of plotting the carbon ‘investment’ (see Fig. 4).

**Conflict of interests** The authors declare that they have no conflict of interest.

**Appendix**

There are (at least) two ways to prove the recurrence formula (4). The “elegant” way uses Pólya’s Enumeration Theorem (Pólya 1937) (also known as Redfield-Pólya’s Theorem) to do so: Here, we first notice that instead of counting rooted trees with *n* nodes and at most three children at every node, we can also count rooted trees with *exactly* three children at every node, and *n* internal (non-leaf) vertices. To see this, we simply complete a rooted tree by attaching up to three leaves to every node of the original tree, so that all vertices except the new leaves have degree three (Fig. 6). Such trees are called rooted *ternary* trees. Then, we note that the symmetry between such trees is due to arbitrarily sorting the children of any node. The mathematical representation of this fact is the symmetric group  $S_3$ . One proceeds by computing the cycle index of  $S_3$ , which is

$$Z(S_3) = \frac{1}{6}(a_1^3 + 3a_1a_2 + 2a_3).$$

It is helpful to consider the *generating function*  $T(z)$  for the number of rooted ternary trees, which is defined as  $T(z) = x_0z^0 + x_1z^1 + x_2z^2 + \dots$ , where  $x_n$  are exactly the numbers introduced above. Pólya’s Enumeration Theorem (Pólya 1937) then tells us that this function fulfils the functional equation

$$T(z) = 1 + \frac{1}{6}z [T(z)^3 + 3T(z)T(z^2) + 2T(z^3)] \tag{8}$$

With the functional equation, we can now calculate the coefficient of any power  $z^n$  in the generating function: for

example, regarding  $T(z^3)$ , the three coefficients must add up to  $n-1$ . Doing so, we directly reach Eq. 4. We omit all further detail and refer the reader to any textbook about generating functions (e.g. Wilf 1994).

Now, we show a direct way to prove Eq. 4. Before doing so, we note that a slightly more complicated way of computing  $x_n$ , which ultimately also results in Eq. 4, was described by Henze and Blair in 1931. We simplify their presentation to calculate  $x_n$  as follows: To any node, we may attach three trees such that subtrees have pairwise different numbers of nodes. We explicitly allow that a tree has zero nodes, in which case we attach nothing—recall that we have defined  $x_0=1$  above. For this case, we do not have to take into account symmetry considerations, since the subtrees must be pairwise different. So, we have

$$\sum_{\substack{i+j+k=n-1 \\ i < j < k}} x_i x_j x_k = \frac{1}{6} \sum_{\substack{i+j+k=n-1 \\ i \neq j, j \neq k, k \neq i}} x_i x_j x_k \tag{9}$$

possibilities of doing so. Next, assume that we attach two trees of the same size *j*, and a third tree of size *i*. There exist  $\binom{x_j+1}{2} = \frac{1}{2}(x_j+1)x_j$  ways to choose two trees of size *j*, since this is a combination with repetition. We calculate the number of tree as

$$\sum_{\substack{i \neq j \\ i+2j=n-1}} x_i \binom{x_j+1}{2} = \frac{1}{2} \left( \sum_{\substack{i \neq j = k \\ i+j+k=n-1}} x_i x_j x_k + \sum_{\substack{i \neq j \\ i+2j=n-1}} x_i x_j \right) \tag{10}$$

Finally, all trees may have size *x<sub>i</sub>*. In this case, there exist

$$\binom{x_i+2}{3} = \frac{1}{6}(x_i+2)(x_i+1)x_i = \frac{1}{6}(x_i^3 + 3x_i^2 + 2x_i) \tag{11}$$

ways to choose three trees of size *i*. We calculate

$$\sum_{3i=n-1} \binom{x_i}{3} = \frac{1}{6} \left( \sum_{i=j=k, i+j+k=n-1} x_i x_j x_k + 3 \sum_{i=j, i+2j=n-1} x_i x_j + 2 \sum_{3i=n-1} x_i \right) \tag{12}$$

Note that if  $n-1$  is not divisible by three, then all of these sums are empty and, by definition, equal zero. Now, we add these three values, and sort them: First, we put all sums of products  $x_i x_j x_k$ . One can easily see that this equals

$\frac{1}{6} \sum_{i+j+k=n-1} x_i x_j x_k$  as desired: For the first sum on the right-hand side of Eq. 10, there are three possibilities of choosing two indices from  $i, j, k$  to be equal, as we can permute the indices. Similarly, we can reproduce the second summand of Eq. 4. Finally, the third summand directly comes from Eq. 12, which completes the proof.

## References

- Alberts B, Johnson A, Walter P, Lewis J, Raff M, Roberts K (2007) Molecular biology of the cell, 5th edn. Taylor & Francis, London
- Balaban AT, Kennedy JW, Quintas LV (1988) The number of alkanes having  $n$  carbons and a longest chain of length  $d$ . *J Chem Educ* 65:304–313
- Berg JM, Tymoczko JL, Stryer L (2002) Biochemistry, 5th edn. Freeman, New York
- Cayley A (1874) On the mathematical theory of isomers. *Phil Mag* 67:444–447
- Colley KJ, Baenziger JU (1987) Identification of the post-translational modifications of the core-specific lectin. The core-specific lectin contains hydroxyproline, hydroxylysine, and glucosylgalactosyl-hydroxylysine residues. *J Biol Chem* 262:10290–10295
- Crick FHC (1968) The origin of the genetic code. *J Mol Biol* 38:367–379
- Fowden L, Smith A (1968) Newly characterized amino acids from *Aesculus californica*. *Phytochemistry* 7:809–819
- Gabius HJ (ed) (2009) The sugar code: fundamentals of glycosciences. Wiley VCH, Weinheim
- Henze HR, Blair CM (1931) The number of structurally isomeric alcohols of the methanol series. *J Am Chem Soc* 53:3042–3046
- Ivanova V, Oriol M, Montes MJ, García A, Guinea J (2001) Secondary metabolites from a *Streptomyces* strain isolated from Livingston Island, Antarctica. *Z Naturforsch C* 56:1–5
- Kannicht C (2002) Posttranslational modifications of proteins: tools for functional proteomics, 1st ed. In: Kannicht C (ed) Methods in molecular biology, vol 194. Humana, Totowa
- Karas V (1954) Systematization of amino acids according to the increasing number of carbon atoms in the main aliphatic chain. *Farm Glas* 10:138–153 (in Croatian)
- Kikuchi T, Kadota S, Hanagaki S, Suehara H, Namba T, Lin CC, Kan WS (1981) Studies on the constituents of orchidaceous plants. I. Constituents of *Nervilia purpurea* Schlechter and *Nervilia aragoana* Gaud. *Chem Pharm Bull* 29:2073–2078
- Kim JK, Samaranyake M, Pradhan S (2009) Epigenetic mechanisms in mammals. *Cell Mol Life Sci* 66:596–612
- Lee BJ, Worland PJ, Davis JN, Stadtman TC, Hatfield DL (1989) Identification of a selenocysteyl-tRNA(Ser) in mammalian cells that recognizes the nonsense codon, UGA. *J Biol Chem* 264:9724–9727
- Leinfelder W, Stadtman TC, Böck A (1989) Occurrence in vivo of selenocysteyl-tRNA (SERUCA) in *Escherichia coli*. Effect of sel mutations. *J Biol Chem* 264:9720–9723
- Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J (2000) Molecular cell biology, 4th edn. Freeman, New York
- Miller EJ, Robertson PB (1973) The stability of collagen cross-links when derived from hydroxylsyl residues. *Biochem Biophys Res Commun* 54:432–439
- Miller SL, Urey HC (1959) Organic compound synthesis on the primitive earth. *Science* 130:245–251
- Ming XF, Rajapakse AG, Carvas JM, Ruffieux J, Yang Z (2009) Inhibition of S6K1 accounts partially for the anti-inflammatory effects of the arginase inhibitor L-norvaline. *BMC Cardiovasc Disord* 9:12–18
- Otter R (1948) The number of trees. *Ann Mathem* 49:583–599
- Pólya G (1937) Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen. *Acta Math* 68:145–254
- Riga E, Perry RN, Barrett J, Johnston MRL (1997) Electrophysiological responses of male potato cyst nematodes, *Globodera rostochiensis* and *G. pallida*, to some chemicals. *J Chem Ecol* 23:417–428
- Seligmann H (2003) Cost-minimization of amino acid usage. *J Mol Evol* 56:151–161
- Srinivasan G, James CM, Krzycki JA (2002) Pyrrolysine encoded by UAG in *Archaea*: charging of a UAG-decoding specialized tRNA. *Science* 296:1459–1462
- Taylor WR (1986) The classification of amino acid conservation. *J Theor Biol* 119:205–218
- Tsantrizos YS, Pischos S, Sauriol F (1996) Structural assignment of the peptide antibiotic LP237-F8, a metabolite of *Tolypocladium geodes*. *J Organ Chem* 61:2118–2121
- Wilf H (1994) Generating functionology, 2nd edn. Academic, Boston

### 2.2 The alternative messages of fungal genomes

In Grützmann et al. 2010<sup>119</sup> we present a comparative study of AS in 26 fungal species. We found that a greater part of fungal genes than previously expected are associated with AS, and that a wide range of gene categories are affected. Intron retention is the most frequent AS type in the studied fungi, and skipped exons are rare in contrast to higher animals.

# The alternative messages of fungal genomes

Konrad Grützmann<sup>1,\*</sup>, Karol Szafranski<sup>2</sup>, Martin Pohl<sup>1</sup>, Matthias Platzer<sup>2</sup>, Stefan Schuster<sup>1</sup>

<sup>1</sup>*Friedrich-Schiller-University, Jena, Germany*

<sup>2</sup>*Genome Analysis, Leibniz Institute for Age Research, Fritz Lipmann Institute, Germany*

\*Speaker and correspondence: konrad.g@uni-jena.de

**Abstract:** Alternative splicing (AS) is a cellular process that increases a cell's coding capacity from a limited set of genes. Although, AS is common in higher plants and animals, its prevalence and abundance in the whole eukaryote domain is unknown. We present a comparative study of AS in 26 fungal species. The data suggest that a greater part of fungal genes than previously expected are alternatively spliced (up to 14%). We find evidence that in 26 of the examined fungi, a wide range of gene categories are affected. Hence, AS is a rather common mechanism in many fungi.

## 1 Introduction

Alternative splicing (AS) is considered a frequent and complex process of eukaryote cells. AS is a mechanism to enlarge a cell's proteome, functions as a layer of gene expression regulation, and increases phenotype variation while maintaining acquired function [Sor07]. Irimia et al. [IRPR07] found that ancient genes show high rates of AS, and date simple forms of AS to ancestors of plants, fungi and animals. There are many studies on AS in animals and higher plants, showing significant levels of AS affected genes in these species ( $\approx 20\%$  in *Arabidopsis thaliana* and *Drosophila melanogaster*,  $\approx 55\%$  in mouse,  $\geq 60\%$  in human, based on ESTs [KMA07]) The question remains how widespread and abundant AS occurs in the eukaryote domain, and how big the impact on different species' lives is.

For fungi, only few studies address this issue on a genome-wide scale. Among some known fungal species, AS occurs with varying but relatively low frequencies (0-5% of genes [IRPR07]). Due to a reduced proteome and genome size, it is tempting to speculate that some biochemical pathways, including splicing, are less complex in fungi than in higher animals. Fungi have smaller introns. They also show extended consensus sequences for certain splicing signals, namely the 5' splice site and the branchpoint region [KDB<sup>+</sup>04]. These features facilitate a structural interpretation of intron sequences, and they suggest low-complex AS patterns, both of which underline the role of fungi as genetic models for (alternative) splicing. With regard to this model role, as well as for a better understanding of the evolutionary dynamics of AS, we undertook a comparative investigation of AS in 26 fungal species.

## 2 Methods

We downloaded gene annotations, genome and transcript sequences from NCBI's GenBank, RefSeq and dbEST databases ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)), from Broad Institute ([www.broadinstitute.org](http://www.broadinstitute.org)) and

from Joint Genome Institute ([www.jgi.doe.gov](http://www.jgi.doe.gov)) for 34 fungi. Mitochondrial and alien strain genome sequences were excluded. After removing cloning adapters, repetitive and dust sequences, we built spliced alignments using exalin [ZG06]. Since EST sequencing is error-prone, exons and introns are only further used if their sequences were well aligned ( $\leq 10\%$  mismatches, no mismatches in 5nt region of splice sites). Other criteria are least length (exons  $\geq 6$ nt, introns  $\geq 50$ nt) and splice sites being from canonical (GT|AG) or known non-canonical (GC|AG, AT|AC) classes [KDB<sup>+</sup>04]. For each transcript, the alignment with best exalin score (min. score 20 bits) was selected.

### 3 Results

The number of available ESTs per species varies in a wide range (from 43 for *P. marneffeii* to 277,147 for *N. crassa*). We counted the number of introns detected by these ESTs after mapping, as described above. Relating it to the number of annotated introns (which are not necessarily the same), we left out species that had less than 5% of annotated introns covered by ESTs. This left 26 species with sufficient EST coverage including three different strains for *P. brasiliensis*). We find that 97-100% of all detected introns harbour the above mentioned allowed splice sites [KDB<sup>+</sup>04]. Non-canonical splice sites are rare among the remaining ones (0-3%), in accord with a previous study [KDB<sup>+</sup>04]. This is a first hint at the accuracy of our mapping and filtering approach.

The sets of reliable introns and exons (possibly from different mature mRNA isoforms) were examined for overlaps that are results of the basic AS modes: a) retained introns, that remain in the mature transcript, b) skipped exons, that are spliced out, c) and d) alternative 3' and 5' exon ends, where close alternative splice sites are used (alt. 3'/5'ss). Also see Figure 1. Thus, AS events prediction is based on EST conflicts only and not on conflicts of ESTs with available annotated intron/exon structures. Since availability and abundance of transcript data for fungi is a problem, an isoform was already regarded as sufficiently supported by only one EST. We counteracted false positive predictions with the above mentioned strict filter criteria on alignments.

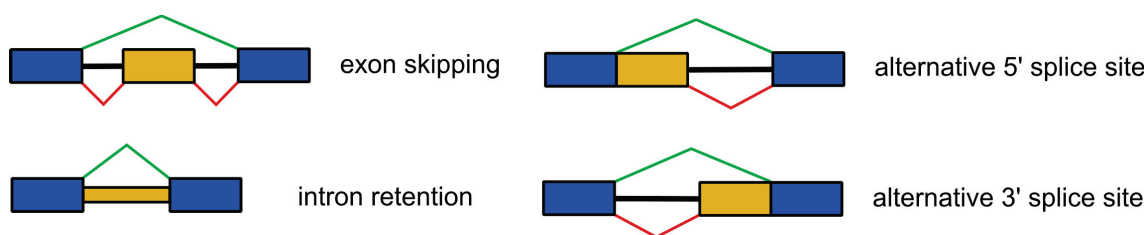


Figure 1: Basic alternative splicing types with constitutive exons (dark grey bars), alternative exons/exon parts (light grey bars), introns (thin black lines) and alternative splicing ways (flexed lines above and below).

Among the 26 studied species we detected a varying number of AS events: from 0 and 2 events for *P. stipitis* and *S. pombe* up to 796 and 870 events for *C. immitis* and *C. neoformans* (Table 1). We find a very similar distribution of the AS types for all species. The most frequent events are intron retention (48% of all AS events on average for all species), followed by alternative 3'ss

phylum	species	genes	introns	mapped ESTs	IR	SE	A5	A3	% AS genes	% est. AS gn.
Asco.	<i>A. niger</i>	11197	24096	42831	167	7	53	46	1.9	5.1
Asco.	<i>C. immitis</i>	10560	25503	58094	364	18	175	239	5.0	11.1
Asco.	<i>G. zeae</i>	11578	25741	19215	44	1	15	17	0.5	3.5
Asco.	<i>M. grisea</i>	14324	19662	76108	123	31	92	148	1.8	4.3
Asco.	<i>P. stipitis</i>	5816	2577	18643	0	0	0	0	0	0
Asco.	<i>S. cerevisiae</i>	6201	0	33628	2	0	2	9	0.2	0.2
Asco.	<i>S. pombe</i>	5238	4757	6329	2	0	0	0	0	0
Basidio.	<i>C. neoformans</i>	6617	36248	68966	506	31	126	207	8.3	14.5
Basidio.	<i>L. bicolor</i>	18264	90677	29746	125	18	35	57	0.01	0.03
Basidio.	<i>U. maydis</i>	6631	4900	34492	18	13	26	49	0.9	2.6
Micro.	<i>E. cuniculi</i>	2029	15	25	0	0	0	0	0	0
Zygo.	<i>R. oryzae</i>	17459	40515	11341	7	0	3	4	0.1	0.3

Table 1: Selection of studied species: phylum (Ascomycota, Basidiomycota, Zygomycota, Microsporidia), species, gene and intron no. from public annotation, no. of mapped ESTs, AS events (intron retention, skipped exons, alternative 5' and 3' splice site), percentage of genes with AS, extrapolated percentage of genes with AS.

(29%), alternative 5'ss (19%) and skipped exons (4%). See Figure 2 for a pie diagram. These results are in agreement with a former study on 14 fungi [MPNG08].

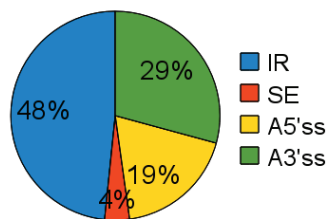


Figure 2: Distribution of the alternative splicing types on average over all investigated species: intron retention, skipped exons, alternative 3' splice sites and 5' splice sites.

We mapped the predicted AS events to genomic regions and found that on average 84.6% of the events overlap with annotated genes. We calculated the percentage of genes presumably affected by AS (table 1). These raw numbers only have weak comparability between the species, as the amount of available EST data is very diverse and the distribution of the transcripts across the genome unknown. Hence, we normalized the AS percentages with the inverse fraction of EST covered introns (7.7-58.1%). This yields an estimation of the AS rate of the complete genomes, for better comparability. On average we find that 4% of all genes of the studied fungi are affected by AS. There is a group with very low rates of AS (0-1% for *S. cerevisiae*, *S. pombe*, *P. stipitis*) and a group with intriguingly high rates (11% and 14% in *C. immitis* and *C. neoformans JEC21*).

The 20% of genes with meaningful description (omitting descriptions like “hypothetical|unknown protein”) are spread over a wide range of functional and structural categories: metabolism (malate dehydrogenase, enolase), gene expression (ribosomal proteins, replication factors), cytoskeleton (actin, tubulin folding), organelles and transport. In *C. neoformans* there is also a group of stress related and other genes, which may hint at an altering environment of the fungus as it appears during host infection.

## 4 Discussion and Conclusion

Of 34 studied fungal species where genome sequences and EST data are publicly available, 26 have acceptable amounts of data in terms of coverage of introns with aligned ESTs. Application of our method with stringent filter criteria results in detection of introns with splice site distributions in agreement with former results [KDB<sup>+</sup>04]. For most of the studied species we indeed find several events of AS, with AS rates per gene of 0-8% (1.6% on average). Careful extrapolation of the amounts of AS events to a hypothetical intron coverage by ESTs of 100% results in an AS rate of up to 11% and 14% of the genes in *C. immitis* and *C. neoformans*, respectively. Thus, the 'model organism' *S. cerevisiae*, having very few AS events, should be considered rather as an exception than as a prototypic fungus in the context of splicing.

As EST data is sparsely and heterogeneously annotated, we could not exclude ESTs from alien species strains directly. We hope to have circumvented this by strict filter criteria. Alien strain ESTs have, e.g., a higher mismatch rate in EST-genome alignments. Nonetheless, this issue could have led to false positive AS event classifications.

The data shown here and data for other non-animal organisms strongly support the view that exon skipping is an AS type that is very specific to higher animals. In view of the elevated AS rates compared to former studies [MPNG08], [IRPR07], and in view of the broad range of gene categories affected (metabolism, gene expression, cytoskeleton and others), we conclude that AS is a rather common phenomenon in many fungi.

## References

- [IRPR07] Manuel Irimia, Jakob Lewin Rukov, David Penny, and Scott William Roy. Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing. *BMC Evol Biol*, 7:188, 2007.
- [KDB<sup>+</sup>04] Doris M Kupfer, Scott D Drabenstot, Kent L Buchanan, Hongshing Lai, Hua Zhu, David W Dyer, Bruce A Roe, and Juneann W Murphy. Introns and splicing elements of five diverse fungi. *Eukaryot Cell*, 3(5):1088–1100, 2004.
- [KMA07] Eddo Kim, Alon Magen, and Gil Ast. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res*, 35(1):125–131, 2007.
- [MPNG08] Abigail M McGuire, Matthew D Pearson, Daniel E Neafsey, and James E Galagan. Cross-kingdom patterns of alternative splicing and splice recognition. *Genome Biol*, 9(3):R50, 2008.
- [Sor07] Rotem Sorek. The birth of new exons: mechanisms and evolutionary consequences. *RNA*, 13(10):1603–1608, 2007.
- [ZG06] Miao Zhang and Warren Gish. Improved spliced alignment from an information theoretic approach. *Bioinformatics*, 22(1):13–20, 2006.

## 2. PUBLICATIONS

---

### 2.3 Fungal alternative splicing is associated with multicellular complexity and virulence – A genome-wide multi-species study

In Grützmann et al. (accepted in *DNA Research*)<sup>120</sup> we present a genome-wide comparative study of AS in 23 fungi of different taxa based on alignments of ESTs to genome sequences. Using a robust random sampling strategy to estimate AS rates, we found that the average rate of AS affected fungal genes is 6.4%, and that Basidiomycetes show higher rates than Ascomycetes. We showed that AS associates with higher multicellular complexity and pathogenicity.



# **Fungal alternative splicing is associated with multicellular complexity and virulence - A genome-wide multi-species study**

Konrad Grützmann<sup>1</sup>, Karol Szafranski<sup>2</sup>, Martin Pohl<sup>1</sup>, Kerstin Voigt<sup>3</sup>, Andreas Petzold<sup>2</sup>, and Stefan Schuster<sup>1</sup>

<sup>1</sup> Department of Bioinformatics, Friedrich Schiller University Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany

<sup>2</sup> Genome Analysis, Leibniz Institute for Age Research - Fritz Lipmann Institute, Jena, Germany

<sup>3</sup> Leibniz Institute for Natural Product Research and Infection Biology - Hans-Knöll-Institute, Jena, Germany

Corresponding Author:

Konrad Grützmann

Department of Bioinformatics, Friedrich Schiller University Jena

Ernst-Abbe-Platz 2, D-07743 Jena, Germany

Phone: +49-3641-949581

Fax: +49-3641-946452

E-Mail: [konrad.g@uni-jena.de](mailto:konrad.g@uni-jena.de)

Running title:

Fungal alternative splicing

## **Abstract**

Alternative splicing (AS) is a cellular process that increases a cell's coding capacity from a limited set of genes. Although AS is common in higher plants and animals, its prevalence in other eukaryotes is mostly unknown. In fungi the involvement of AS in gene expression and its effect on multicellularity and virulence is of great medical and economic interest. We present a genome-wide comparative study of AS in 23 informative fungi of different taxa, based on alignments of public transcript sequences. Random sampling of ESTs allow for robust and comparable estimations of AS rates. We find that a greater fraction of fungal genes than previously expected is associated with AS. We estimate that on average, 6.4% of the annotated genes are affected by AS, with *Cryptococcus neoformans* showing an extraordinary rate of 18%. The investigated Basidiomycota show higher average AS rates (8.6%) than the Ascomycota (6.0%), although not significant. We find that multicellular complexity and younger evolutionary age associate with higher AS rates. Furthermore, AS affects genes involved in pathogenic lifestyle, particularly in functions of stress response and dimorphic switching. Together, our analysis strongly supports the view that AS is a rather common phenomenon in fungi and associates with higher multicellular complexity.

**Keywords:** Alternative Splicing; Fungal Genomes; Transcriptome Analysis; Multicellular Complexity; Retained Intron

## ***Introduction***

Via alternative splicing (AS) different mRNA isoforms are produced from one single gene. This diversification is one explanation for the discrepancy between the relatively low gene numbers of higher eukaryotes on the one hand and their cellular complexity on the other hand. AS affects binding properties, intracellular localization, enzymatic activity and many more properties of proteins<sup>1</sup>.

Examples of regulated pathways are sex determination in *Drosophila melanogaster*<sup>2</sup>, neuronal differentiation in rat<sup>3</sup> and auto-regulation of LAMMER kinases, which take part in splicing factor activation<sup>4</sup>. Not only the mere presence of an isoform, but also the exact splice isoform ratio can influence the phenotype of cells and can be regulated in a tissue dependent manner<sup>5</sup>.

Expressed sequence tags (ESTs) have widely been used to detect AS and quantify the transcript diversity arising from AS. For example, AS estimates for animals range from 53% of the multi-exon genes in human, 53% in mouse, 24% in rat, 22% in chicken, 19% in fruit fly to 6% in roundworm<sup>6</sup>. Interestingly, despite their relatedness, the estimates for mouse and rat differ remarkably. A possible reason for this are too few transcript data that limit the detection of AS events. Therefore, an approach was suggested that corrects for the amounts of transcripts and yields similar AS rates of approximately 31% for mouse and rat<sup>7</sup>. Finally, from deep transcriptome sequencing an AS rate of > 90% was estimated for humans<sup>8</sup>. These findings support the view that sensitive methods will ultimately detect splicing variants for every multi-exon gene<sup>9</sup>.

The basic alternative splicing types are the following: In exon skipping (SE, cassette exon), the exon can be spliced out of the transcript together with its flanking introns<sup>10</sup>. Alternative 5' splice site (A5'SS, alternative donor) selection<sup>11</sup> and alternative 3' splice site (A3'SS, alternative acceptor) selection<sup>12</sup> result in longer exons and corresponding isoforms<sup>13</sup>. Intron retention describes a mechanism where an intron can remain in the mature mRNA<sup>14</sup>. Previous studies showed that eukaryote species do not have equal distributions of these AS types. Cassette exons predominantly occur in animals whereas intron retention is more frequent in other taxa<sup>15</sup>.

Fungi, especially *Saccharomyces cerevisiae*, have been used extensively as a reduced and easily manageable model system in biological research. There are many fungi that cause human and plant diseases (Table S1), which provoke worldwide costs of several billion dollars a year. Other fungi are used for industrial fermentation and production of food and feed additives or are crucial in the degradation of xenobiotics and in the conversion of cellulose into biofuels (Table S1). Fungi have compact genomes (the majority 10-90 Mbp) and genes with small introns. They also show extended consensus sequences for the 5' SS and the branchpoint region<sup>16</sup>. These features facilitate a structural

interpretation of intron sequences, and they suggest low-complex AS patterns, both of which make fungi attractive models for mechanistic studies of (alternative) splicing.

A few studies estimated fungal AS rates on a genome-wide scale in comparative manner. Varying but relatively low AS frequencies were discovered in fungi and microsporidia (0-5% of genes in *S. cerevisiae*, *Schizosaccharomyces pombe*, *Encephalitozoon cuniculi* and *Cryptococcus neoformans*)<sup>17</sup>. A correspondence between intron numbers per gene and AS numbers was found for the studied species. Also, AS seems to affect genes of different functions. Genes associated with regulation have higher AS levels. Evolutionarily old genes were found to be affected more often<sup>17</sup>. In another comparative study of 14 fungi among other eukaryotes, also varying amounts of AS were observed. Yeasts showed nearly no events, and around 1000 AS instances were found for *C. neoformans* and *Coccidioides posadasii*, each<sup>15</sup>. Studies of single fungal species show results from only a few AS events in *Magnaporthe grisea*<sup>18</sup> to rates of 8.6% in *Aspergillus oryzae*<sup>19</sup> and 4.2% in *C. neoformans*<sup>20</sup>. Remarkably, Ho et al. estimate an AS rate for *Ustilago maydis* of 26% in a subset of multi-exon genes that have support by at least two ESTs<sup>21</sup>.

So far, AS research was mainly focused on animals and plants. With this study, we give a comprehensive report on fungi as the third eukaryote crown group. The comparability of the previous results on fungal AS is hampered due to the application of different biochemical and computational strategies. Thus, we undertook a systematic genome-wide comparative analysis of AS in 23 informative fungal species. The basis of our analysis are alignments of transcript sequences to genome sequences, and an AS rate estimation similar to that of Kim and coworkers<sup>7</sup>.

## **Materials and Methods**

### **Data sources and preparation**

We downloaded chromosomal sequences, reference transcripts and gene annotations of 25 species (26 different strains) from NCBI's GenBank, RefSeq and EntrezGene databases, respectively<sup>22</sup>. These data were complemented with most up-to-date sequences and annotations of *Pichia stipitis* (Joint Genome Institute<sup>23</sup>) and *S. cerevisiae* (Saccharomyces genome database<sup>24</sup>). Genome sequences and annotations of further three species (five strains) were from the Broad Institute (*Fusarium oxysporum*, *Paracoccidioides brasiliensis* Pb01, Pb03 and Pb18, *Rhizopus oryzae*<sup>25</sup>) and for further three species from the Joint Genome Institute (*Phanerochaete chrysosporium*, *Trichoderma reesei*, *Mycosphaerella graminicola*<sup>23</sup>). ESTs for all species were downloaded from NCBI's dbEST database, except for *A. benhamiae*, where Roche 454 data are from NCBI's SRA<sup>22</sup>. Four species were excluded from the

analysis because there were less than 200 ESTs. This yielded 27 species (30 strains) with sufficient data (Table 1). We masked low complexity repeats from the genome sequences using the program RepeatMasker (Smit et al., unpublished). We removed sequence contamination, low quality and low-complexity sequences from the ESTs using SeqClean (unpublished, “The Gene Index Project” of Harvard University). 454 reads were additionally cleaned for adapter stretches using in-house software.

### **Transcriptome-genome alignments and splice site conservation**

Spliced transcript-genome alignments were built in two steps: ESTs were mapped with Blat<sup>26</sup> to obtain first rough guide-alignments. The best Blat hits were further splice-aligned with exalin<sup>27</sup>. To use SS information as additional input for exalin, we prepared a scoring matrix based on SS consensi from *Neurospora crassa* as suggested by Zhang et al. Since SSs are conserved among fungi, this model was used for the analysis of all species. Alignments were filtered for minimal score (20 bits), mismatches ( $\leq 10\%$ , no mismatches in 5nt region of SSs) and minimum length of exons and introns (6nt and 40nt, respectively). Only alignments with SSs from canonical (GT|AG) or well-accepted non-canonical (GC|AG, AT|AC) classes<sup>16</sup> were considered for further analysis.

SS sequence conservation was calculated as information content per position<sup>28</sup>. To this end, we extracted the sequence from -4nt to +7nt from the exon-intron boundary, and the region from -4nt to +4nt from the intron-exon boundary. We used the upstream boundary of alternative 3' SSs and the downstream boundary of alternative 5' SSs.

### **Detection of alternative splicing**

Custom Perl scripts were used to analyse filtered transcript-genome alignments for four alternative splicing (AS) events: exon skipping (cassette exon), alternative 5' SS and alternative 3' SS selection, and intron retention. Using splice positions (genomic starts and ends of exons and introns) we compared the positions between all exons and introns to find overlaps and identify the basic AS types. AS events were predicted based on EST discrepancies only, not on discrepancies between ESTs and annotations. To account for the limited sequence data used in our analysis, one EST was considered sufficient to support an mRNA isoform. Constitutively spliced exons and introns were defined as not having support of AS at a minimum coverage of 10 ESTs.

### **Random sampling of transcripts and per-gene alternative splicing rates**

Random sampling was done for each genomic location with  $n \geq 2$  aligned transcripts. We randomly drew a defined number of transcripts and estimated the AS rate. To do so, AS events were assigned to

genes based on mapping coordinates, and the number of AS affected genes was divided by the overall number of genes detected by random sampling. Then, we multiplied the AS rates with the number of genes having introns (potential AS candidates) divided by the number of all genes. This yielded whole genome AS rate estimations. This procedure was repeated 20 times to calculate a mean AS rate estimation. The procedure was done with different sampling depths, drawing two to ten ESTs per locus. Due to a low EST coverage, loci with a higher coverage than 10 ESTs are rare for most analyzed species. Thus, to avoid a bias towards highly expressed genes, results from lower sampling depths were kept in sampling repeats with higher sampling depths. That is, sampling depth  $i$  means to draw at most  $i$  ESTs from a locus. Pearson's product-moment and its corresponding significance test was used to assess correlation between AS rates and number of mapped ESTs (based on Student's t-test, assuming normal distribution of the data, R version 2.12.1<sup>29</sup>). Four species were excluded from correlation analysis because less than 5% of their annotated genes were covered by the sampled ESTs (Table 1 and S2): *Aspergillus oryzae*, *Chaetomium globosum*, *Fusarium oxysporum*, *Sclerotinia sclerotiorum*.

### Functional gene annotations and enrichment statistics

Genes were searched for protein domain motifs using HMMER3<sup>30</sup> (e-value<0.01) together with the Pfam database (release 24). Alternatively, Pfam domain annotations were downloaded from the Broad Institute (*Fusarium spec.*, *Paracoccidioides spec.*, *R. oryzae*). Associations between Pfam domains and AS were tested using the following model: per species, all genes that have at least two EST hits are taken into consideration with its Pfam assignment and number of introns. Thereof, all introns are assumed to have an equal, species-specific probability  $p$  to be alternatively spliced, as averaged from the empirical data. The probability  $P(g \in AS)$  that a gene with  $n$  introns is alternatively spliced is calculated as  $P(g \in AS) = 1 - (1-p)^n$ .

Then, the expected number of alternatively spliced genes coding a certain Pfam is calculated by cumulation:  $Exp(n_{AS, Pfam}) = \sum P(g_i \in AS)$ .

The distribution of  $n_{AS, Pfam}$  was obtained from Monte Carlo simulation of the cumulation terms ( $n=10^6$ ). Binomial rather than hypergeometric simulation of  $P(g_i \in AS)$  simplified the calculations and yielded a slightly wider distribution, resulting in conservative estimates of the distribution quantiles. Correction for multiple testing was done using the Bonferroni method.

## **Results**

### **Mining of introns and splicing signals**

The number of available ESTs per species varies in a wide range (1,557-1,040,774). To detect alternative splicing, at least two transcripts per locus are needed, i.e. one for each of at least two splicing isoforms. We find that, depending on the species, 0-100% of the annotated introns are overlapped by at least two ESTs (25% on average over all fungi; supp. Table S2), and 1-86% of genes are overlapped by at least two ESTs (28% on average; Table 1 and S2). Per species, 98-100% of the detected introns per species harbor typical SSs (GT|AG), whereas non-canonical SSs (GC|AG, AT|AC) are rare (0-2%), in accord with a previous study on fungi<sup>16</sup>. The sets of reliable genomic intron and exon coordinates were subsequently examined for AS events.

### **Whole genome alternative splicing rates**

Numbers of detected AS events strongly depend on numbers of available ESTs (Figure S1A; Pearson correlation coefficient  $r = 0.82$ ,  $p$ -value  $1.8 \cdot 10^{-6}$ ). A very high coverage of introns with ESTs, especially when using next generation transcriptome sequencing, can reveal even very rare events that may partly represent splicing noise of the cell. This can lead to overestimation of AS propensity of a species. On the other hand, an uneven genomic distribution of transcripts leads to an under-sampling of the genome-wide splice isoforms. To circumvent these pitfalls, we applied a random sampling strategy, similar to the one of Kim et al.<sup>7</sup> to obtain AS rate estimations that are independent of EST amounts and distributions. We left out species where less than 5% of multi-exon genes were covered by the sampled ESTs for estimation of whole genome AS rates (last column Table S2). For them we do not expect the estimations to be reliable enough. We mapped the AS events that were recovered by random sampling to genomic locations of annotated genes to calculate AS rates per gene. We found that the correlation between these AS rates and the EST numbers is clearly reduced ( $r = 0.16$ ,  $p$ -value = 0.46, Figure S1B). Thus, random sampling gives AS rate estimates that are comparable between species.

The more ESTs were sampled from a genomic location (sampling depth) the higher is the chance of finding AS events (Figure S2). We decided to sample up to 10 ESTs per locus to reduce the chance of sampling rare events and, thus, overestimation of AS capacities. The reduced gains of AS rates with higher sampling depth support this decision (decreasing slopes of curves in Figure S2). Thus, the following results refer to a sampling depth of 10 ESTs, if not stated differently.

6.4% of fungal genes are affected by AS when averaging on species level (Table 1 and S2). Excluding

ascomycetous yeasts (*P. stipitis*, *S. cerevisiae* and *S. pombe* 0.26% AS affected genes), the rate is 7.3%. *Coccidioides immitis* and *C. neoformans* show outstanding AS rates of 13% and 18%/20% (strains JEC21 and B-3501A). The relative proportions of the AS types averaged over all species, in the order of frequency are: intron retention 61%, alternative 3' SSs 23%, alternative 5' SSs 13%, and skipped exons 3% (Figure 1A). We only took into account strain B-3501A and Pb01 from *C. neoformans* and *Paracoccidioides brasiliensis*, respectively, for mean value calculation.

### **Validation of retained introns**

An alternative explanation for detected retained introns (RIs) would be the presence of unprocessed pre-mRNA in the sequenced samples or a contamination with DNA. First of all, the EST libraries used for this study were all prepared from total RNA and enriched for poly(A)-mRNA. This makes DNA contamination very unlikely. To further validate the detected RIs, we assessed the number of RI-supporting ESTs that have been already processed in the following way. For each species, we counted the number of RIs where at least one EST of the isoform that harbors the RI supports a spliced intron at another EST position. For all species with RIs, between 74% and 100% (average 96%) of those isoforms contain a processed intron (details see supplementary Table S3). This clearly indicates that most RIs are authentic RNA events.

### **Correlations of alternative splicing rates and genomic features**

We calculated the correlations between genome and splicing quantities. We find a strong correlation ( $r = 0.73$ ,  $p\text{-value } 8.8 \cdot 10^{-5}$ ) between the number of EST-covered introns and the number of retained introns across the species. This hints at fungal introns to have a certain chance *per se* to be retained in an alternative manner. By contrast, there is only a slight and barely significant correlation of the extrapolated genome-wide AS numbers with gene numbers ( $r = 0.41$ ,  $p = 0.0502$ ) and with genome size in nucleotides ( $r = 0.53$ ,  $p = 0.009$ ). Further, there is no correlation of the AS rate per gene with gene numbers ( $r = -0.05$ ,  $p = 0.82$ ) nor with genome size ( $r = 0.12$ ,  $p = 0.57$ ). Nonetheless, the small genome sizes (in base pairs and gene numbers) of the yeasts *S. cerevisiae*, *S. pombe* and *P. stipitis* coincide with their clearly reduced AS propensity.

### **Intron retention is the major AS type in fungi**

Intron retention makes up two thirds of the AS events in the investigated fungi. This also holds for each fungal group separately. We investigated properties of affected introns and their aberration from constitutively spliced ones. We find that RIs are shorter (89nt) than constitutively spliced introns (93nt), on average across all species, though not significant (Mann-Whitney U test,  $p=0.211$ ,  $n =$



5,665/23,268). Neither constitutively spliced nor retained introns tend to preserve the reading frame. That is, in both sets, intron lengths are distributed evenly over the three possible remainders of division by three. Constitutively spliced introns: remainder zero - 32%, remainder one - 34%, remainder two - 34%; retained introns: 33%, 34% and 33%, respectively.

### **Varying alternative splice propensity is taxon-dependent**

We summarized and averaged the resampled AS rates into to different fungal taxa (see a species tree in Figure 2). On average in Basidiomycota more genes are affected by AS (8.6%) than in Ascomycota (7.2% w/o ascomycetous yeasts, Mann-Whitney U test, not significant,  $n=5/14$ ). Without the species showing outlying AS rates (*C. immitis*, *P. brasiliensis*, *C. neoformans*), the rates for Basidiomycota (6.1%) and Ascomycota (4.9%) are still different. Basidiomycota and Ascomycota have very similar AS type proportions, with Basidiomycota showing slightly more RIs and less alternative 5' SSs (Figure 1). In both cases RIs make up around two thirds of all AS events while skipped exons are only marginally present. The ascomycetous yeasts of our study (*P. stipitis*, *S. cerevisiae* and *S. pombe*) show an AS rate of 0.26% on average, which is significantly lower than the rate of the other Ascomycota (Mann-Whitney U test, p-value 0.003,  $n = 3/14$ ). An explanation for this difference may be deviations in structural gene properties that influence splicing. We find that Basidiomycota have on average shorter constitutively spliced introns (86nt) than Ascomycota (96nt, Mann-Whitney U test,  $p < 2.2 \cdot 10^{-16}$ ,  $n = 5,205/17,936$ ), and also shorter retained introns (72nt vs. 95nt,  $p < 2.2 \cdot 10^{-16}$ ,  $n = 1,545/4,093$ ). Considering the ascomycetous yeasts separately, they show on average 326 nt long constitutively spliced introns, and 132nt long retained introns, though it should be noted that yeast retained intron data is only based on five introns. In contrast, the one Mucoromycotina (formerly Zygomycota) *R. oryzae* has very short constitutively spliced (61nt) and retained introns (54nt). To further support the idea of the influence of gene properties on taxon-dependent AS frequencies, we compared the average conservation of SS motifs (Figures S3A and B). Sequence conservation in terms of information content can be considered as a proxy for SS fidelity. We find that ascomycetous retained as well as constitutively spliced introns show higher SS conservation than the corresponding basidiomycetous ones (not significant, Mann-Whitney U test, all  $p > 0.08$ ). The one Mucoromycotina, *R. oryzae*, has higher SS conservation in both types of introns than the Basidiomycota, yet can not clearly be distinguished from Ascomycota in this respect. Yeasts show the highest SS conservation. However, the number of sampled yeasts and Basidiomycota are very small so that only 5' SSs of yeast RIs are significantly more highly conserved than 5' SSs of basidiomycetous RIs ( $p=0.036$ ,  $n=2$  yeasts<sup>a</sup>,

---

<sup>a</sup> Yeast *P. stipitis* contributes no RIs.

5 Basidiomycota).

### **Functional characterization of alternative splicing**

To study the function of fungal AS we analyzed annotated and predicted Pfam domains for all genes and their relations to the AS rate of the gene families. We pooled all data and asked if particular Pfam domains are associated with higher AS rates. In a neutral model, AS is homogeneously distributed over all introns. Based on this model, we calculated the expected fraction of AS-associated genes per Pfam domain and compared it to the observed AS fraction. Together, six significantly AS-enriched Pfam gene families were identified (Table S4). Two are ribosomal genes (PF01479, PF01599), and two are genes involved in thiamine biosynthesis (PF09084, PF01946). We note that these gene families show particularly high expression rates (on average, EST coverage is 34, compared to 0.6 for a non-AS-enriched control group). Since ESTs are the primary evidence for AS, and the detection rate of AS increases with EST coverage, it is well possible that the high expression rates alone account for the Pfam-AS association in these gene groups.

Apart from these, the other two significantly AS-enriched Pfam gene families are fungi-specific (PF08520, PF12586) with unknown domain function. Remarkably, domain PF12586 occurs only in *Cryptococcus*. The next Pfam gene family with a known function, though below global significance ( $P=0.35$  with Bonferroni correction), is PF03073 and comprises integral membrane proteins that act as negative regulators of gene expression in response to oxygen or light (Table S5).

### **Alternative splicing is associated with dimorphic switch and pathogenicity**

Comparing AS rates of pathogenic and non-pathogenic fungi, we found interesting aspects: the rate in pathogenic species is higher (7.6%) than in non-pathogenic species (5.1%). Considering only human pathogens, the rate of 10.7% is even more striking, yet the differences are not significant (Mann-Whitney U test, all p-values above 0.09,  $n=11$  non-pathogenic, 6 human pathogenic).

The Pfam domain descriptions of the AS affected genes pointed to an involvement in stress response to an altering environment as it occurs during host infection: heat shock proteins, chaperone/chaperonin. These proteins mediate stress response, for example thermotolerance in mammalian hosts<sup>32</sup>. Furthermore, AS affected genes are often related to availability of copper, which is typical when penetrating human host tissue: multicopper oxidase and CTR copper transporter family. Glucuronoxylomannan, the predominant capsular polysaccharide in *C. neoformans*, experiences a structural change during dimorphic switching. Thus, the capsule surface changes, which results in a reduced recognition by the host's immune system<sup>33</sup>. We identified homologs of the proteins involved in production and modification of glucuronoxylomannan for all investigated fungi

via sequence similarity using BLASTP. Four of these proteins do show AS association, namely retained introns, two in *C. neoformans* JEC21 and two in *C. neoformans* B-3501A. Three are hypothetical proteins harboring a glycosyltransferase GTB or CAP59 mtransfer region. Two are annotated as mannosyltransferase 1 (Table S6). However, the predicted homologs are not significantly enriched in AS association (hypergeometric test,  $p > 0.1$ ).

Another virulence factor is the adaptation of a fungus to the altered environment of the host tissue. Up-regulation of oxidative and heat shock stress associated genes as *tps1*, *hsp30* and *ddr48* in *P. brasiliensis* P01 likely convey to cope with this micro-niche climate<sup>34</sup>. The identified homologs of these three genes are frequently affected by AS in pathogenic fungi (15 cases), and five times in non-pathogenic fungi (Table S7). Among the Tps1 homologs are genes from *C. neoformans* B-3501A and JEC21, one of *P. anserina* and one of *T. reesei*, an alpha,alpha-trehalose-phosphate synthase Tps1 subunit of *L. bicolor* and a hypothetical protein similar to alpha,alpha-trehalose-phosphate synthase subunit TPS3 of *N. crassa*. Some of the genes are affected by multiple AS events. The AS-associated Hsp30 homologs are three chaperones/small heat shock proteins from *C. immitis*, *L. bicolor* and *U. maydis*. Finally, Ddr48-homologs with AS-association are hypothetical/predicted proteins of *C. immitis*, *A. capsulatus*, *C. neoformans* and *M. graminicola*, two of which have a predicted function: “similar to potential stress response protein”, and “Glycosyltransferase GTB type”. These stress response related proteins are significantly enriched in AS association (hypergeometric test,  $p = 0.00022$ ).

## **Discussion**

### **Alternative splicing rate estimation**

We here present a comparative genome-wide survey of AS in the fungal kingdom. We based our survey mainly on Sanger-sequenced EST data (one species' ESTs are from 454 sequencing) and corresponding annotated genomes. In current AS studies, often next generation transcriptome sequence data with millions of short ESTs are used. However, though for some fungi, these data are available, the other prerequisite of having a well annotated genome is rarely fulfilled. According to our results, next generation sequencing technologies with EST lengths of above 200nt (met by Roche 454 as well as Illumina/Solexa platforms) should be well feasible for detection of the basic AS types in fungi. This is because the read length is clearly longer than the average fungal intron and exon lengths (constitutively spliced introns 93nt, exons 132nt).

The alignment of transcript sequences to genomes is currently the most effective way to detect

alterations of mature mRNA at large scale. However, Fox-Walsh and Hertel argued that every multi-exon gene has a certain AS frequency, and the detection of an alternative isoform is a matter of sensitivity of the method applied<sup>9</sup>. Thus, we here use a random sampling approach similar to the one by Kim et al.<sup>7</sup>. This universal normalization approach led to AS rate estimates that are independent of the number and distribution of the ESTs, and thus, are more comparable across species. We found AS events in every of the 27 studied fungi except one (*P. stipitis*), with an average rate of 6.4% of genes. Thus, we suppose that AS is a common phenomenon in the fungal kingdom. *C. imitidis* and *C. neoformans* show outstanding AS rates of 13% and 18%, the latter being about three times more than anticipated in earlier studies<sup>17,20</sup>. While successively increasing the sampling depth from two to ten, we found that the relative proportions of the AS rates between most species remain constant (Figure S2). This underpins the reliability of our normalization method. Because many loci have a lower EST coverage than the sampling depth of 10, our analysis yielded rather conservative estimates. It is likely that with deep transcriptome sequencing more fungal AS events will be found. Even when excluding very rare events, this may elevate the AS rates. This trend was seen for human and other mammals already<sup>8</sup>, and can be supported by the finding that for *A. benhamiae* and *N. crassa*, both of which have high EST coverage, AS rates clearly kept rising at higher sampling depths (Fig. S2), opposed to most of the other species.

Finally, in a recent study on fission yeasts, 433 AS events in overall 5144 genes were found in *S. pombe*<sup>35</sup>. While considering scaling effects due to sequencing depth, our results agree well with these findings in that the AS rate is very low compared to that in non-yeast Ascomycota (see Supplementary Calculation S1). This validates the comparability of our normalized AS rate results.

### **Fungal introns have an innate propensity to be retained**

We found that the trend of relative AS type distribution was the same in all the investigated fungal species. Intron retention made up the most prevalent of the investigated types (61% of the events). Contrarily, skipped exons were very rare (3%) and alternative 3' (23%) and 5' SSs (13%) comprised a third of the events. These results are in general agreement with previous findings on fungal AS<sup>15</sup> and are similar to trends in plants<sup>7,15</sup>. By contrast, skipped exons are more common than retained introns in invertebrates and even more frequent in vertebrates<sup>7</sup>.

The more introns a species genome harbors the more splicing needs to take place. The question is if this also increases the chance to have alternatively spliced introns *per se*. Indeed, we found a strong correlation of genome-wide intron numbers and numbers of retained introns. Thus, fungal introns seem to have an innate chance to be alternatively spliced. Similarly, Irimia et al. found a

correspondence between AS and intron number per gene in 12 eukaryotes<sup>17</sup>.

We found that fungal RIs are shorter than constitutively spliced introns. Also, on the species level, there is a correspondence between intron lengths and their propensity to be alternatively spliced. Together this hints at an involvement of the intron length in the recognition of introns. The intron definition mechanism is a model proposed to explain this same effect in plants. Splicing factors bind to the recognition sites on the RNA, and “bridge” across the intron by mutual binding. Thus, failed recognition of one SS typically results in intron retention<sup>15</sup>. This is in contrast to metazoan splicing, where splicing factors are assumed to form stable complexes across exons (exon definition mechanism) and where failed SS recognition typically results in exon skipping. It explains why metazoan introns tend to be much longer (e.g. 3,413nt in human<sup>36</sup>) but are rarely retained. Thus, we propose that the intron definition mechanism is prevalent in fungi similar to plants.

Finally, there is a hypothesis that connects SS conservation with splicing propensity, saying that strict adherence to the SS motif promotes the splicing machinery to bind more reliably to the SS and thus decreases the chance of AS<sup>37</sup>. McGuire et al. find weaker (i.e., less conserved) SSs at RIs compared to constitutively spliced introns in all their investigated species<sup>15</sup>. Here, when comparing introns (both retained and normally spliced ones) between the taxa, we find that higher SS conservation correlates with lower AS rates, which supports the hypothesis.

### **Fungal retained introns are authentic and likely trigger NMD**

There is a debate if RIs are authentic AS events or represent incompletely spliced pre-mRNA. Contamination with genomic DNA is very unlikely since the construction of EST libraries relies on affinity-based poly(A)<sup>+</sup> mRNA enrichment. From the analysis of fungal RIs, we found no tendency to preserve the reading frame, similar to results on non-fungal species in a previous study<sup>15</sup>. This may support the hypothesis of spurious intron retention. However, we have several arguments against it. For the majority of EST libraries analyzed here, cDNA was produced by poly(A)-tail capture, ensuring that ESTs derive from fully transcribed mRNAs. The current consensus is that intron splicing occurs predominantly cotranscriptionally<sup>38</sup>, corroborated by findings that the nascent mRNA can recruit multiple spliceosomes simultaneously<sup>39</sup>. Though the exact kinetics of RNA processing and export are unknown, intron splicing is likely finished shortly after transcription. This supports the hypothesis that if a detected multi-intron mRNA was spliced at one intron, it has already been spliced at the other introns, too. In fact, averaged over all species, 96% of the transcript isoforms that support an RI contain a processed intron at another position, as was similarly reported for RIs in *A. thaliana*<sup>40</sup>. In these cases, the completed splicing of co-transcribed introns indicates that the molecules have

passed spliceosomal processing and that RIs likely represent authentic events on mRNA. However, it is possible that RI-containing mRNAs had not left the nucleus, awaiting a later processing cycle or degradation. Nevertheless, even if this is true, these cases illustrate inherent differences in splicing efficiency.

It was argued that despite a weak selection for coding potential, splice variants having RIs unlikely yield functional proteins<sup>15</sup>. While we suppose that most RIs are authentic AS events, the isoforms with a frame-shifting RI unlikely yield productive, protein-coding mRNAs. However, we hypothesize that fungal RIs may in part be a means for post-transcriptional regulation via nonsense-mediated mRNA decay (NMD), in which transcripts containing premature termination codons (PTCs) are degraded<sup>41</sup>. This is because RI sequences with frame shifts probably introduce PTCs (15 randomly drawn triplets pose a chance of >50% to contain a stop codon). Most of the NMD-related components<sup>41</sup> are conserved in most of the fungi present in NCBI's HomoloGene database (Table S8). *S. cerevisiae* has an NMD machinery which is, however, not essential. Most RIs of the yeast *Yarrowia lipolytica*, contain PTCs and there is evidence that corresponding RNA is degraded by NMD<sup>42</sup>. Finally, first evidence for functional NMD were found in *N. crassa*<sup>43</sup>. As long as experimental data for a functional relevance of RIs is missing, we note that RIs qualify as mediators for a splicing-dependent mechanism of gene expression regulation, based on structure as well as on statistical association with functional categories (see below).

### **Does alternative splicing facilitate multicellular complexity?**

The complexity of the (multi-)cellular structure has long since been an important feature to classify fungi into sub-taxa<sup>44</sup>. Typical instances of diverse complexity are, being yeast or mold, and characteristics of sexual structures. There are predominantly single-celled yeasts, namely *S. pombe*, *S. cerevisiae* and *P. stipitis*, within the phylum of the Ascomycota, whose most complex yeast form is a four-spore ascus. The Mucoromycotina *R. oryzae* forms simple zygospores during sexual reproduction, but differentiated multicellular sporangia for asexual reproduction. Filamentous Ascomycetes produce more complex thalli, as e.g., ascocarps (apothecium, cleistothecium, perithecium). Finally, Basidiomycota, probably the most recent “crown group” of fungi, develop complex fruiting bodies<sup>44</sup>. We here find that the average AS rate of the mentioned taxa correlates with this order of complexity: *Saccharomycotina* and *Taphrinomycotina* (0.26% per genes AS rate), Mucoromycotina (2.3%), Pezizomycotina (7.2%, Ascomycota excluding yeasts) and Basidiomycota (8.6%). We speculate that AS contributes to multicellular complexity of the fungi.

We find that the fungi with the smallest genomes show nearly no AS. These are the ascomycetous



yeasts *S. cerevisiae*, *S. pombe* and *P. stipitis*. This is consistent with an earlier study on *S. cerevisiae* and *S. pombe*<sup>17</sup>. A major reason for this is probably the reduced proportion of intron containing genes, e.g., 5% of *S. cerevisiae* genes vs. 86% in *C. immitis*, since Hemiascomycetes (Saccharomycetes) experienced intron loss during the course of evolution<sup>45</sup>. However, from a certain genome size on, neither AS rate nor absolute AS number show any correlation. And, to an extreme, *C. neoformans* has only ca. 6600 genes but the highest found AS rate (18%).

The composition of the splicing machinery can give another perspective in understanding the differences in AS capability. The core components of the spliceosome, i.e. the five snRNPs and essential dynamic factors like Prp8 or Slu7, are generally conserved in eukaryotes. However, the small subunit of U2af (U2AF35 in human), involved in recognition of the 3' SS, is absent in *S. cerevisiae*. The family of serine/arginine-rich (SR) proteins comprises many known splicing regulators, and it was proposed that a higher SR protein diversity increases the AS complexity<sup>46</sup>. Our results do support this hypothesis: among yeasts, which have the lowest AS rates, *S. cerevisiae* has no SR proteins, only an SR-like homolog Npl3<sup>47</sup>, and *S. pombe* has only two SR proteins<sup>48</sup>. On the other hand, many of the other species of our study were found to have many SR and SR-related proteins<sup>49</sup>, in accordance with their higher AS rates.

We used Pfam domain annotations to analyze the possible functional associations of AS. The most significantly AS-enriched Pfam-coding gene families are ribosomal or do function in thiamine biosynthesis. However, these findings should be taken with caution since the expression level (i.e. EST coverage) is about 50-fold higher than average. Other AS-enriched gene families with moderate gene expression levels do code for fungi-specific protein domains of unknown function. This may indicate that AS is associated with enhanced evolutionary dynamics in these gene families, consistent with a supportive role of AS in gene evolution<sup>50</sup>.

Taking together the relatively low fraction of AS-associated gene families and the gene expression bias among the few candidates, we conclude that a homogenous distribution model is currently a sufficient explanation for the occurrence of AS among the EST-covered genes. However, we anticipate that increasing EST sequencing depths, and a saturation of a major fraction of genes, will allow more detailed insights into the functional association of AS in fungi.

Both, the elevated AS rates as well as the greater amount of splicing regulators of the more complex fungi suggest the hypothesis that AS may facilitate multicellular complexity. Furthermore, we found that AS is involved in another elaborate trait of certain fungi, namely virulence.

## Alternative splicing likely regulates virulence of pathogenic fungi

A first hint of AS involvement in pathogenicity was given by mere comparison of average AS rates. Human pathogenic fungi show on average a twice as high AS rate (10.7%) than non-pathogenic fungi (5.1%, neither plant nor human pathogenic). This is corroborated by the keywords of AS-associated Pfam domains which indicate enrichment for stress response functions. Moreover, a direct search for homologs of *P. brasiliensis*' *tps1*, *hsp30* and *ddr48* genes that convey cell rescue of this fungus while facing oxidative and heat shock stress in the human body<sup>34</sup>, yielded many AS-associated genes in human and plant pathogenic fungi. Hence, it is likely that AS is involved in gene expression regulation during the adaptation to the environmental conditions in the host.

The dimorphic switch is another virulence factor, and a key of persistent virulence<sup>33</sup>. During host penetration, a fungus can either switch to filamentous growth (e.g. *C. albicans*, *A. fumigatus*), or switch from filamentous to uni-cellular growth (e.g. *P. brasiliensis* Pb01, *C. immitis*). The dimorphic switch is only poorly understood. However, several contributing compounds have been identified. *C. neoformans*' glucuronoxylomannan (GMX), a capsular polysaccharide, is crucial for switching, as it alters the capsule surface. This increases the resistance against host immune system by hampering antibody and complement mediated phagocytosis<sup>33</sup>. We found two homologs of GMX production and modification proteins in *C. neoformans* (in B-3501A and JEC21), each containing a RI. Of the 19 predicted homologs of *tps1*, *hsp30*, *ddr48* and GMX-related genes, five have AS association in four non-pathogenic fungi (Tab S6 and S7).

An association of AS with pathogenicity has been found in former studies already. The *UrRm75* gene in *U. maydis*, involved in dimorphism and virulence, contains four introns and has an alternative 3' SS<sup>51</sup>. A putative heat shock protein and a putative alpha,alpha-trehalosephosphate synthase (both stress response-associated) were predicted to be affected by AS in *C. neoformans*<sup>20</sup>. Transcripts of cryptococcal intersectin 1 undergo AS and its disruption affects production of several virulence factors in *C. neoformans*<sup>52</sup>. In many fungi, Ste12-like transcription factors play essential roles in invasive growth and pseudohyphal development, and their gene transcripts are affected by AS within a conserved exon-intron structure<sup>53</sup>. Summarizing, gene regulation via AS likely facilitates virulence of pathogenic fungi on various levels.

## Acknowledgements

The authors thank Matthias Platzer from the Fritz Lipmann Institute (Jena), Ina Weiß from the chair of bioinformatics Jena, Michael Hiller from MPI of Molecular Cell Biology and Genetics (Dresden), and Kerstin Hoffmann from the Jena Microbial Resource Collection for helpful discussions. Further,



we thank Igor Grigoriev from JGI and Lucia Alvarado-Balderrama from Broad Institute for the permission to use and publish data of these institutes.

## References

1. Stamm S., Ben-Ari S., Rafalska I., et al. 2005, Function of alternative splicing, *Gene*, 344, 1–20.
2. Black D.L. 2003, Mechanisms of alternative pre-messenger RNA splicing, *Annu Rev Biochem*, 72, 291–336.
3. Lee H., Dean C., and Isacoff E. 2010, Alternative splicing of neuroligin regulates the rate of presynaptic differentiation, *J Neurosci*, 30, 11435–11446.
4. Stamm S. 2008, Regulation of alternative splicing by reversible protein phosphorylation, *J Biol Chem*, 283, 1223–1227
5. Kemper K., Tol M.J.P.M., and Medema J.P. 2010, Mouse tissues express multiple splice variants of prominin-1, *PLoS One*, 5, e12325
6. Kim N., Alekseyenko A.V., Roy M., and Lee C. 2007, ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species, *Nucl Acids Res*, 35, D93-D98
7. Kim E., Magen A., and Ast G. 2007, Different levels of alternative splicing among eukaryotes, *Nucleic Acids Res*, 35, 125–131
8. Pan Q., Shai O., Lee L.J., Frey B.J., and Blencowe B.J. 2008, Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing, *Nat Genet*, 40, 1413–1415
9. Fox-Walsh K.L. and Hertel K.J. 2009, Splice-site pairing is an intrinsically high fidelity process, *Proc Natl Acad Sci U S A*, 106, 1766–1771
10. Blencowe B.J. 2006, Alternative splicing: new insights from global analyses, *Cell*, 126, 37–47
11. Hiller M., Huse K., Szafranski K., Rosenstiel P., Schreiber S., Backofen R., and Platzer M. 2006, Phylogenetically widespread alternative splicing at unusual GYNGYN donors, *Genome Biol*, 7, R65
12. Hiller M., Huse K., Szafranski K., et al. 2004, Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity, *Nat Genet*, 36, 1255–1257
13. Sinha R., Lenser T., Jahn N., et al. 2010, Tassdb2 - a comprehensive database of subtle alternative splicing events, *BMC Bioinformatics*, 11, 216
14. Sakabe N.J. and de Souza S.J. 2007, Sequence features responsible for intron retention in human, *BMC Genomics*, 8, 59
15. McGuire A.M., Pearson M.D., Neafsey D.E., and Galagan J.E. 2008, Cross-kingdom patterns of alternative splicing and splice recognition, *Genome Biol*, 9, R50
16. Kupfer D.M., Drabentstot S.D., Buchanan K.L., et al. 2004, Introns and splicing elements of five

- diverse fungi, *Eukaryot Cell*, 3, 1088–1100
17. Irimia M., Rukov J.L., Penny D., and Roy S.W. 2007, Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing, *BMC Evol Biol*, 7, 188
  18. Ebbole D.J., Jin Y., Thon M., Pan H., Bhattarai E., Thomas T., and Dean R. 2004, Gene discovery and gene expression in the rice blast fungus, *Magnaporthe grisea*: analysis of expressed sequence tags, *Mol Plant Microbe Interact*, 17, 1337–1347
  19. Wang B., Guo G., Wang C., et al. 2010, Survey of the transcriptome of *Aspergillus oryzae* via massively parallel mRNA sequencing, *Nucleic Acids Res*, 38, 5075–5087
  20. Loftus B.J., Fung E., Roncaglia P., et al. 2005, The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*, *Science*, 307, 1321–1324
  21. Ho E.C.H., Cahill M.J., and Saville B.J. 2007, Gene discovery and transcript analyses in the corn smut pathogen *Ustilago maydis*: expressed sequence tag and genome sequence comparison, *BMC Genomics*, 8, 334
  22. NCBI - National Center of Biotechnology Information 2011, <http://www.ncbi.nlm.nih.gov>.
  23. DOE Joint Genome Institute 2011, A DOE Office of Science User Facility of Lawrence Berkeley National Laboratory. <http://www.jgi.doe.gov/>.
  24. Dwight S.S., Balakrishnan R., Christie K.R., et al. 2004, Saccharomyces genome database: underlying principles and organisation, *Brief Bioinform*, 5, 9–22.
  25. *Fusarium oxysporum*, *Paracoccidioides brasiliensis*, and *Rhizopus oryzae* Sequencing Projects, Broad Institute of MIT and Harvard 2011, <http://www.broadinstitute.org>.
  26. Kent W.J. 2002, BLAT—the BLAST-like alignment tool, *Genome Res*, 12, 656–664
  27. Zhang M. and Gish W. 2006, Improved spliced alignment from an information theoretic approach, *Bioinformatics*, 22, 13–20
  28. Schneider T.D. and Stephens R.M. 1990, Sequence logos: a new way to display consensus sequences, *Nucleic Acids Res*, 18, 6097–6100.
  29. R Development Core Team 2010, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
  30. Eddy S.R. 2008, A probabilistic model of local sequence alignment that simplifies statistical significance estimation, *PLoS Comput Biol*, 4, e1000069
  31. James T.Y., Kauff F., Schoch C.L., et al. 2006, Reconstructing the early evolution of fungi using a six-gene phylogeny, *Nature*, 443, 818–822
  32. Abad A., Fernández-Molina J.V., Bikandi J., et al. 2010, What makes *Aspergillus fumigatus* a

- successful pathogen? Genes and molecules involved in invasive aspergillosis, *Rev Iberoam Micol*, 27,155-182
33. Jain N. and Fries B.C. 2008, Phenotypic switching of *Cryptococcus neoformans* and *Cryptococcus gattii*, *Mycopathologia*, 166, 181–188
34. Borges C.L., Bailão A.M., Báo S.N., Pereira M., Parente J.A., and de Almeida Soares C.M. 2011, Genes potentially relevant in the parasitic phase of the fungal pathogen *Paracoccidioides brasiliensis*, *Mycopathologia*, 171, 1–9
35. Rhind N., Chen Z., Yassour M., et al. 2011, Comparative functional genomics of the fission yeasts, *Science*, 332, 930–936
36. Deutsch M. and Long M. 1999, Intron-exon structures of eukaryotic model organisms, *Nucleic Acids Res*, 27, 3219–3228.
37. Kim E., Goren A., and Ast G. 2008, Alternative splicing: current perspectives, *Bioessays*, 30, 38–47
38. Dujardin G., Lafaille C., Petrillo E., et al. 2012, Transcriptional elongation and alternative splicing, *Biochim Biophys Acta*
39. Brody Y., Neufeld N., Bieberstein N., et al. 2011, The in vivo kinetics of RNA polymerase II elongation during co-transcriptional splicing, *PLoS Biol*, 9, e1000573
40. English A.C., Patel K.S., and Loraine A.E. 2010, Prevalence of alternative splicing choices in *Arabidopsis thaliana*, *BMC Plant Biol*, 10, 102, doi:10.1186/1471-2229-10-102.
41. Bhuvanagiri M., Schlitter A.M., Hentze M.W., and Kulozik A.E. 2010, NMD: RNA biology meets human genetic medicine, *Biochem J*, 430, 365–377
42. Mekouar M., Blanc-Lenfle I., Ozanne C., et al. 2010, Detection and analysis of alternative splicing in *Yarrowia lipolytica* reveal structural constraints facilitating nonsense-mediated decay of intron-retaining transcripts, *Genome Biol*, 11, R65
43. Hood H.M., Neafsey D.E., Galagan J., and Sachs M.S. 2009, Evolutionary roles of upstream open reading frames in mediating gene regulation in fungi, *Annu Rev Microbiol*, 63, 385–409
44. Carlile M.J., Watkinson S.C., and Gooday G.W. 2001, *The Fungi*, Academic Press, 2 edition.
45. Stajich J.E., Dietrich F.S., and Roy S.W. 2007, Comparative genomic analysis of fungal genomes reveals intron-rich ancestors, *Genome Biol*, 8, R223
46. Busch A. and Hertel K.J. 2012, Evolution of SR protein and hnRNP splicing regulatory factors, *Wiley Interdiscip Rev RNA*, 3, 1–12
47. Fabrizio P., Dannenberg J., Dube P., Kastner B., Stark H., Urlaub H., and Lührmann R. 2009, The evolutionarily conserved core design of the catalytic activation step of the yeast spliceosome, *Mol*

*Cell*, 36, 593–608

48. Tang Z., Käufer N.F., and Lin R.J. 2002, Interactions between two fission yeast serine/arginine-rich proteins and their modulation by phosphorylation, *Biochem J*, 368, 527–534
49. Califice S., Baurain D., Hanikenne M., and Motte P. 2012, A single ancient origin for prototypical serine/arginine-rich splicing factors, *Plant Physiol*, 158, 546–560
50. Xing Y, Lee C. 2006, Alternative splicing and RNA selection pressure - evolutionary consequences for eukaryotic genomes, *Nat Rev Genet*, 7, 499-509
51. Rodríguez-Kessler M., Baeza-Montañez L., García-Pedrajas M.D., Tapia-Moreno A., Gold S., Jiménez Bremond J.F., and Ruiz-Herrera J. 2012, Isolation of *UmRrm75*, a gene involved in dimorphism and virulence of *Ustilago maydis*, *Microbiol Res*, 167, 270–282
52. Shen G., Whittington A., Song K., and Wang P. 2010, Pleiotropic function of intersectin homologue Cin1 in *Cryptococcus neoformans*, *Mol Microbiol*, 76, 662–676
53. Hoi J.W.S. and Dumas B. 2010, Ste12 and ste12-like proteins, fungal transcription factors regulating development and pathogenicity, *Eukaryot Cell*, 9, 480–485

## Tables

Table 1: Annotation, EST mapping and alternative splicing data of the studied species.

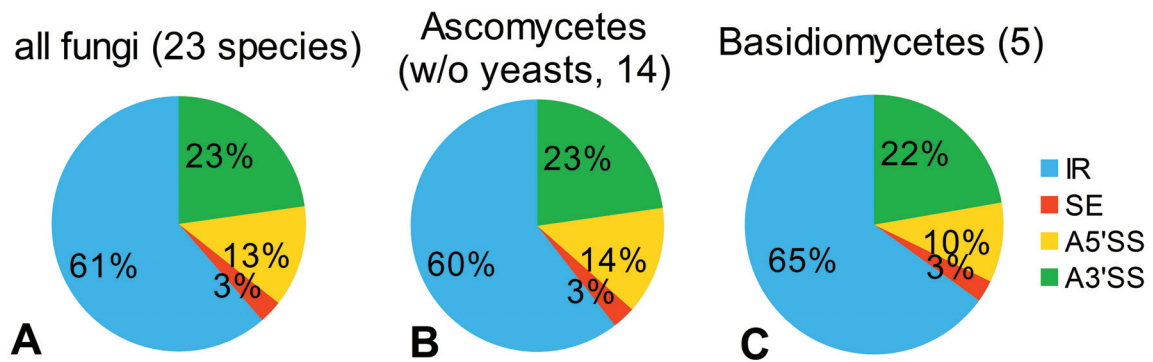
taxon <sup>1</sup>	species	Lifestyle <sup>2</sup>	annotated genes	annotated introns	number of available reads	% filtered and mapped reads	% genes covered with retained introns	% genes covered with skipped introns	alternative 5' intron ends	alternative 3' intron ends	% genes w. any type of AS	
A	<i>Ajellomyces capsulatus</i>	HP	9314	16275	26389	55	11	51	2	5	15	6.5
A	<i>Arthroderma benhamiae</i>	HP	7984	10332	1040774	86	86	1381	68	292	445	8.2
A	<i>Chaetomium globosum</i>	HP	11048	17396	1557	34	1	1	0	0	0	
A	<i>Coccidioides immitis</i>	HP	10440	17815	62729	93	49	664	18	152	225	13.4
A	<i>Paracoccidioides brasiliensis</i> Pb01	HP	9132	28179	41463	75	33	235	23	67	110	15.4
A	<i>Paracoccidioides brasiliensis</i> Pb03	HP	7875	19575	41463	71	35	134	16	31	52	10
A	<i>Paracoccidioides brasiliensis</i> Pb18	HP	8741	24498	41463	71	32	134	15	31	52	10.5
A	<i>Aspergillus nidulans</i>	NP	9541	16797	16848	89	15	81	1	11	14	7.3
A	<i>Aspergillus niger</i>	NP	10597	17668	46938	91	28	323	7	37	43	9.5
A	<i>Aspergillus oryzae</i>	NP	12823	20916	9051	94	9	70	2	5	12	
A	<i>Neurospora crassa</i>	NP	9841	14323	277147	83	52	511	57	128	164	8.8
A	<i>Pichia stipitis</i>	NP	5807	2580	19621	95	21	0	0	0	0	0
A	<i>Podospora anserina</i>	NP	10257	11261	51862	92	30	194	5	43	83	4.8
A	<i>Saccharomyces cerevisiae</i>	NP	5781	332	34915	97	41	2	0	2	7	0.18
A	<i>Schizosaccharomyces pombe</i>	NP	5073	3878	8123	78	10	3	0	0	0	0.6
A	<i>Trichoderma reesei</i>	NP	9143	18802	44964	76	40	66	2	18	22	2.5
A	<i>Botryotinia fuckeliana</i>	PP	16389	22334	10982	58	5	19	2	5	3	2.7
A	<i>Fusarium oxysporum</i>	PP	17735	30161	9248	67	3	33	0	4	5	
A	<i>Gibberella zeae</i>	PP	23218	38261	21355	91	14	75	1	9	16	5.9
A	<i>Magnaporthe grisea</i>	PP	14010	18795	88292	86	35	222	31	62	128	7.9
A	<i>Mycosphaerella graminicola</i>	PP	10952	17661	32194	83	33	140	9	29	55	6.1
A	<i>Phaeosphaeria nodorum</i>	PP	15983	21371	15973	79	9	20	1	2	7	2.4
A	<i>Sclerotinia sclerotiorum</i>	PP	14446	20240	1844	74	1	2	0	1	0	
B	<i>Cryptococcus neoformans</i> B-3501A	HP	6583	15244	74724	92	69	900	31	106	229	18.2
B	<i>Cryptococcus neoformans</i> JEC21	HP	6604	15554	74724	92	70	945	31	120	244	19.9
B	<i>Coprinopsis cinerea</i>	NP	13544	30180	15777	84	15	173	4	15	36	8.6
B	<i>Laccaria bicolor</i>	NP	18216	36757	34345	87	21	253	18	35	74	5.9
B	<i>Phanerochaete chrysosporium</i>	NP	10048	48688	12869	97	18	186	5	21	51	7.7
B	<i>Ustilago maydis</i>	PP	6522	4279	39308	88	50	34	13	14	36	2.3
M	<i>Rhizopus oryzae</i>	HP	17459	40515	13313	85	9	26	0	4	11	2.3
										<b>mean</b>	<b>6.4</b>	

Note, for *P. brasiliensis* and *C. neoformans* the same EST data was used for all strains, and hence, the same EST statistics come about. 454 transcript sequences are used for *A. benhamiae*, and classical EST data for all other species. AS rates in the last column are from random sampling.

<sup>1</sup>Taxa are Ascomycota, Basidiomycota and Mucoromycotina. Yeasts are underlined.

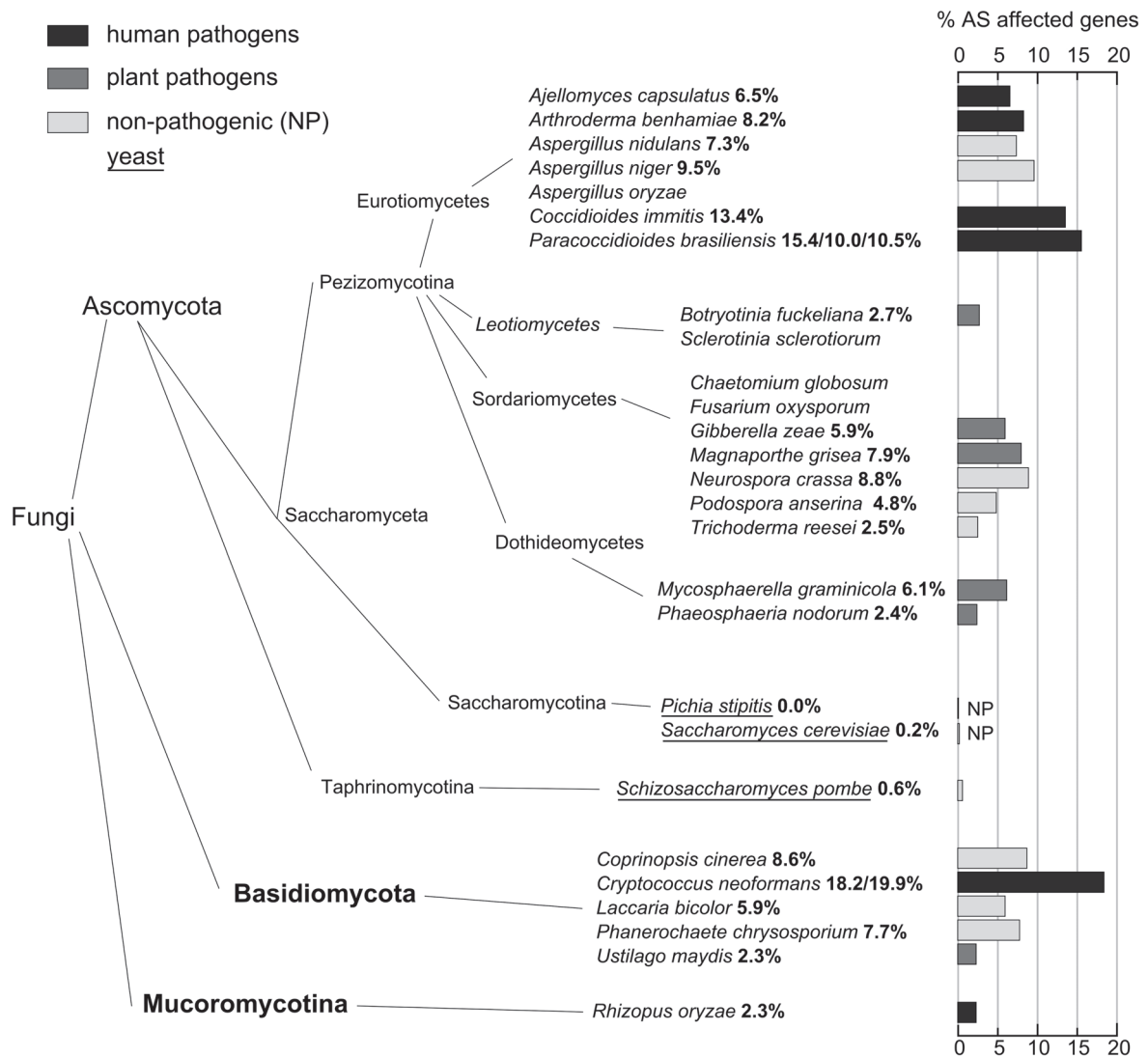
<sup>2</sup>Lifestyle: non-pathogenic (NP), plant pathogenic (PP), human pathogenic (HP).

## Figures



**Figure 1 - Alternative splice type distribution per taxon from random sampling approach.**

Pie portions: Intron Retention, Skipped Exons, Alternative 5' Splice Sites and 3' Splice Sites. Only the 23 informative fungi are considered, i.e. those where AS rates could be estimated (cf. Table 1). Only non-yeasts are considered in chart 1B (17 Ascomycetes - 3 yeasts = 14).



**Figure 2 - Species tree.**

This phylogenetic tree shows the evolutionary relationship between the analyzed species, based on James et al.<sup>31</sup> Percentages and bars next to the species represent the estimated AS rates per gene. AS rates for each strain are shown in case of species with more than one analyzed strain. Species' lifestyles are color-coded: human pathogens – black, plant pathogens – dark gray, non-pathogenic fungi – light gray. Yeasts are underlined.



## **SUPPLEMENTARY MATERIAL**

### **Supplementary Tables**

Table S1: **Phylum, anamorphic and teleomorphic synonyms and relevance of the investigated fungi.**

phylum	species	anamorph / teleomorph	relevance
<i>Ascomycota</i>	<i>Ajellomyces capsulatus</i>	<i>Histoplasma capsulatum</i> (anamorph)	systemic mycosis (histoplasmosis) in humans (1)
<i>Ascomycota</i>	<i>Arthroderma hamiae</i>	<i>Trichophyton erinacei</i> (anamorph)	superficial mycoses of keratinized host structures in humans and animals (2)
<i>Ascomycota</i>	<i>Aspergillus nidulans</i>	<i>Emericella nidulans</i> (teleomorph)	important model organism in genetics (3)
<i>Ascomycota</i>	<i>Aspergillus niger</i>		production of citric acid and numerous commercial enzymes (3)
<i>Ascomycota</i>	<i>Aspergillus oryzae</i>		food fermentation (3)
<i>Ascomycota</i>	<i>Botryotinia fuckeliana</i>	<i>Botrytis cinerea</i> (anamorph)	plant pathogen, causes gray mold disease on more than 200 host species (4)
<i>Ascomycota</i>	<i>Chaetomium globosum</i>		decomposition of cellulose-rich materials, causes skin and nail infections in humans, rarely cerebral and systemic infections, can act as allergen <sup>1</sup>
<i>Ascomycota</i>	<i>Coccidioides immitis</i>		causes systemic mycosis, coccidioidomycosis (Valley Fever) (5)
<i>Ascomycota</i>	<i>Fusarium oxysporum</i>		plant pathogen with broad host range, e.g. vegetables, cotton, banana, date and oil palm (6)
<i>Ascomycota</i>	<i>Gibberella zeae</i>	<i>Fusarium graminearum</i> (anamorph)	plant pathogen, head blight in wheat and other small grains, caused with other fusarium species more than 1 bn. US\$ damage in US agriculture in 1993 (7)
<i>Ascomycota</i>	<i>Magnaporthe grisea</i>	<i>Pyricularia grisea</i> (anamorph)	plant pathogen, causes rice blast (8)
<i>Ascomycota</i>	<i>Mycosphaerella graminicola</i>	<i>Septoria tritici</i> (anamorph)	major wheat pathogen worldwide, causes leaf blotch disease (9)
<i>Ascomycota</i>	<i>Neurospora crassa</i>	<i>Chrysonilia crassa</i> (anamorph)	non-pathogenic model organism for biological research <sup>2</sup>
<i>Ascomycota</i>	<i>Paracoccidioides brasiliensis</i>		human pathogen, causes paracoccidioidomycosis (South American blastomycosis) (10)
<i>Ascomycota</i>	Pb01/Pb03/Pb18	<i>Stagonospora nodorum</i> (anamorph)	plant pathogen, causes leaf blotch and glume blotch diseases on wheat (11)
<i>Ascomycota</i>	<i>Pichia stipitis</i> (yeast)		cellulose fermentation to produce ethanol (12)

<sup>1</sup>www.broadinstitute.de

<sup>2</sup>http://www.nih.gov/science/models/neurospora/

<i>Ascomycota</i>	<i>Podospora anserina</i>	non-pathogenic model organism (13)
<i>Ascomycota</i>	<i>Saccharomyces cerevisiae</i> (yeast)	budding yeast, food fermentation, important model organism in biology
<i>Ascomycota</i>	<i>Schizosaccharomyces pombe</i> (yeast)	fission yeast, non-pathogenic model organism in biology (14), beer brewing
<i>Ascomycota</i>	<i>Sclerotinia sclerotiorum</i>	necrotrophic plant pathogen (white mold), requires senescent tissues to establish an infection (15)
<i>Ascomycota</i>	<i>Trichoderma reesei</i>	<i>Hypocrea jecorina</i> (teleomorph) non-pathogenic model organism, industrial production of enzymes, food and feed additives (16)
<i>Basidiomycota</i>	<i>Coprinopsis cinerea</i>	edible mushroom (gray shag) (17)
<i>Basidiomycota</i>	<i>Filobasidiella neoformans</i> B-3501A/JEC21	human and animal pathogen (18)
<i>Basidiomycota</i>	<i>Laccaria bicolor</i>	edible symbiotic mushroom living on plant roots (19)
<i>Basidiomycota</i>	<i>Phanerochaete chrysosporium</i>	production of lignin-degrading peroxidases (20)
<i>Basidiomycota</i>	<i>Ustilago maydis</i>	plant pathogen, causes corn smut in maize, important model organism (21)
<i>Mucoromycotina</i> formerly <i>Zygomycota</i>	<i>Rhizopus oryzae</i>	opportunistic human pathogen, causes zygomycosis (22), industrial production of lipases (23)

Table S2: Annotation, EST mapping and alternative splicing data of the studied species.

taxon	lifestyle <sup>1</sup>	species	annotated genes	annotated introns	number of available reads	average read length	% filtered and mapped reads	% introns covered with $\geq 2$ reads	skipped exons	alternative 3' intron ends	% genes w. any type of AS	% genes w. intron retention	% genes w. cassette exon	% genes w. altern. 5' splice site	% genes w. altern. 3' splice site	% detected multi-exon genes				
Ascomycota	HP	<i>Ajellomyces capsulatus</i>	9314	16275	26389	401	55	7	11	51	2	5	15	6.5	5.0	0.3	0.5	1.5	704	8
Ascomycota	HP	<i>Arthroderma benhamiae</i>	7984	10332	1040774	297	86	100	86	1381	68	292	445	8.2	5.7	0.2	1.4	2.3	7251	100
Ascomycota	HP	<i>Chaetomium globosum</i>	11048	17396	1557	425	34	0	1	1	0	0	0	0	0	0	0	0	0	0
Ascomycota	HP	<i>Coelidioides immitis</i>	10440	17815	62729	758	93	48	49	664	18	152	225	13.4	9.4	0.3	2.6	3.7	4270	48
Ascomycota	HP	<i>Paracoccidioides brasiliensis</i> Pb01	9132	28179	41463	560	75	10	33	235	23	67	110	15.4	9.2	0.9	3.0	4.9	1757	22
Ascomycota	HP	<i>Paracoccidioides brasiliensis</i> Pb03	7875	19575	41463	560	71	12	35	134	16	31	52	10.0	6.3	0.7	1.6	2.7	1531	23
Ascomycota	HP	<i>Paracoccidioides brasiliensis</i> Pb18	8741	24498	41463	560	71	10	32	134	15	31	52	10.5	6.6	0.7	1.5	2.8	1534	20
Ascomycota	NP	<i>Aspergillus nidulans</i>	9541	16797	16848	411	89	8	15	81	1	11	14	7.3	6.0	0.1	0.9	1.2	801	10
Ascomycota	NP	<i>Aspergillus niger</i>	10597	17668	46938	634	91	32	28	323	7	37	43	9.5	7.6	0.2	1.1	1.3	2686	30
Ascomycota	NP	<i>Aspergillus oryzae</i>	12823	20916	9051	729	94	1	9	70	2	5	12	5	12	0	0	0	0	0
Ascomycota	NP	<i>Neurospora crassa</i>	9841	14323	277147	675	83	50	52	511	57	128	164	8.8	5.5	0.5	1.6	2.3	4052	50
Ascomycota	NP	<i>Pichia stipitis</i> (yeast)	5807	2580	19621	762	95	5	21	0	0	0	0	0	0	0	0	0	0	0
Ascomycota	NP	<i>Podospora anserina</i>	10257	11261	51862	859	92	44	30	194	5	43	83	4.8	3.1	0.1	0.8	1.4	2815	40
Ascomycota	NP	<i>Saccharomyces cerevisiae</i> (yeast)	5781	332	34915	509	97	55	41	2	0	2	7	0.18	0.05	0.00	0.03	0.11	223	67
Ascomycota	NP	<i>Schizosaccharomyces pombe</i> (yeast)	5073	3878	8123	262	78	5	10	3	0	0	0	0	0	0	0	0	0	0
Ascomycota	NP	<i>Trichoderma reesei</i>	9143	18802	44964	691	76	26	40	66	2	18	22	2.5	1.6	0.1	0.6	0.6	2602	35
Ascomycota	PP	<i>Botryotinia fuckeliana</i>	16389	22334	10982	856	58	5	5	19	2	5	3	2.7	1.8	0.2	0.6	0.2	580	5
Ascomycota	PP	<i>Fusarium oxysporum</i>	17735	30161	9248	408	67	2	3	33	0	4	5	0	0	0	0	0	0	0
Ascomycota	PP	<i>Gibberella zeae</i>	23218	38261	21355	545	91	5	14	75	1	9	16	5.9	4.7	0.1	0.6	0.9	1091	6
Ascomycota	PP	<i>Magnaporthe grisea</i>	14010	18795	88292	691	86	26	35	222	31	62	128	7.9	4.6	0.4	1.3	2.7	2994	28
Ascomycota	PP	<i>Mycosphaerella graminicola</i>	10952	17661	32194	705	83	20	33	140	9	29	55	6.1	3.5	0.2	0.9	1.8	2004	26
Ascomycota	PP	<i>Phaeosphaeria nodorum</i>	15983	21371	15973	536	79	8	9	20	1	2	7	2.4	1.6	0.1	0.2	0.7	896	7
Ascomycota	PP	<i>Sclerotinia sclerotiorum</i>	14446	20240	1844	681	74	1	1	2	0	1	0	0	0	0	0	0	0	0
Basidiomycota	HP	<i>Cryptococcus neoformans</i> B-3501A	6583	15244	74724	613	92	100	69	900	31	106	229	18.2	14.0	0.6	2.2	4.7	3985	62
Basidiomycota	HP	<i>Cryptococcus neoformans</i> JEC21	6604	15554	74724	613	92	100	70	945	31	120	244	19.9	15.2	0.6	2.8	5.3	3923	60
Basidiomycota	NP	<i>Coprinopsis cinerea</i>	13544	30180	15777	717	84	19	15	173	4	15	36	8.6	6.8	0.2	0.7	1.9	1694	14
Basidiomycota	NP	<i>Laccaria bicolor</i>	18216	36757	34345	566	87	26	21	253	18	35	74	5.9	4.4	0.1	0.7	1.4	3686	23
Basidiomycota	NP	<i>Phanerochaete chrysosporium</i>	10048	48688	12869	707	97	12	18	186	5	21	51	7.7	5.7	0.2	1.0	1.9	1818	21
Basidiomycota	PP	<i>Ustilago maydis</i>	6522	4279	39308	457	88	23	50	34	13	14	36	2.3	0.9	0.3	0.4	1.1	872	35
Mucoromycotina	HP	<i>Rhizopus oryzae</i>	17459	40515	13313	519	85	6	9	26	0	4	11	2.3	0.9	0.3	0.4	1.1	1160	9
											<b>mean</b>	<b>6.4</b>	<b>4.4</b>	<b>0.2</b>	<b>0.9</b>	<b>1.6</b>	<b>2096</b>	<b>30</b>		

Note, for *P. brasiliensis* and *C. neoformans* the same EST data was used for all strains, and hence, the same EST statistics come about. 454 transcript sequences are used for *A. benhamiae*, and classical EST data for all other species.

<sup>1</sup>Lifestyle: non-pathogenic (NP), plant pathogenic (PP), human pathogenic (HP).

Table S3: **Results of intron retention validation.** The third column shows numbers of intron retention events where at least one read that contains the retained intron has been spliced at another position.

species	retained introns	validated retained introns	% validated
<i>Ajellomyces capsulatus</i>	51	47	92.2
<i>Arthroderma benhamiae</i>	1381	1368	99.1
<i>Aspergillus nidulans</i>	81	78	96.3
<i>Aspergillus niger</i>	323	321	99.4
<i>Aspergillus oryzae</i>	70	67	95.7
<i>Botryotinia fuckeliana</i>	19	14	73.7
<i>Chaetomium globosum</i>	1	1	100
<i>Coccidioides immitis</i>	664	661	99.5
<i>Coprinopsis cinerea</i>	173	171	98.8
<i>Filobasidiella neoformans</i> B-3501A	900	875	97.2
<i>Filobasidiella neoformans</i> JEC21	945	925	97.9
<i>Fusarium oxysporum</i>	33	30	90.9
<i>Gibberella zeae</i>	75	72	96
<i>Laccaria bicolor</i>	253	249	98.4
<i>Magnaporthe grisea</i>	222	211	95
<i>Mycosphaerella graminicola</i>	140	134	95.7
<i>Neurospora crassa</i>	511	491	96.1
<i>Paracoccidioides brasiliensis</i> Pb01	235	208	88.5
<i>Paracoccidioides brasiliensis</i> Pb03	134	117	87.3
<i>Paracoccidioides brasiliensis</i> Pb18	134	121	90.3
<i>Penicillium marneffeii</i>	1	1	100
<i>Phaeosphaeria nodorum</i>	20	19	95
<i>Phanerochaete chrysosporium</i>	186	184	98.9
<i>Podospora anserina</i>	194	193	99.5
<i>Rhizopus oryzae</i>	26	25	96.2
<i>Saccharomyces cerevisiae</i> (yeast)	2	2	100
<i>Schizosaccharomyces pombe</i> (yeast)	3	3	100
<i>Sclerotinia sclerotiorum</i>	2	2	100
<i>Trichoderma reesei</i>	66	63	95.5
<i>Ustilago maydis</i>	34	34	100
		<b>mean</b>	<b>95.77</b>
		<b>min</b>	<b>73.7</b>
		<b>max</b>	<b>100</b>

Table S4: Pfam domains positively associated with alternative splicing.

Pfam accession	Pfam description	P value*	Avg. EST
PF01479	Ribosomal S4 domain	< 0.0007	21.8
PF09084	NMT1/THI5 like; involved in thiamine biosynthesis	0.0014	37.1
PF08520	fungal protein of unknown function (DUF)	0.0048	10.5
PF01946	Thi4 family; involved in thiamine biosynthesis	0.0089	30.3
PF01599	Ribosomal protein S27a	0.015	47.1
PF12586	protein of unknown function (DUF), <i>Cryptococcus</i>	0.045	2.0

\* P-values include the Bonferroni correction

Table S5: **Pfam domains slightly positively associated with alternative splicing.**

Pfam accession	Pfam description	P value*	Avg. EST
PF01248	Ribosomal protein L7Ae / L30e / S12e / Gadd45 family	0.069	8.0
PF00163	Ribosomal protein S4 / S9 N-terminal domain	0.12	27.6
PF00900	Ribosomal family S4e	0.28	20.7
PF03073	TspO/MBR family; integral membrane protein that acts as a negative regulator of gene expression in response to oxygen/light	0.35	2.8
PF02453	Reticulon, know as neuroendocrine-specific protein (NSP), associated with the endoplasmic reticulum	0.42	7.5
PF00428	60S acidic ribosomal protein	0.42	13.4

\* P-values include the Bonferroni correction

Table S6: **Glucuronoxylomannan-related genes affected by AS.** Sequences were downloaded from NCBI's protein database using the search "(glucuronoxylomannan) AND "fungi"[porgn:txid4751]".

species	protein ID	definition	note
<i>Cryptococcus</i> B-3501A	<i>neoformans</i> 134108310	hypothetical protein CNBB3380	"Glycosyltransferase.GTB_type"
<i>Cryptococcus</i> JEC21	<i>neoformans</i> 58263500	hypothetical protein	"Glycosyltransferase.GTB_type"
<i>Cryptococcus</i> B-3501A	<i>neoformans</i> 134107349	hypothetical protein CNBA6810	"CAP59_mtransfer", "Cryptococcal mannosyltransferase 1; pfam11735"
<i>Cryptococcus</i> JEC21	<i>neoformans</i> 58259209	capsular associated protein	"CAP59_mtransfer", "Cryptococcal mannosyltransferase 1; pfam11735"



Table S7: **AS affected genes involved in stress response.** Relation is based on blast similarity to *P. brasiliensis* Pb01 genes TPS1 (NCBI accession EEH35656), HSP30 (EEH37950), and DDR48 (EEH33596).

species	protein ID	definition	note
<b>TPS1-relation</b>			
<i>Cryptococcus neoformans</i> B-3501A	134108248	hypothetical protein CNBB3070	"Glycosyltransferase.GTB_type", "GT1_TPS", "Trehalose-6- Phosphate Synthase (TPS)"
<i>Cryptococcus neoformans</i> B-3501A	134116029	hypothetical protein CNBI3400	"Glycosyltransferase.GTB_type", "GT1_TPS", "Trehalose-6- Phosphate Synthase (TPS)"
<i>Cryptococcus neoformans</i> JEC21	58264051	trehalose- phosphatase	"Glycosyltransferase.GTB_type", "GT1_TPS", "Trehalose-6- Phosphate Synthase (TPS)"
<i>Cryptococcus neoformans</i> JEC21	58270702	alpha,alpha- trehalose-phosphate synthase (UDP- forming)	"Glycosyltransferase.GTB_type", "GT1_TPS", "Trehalose-6- Phosphate Synthase (TPS)"
<i>Laccaria bicolor</i>	170098941	alpha,alpha- trehalose-phosphate synthase subunit	"alpha,alpha-trehalose-phosphate synthase TPS1 subunit", Glycosyl- transferase family 20; pfam00982"
<i>Neurospora crassa</i>	164428605	hypothetical protein NCU00793	"hypothetical protein", "similar to alpha,alpha-trehalose phos- phate synthase subunit TPS3", "Glyco.transf_20"
<i>Podospora anserina</i>	171695706	hypothetical protein	"Glyco.transf_20", "GT1_TPS", "Trehalose-6-Phosphate Synthase (TPS)"
<i>Trichoderma reesei</i>	48707	alpha,alpha- trehalose-phosphate synthase	catalyzes UDP-glucose + D-glucose- 6-phosphate = UDP + alpha,alpha- trehalose-6-phosphate
<b>HSP30-relation</b>			
<i>Coccidioides immitis</i>	119194749	30 kDa heat shock protein	"IbpA", "Molecular chaperone (small heat shock protein)"
<i>Laccaria bicolor</i>	170101017	predicted protein	"ACD_sHsps-like", "Alpha- crystallin domain (ACD) of alpha-crystallin-type small(s) heat shock proteins (Hsps)"
<i>Ustilago maydis</i>	71019595	hypothetical protein UM03881.1	"IbpA", Molecular chaperone (small heat shock protein)"
<b>DDR48-relation</b>			
<i>Ajellomyces capsulatus</i>	154277766	hypothetical protein HCAG_05184	"similar to potential stress response protein", "PTZ00110", "helicase; Provisional"
<i>Coccidioides immitis</i>	119192856	predicted protein	"hypothetical protein"
<i>Cryptococcus neoformans</i> B-3501A	134108310	hypothetical protein CNBB3380	"Glycosyltransferase.GTB_type"
<i>Mycosphaerella graminicola</i>	103686	-	-

Table S8: NMD-related components and their homologs in NCBI HomoloGene database. Note, HomoloGene contains data of only six fungi (*S. cerevisiae*, *K. lactis*, *S. pombe*, *M. oryzae*, *N. crassa*, *E. gossypii*), hence the limited coverage of species.

NMD factor	Homolo- Gene ID	gene	description	species
<b>UPF1</b>	2185	NAM7	Nam7p	<i>S.cerevisiae</i>
		KLLA0B06435g	hypothetical protein	<i>K.lactis</i>
		upf1	ATP-dependent RNA helicase Upf1	<i>S.pombe</i>
		MGG_00976	regulator-nonsense transcripts 1	<i>M.oryzae</i>
		NCU04242	ATP-dependent helicase NAM7	<i>N.crassa</i>
<b>UPF2</b>	6101	AGOS_ABR022C	ABR022Cp	<i>E.gossypii</i>
		KLLA0D13156g	hypothetical protein	<i>K.lactis</i>
		NMD2	Nmd2p	<i>S.cerevisiae</i>
		upf2	nonsense-mediated decay protein Upf...	<i>S.pombe</i>
		MGG_06063	nonsense-mediated mRNA decay factor	<i>M.oryzae</i>
<b>UPF3</b>	11307	SPAC13G7.03	hypothetical protein	<i>N.crassa</i>
		39098	UPF3	<i>S.pombe</i>
	126047	KLLA0D03718g	UPf3p	<i>S.cerevisiae</i>
		AGOS_AER204W	hypothetical protein	<i>K.lactis</i>
		AER204Wp	<i>E.gossypii</i>	
<b>CBP80</b>	68864	MGG_03912	hypothetical protein	<i>M.oryzae</i>
		NCU03435	hypothetical protein	<i>N.crassa</i>
		STO1	Sto1p	<i>S.cerevisiae</i>
		KLLA0F17523g	hypothetical protein	<i>K.lactis</i>
		AGOS_AFR218W	AFR218Wp	<i>E.gossypii</i>
<b>CBP20</b>	103828	SPAC6G10.07	nuclear cap-binding complex large s...	<i>S.pombe</i>
		MGG_12123	cap binding protein	<i>M.oryzae</i>
		NCU04187	hypothetical protein	<i>N.crassa</i>
		CBC2	Cbc2p	<i>S.cerevisiae</i>
		KLLA0B10472g	hypothetical protein	<i>K.lactis</i>
<b>Y14</b>	3744	AGOS_AFL050W	AFL050Wp	<i>E.gossypii</i>
		SPBC13A2.01c	nuclear cap-binding complex small s...	<i>S.pombe</i>
		MGG_06296	nuclear cap-binding protein subunit...	<i>M.oryzae</i>
		NCU00210	nuclear cap binding protein subunit...	<i>N.crassa</i>
		SPAC23A1.09	RNA-binding protein	<i>S.pombe</i>
<b>BTZ</b>	127412	MGG_03740	RNA-binding protein 8A	<i>M.oryzae</i>
		NCU03226	hypothetical protein	<i>N.crassa</i>
		MGG_00982	hypothetical protein	<i>M.oryzae</i>
<b>eIF4AIII</b>	5602	NCU04270	hypothetical protein	<i>N.crassa</i>
		FAL1	Fal1p	<i>S.cerevisiae</i>
		KLLA0A10659g	hypothetical protein	<i>K.lactis</i>
		AGOS_AER408W	AER408Wp	<i>E.gossypii</i>
		SPAC1F5.10	ATP-dependent RNA helicase, eIF4A r...	<i>S.pombe</i>
<b>MAGOH</b>	127412	MGG_04885	ATP-dependent RNA helicase fal-1	<i>M.oryzae</i>
		NCU01234	eukaryotic initiation factor 4A-12	<i>N.crassa</i>
		MGG_00982	hypothetical protein	<i>M.oryzae</i>
		NCU04270	hypothetical protein	<i>N.crassa</i>
		SPBC3B9.08c	protein mago nashi	<i>S.pombe</i>
	56794	MGG_06859	mago nashi like 2	<i>M.oryzae</i>
		NCU04405	mago nashi protein	<i>N.crassa</i>

<b>SMG1</b>	44191	smg1 MGG_10740 NCU09880	Sm snRNP core protein Smg1 sm snRNP core protein Smg1 hypothetical protein	<i>S. pombe</i> <i>M. oryzae</i> <i>N. crassa</i>
<b>PYM</b>	no entry for fungi			
<b>hNAG</b>	no entry for fungi			
<b>DHX34</b>	no entry for fungi			

## Supplementary Calculation S1

We downloaded all reads of *S. pombe* of the study on fission yeasts (24) from NCBI's short read archive. We converted the reads to fastq format with fastq-dump 2.0.5 using standard parameters (25). To estimate the number of useful reads we filtered them with the "lite" version of PRINSEQ (26) with the following strict filter settings<sup>3</sup>:

```
prinseq-lite.pl -fastq in.fastq -trim_qual_left 25 -trim_qual_right 25 -min_len 50
  -min_qual_mean 25 -ns_max_p 1 -noniupac -lc_method entropy -lc_threshold 70
  -out_format 3 -out_good good_reads.fastq -out_bad bad_reads.fastq
```

This left 261,459,213 of the overall 307,223,097 reads (85%). The more reads there are available, the more AS events can be detected and the higher is the "raw" AS rate of a species, i.e. dividing all detected AS events by the number of genes. On the other hand, random sampling normalizes the rate. This is why we found a slightly negative correlation (-0.31) of the number of reads with the ratio between the AS rate from random sampling and the raw AS rate. For the species with the most EST data in our study (*A. benhamiae*, 1,040,774 NGS reads), this ratio is 0.3, that means, the AS rate from random sampling (8.2%) is around one third of the raw AS rate (27.4%). Similar, we assume the raw AS rate of *S. pombe* from the Rhind et al. study (8.4% = 433 AS events/5144 genes) is a strong over-estimation of the real rate. We suppose that the rate from random sampling would be much smaller than 2.5% (0.3 x 8.4%). This is because the use of over 250 times more reads than for *A. benhamiae* likely has revealed many AS events with a strong expression bias towards one isoform, which are not accounted for in random sampling. Thus, the AS rate of *S. pombe* is clearly lower the mean AS rate of non-yeast Ascomycota (7.2%).

---

<sup>3</sup>For explanation of PRINSEQ parameters see <http://prinseq.sourceforge.net/manual.html>

## Supplementary Figures

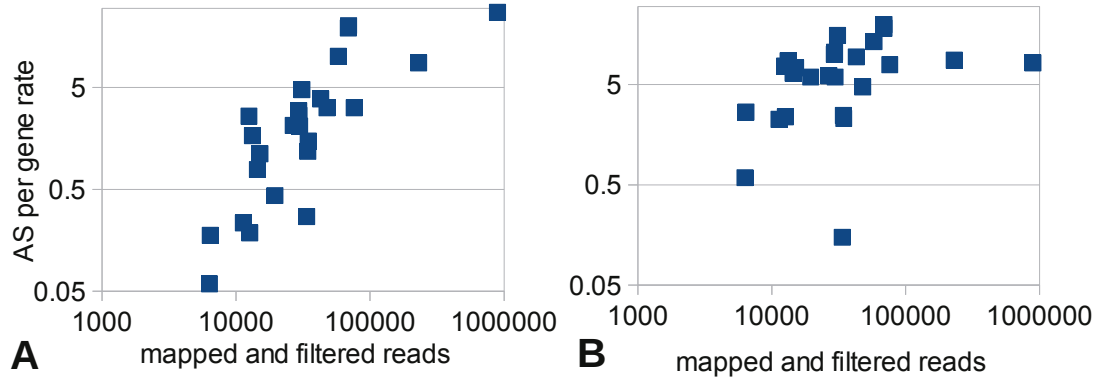


Figure S1: **Dependence of alternative splicing rates on read amounts.** Each point in the diagrams represents data of one species. (A) AS rate from dividing the number of AS events by the annotated gene number. (B) AS rate as average from repeated random sampling of transcripts. Note the logarithmic scaling of the axes.

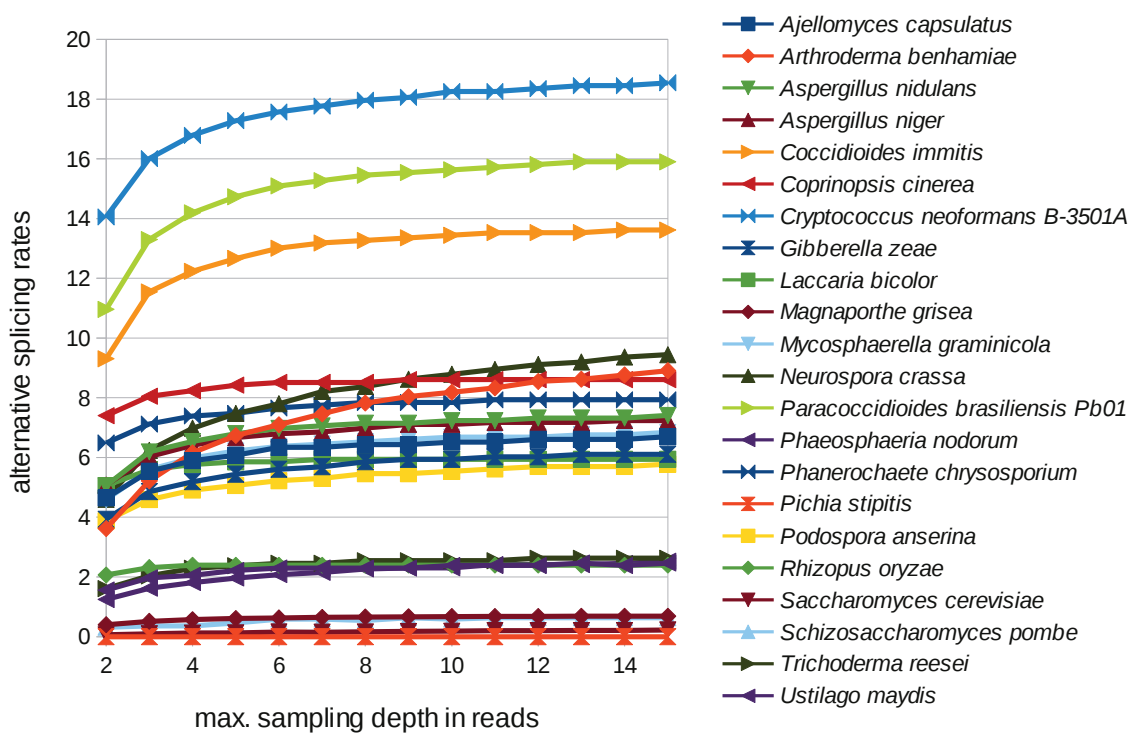
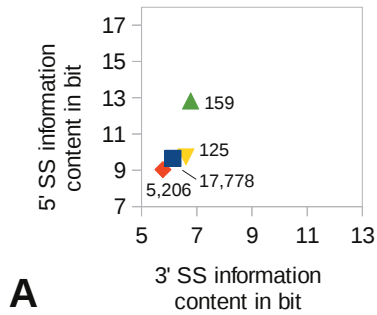


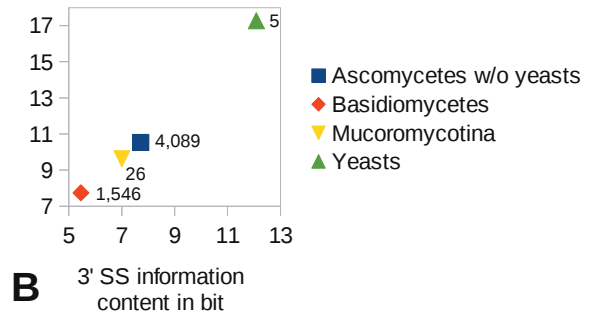
Figure S2: AS rates versus maximal randomly sampled reads per locus.

constitutively spliced introns



**A**

retained introns



**B**

Figure S3: **Mean splice site conservation** Splice site (SS) conservation is calculated as information content in bits as follows: For each intron and SS type (3' and 5'), the SS regions of all introns were stacked. The logarithm of each base's frequency times the frequency is added and summed over all sequence positions. Numbers of underlying introns are noted besides the chart symbols.

## References

- [1] Aidé MA (2009) Chapter 4–histoplasmosis. *J Bras Pneumol* 35: 1145–1151.
- [2] Burmester A, Shelest E, Gloeckner G, Heddergott C, Schindler S, et al. (2011) Comparative and functional genomics provide insights into the pathogenicity of dermatophytic fungi. *Genome Biol* 12: R7.
- [3] Bennett JW (2009) *Aspergillus: a primer for the novice*. *Med Mycol* 47 Suppl 1: S5–12.
- [4] Choquer M, Fournier E, Kunz C, Levis C, Pradier JM, et al. (2007) *Botrytis cinerea* virulence factors: new insights into a necrotrophic and polyphageous pathogen. *FEMS Microbiol Lett* 277: 1–10.
- [5] de Deus Filho A (2009) Chapter 2: *coccidioidomycosis*. *J Bras Pneumol* 35: 920–930.
- [6] Michielse CB, Rep M (2009) Pathogen profile update: *Fusarium oxysporum*. *Mol Plant Pathol* 10: 311–324.
- [7] McMullen M, Jones R, Gallenberg D (1997) Scab of wheat and barley: A re-emerging disease of devastating impact. *Plant Disease* 81: 1340 - 1348.
- [8] Ribot C, Hirsch J, Balzergue S, Tharreau D, Nottéghem JL, et al. (2008) Susceptibility of rice to the blast fungus, *magnaporthe grisea*. *J Plant Physiol* 165: 114–124.
- [9] Bowler J, Scott E, Tailor R, Scalliet G, Ray J, et al. (2010) New capabilities for *mycosphaerella graminicola* research. *Mol Plant Pathol* 11: 691–704.
- [10] Ray CG, Ryan KJ, editors (2004) *Sherris Medical Microbiology*. MCGRAW-HILL, MEDICAL PUBLISHING DIVISION, 4 edition.
- [11] Stukenbrock EH, Banke S, McDonald BA (2006) Global migration patterns in the fungal wheat pathogen *phaeosphaeria nodorum*. *Mol Ecol* 15: 2895–2904.
- [12] Agbogbo FK, Coward-Kelly G (2008) Cellulosic ethanol production using the naturally occurring xylose-fermenting yeast, *pichia stipitis*. *Biotechnol Lett* 30: 1515–1524.
- [13] Paoletti M, Saupe SJ (2008) The genome sequence of *podospora anserina*, a classic model fungus. *Genome Biol* 9: 223.
- [14] Olsson I, Bjerling P (2011) Advancing our understanding of functional genome organisation through studies in the fission yeast. *Curr Genet* 57: 1–12.
- [15] Hegedus DD, Rimmer SR (2005) *Sclerotinia sclerotiorum*: when "to be or not to be" a pathogen? *FEMS Microbiol Lett* 251: 177–184.
- [16] Schuster A, Schmoll M (2010) Biology and biotechnology of *trichoderma*. *Appl Microbiol Biotechnol* 87: 787–799.
- [17] McKnight KH, McKnight VB, Peterson RT (1998) *A Field Guide to Mushrooms: North America*. Houghton Mifflin Harcourt.
- [18] Sidrim JJC, Costa AKF, Cordeiro RA, Brillhante RSN, Moura FEA, et al. (2010) Molecular methods for the diagnosis and characterization of *cryptococcus*: a review. *Can J Microbiol* 56: 445–458.
- [19] Martin F, Nehls U (2009) Harnessing ectomycorrhizal genomics for ecological insights. *Curr Opin Plant Biol* 12: 508–515.



- [20] Singh D, Chen S (2008) The white-rot fungus *phanerochaete chrysosporium*: conditions for the production of lignin-degrading enzymes. *Appl Microbiol Biotechnol* 81: 399–417.
- [21] Brefort T, Doehlemann G, Mendoza-Mendoza A, Reissmann S, Djamei A, et al. (2009) *Ustilago maydis* as a pathogen. *Annu Rev Phytopathol* 47: 423–445.
- [22] Gonzalez CE, Rinaldi MG, Sugar AM (2002) Zygomycosis. *Infect Dis Clin North Am* 16: 895–914, vi.
- [23] Coenen TM, Aughton P, Verhagen H (1997) Safety evaluation of lipase derived from *rhizopus oryzae*: summary of toxicological data. *Food Chem Toxicol* 35: 315–322.
- [24] Rhind N, Chen Z, Yassour M, Thompson DA, Haas BJ, et al. (2011) Comparative functional genomics of the fission yeasts. *Science* 332: 930–936.
- [25] NCBI (2009). SRA Handbook [Internet]. URL <http://www.ncbi.nlm.nih.gov/books/NBK47540/>.
- [26] Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27: 863–864.

## 2. PUBLICATIONS

---

### 2.4 Mutually exclusive spliced exons show non-adjacent and grouped patterns

In Pohl et al. 2009<sup>121</sup> we present an analysis of human and mouse mutually exclusive spliced exons based on mappings of transcript sequences to genomic sequences. We detected more than 1000 MXEs per species, and report, to our knowledge for the first time, a genome-wide frequent presence of non-adjacent and cluster-spliced MXEs. As these special types comprise more than 95% of the detected events, the suitability of existing regulatory models of MXE splicing for mammals are questioned.

## Mutually exclusive spliced exons show non-adjacent and grouped patterns

Martin Pohl<sup>1,4</sup>, Dirk Holste<sup>2</sup>, Ralf Bortfeldt<sup>3</sup>, Konrad Grützmann<sup>1</sup> and Stefan Schuster<sup>1</sup>

<sup>1</sup>Department of Bioinformatics, Friedrich-Schiller University of Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany

<sup>2</sup>Austrian Institute of Technology, Donau-City-Straße 1, A-1220 Vienna, Austria

<sup>3</sup>Breeding Biology and Molecular Genetics, Humboldt-University, Invalidenstrasse 42, 10115 Berlin, Germany

<sup>4</sup>To whom correspondence should be addressed. E-mail: [m.pohl@uni-jena.de](mailto:m.pohl@uni-jena.de)

### Abstract

The deciphering of the human genome provides the base for many bioinformatic methods researching genome related matters. Splicing of mRNA is one of the processes which depend on the genomic sequence. We present an analysis of the results from a method that detects mutually exclusive spliced exons in the genome. The method is based on transcript data mapped to the genomic sequence. Applying it to human as well as mouse genome we detected more than 1000 mutually exclusive spliced exon pairs per species. The events we analysed broaden the view on mutually exclusive regulated splicing and reveal some unexpected characteristics.

### 1 Introduction

Alternative splicing (AS) of pre-mRNAs plays many different roles and in higher eukaryotes it is one regulatory mechanism for tissue-specific protein variability [Bla00; Gra01]. Estimations of the scope of AS are often different. Most recent ones (based on short cDNA reads/mRNA-Seq data) show almost every (95%) human multi-exon gene to undergo AS [Wan+08]. Traditionally, four main AS patterns are considered: exon skipping, alternative acceptor or donor site, intron retention [Fer+07; Kim+08]. Among the different types of AS, mutually exclusive exons (MXEs) constitute a comparatively rare, but very intriguing type [Sam+08]. In its simple form, two internal exons are spliced in such a way that one exon is excised, while the other one is kept

within the mature mRNA (and vice versa), thereby linking the regulation of different exons. So far, MXEs have been usually assumed to be genomic adjacent [Hol+06; Nag+06; Zhe+05]. One hypothesis about its functional role is the selection of context specific sites within protein domains. Splicing of MXEs has not been analysed in as much detail as more frequent types of AS, e.g., exon-skipping or alternative exon boundaries [Bor+08; Hil+04; ML03]. Nonetheless, such AS events contribute to our understanding of this still largely unclear phenomenon. Several regulatory mechanisms for MXEs have been proposed, including spliceosome incompatibility, steric hindrance, docker-selector mechanism, or the regulation by splicing factors coupled with nonsense-mediated decay [Smi05]. In this study we used a computational approach to infer about two thousand MXEs across more than 20,000 of the mapped genes for human and mouse genome, respectively. We then compared and contrasted our data for their compatibility with known AS patterns as well as genomic features predicted by existing models for the regulation of MXEs.

## 2 Results

All analyses are based on annotations of protein coding genomic regions retrieved from Ensembl (<http://www.ensembl.org/>). Onto these regions, we mapped transcripts from the UCSC database (<http://genome.ucsc.edu/>; 4.8 million ESTs and 150,000 Fl-mRNAs). For the mapping we relied on the transcript to genome alignments provided by UCSC [Kuh+09]. We considered a pair of exons to be mutually exclusive if all transcripts, mapped such that they span the genomic region of both exons, exactly one of the exons is included (for further details on mapping, filtering and constraints please contact authors). For about 20,000 human genes, we inferred 1,300 MXE pairs (3.4% of all genes). In a similar study, we inferred 1,200 MXE pairs (3.3% of all genes) for over 21,000 mouse genes. Comparing the detected pattern and their features (genomic, transcriptional) with existing regulatory models, we found for both studies that nearly all MXEs were not adjacent with respect to their genomic location (99%, cf. Fig. 1.1), which is in contrast to the predicted patterns of existing models. Furthermore, while we could not infer complex patterns, such as docker-selector sites in the fruitfly gene *Dscam* [Ana+06], we found MXEs to be involved in mutually exclusive splicing of more than one exon simultaneously (cluster-spliced exons, cf. Fig. 1.1). None of the

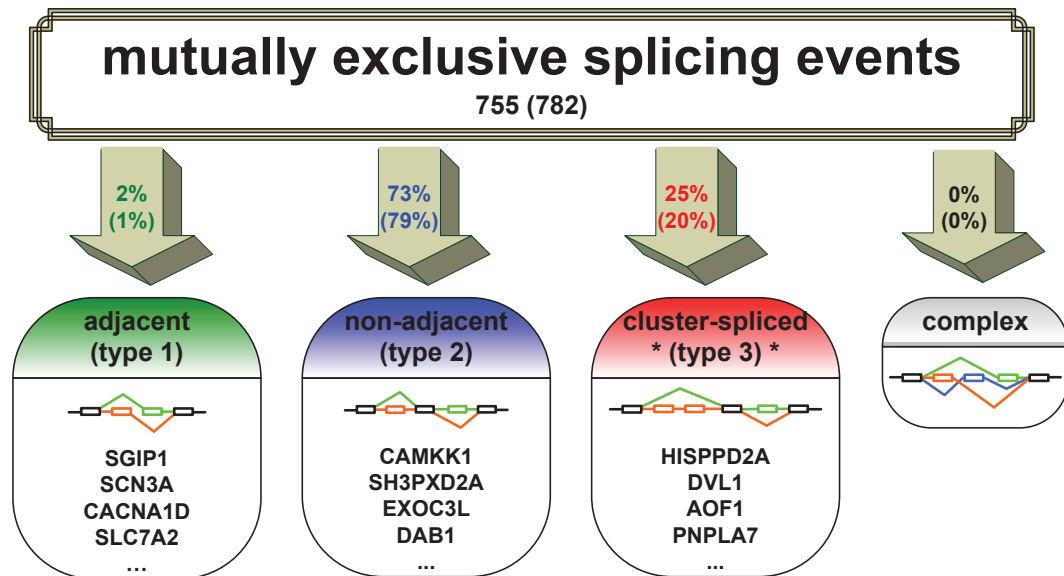


Figure 1.1: Representation of human (mouse) MXEs inferred within this study. We divided MXE events into four distinct groups based on known regulatory mechanisms and detected splicing pattern. Examples of splicing pattern are shown at the bottoms of each type. Left hand side: three quarter of all inferred mutually exclusive splicing events comprised single alternative exons, split into adjacent (type 1) and non-adjacent (type 2) patterns. Middle: one quarter of inferred events featured the newly introduced mutually exclusive splicing of exon clusters (type 3). Right hand side: we did not observe more complex patterns, e.g., like the selection of an exon, singled-out from a cluster of possible MXEs in the *Dscam* gene of the fruitfly.

inferred MXEs obsessed the /AT and AC\splice site motifs for the minor spliceosome. At first glance, checking for orthologous genes we could infer for 21% (11%) of human genes with adjacent events (non-adjacent) an event for mouse as well. But a closer look at exon specific sequence conservation remains for a more substantive reasoning on conservation of events. While we could find some previously reported events as for voltage-gated sodium and calcium channel protein alpha subunits or human glycine receptor a2 gene (SCNA, CACNA1d, GLRA2) there is as well a number of reported events not among our results like KCNMA1, TCL6 [Cop04; Sam+08]. This is in particu-

lar due to that mutually exclusive properties are frequently defined on tissue specific or transcript pair context while our approach requires global mutual exclusivity. Frame shift analysis revealed higher frame conservation for adjacent than for non-adjacent MXEs. For the latter ones we also found a variable number of enclosed constitutive exons reaching up to 67 with a strong bias towards two.

### 3 Conclusions

One central outcome of this study is that non-adjacent exons constitute the most frequent MXE event. Consequently, the inferred AS patterns, determined by transcript alignments, do not further substantiate existing models and hence challenge their suitability as widely functioning mechanisms for mutually exclusive splicing in higher eukaryotes, e.g., mammals. The model incorporating spliceosome incompatibility found no evidence in this study, the model incorporating steric hindrance is most conceivable for small introns intervening MXEs, while the model incorporating docker-selector sites is not conceivable for non-adjacent MXEs. The regulation by splicing-factors accompanied by the NMD mechanism remains a candidate for adjacent MXEs and ought to be validated in future studies. Regulation of the new subtype of cluster-spliced MXEs cannot be explained well by current models and new hypotheses for AS regulation are necessary. Evolutionary origin, effects on evolvability [Che+06], splice site recognition as well as open questions raised by our results will be discussed.

### References

- [Ana+06] D. Anastassiou et al. “Variable window binding for mutually exclusive alternative splicing”. In: *Genome Biology* 7.1 (2006), R2. DOI: 10.1186/gb-2006-7-1-r2.
- [Bla00] D. L. Black. “Protein diversity from alternative splicing: a challenge for bioinformatics and postgenome biology”. In: *Cell* 103.3 (2000), pp. 367–370. DOI: 10.1016/S0092-8674(00)00128-8.
- [Bor+08] R. Bortfeldt et al. “Comparative analysis of sequence features involved in the recognition of tandem splice sites”. In: *BMC Genomics* 9.1 (2008), p. 202. DOI: 10.1186/1471-2164-9-202.

- [Che+06] F.-C. Chen et al. “Alternatively and Constitutively Spliced Exons Are Subject to Different Evolutionary Forces”. In: *Molecular Biology and Evolution* 23.3 (2006), pp. 675–682. DOI: 10.1093/molbev/msj081.
- [Cop04] R. R. Copley. “Evolutionary convergence of alternative splicing in ion channels”. In: *Trends in Genetics* 20.4 (2004), pp. 171–176. DOI: 10.1016/j.tig.2004.02.001.
- [Fer+07] E. N. Ferreira et al. “Alternative splicing: a bioinformatics perspective”. In: *Molecular BioSystems* 3.7 (2007), pp. 473–477. DOI: 10.1039/b702485c.
- [Gra01] B. R. Graveley. “Alternative splicing: increasing diversity in the proteomic world”. In: *Trends in Genetics* 17.2 (2001), pp. 100–107. DOI: 10.1016/S0168-9525(00)02176-4.
- [Hil+04] M. Hiller et al. “Widespread occurrence of alternative splicing at NAG-NAG acceptors contributes to proteome plasticity”. In: *Nature Genetics* 36.12 (2004), pp. 1255–1257. DOI: 10.1038/ng1469.
- [Hol+06] D. Holste et al. “HOLLYWOOD: a comparative relational database of alternative splicing”. In: *Nucleic Acids Research* 34.suppl\_1 (2006), pp. D56–62. DOI: 10.1093/nar/gkj048.
- [Kim+08] E. Kim et al. “Alternative splicing: current perspectives”. In: *BioEssays* 30.1 (2008), pp. 38–47. DOI: 10.1002/bies.20692.
- [Kuh+09] R. M. Kuhn et al. “The UCSC Genome Browser Database: update 2009”. In: *Nucleic Acids Research* 37.suppl\_1 (2009), pp. D755–761. DOI: 10.1093/nar/gkn875.
- [ML03] B. Modrek and C. J. Lee. “Alternative Splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss”. In: *Nature Genetics* 34.2 (2003), pp. 177–80. DOI: 10.1038/ng1159.
- [Nag+06] H. Nagasaki et al. “Automated classification of alternative splicing and transcriptional initiation and construction of visual database of classified patterns”. In: *Bioinformatics* 22.10 (2006), pp. 1211–1216. DOI: 10.1093/bioinformatics/bt1067.
- [Sam+08] M. Sammeth et al. “A General Definition and Nomenclature for Alternative Splicing Events”. In: *PLoS Computational Biology* 4.8 (2008), e1000147. DOI: 10.1371/journal.pcbi.1000147.
- [Smi05] C. W. J. Smith. “Alternative Splicing — When Two’s a Crowd”. In: *Cell* 123.1 (2005), pp. 1–3. DOI: 10.1016/j.cell.2005.09.010.

- [Wan+08] E. T. Wang et al. “Alternative isoform regulation in human tissue transcriptomes”. In: *Nature* 456.7221 (2008), pp. 470–476. DOI: 10.1038/nature07509.
- [Zhe+05] C. L. Zheng et al. “MAASE: An alternative splicing database designed for supporting splicing microarray applications”. In: *RNA* 11.12 (2005), pp. 1767–1776. DOI: 10.1261/rna.2650905.



## 2.5 Alternative splicing of mutually exclusive exons – A review

In Pohl et al. 2013<sup>122</sup> we review the current knowledge of MXE splicing. Emphasis is put on bioinformatics methods to detect MXEs, as well as on definitions and nomenclatures of this AS type. Molecular mechanisms of MXE splicing are discussed, especially in the light of adjacent, non-adjacent, grouped, and cluster-spliced subtypes of MXEs.



Contents lists available at ScienceDirect

BioSystems

journal homepage: [www.elsevier.com/locate/biosystems](http://www.elsevier.com/locate/biosystems)

Review Article

## Alternative splicing of mutually exclusive exons—A review

Martin Pohl<sup>a,\*</sup>, Ralf H. Bortfeldt<sup>b</sup>, Konrad Grützmann<sup>a</sup>, Stefan Schuster<sup>a</sup><sup>a</sup> Department of Bioinformatics, Friedrich Schiller University of Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany<sup>b</sup> Department of Crop and Animal Sciences, Humboldt University Berlin, Invalidenstrasse 42, 10115 Berlin, Germany

## ARTICLE INFO

Article history:  
Received 30 May 2013  
Accepted 3 July 2013

Keywords:  
Alternative splicing  
Exon clusters  
Mutually exclusive exons  
Non-sense mediated decay  
Splicing mechanisms

## ABSTRACT

Alternative splicing (AS) of pre-mRNAs in higher eukaryotes and several viruses is one major source of protein diversity. Usually, the following major subtypes of AS are distinguished: exon skipping, intron retention, and alternative 3' and 5' splice sites. Moreover, mutually exclusive exons (MXEs) represent a rare subtype. In the splicing of MXEs, two (or more) splicing events are not independent anymore, but are executed or disabled in a coordinated manner. In this review, several bioinformatics approaches for analyzing MXEs are presented and discussed. In particular, we revisit suitable definitions and nomenclatures, and bioinformatics tools for finding MXEs, adjacent and non-adjacent MXEs, clustered and grouped MXEs. Moreover, the molecular mechanisms for splicing MXEs proposed in the literature are reviewed and discussed.

© 2013 Elsevier Ireland Ltd. All rights reserved.

## Contents

1. Introduction .....	31
1.1. Molecular biology background .....	31
1.2. Bioinformatics resources for analyzing MXE splicing .....	32
2. Nomenclatures and definitions .....	33
2.1. Established approaches .....	33
2.2. Adjacent MXEs and non-adjacent dependencies .....	34
2.3. A Boolean nomenclature .....	34
3. Detection .....	34
4. Mechanisms leading to mutual exclusion of exons .....	35
5. Evolutionary conservation .....	36
6. Conclusions .....	36
Acknowledgements .....	36
References .....	36

## 1. Introduction

## 1.1. Molecular biology background

In addition to the genetic code, several other codes are used by the living cell at the molecular level, for example, the calcium

oscillation code and the code used for signaling among plants by volatile chemicals. In eukaryotes one of these is the splicing code, by which the cell decides which sequence parts are finally used (Choudhary and Krithivasan, 2007; Barbieri, 2008; Barash et al., 2010; Reddy et al., 2012).

In the post-genomic era, alternative splicing (AS) of pre-mRNAs in higher eukaryotes got in the focus of research as one major source of protein diversity (Black, 2000; Graveley, 2001; Kim et al., 2008; Nilsen and Graveley, 2010; Chen et al., 2012a). AS was discovered in adenoviruses (Berget et al., 1977) and also occurs in several other viruses such as cytomegalovirus (Gatherer et al., 2011). Protein variability contributes to a high complexity of higher eukaryotes while keeping the numbers of genes relatively low. AS is a means to change proteins, in dependence on gender, developmental stage or environmental conditions and can affect binding

*Abbreviations:* AS, alternative splicing or alternatively spliced; MXE, mutually exclusive exon; EST, expressed sequence tag; NMD, nonsense mediated mRNA decay; ORF, open reading frame; *Dscam*, *Drosophila* Down Syndrome Cell Adhesion Molecule.

\* Corresponding author. Tel.: +49 3641 949581.

E-mail addresses: [m.pohl@uni-jena.de](mailto:m.pohl@uni-jena.de) (M. Pohl), [ralf.bortfeldt@agrar.hu-berlin.de](mailto:ralf.bortfeldt@agrar.hu-berlin.de) (R.H. Bortfeldt), [konrad.g@uni-jena.de](mailto:konrad.g@uni-jena.de) (K. Grützmann), [stefan.schu@uni-jena.de](mailto:stefan.schu@uni-jena.de) (S. Schuster).

0303-2647/\$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.  
<http://dx.doi.org/10.1016/j.biosystems.2013.07.003>

properties, intracellular localization, enzymatic activity and many more properties of proteins (Stamm et al., 2005; Yap and Makeyev, 2013). Estimations raised from one third up to 95% of human genes affected by AS, with other mammals showing similar high AS levels (Florea, 2006; Pan et al., 2008; Wang et al., 2008). Alternative splicing and splicing in general is a major problem in gene finding in eukaryotes because it may disrupt ORFs (Pohl et al., 2012).

The potential for variability is enormous. For instance, the human calcium-activated potassium channel subunit alpha-1 gene and the three neurexin genes could potentially generate 500 and more than 2000 different protein isoforms, respectively, by different ways of splicing (Black, 1998; Tabuchi and Südhof, 2002). The *Drosophila* Down Syndrome Cell Adhesion Molecule gene (*Dscam*) has several sets of cassette exons with one of them involving 48 alternative exons among which one is selected (Graveley, 2005; Anastassiou et al., 2006; Meijers et al., 2007; Olson et al., 2007; Hemani and Soller, 2012; Wang et al., 2012). This leads to 38,016 theoretical splicing variants.

AS is thought to lower selective pressure on gene sequences allowing a higher trial and error rate by mutations in one of the isoforms without compromising the acquired functionality of the other isoform (Boué et al., 2003; Chen et al., 2006; Noh et al., 2006). The apparent evolutionary advantages of AS require, however, significant energetic and metabolic costs because the spliceosome, which performs the splicing reaction, is a large complex of proteins and RNA including up to several hundreds of constituents (Jurica and Moore, 2003; Kielbassa et al., 2009; Bortfeldt et al., 2010; Hoskins et al., 2011). Given the enormous effort to assemble such complicated molecular machinery it can be assumed that the benefit of transcript flexibility outweighs the biochemical costs. In contrast, some organisms such as many plants, seem to have achieved their level of protein variability mainly by gene duplications i.e., an increase in genome length (Kopelman et al., 2005).

The ability to cope with stress is widely enhanced via transcriptome plasticity (Mastrangelo et al., 2012). Moreover, the involvement and prevalence of AS in many diseases is becoming increasingly clear. Hence, protein variability as generated by alternative splicing is of great medical and biotechnological importance because different isoforms are often associated with diseases such as cancer (Hernandez-Lopez and Graham, 2012) or with the distinction between intracellular and extracellular enzymes (Andreassi and Riccio, 2009). This renders AS and its regulation a potential therapeutic target (Mount and Pandey, 2005; Garcia-Blanco, 2006; He et al., 2009; Tazi et al., 2009, 2010; Douglas and Wood, 2011; Germann et al., 2012; Hernandez-Lopez and Graham, 2012; Sanchez-Pla et al., 2012).

Several attempts for general AS annotations have been presented (Xing et al., 2004; Nagasaki et al., 2006; Sammeth et al., 2008; Kroll et al., 2012). Among the well-known subtypes of AS are exon skipping (Sorek et al., 2004b), intron retention (Wang et al., 2006), alternative 5' splice sites (Dou et al., 2006; Bortfeldt et al., 2008; Hiller and Platzer, 2008), alternative 3' splice sites (Bortfeldt et al., 2008; Hiller and Platzer, 2008). A less abundant subtype of AS is represented by mutually exclusive exon (MXE) splicing.

MXEs are characterized by splicing of exons in a coordinated manner such that two or more splicing events are not independent. As the name "mutually exclusive" indicates, exactly one out of two exons (or one group out of two exon groups) is retained, while the other one is spliced out. Sammeth (2009) applies the term in a less strict way, allowing the case that none or all of the exons under consideration are retained. In contrast to other variants of alternative splicing, mutually exclusive splicing can leave the size of the protein unchanged provided that the exchanged sequence is of the same length and does not introduce a premature stop codon. Depending on the similarity of exchanged exon sequences, minor changes as in subtle alternative 5' and 3' splicing events or major changes

of whole protein domains as in exon skipping are possible. In case of minor protein sequence changes, MXEs may provide an advantage to many types of proteins, such as ion channels, because the spatial structure is preserved, while the protein exhibits an altered function (Birzele et al., 2008a). Interestingly, another RNA processing mechanism, RNA editing, can also occur in a mutually exclusive manner as shown for the *TPH2* gene (Grohmann et al., 2010) resulting in a similar effect as mutually exclusive exon splicing.

A common assumption is that MXEs have originated from exon duplication and, hence, are highly similar (Letunic et al., 2002; Copley, 2004; Sorek, 2009; Pillmann et al., 2011). Accordingly, some authors (Stephan et al., 2007; Pillmann et al., 2011) define MXEs based on similar length and sequence. In our opinion, these criteria are not necessary. The term "mutually exclusive" only implies that exons do not occur together but does not refer to length, sequence or exon numbers. In general, also a group (cluster) of exons can be mutually exclusive with respect to another group (cluster) of exons. Such cases should be distinguished from exon cassettes where exactly one out of several exons is retained in the mature transcript, such as in the *Dscam* gene in *Drosophila*. However, the terminology is not used consistently among researchers, MXE were previously also termed as "exon clusters" (Pillmann et al., 2011) or "cassette exons" (Stephan et al., 2007).


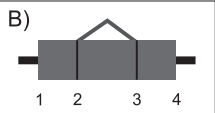
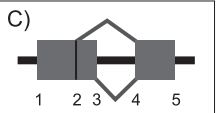
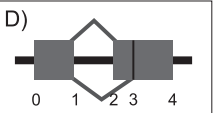
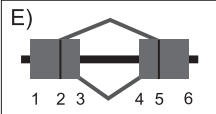



MXEs turned out to be very promising candidates for generation of highly diverse but specific processes (Anastassiou et al., 2006; Soom et al., 2008). The alternative selection of exons enables the encoding of a whole class of proteins with similar scaffold and similar length but with highly specific functionality. Beside the above-mentioned *Drosophila Dscam* gene, examples of biological relevance are provided by the voltage dependence of ion channels (Soom et al., 2008) and calcium sensitivity of muscle proteins in higher animals (Waites et al., 1992). Like other AS types, MXEs proved to be of medical relevance, e.g., at regulation of expression levels of the mammalian pyruvate kinase M isoforms (Chacko and Ranganathan, 2009b; Chen et al., 2012b). Examples of MXEs have been described in human (Soom et al., 2008), mouse (Chacko and Ranganathan, 2009a), rat (Gustafson et al., 1993), chicken (Waites et al., 1992; Chacko and Ranganathan, 2009a), cow (Chacko and Ranganathan, 2009b), nematode (Johnson et al., 2003) and other species.

## 1.2. Bioinformatics resources for analyzing MXE splicing

As biochemical analyses are expensive and time consuming, computational approaches have attracted an ever increasing interest. Accordingly, AS is an important topic in bioinformatics (Dou et al., 2006; Zavolan and van Nimwegen, 2006; Hiller et al., 2007; Bortfeldt et al., 2008; Hiller and Platzer, 2008; Sammeth et al., 2008; Busch and Hertel, 2012; Chen et al., 2012a; Sanchez-Pla et al., 2012). To date many resources on AS emerged thanks to the growing amount of sequence and alignment data, in spite of incompleteness and considerable noise within the data (Black, 2003; Lareau et al., 2004; Chen et al., 2012a). Relevant databases that emerged in the context of MXE are MAASE (Zheng et al., 2005), HOLLYWOOD (Holste et al., 2006), ASAP II (Kim et al., 2007), ECGene (Lee et al., 2007), Ensembl (including former ASD/ATD/ASTD/AEdb projects (Koscielny et al., 2009), SPLOOCE (Kroll et al., 2012).

Also, the assembly of the spliceosome has been described by bioinformatics approaches (Kielbassa et al., 2009; Bortfeldt et al., 2010; Hoskins et al., 2011). Different types of the spliceosome were suggested to produce MXE splicing patterns (see Section 4). Beside the major spliceosome, a minor spliceosome can process splice sites that have distinct consensus sequences and are incompatible with the major spliceosome (Will and Lührmann, 2005).

In this review, we discuss several bioinformatics approaches for analyzing MXE splicing. In particular, we will focus on appropriate

AS pattern				
Descriptive name	Exon skipping	Intron retention	Alt. 5' end	Alt. 3' end
Bit array	10101 10001	101 111	1001 1101	1011 1001
Number vector	(21, 17)	(5, 7)	(9, 13)	(11, 9)
Sammeth et al.	$1^2-3^4, 1^4$	$1-2^3-4^1, 1-4^1$	$1-2^4, 1-3^4$	$1^2-4, 1^3-4$
SPLOOCE	e-s-e	ere	f-e	e-t
Boolean	$E_{0,1} \wedge (E_{2,3} \oplus \emptyset) \wedge E_{4,5}$	$(E_{1,2} \wedge E_{3,4}) \oplus E_{1,4}$	$(E_{1,2} \oplus E_{1,3}) \wedge E_{4,5}$	$E_{0,1} \wedge (E_{2,4} \oplus E_{3,4})$
AS pattern				
Descriptive name	Twintron	Mutual exclusion	MXE sets	non-adjacent MXE
Bit array	10001 11011	1010001 1000101	101000001 100010001	not considered
Number vector	(17, 27)	(81, 69)	(321, 273)	---
Sammeth et al.	$1-2^5-6, 1^3-4^6$	$1^2-3^6, 1^4-5^6$	$1^2-3^8, 1^4-5^8, \dots$	not considered
SPLOOCE	f-t	e-S-s-e	not considered	e-S-e-s-e
Boolean	$(E_{1,2} \wedge E_{5,6}) \oplus (E_{1,3} \wedge E_{4,6})$	$E_{0,1} \wedge (E_{2,3} \oplus E_{4,5}) \wedge E_{6,7}$	$E_{0,1} \wedge (E_{2,3} \oplus E_{4,5} \oplus E_{6,7}) \wedge E_{8,9}$	$E_{0,1} \wedge E_{4,5} \wedge E_{8,9} \wedge (E_{2,3} \oplus E_{6,7})$

**Fig. 1.** Graphical and symbolic representations of different AS patterns. Gray boxes, exons; thick solid horizontal lines, introns; flexed lines, splice junctions. Explanation, see text.

definitions, present bioinformatics tools for finding MXEs, outline Boolean approaches, and compare adjacent with non-adjacent MXEs and clustered with non-clustered MXEs. Moreover, we will discuss molecular mechanisms leading to MXEs.

## 2. Nomenclatures and definitions

### 2.1. Established approaches

A widely used graphical representation of AS events shows the alignments of transcripts as boxes representing exons connected by individual links for each isoform (Fig. 1). Early in the analysis of AS, it became clear that standardization of the nomenclature for AS forms is important (Zavolan and van Nimwegen, 2006). Since then, some attempts have been made without leading to a broadly used and accepted nomenclature. Nagasaki et al. (2006) introduced number vectors based on bit arrays where exonic regions are denoted with '1' and intronic regions are denoted with '0'. Sammeth et al. (2008) suggested a general nomenclature for AS by representing transcripts as sequence of splice sites connected by symbols for exons and exon–exon junctions. Recently, Kroll et al. (2012) introduced a character-based syntax to describe results achieved with the analysis of bit arrays by regular expressions. An overview of these notations is given in Fig. 1.

Even more notations have been suggested in the literature. Malko et al. (2006) used strings of one-character-codes for basic alternative events. Based on the exon–intron structure of isoforms, Riva and Pesole (2009) computed unique signature strings as a basis for an unambiguous nomenclature which facilitates database searches.

One drawback of such nomenclatures is that transcripts are considered and represented only individually, such that possible dependencies between splicing events remain hidden. The

resulting codes describing the splice patterns make it presently cumbersome to detect such dependencies because they have to be decoded and compared first. Additionally, for a more comprehensive splicing picture more than two transcript variants must be considered. One solution to this is the search for subgraphs within splicing graph representations of transcript isoforms (Sammeth, 2009).

In the strict definition, MXEs should be perfectly mutually exclusive, as the name suggests. In many studies, only two transcripts are considered, e.g. in two tissues or developmental stages. However, MXEs found in this manner need not be MXEs when taking more abundant transcript data into account. Thus, the term MXE is relative with respect to the abundance of known transcripts at a gene locus, which in turn depends on how many different conditions are studied. This is the case in the example of TCL6, where specific tissues show indeed exclusive patterns while on the basis of all known transcripts the pattern is lost (Sammeth et al., 2008). We suppose that this also applies to KCNMA1 (Soom et al., 2008). The more detailed analysis by Nilsen and Graveley (2010) shows that on the basis of more transcripts, the exons of KCNMA1 are not perfectly exclusive, because there are transcripts containing both mutually exclusive exons.

Summarizing these considerations, it might be worthwhile relaxing the strict definition of MXEs in that a certain percentage of non-exclusive events are allowed. In living organisms in general, many exceptions and deviations occur, for example in the number of teeth in humans. The clover plant usually has three leaves with a rare deviation showing four leaves. These deviations do not prevent the individual to be classified in the general type (e.g., human, clover, etc.). Analogously, AS events could be defined as context specific MXEs, allowing a certain percentage of cases, that are not mutually exclusive when compared across different conditions, e.g., developmental stages, tissues or disease states. However,

it is difficult to define a biologically well-founded value for such a threshold.

### 2.2. Adjacent MXEs and non-adjacent dependencies

In a genome wide study of MXE events in humans, we detected a number of non-adjacent MXEs (Pohl et al., 2009). This contradicts the intuitive assumption that MXEs – which are expected to originate from exon duplication – should usually be in direct genomic neighborhood (Letunic et al., 2002; Copley, 2004). Glauser et al. (2011) found dependencies between individual sites even in the case of non-adjacent alternatively spliced exons in *Caenorhabditis elegans slo-1*, which is in line with our findings. This is also supported by results presented in the ASTALAVISTA study (Foissac and Sammeth, 2007). Non-adjacent, mutually dependent exons have also been reported recently by Kroll et al. (2012) (see the pattern “-s-E-s-” in Table 3 in that reference). That indicates that only approaches and nomenclatures considering mutual (including long-ranging) dependencies among exons that include a constitutive exon in between will have a chance of success in predicting the full splicing picture, all the more as the line between different AS types is not clear-cut (Sammeth et al., 2008). For example, distal exclusions are a combination of skipped exons, and “Twintrons” are formed by co-occurring alternative 5' and 3' ends (Fig. 1).

### 2.3. A Boolean nomenclature

In case of clustered and/or non-adjacent MXEs, it may be difficult to grasp the dependencies among the exons. Then, it is helpful to formalize the notation. For example, if either an exon A is used or a cluster of two exons B and C, we may write:  $A \oplus (B \wedge C)$  (with  $\oplus$  denoting the exclusive disjunction, XOR). However, it is more convenient to attach indices to splice sites rather than to exons, because the possible overlap of exons (such as in the cases of intron retention or alternative ends) is easier to recognize in that way.

Denoting the splice sites in the above example by 1–6, the notation is  $E_{1,2} \oplus (E_{3,4} \wedge E_{5,6})$ , where  $E_{ij}$  stands for an exon between splice sites  $i$  and  $j$ . The usual case of one pair of MXEs would then be written as  $E_{1,2} \oplus E_{3,4}$ .

Also alternative 5' and 3' ends can now be described in a unique way by this notation. Assume, for example, that B includes A and three further nucleotides at the 3' end, then  $A \oplus B$  describes a tandem donor splice site AS event, not an MXE. Applying splice site indices we obtain the distinct  $E_{1,2} \oplus E_{1,3}$  for the alternative exon end (Fig. 1C–E).

When these Boolean expressions get rather long and complex, it is useful to condense them to the so-called disjoint normal forms. For example, the above-mentioned case of non-adjacent MXEs might be written first as  $(E_{1,2} \wedge E_{3,4}) \oplus (E_{3,4} \wedge E_{5,6})$ . This can be simplified to the expression  $(E_{1,2} \oplus E_{5,6}) \wedge E_{3,4}$ .

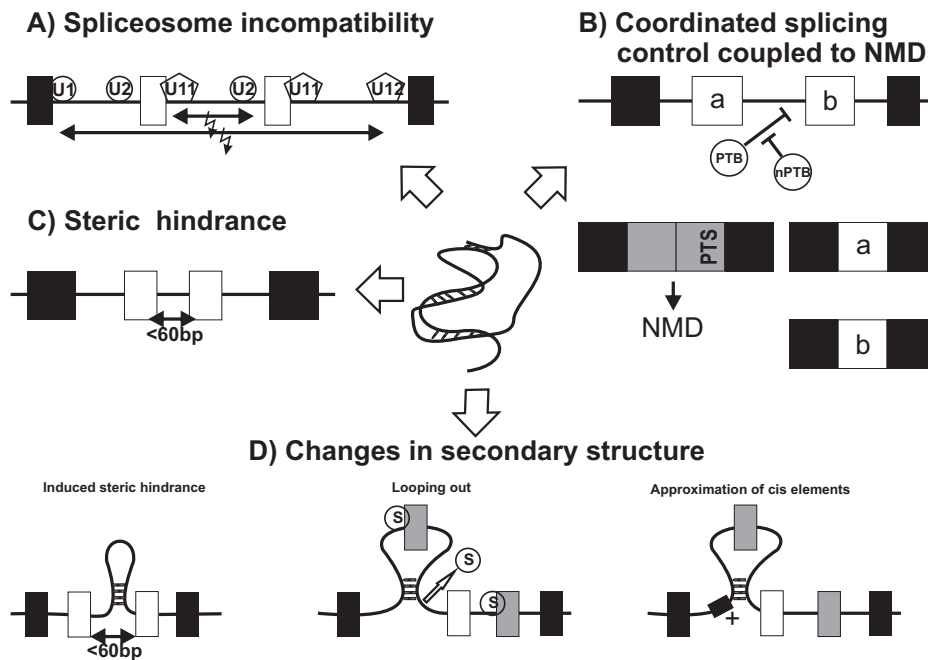
It is worth noting that not only MXEs, but also other AS events can be described in this way. For example, the notation  $(E_{1,2} \wedge E_{5,6}) \oplus (E_{1,2} \wedge E_{3,4} \wedge E_{5,6})$  describes the usual exon skipping event (Fig. 1A). Interestingly, there are cases that are neither perfect MXEs nor classical exon skipping, but more complex dependencies. Sammeth (2009) considered such dependencies earlier, using another notation, and included more complex dependencies into the class of MXEs (Fig. 6 in Sammeth, 2009), while we advocate to use this term for the “exclusive or” case only (Fig. 1F). For example,  $E_{1,2} \vee E_{3,4}$  describes the situation where exon A or B can be skipped but also the case may occur where both exons are used. It excludes, however, that both exons are skipped. Conversely, there may be the case  $\neg(E_{1,2}) \vee \neg(E_{3,4})$ . Another notation for this case is  $\emptyset \wedge (E_{1,2} \oplus E_{3,4})$ .

In doing so, Sammeth et al. (2008) proposed a Boolean-like notation in which the splice sites rather than exons are used as elementary units, denoted by numbers. For example, “1-2^, 3-4^” describes two mutually exclusive exons (Fig. 1F), where the hyphens stand for exons, the caret stands for introns and the comma separates isoforms. Such annotation will perfectly describe each isoform, but dependencies among splicing events are included only indirectly. This holds for other approaches like binary array based number vectors (Nagasaki et al., 2006) or isoform signatures (Riva and Pesole, 2009). Dependencies that span constitutive positions (Glauser et al., 2011) will get lost. The intuitive description used by SPLOOCE (Kroll et al., 2012) is inappropriate for cases with more than two dependent isoforms (Fig. 1G). Our suggestion is more general in that it describes more than two possibilities (e.g.  $E_{1,2} \oplus E_{3,4} \oplus E_{5,6}$ ) at once and enables the explicit inclusion of mutual dependencies between splicing events.

## 3. Detection

The detection of AS can be well distinguished by the three data sources they depend on, namely microarrays, RNA sequence alignments and mere DNA or RNA sequences. Microarrays as a long consolidated methodology still remain useful and accurate for transcriptomic analysis with low input requirements, while RNA-seq technology complements and extends microarray measurements for novel discoveries. Sequence based approaches utilize existing knowledge about splicing and its regulation, e.g. sequence and secondary structure motifs, for investigation of pure genomic sequence data (Hallegger et al., 2010; Raghavachari et al., 2012; Sanchez-Pla et al., 2012).

- *Microarrays* are useful for identification and expression analysis of predicted exons (Exon arrays), in scanning areas for reported and novel exon usage (tiling arrays) and for measurement of known exon junctions (splicing arrays). Exon junctions are sequences after joining at splice sites. Microarray approaches proved to be useful although being strongly dependent on previous knowledge on gene transcription (Castle et al., 2008; Gonzalez-Porta et al., 2012; Raghavachari et al., 2012; Sanchez-Pla et al., 2012)
- *Alignment based analysis* is widely applied in identification of novel splicing events de novo (Sacomoto et al., 2012) or by alignment to genomes (Wang et al., 2008, 2010; Zhou et al., 2012) and in combination with a variety of data reduction techniques like number strings (Sammeth et al., 2008), isoform signatures, Boolean arrays (Nagasaki et al., 2005, 2006) and regular expressions (Kroll et al., 2012)
- *De novo sequence analysis*
  - *Similarity based search* depends on assumptions from knowledge about former described splice events like associated sequence motifs. One assumption is that mutually exclusive exons encode regions of the same structural part of the protein product. This precondition provides restrictions to the search for candidate exons concerning their length, splice site conservation and reading frame preservation, and overall homology. Mutually exclusive exons that are not homologous and not of about the same length will not be found (Sorek et al., 2004b; Stephan et al., 2007; Pillmann et al., 2011).
  - *RNA secondary structure prediction* applies the principles of sequence similarity search on the structural level, such shifting the focus from sequence level to the regulatory relevant structural formation (Washietl et al., 2005; Raker et al., 2009; Reuter and Mathews, 2010; Yang et al., 2011).



**Fig. 2.** Various regulatory mechanisms for MXE splicing as proposed in the literature. (A) Spliceosome incompatibility: either one exon is excluded by the U1 spliceosome or another exon is excluded by the U12 spliceosome. (B) Coordinated splicing control coupled to NMD: exons are regulated by *trans*-acting splicing factors (e.g. PTB), which suppress splicing of one exon while enhancing splicing of the other one. Mis-spliced isoforms that contain both exons incorporate premature termination signal (PTS) and are degraded by non-sense mediated mRNA decay (NMD). (C) Steric hindrance: for introns shorter than 60 bp the spliceosome uses one splice site while the other one cannot be used because it is too close. (D) Change in secondary structure: MXE splicing is brought about by formation of loops in the pre-mRNA. S: suppressor. For further explanation, see text.

#### 4. Mechanisms leading to mutual exclusion of exons

Several general mechanisms for realizing mutually exclusive splicing have been proposed (Smith, 2005; Nilsen and Graveley, 2010; Jin et al., 2011; Pervouchine et al., 2012; Hemani and Soller, 2012) (Fig. 2):

- *Spliceosome incompatibility* (Fig. 2A): combinations of alternative splice sites can imply that the mutually exclusive exons are recognized and spliced by different spliceosomes, i.e. the U1 or U12 spliceosome (Burge et al., 1998; Letunic et al., 2002; Will and Lührmann, 2005). Beside the major spliceosome, a minor spliceosome, in which the subunits U1 and U2 are replaced by U11 and U12, can process splice sites that have distinct consensus sequences and are incompatible with the major spliceosome (Will and Lührmann, 2005). Thus, each type of the two different spliceosomes is compatible with only one of the MXEs.
- *Coordinated splicing control coupled to NMD* (Fig. 2B): if exons are regulated by *trans*-acting splicing factors, which suppress splicing of one exon while enhancing splicing of the other one, the mature mRNA can be mutually exclusive (Jones et al., 2001; Spellman et al., 2005). To avoid mis-splicing in the absence of splicing factors, isoforms that contain both exons incorporate premature stop codons and are potentially degraded by non-sense mediated mRNA decay (NMD), as was suggested also for subtle tandem donor splicing (Bortfeldt et al., 2008). This mechanism has been described for the growth factor receptor FGFR2 (Jones et al., 2001),  $\alpha$ -tropomyosin (Spellman et al., 2005), and the channel CaV1.2 (Tang et al., 2011).
- *Steric hindrance* (Fig. 2C): for introns shorter than 60 bp the spliceosome cannot perform the necessary structural

arrangements, hence, one splice site is used and the other one is skipped due to steric hindrance (Smith and Nadal-Ginard, 1989; Mullen et al., 1991; Kennedy and Berget, 1997). Thus, by skipping one exon an intron is generated that is long enough to become properly spliced.

- Formation of secondary structure elements (Fig. 2D):
  - *Induced steric hindrance*: loop formation within long introns may bring splice sites in close proximity preventing splicing in a similar manner as steric hindrance at short introns (Jin et al., 2011).
  - *Docker-selector pairing, Looping out*: a group of exons, each possessing a similar upstream selector site, can compete in forming a secondary structure by binding, forming intronic RNA pairings, to a complementary docker site upstream of all these exons (Miriami et al., 2003; Graveley, 2005; Anastassiou et al., 2006; Yang et al., 2011). Each exon is normally bound by a repressor, which inhibits the inclusion of the exon. Base pairing of the docker with one selector site activates the exon downstream adjacent to the selector by releasing the repressor. As only one loop can be formed there will be one exon included exclusively. This is known from *Drosophila* as docker-selector principle (Schmucker et al., 2000; Graveley, 2005; Anastassiou et al., 2006; Olson et al., 2007). This mechanism may have occurred by exon duplication and subsequent mutations.
  - *Approximation of cis elements*: RNA pairing directs control elements flanking an exon into close physical distance, thus forming a splicing activating complex (Muh et al., 2002; Yang et al., 2011).

Many of these mechanisms suggest that MXEs are adjacent to each other in the genome. Moreover, it has been proposed



that regulatory microRNAs and epigenetic mechanisms such as chromatin structure and nucleosome organization play a role in exon choice (Luco and Misteli, 2011; Huang et al., 2012). One may argue that this implies a splicing code at a higher level.

Findings about non-adjacent and clustered MXEs are of interest in view of the relevance of the above-mentioned mechanisms of MXE splicing proposed in the literature (Fig. 2). Spliceosome incompatibility seems unlikely in the case of non-adjacent MXEs. This is because the entire region between the two MXEs would have to be removed (a contradiction to the assumption that it includes constitutive exons), unless the constitutive exons would be recognized by both spliceosomes. Steric hindrance is prevented if constitutive exons exist between the MXEs because these MXEs would then be sufficiently distant. Also the mechanism via NMD and *trans*-acting splicing factors is unlikely, since such a mechanism would require the following conditions: (a) retention of both alternative exons should cause a frame shift leading to NMD; (b) removal of both alternative exons should cause a frame shift leading to NMD (conditions a and b would guarantee mutual exclusivity); (c) splicing of the first alternative exon should not cause a frame shift in order to preserve functionality of the constitutive exons in between the alternative exons. However, these conditions contradict each other, since they imply that the second alternative exon would cause a frame shift both when absent and present. The docker-selector principle would not work either because constitutive exons must not be situated in the loop occurring in that mechanism.

As mentioned in the Introduction, the *Dscam* gene has a set involving 48 alternative exons (Graveley, 2005). This led to hypotheses about the possibility of more than two mutually exclusive exons (or exon groups) within one and the same gene in the human genome. Such cases would be an indication of docker-selector pairing or related mechanisms. However, in men and mice, we found no more than two mutually exclusive parties (Pohl et al., 2009). Such cases had not been detected by other groups either, nor have they been considered in splicing databases for the human genome. Therefore, databases like ASAP or HOLLYWOOD are designed for mutually exclusive splicing with exactly two exclusive exons and do not cover mutually exclusive exon groups. Nonetheless, the docker-selector mechanism cannot be excluded as hypothetical splicing mechanism for adjacent MXEs. As for non-adjacent MXEs, in our opinion, the existence of a hitherto unknown mechanism has to be considered.

## 5. Evolutionary conservation

In general, functional AS events conserved across species tend to preserve the reading frame (Sorek et al., 2004a; Kim et al., 2008). Events shifting the reading frame imply NMD and are hardly conserved between humans and mice (Zhang et al., 2009). By utilizing AS, evolution may currently be working on adoption of further functions for the concerned genes.

In future investigations, it would be interesting to analyze the difference in conservation between more frequent and less frequent MXE isoforms. From an evolutionary point of view, one can expect that one of the two MXE isoforms shows a high sequence conservation on the DNA level to maintain the biological function of the protein (Boué et al., 2003). The other form (probably the minor form) may be less conserved, to serve as a “playground of evolution”. Similar observations were made for skipped exons (Xing and Lee, 2006) and subtle tandem donor AS (Bortfeldt et al., 2008). Note that a partially low EST coverage might have prevented the detection of some orthologous isoforms possibly serving this purpose. Since non-adjacent MXEs are less conserved and tend to shift the reading frame, they are more likely to be a “playground of evolution” than adjacent MXEs. Analogous hypotheses have been put

forward for intron retention and alternative ends (Sorek, 2007; Tarrío et al., 2008). In contrast, in situations where both MXEs are equally important, as it appears to be the case for ion channels, the degree of conservation is not likely to differ. In fact, the instances of adjacent MXEs within the human genome we found within our studies, were predominantly assigned to ion channels (Pohl et al., 2009).

A following step would be to investigate the impact of MXE splicing on protein function. To do so, it is useful to compare the protein structures resulting from the alternative transcripts (Birzele et al., 2008a,b). However, at present for many proteins no structural information is available.

## 6. Conclusions

Here we have reviewed approaches to analyze a rare subtype of alternative splicing (AS), termed mutually exclusive exon (MXE) splicing. We have discussed various approaches and nomenclatures in this context. MXE splicing is the only type of AS that can maintain the size of the protein introducing a quasi-exchange, provided that the exons are (nearly) of the same length.

It is often assumed that MXEs – which are expected to originate from exon duplication – should usually be in direct genomic neighborhood (Letunic et al., 2002; Copley, 2004). However, there is no reason to exclude non-adjacent MXEs. Hence, only approaches and nomenclatures considering mutual (perhaps long-ranging) dependencies within complete genes will have a chance of success in deciphering the full splicing picture.

In many studies often only two transcripts (e.g. in two tissues or developmental stages) are considered. MXEs found in this manner need not to be MXEs when taking more abundant transcript data into account. Thus, the term MXE is relative; it depends on how many different conditions are studied and the regulation mechanism might not work in a perfectly strict manner (Tang et al., 2011). Summarizing these considerations, it may be worth relaxing the rigid definition of MXEs in that a certain fraction of non-exclusive events is allowed. Nevertheless, as many different tissues as possible should be analyzed as this for example allowed the detection of switch like skipped exons (Wang et al., 2008). In clustered MXEs, a strict dependence between the exons within each cluster occurs. Such a dependence can also occur in other types of AS. For example, in some events of exon skipping, a strict correlation between two consecutive exons was found (Sammeth et al., 2008).

The analysis of MXEs by various bioinformatics techniques is likely to be very helpful in diagnosing diseases and tailoring pharmaceuticals in personalized medicine. For example, the knowledge of receptor variants can help identify new drug targets. The knowledge of variants of exoenzymes degrading polymers can be useful in optimizing biotechnological processes.

## Acknowledgements

We thank Ines Heiland, Stefan Heinemann, Dirk Holste and Günther Theißen for stimulating discussions and Ina Weiß for valuable assistance in the literature search. S.S. acknowledges financial support by the German Ministry of Education and Research (BMBF) in the Virtual Liver program.

## References

- Anastassiou, D., Liu, H., Varadan, V., 2006. Variable window binding for mutually exclusive alternative splicing. *Genome Biol.* 7, R2.
- Andreassi, C., Riccio, A., 2009. To localize or not to localize: mRNA fate is in 3'UTR ends. *Trends Cell Biol.* 19, 465–474.
- Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., Frey, B.J., 2010. Deciphering the splicing code. *Nature* 465, 53–59.

- Barbieri, M., 2008. Biosemiotics: a new understanding of life. *Die Naturwissenschaften* 95, 577–599.
- Berget, S.M., Moore, C., Sharp, P.A., 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. U.S.A.* 74, 3171–3175.
- Birzele, F., Csaba, G., Zimmer, R., 2008a. Alternative splicing and protein structure evolution. *Nucleic Acids Res.* 36, 550–558.
- Birzele, F., Küffner, R., Meier, F., Oefinger, F., Potthast, C., Zimmer, R., 2008b. ProSAS: a database for analyzing alternative splicing in the context of protein structures. *Nucleic Acids Res.* 36, D63–D68.
- Black, D.L., 1998. Splicing in the inner ear: a familiar tune, but what are the instruments? *Neuron* 20, 165–168.
- Black, D.L., 2000. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* 103, 367–370.
- Black, D.L., 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* 72, 291–336.
- Bortfeldt, R., Schindler, S., Szafranski, K., Schuster, S., Holste, D., 2008. Comparative analysis of sequence features involved in the recognition of tandem splice sites. *BMC Genomics* 9, 202.
- Bortfeldt, R.H., Schuster, S., Koch, I., 2010. Exhaustive analysis of the modular structure of the spliceosomal assembly network: a Petri net approach. In *Silico Biol.* 10, 0007.
- Boué, S., Letunic, I., Bork, P., 2003. Alternative splicing and evolution. *Bioessays* 25, 1031–1034.
- Burge, C.B., Padgett, R.A., Sharp, P.A., 1998. Evolutionary fates and origins of U12-type introns. *Mol. Cell* 2, 773–785.
- Busch, A., Hertel, K.J., 2012. Extensive regulation of NAGNAG alternative splicing: new tricks for the spliceosome? *Genome Biol.* 13, 143.
- Castle, J.C., Zhang, C., Shah, J.K., Kulkarni, A.V., Kalsotra, A., Cooper, T.A., Johnson, J.M., 2008. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nat. Genet.* 40, 1416–1425.
- Chacko, E., Ranganathan, S., 2009a. Comprehensive splicing graph analysis of alternative splicing patterns in chicken, compared to human and mouse. *BMC Genomics* 10 (Suppl. 1), S5.
- Chacko, E., Ranganathan, S., 2009b. Genome-wide analysis of alternative splicing in cow: implications in bovine as a model for human diseases. *BMC Genomics* 10 (Suppl. 3), S11.
- Chen, F.-C., Wang, S.-S., Chen, C.-J., Li, W.-H., Chuang, T.-J., 2006. Alternatively and constitutively spliced exons are subject to different evolutionary forces. *Mol. Biol. Evol.* 23, 675–682.
- Chen, L., Tovar-Corona, J.M., Urrutia, A.O., 2012a. Alternative splicing: a potential source of functional innovation in the eukaryotic genome. *Int. J. Evol. Biol.* 2012, 596274.
- Chen, M., David, C.J., Manley, J.L., 2012b. Concentration-dependent control of pyruvate kinase M mutually exclusive splicing by hnRNP proteins. *Nat. Struct. Mol. Biol.* 19, 346–354.
- Choudhary, A., Krithivasan, K., 2007. Network of evolutionary processors with splicing rules and permitting context. *BioSystems* 87, 111–116.
- Copley, R.R., 2004. Evolutionary convergence of alternative splicing in ion channels. *Trends Genet.* 20, 171–176.
- Dou, Y., Fox-Walsh, K.L., Baldi, P.F., Hertel, K.J., 2006. Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. *RNA* 12, 2047–2056.
- Douglas, A.G., Wood, M.J., 2011. RNA splicing: disease and therapy. *Brief. Funct. Genomics* 10, 151–164.
- Florea, L., 2006. Bioinformatics of alternative splicing and its regulation. *Brief. Bioinform.* 7, 55–69.
- Foissac, S., Sammeth, M., 2007. ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Res.* 35, W297–W299.
- García-Blanco, M.A., 2006. Alternative splicing: therapeutic target and tool. *Prog. Mol. Subcell Biol.* 44, 47–64.
- Gatherer, D., Seiraffian, S., Cunningham, C., Holton, M., Dargan, D.J., Baluchova, K., Hector, R.D., Galbraith, J., Herzyk, P., Wilkinson, G.W., Davison, A.J., 2011. High-resolution human cytomegalovirus transcriptome. *Proc. Natl. Acad. Sci. U.S.A.* 108, 19755–19760.
- Germann, S., Grataadou, L., Dutertre, M., Auboeuf, D., 2012. Splicing programs and cancer. *J. Nucleic Acids* 2012, 269570.
- Glauser, D.A., Johnson, B.E., Aldrich, R.W., Goodman, M.B., 2011. Intragenic alternative splicing coordination is essential for *Caenorhabditis elegans slo-1* gene function. *Proc. Natl. Acad. Sci. U.S.A.* 108, 20790–20795.
- Gonzalez-Porta, M., Calvo, M., Sammeth, M., Guigo, R., 2012. Estimation of alternative splicing variability in human populations. *Genome Res.* 22, 528–538.
- Graveley, B.R., 2001. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* 17, 100–107.
- Graveley, B.R., 2005. Mutually exclusive splicing of the insect *Dscam* pre-mRNA directed by competing intronic RNA secondary structures. *Cell* 123, 65–73.
- Grohmann, M., Hammer, P., Walther, M., Paulmann, N., Büttner, A., Eisenmenger, W., Baghai, T.C., Schüle, C., Rupprecht, R., Bader, M., Bondy, B., Zill, P., Priller, J., Walther, D.J., 2010. Alternative splicing and extensive RNA editing of human *TPH2* transcripts. *PLoS ONE* 5, e8956.
- Gustafson, T.A., Clevinger, E.C., O'Neill, T.J., Yarowsky, P.J., Krueger, B.K., 1993. Mutually exclusive exon splicing of type III brain sodium channel alpha subunit RNA generates developmentally regulated isoforms in rat brain. *J. Biol. Chem.* 268, 18648–18653.
- Halleger, M., Llorian, M., Smith, C.W.J., 2010. Alternative splicing: global insights. *FEBS J.* 277, 856–866.
- He, C., Zhou, F., Zuo, Z., Cheng, H., Zhou, R., 2009. A global view of cancer-specific transcript variants by subtractive transcriptome-wide analysis. *PLoS ONE* 4, e4732.
- Hemani, Y., Soller, M., 2012. Mechanisms of *Drosophila Dscam* mutually exclusive splicing regulation. *Biochem. Soc. Trans.* 40, 804–809.
- Hernandez-Lopez, H.R., Graham, S.V., 2012. Alternative splicing in human tumour viruses: a therapeutic target? *Biochem. J.* 445, 145–156.
- Hiller, M., Nikolajewa, S., Huse, K., Szafranski, K., Rosenstiel, P., Schuster, S., Backofen, R., Platzer, M., 2007. TassDB: a database of alternative tandem splice sites. *Nucleic Acids Res.* 35, D188–D192.
- Hiller, M., Platzer, M., 2008. Widespread and subtle: alternative splicing at short-distance tandem sites. *Trends Genet.* 24, 246–255.
- Holste, D., Huo, G., Tung, V., Burge, C.B., 2006. HOLLYWOOD: a comparative relational database of alternative splicing. *Nucleic Acids Res.* 34, D56–D62.
- Hoskins, A.A., Friedman, L.J., Gallagher, S.S., Crawford, D.J., Anderson, E.G., Wombacher, R., Ramirez, N., Cornish, V.W., Gelles, J., Moore, M.J., 2011. Ordered and dynamic assembly of single spliceosomes. *Science* 331, 1289–1295.
- Huang, H., Yu, S., Liu, H., Sun, X., 2012. Nucleosome organization in sequences of alternative events in human genome. *BioSystems* 109, 214–219.
- Jin, Y., Yang, Y., Zhang, P., 2011. New insights into RNA secondary structure in the alternative splicing of pre-mRNAs. *RNA Biol.* 8, 450–457.
- Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., Shoemaker, D.D., 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 302, 2141–2144.
- Jones, R.B., Wang, F., Luo, Y., Yu, C., Jin, C., Suzuki, T., Kan, M., McKeehan, W.L., 2001. The nonsense-mediated decay pathway and mutually exclusive expression of alternatively spliced FGFR2IIb and -IIc mRNAs. *J. Biol. Chem.* 276, 4158–4167.
- Jurica, M.S., Moore, M.J., 2003. Pre-mRNA splicing: awash in a sea of proteins. *Mol. Cell* 12, 5–14.
- Kennedy, C.F., Berget, S.M., 1997. Pyrimidine tracts between the 5' splice site and branch point facilitate splicing and recognition of a small *Drosophila* intron. *Mol. Cell Biol.* 17, 2774–2780.
- Kielbassa, J., Bortfeldt, R., Schuster, S., Koch, I., 2009. Modeling of the U1 snRNP assembly pathway in alternative splicing in human cells using Petri nets. *Comput. Biol. Chem.* 33, 46–61.
- Kim, E., Goren, A., Ast, G., 2008. Alternative splicing: current perspectives. *Bioessays* 30, 38–47.
- Kim, N., Alekseyenko, A.V., Roy, M., Lee, C., 2007. The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. *Nucleic Acids Res.* 35, D93–D98.
- Kopelman, N.M., Lancet, D., Yanai, I., 2005. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nat. Genet.* 37, 588–589.
- Koscielny, G., Texier, V.L., Gopalakrishnan, C., Kumanduri, V., Riethoven, J.-J., Nardone, F., Stanley, E., Fallsehr, C., Hofmann, O., Kull, M., Harrington, E., Boué, S., Eyraes, E., Plass, M., Lopez, F., Ritchie, W., Mouchadel, V., Ara, T., Pospisil, H., Herrmann, A., Reich, J.G., Guigó, R., Bork, P., von Knebel Doeberitz, M., Vilo, J., Hide, W., Apweiler, R., Thanaraj, T.A., Gautheret, D., 2009. ASTD: the alternative splicing and transcript diversity database. *Genomics* 93, 213–220.
- Kroll, J.E., Galante, P.A., Ohara, D.T., Navarro, F.C., Ohno-Machado, L., de Souza, S.J., 2012. SPOOCE: a new portal for the analysis of human splicing variants. *RNA Biol.* 9, 1339–1343.
- Lareau, L.F., Green, R.E., Bhatnagar, R.S., Brenner, S.E., 2004. The evolving roles of alternative splicing. *Curr. Opin. Struct. Biol.* 14, 273–282.
- Lee, Y., Lee, Y., Kim, B., Shin, Y., Nam, S., Kim, P., Kim, N., Chung, W.-H., Kim, J., Lee, S., 2007. ECgene: an alternative splicing database update. *Nucleic Acids Res.* 35, D99–D103.
- Letunic, I., Copley, R.R., Bork, P., 2002. Common exon duplication in animals and its role in alternative splicing. *Hum. Mol. Genet.* 11, 1561–1567.
- Luco, R.F., Misteli, T., 2011. More than a splicing code: integrating the role of RNA, chromatin and non-coding RNA in alternative splicing regulation. *Curr. Opin. Genet. Dev.* 21, 1–7.
- Malko, D.B., Makeev, V.J., Mironov, A.A., Gelfand, M.S., 2006. Evolution of exon-intron structure and alternative splicing in fruit flies and malarial mosquito genomes. *Genome Res.* 16, 505–509.
- Mastrangelo, A.M., Marone, D., Laido, G., Leonardis, A.M.D., Vita, P.D., 2012. Alternative splicing: enhancing ability to cope with stress via transcriptome plasticity. *Plant Sci.* 185–186, 40–49.
- Meijers, R., Puettmann-Holgado, R., Skinotis, G., huan Liu, J., Walz, T., huai Wang, J., Schmucker, D., 2007. Structural basis of *Dscam* isoform specificity. *Nature* 449, 487–491.
- Miriami, E., Margalit, H., Sperling, R., 2003. Conserved sequence elements associated with exon skipping. *Nucleic Acids Res.* 31, 1974–1983.
- Mount, D.W., Pandey, R., 2005. Using bioinformatics and genome analysis for new therapeutic interventions. *Mol. Cancer Ther.* 4, 1636–1643.
- Muh, S.J., Hovhannisyann, R.H., Carstens, R.P., 2002. A non-sequence-specific double-stranded RNA structural element regulates splicing of two mutually exclusive exons of fibroblast growth factor receptor 2 (FGFR2). *J. Biol. Chem.* 277, 50143–50154.
- Mullen, M.P., Smith, C.W., Patton, J.G., Nadal-Ginard, B., 1991. Alpha-tropomyosin mutually exclusive exon selection: competition between branchpoint/polypyrimidine tracts determines default exon choice. *Genes Dev.* 5, 642–655.
- Nagasaki, H., Arita, M., Nishizawa, T., Suwa, M., Gotoh, O., 2005. Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes. *Gene* 364, 53–62.



- Nagasaki, H., Arita, M., Nishizawa, T., Suwa, M., Gotoh, O., 2006. Automated classification of alternative splicing and transcriptional initiation and construction of visual database of classified patterns. *Bioinformatics* 22, 1211–1216.
- Nilsen, T.W., Graveley, B.R., 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* 463, 457–463.
- Noh, S.-J., Lee, K., Paik, H., Hur, C.-G., 2006. TISA: tissue-specific alternative splicing in human and mouse genes. *DNA Res.* 13, 229–243.
- Olson, S., Blanchette, M., Park, J., Savva, Y., Yeo, G.W., Yeakley, J.M., Rio, D.C., Graveley, B.R., 2007. A regulator of *Dscam* mutually exclusive splicing fidelity. *Nat. Struct. Mol. Biol.* 14, 1134–1140.
- Pan, Q., Shai, O., Lee, L.J., Frey, B.J., Blencowe, B.J., 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 41, 762.
- Pervouchine, D.D., Khrameeva, E.E., Pichugina, M.Y., Nikolaienko, O.V., Gelfand, M.S., Rubtsov, P.M., Mironov, A.A., 2012. Evidence for widespread association of mammalian splicing and conserved long-range RNA structures. *RNA* 18, 1–15.
- Pillmann, H., Hatje, K., Odronitz, F., Hammesfahr, B., Kollmar, M., 2011. Predicting mutually exclusive spliced exons based on exon length, splice site and reading frame conservation, and exon sequence homology. *BMC Bioinform.* 12, 270.
- Pohl, M., Holste, D., Bortfeldt, R., Grützmann, K., Schuster, S., 2009. Mutually exclusive spliced exons show non-adjacent and grouped patterns. In: Grosse, I., Neumann, S., Posch, S., Schreiber, F., Stadler, P. (Eds.), *German conference on bioinformatics. Gesellschaft für Informatik e.V., Halle*, pp. 19–24.
- Pohl, M., Theissen, G., Schuster, S., 2012. GC content dependency of open reading frame prediction via stop codon frequencies. *Gene* 511, 441–446.
- Raghavachari, N., Barb, J., Yang, Y., Liu, P., Woodhouse, K., Levy, D., O'Donnell, C.J., Munson, P.J., Kato, G.J., 2012. A systematic comparison and evaluation of high density exon arrays and RNA-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease. *BMC Med. Genomics* 5, 28.
- Raker, V.A., Mironov, A.A., Gelfand, M.S., Pervouchine, D.D., 2009. Modulation of alternative splicing by long-range RNA structures in *Drosophila*. *Nucleic Acids Res.* 37, 4533–4544.
- Reddy, A.S., Rogers, M.F., Richardson, D.N., Hamilton, M., Ben-Hur, A., 2012. Deciphering the plant splicing code: experimental and computational approaches for predicting alternative splicing and splicing regulatory elements. *Frontiers Plant Sci.* 3, 18.
- Reuter, J.S., Mathews, D.H., 2010. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinform.* 11, 129.
- Riva, A., Pesole, G., 2009. A unique, consistent identifier for alternatively spliced transcript variants. *PLoS ONE* 4, e7631.
- Sacomoto, G.A., Kielbassa, J., Chikhi, R., Uricaru, R., Antoniou, P., Sagot, M.F., Peterlongo, P., Lacroix, V., 2012. KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. *BMC Bioinform.* 13 (Suppl. 6), S5.
- Sammeth, M., 2009. Complete alternative splicing events are bubbles in splicing graphs. *J. Comput. Biol.* 16, 1117–1140.
- Sammeth, M., Foissac, S., Guigó, R., 2008. A general definition and nomenclature for alternative splicing events. *PLoS Comput. Biol.* 4, e1000147.
- Sanchez-Pla, A., Reverter, F., Ruiz de Villa, M.C., Comabella, M., 2012. Transcription: mRNA and alternative splicing. *J. Neuroimmunol.* 248, 23–31.
- Schmucker, D., Clemens, J.C., Shu, H., Worby, C.A., Xiao, J., Muda, M., Dixon, J.E., Zipursky, S.L., 2000. *Drosophila Dscam* is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* 101, 671–684.
- Smith, C.W., Nadal-Ginard, B., 1989. Mutually exclusive splicing of alpha-tropomyosin exons enforced by an unusual lariat branch point location: implications for constitutive splicing. *Cell* 56, 749–758.
- Smith, C.W.J., 2005. Alternative splicing – when two's a crowd. *Cell* 123, 1–3.
- Soom, M., Gessner, G., Heuer, H., Hoshi, T., Heinemann, S.H., 2008. A mutually exclusive alternative exon of *slol1* codes for a neuronal BK channel with altered function. *Channels (Austin)* 2, 278–282.
- Sorek, R., 2007. The birth of new exons: mechanisms and evolutionary consequences. *RNA* 13, 1603–1608.
- Sorek, R., 2009. When new exons are born. *Heredity* 103, 279–280.
- Sorek, R., Shamir, R., Ast, G., 2004a. How prevalent is functional alternative splicing in the human genome? *Trends Genet.* 20, 68–71.
- Sorek, R., Shemesh, R., Cohen, Y., Basechess, O., Ast, G., Shamir, R., 2004b. A non-EST-based method for exon-skipping prediction. *Genome Res.* 14, 1617–1623.
- Spellman, R., Rideau, A., Matlin, A., Gooding, C., Robinson, F., McGlinchy, N., Grellscheid, S.N., Southby, J., Wollerton, M., Smith, C.W.J., 2005. Regulation of alternative splicing by PTB and associated factors. *Biochem. Soc. Trans.* 33, 457–460.
- Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T.A., Sorek, H., 2005. Function of alternative splicing. *Gene* 344, 1–20.
- Stephan, Möller, M., Wiehe, F., Kleffe Jr., T., 2007. Self-alignments to detect mutually exclusive exon usage. In *Silico Biol.* 7, 613–621.
- Tabuchi, K., Südhof, T.C., 2002. Structure and evolution of neurexin genes: insight into the mechanism of alternative splicing. *Genomics* 79, 849–859.
- Tang, Z.Z., Sharma, S., Zheng, S., Chawla, G., Nikolic, J., Black, D.L., 2011. Regulation of the mutually exclusive exons 8a and 8 in the *CaV1.2* calcium channel transcript by polypyrimidine tract-binding protein. *J. Biol. Chem.* 286, 10007–10016.
- Tarrio, R., Ayala, F.J., Rodríguez-Trelles, F., 2008. Alternative splicing: a missing piece in the puzzle of intron gain. *Proc. Natl. Acad. Sci. U.S.A.* 105, 7223–7228.
- Tazi, J., Bakkour, N., Marchand, V., Ayadi, L., Aboufirassi, A., Branlant, C., 2010. Alternative splicing: regulation of HIV-1 multiplication as a target for therapeutic action. *FEBS J.* 277, 867–876.
- Tazi, J., Bakkour, N., Stamm, S., 2009. Alternative splicing and disease. *Biochim. Biophys. Acta* 1792, 14–26.
- Wailes, G.T., Graham, I.R., Jackson, P., Millake, D.B., Patel, B., Blanchard, A.D., Weller, P.A., Eperon, I.C., Critchley, D.R., 1992. Mutually exclusive splicing of calcium-binding domain exons in chick alpha-actinin. *J. Biol. Chem.* 267, 6263–6271.
- Wang, E.T., Sandberg, R., Luo, S., Khrebukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., Burge, C.B., 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476.
- Wang, L., Xi, Y., Yu, J., Dong, L., Yen, L., Li, W., 2010. A statistical method for the detection of alternative splicing using RNA-seq. *PLoS ONE* 5, e8529.
- Wang, X., Li, G., Yang, Y., Wang, W., Zhang, W., Pan, H., Zhang, P., Yue, Y., Lin, H., Liu, B., Bi, J., Shi, F., Mao, J., Meng, Y., Zhan, L., Jin, Y., 2012. An RNA architectural locus control region involved in *Dscam* mutually exclusive splicing. *Nat. Commun.* 3, 1255.
- Wang, Z., Xiao, X., Nostrand, E.V., Burge, C.B., 2006. General and specific functions of exonic splicing silencers in splicing control. *Mol. Cell* 23, 61–70.
- Washietl, S., Hofacker, I.L., Lukasser, M., Huttenhofer, A., Stadler, P.F., 2005. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.* 23, 1383–1390.
- Will, C.L., Lührmann, R., 2005. Splicing of a rare class of introns by the U12-dependent spliceosome. *Biol. Chem.* 386, 713–724.
- Xing, Y., Lee, C., 2006. Alternative splicing and RNA selection pressure—evolutionary consequences for eukaryotic genomes. *Nat. Rev. Genet.* 7, 499–509.
- Xing, Y., Resch, A., Lee, C., 2004. The multiassembly problem: reconstructing multiple transcript isoforms from EST fragment mixtures. *Genome Res.* 14, 426–441.
- Yang, Y., Zhan, L., Zhang, W., Sun, F., Wang, W., Tian, N., Bi, J., Wang, H., Shi, D., Jiang, Y., Zhang, Y., Jin, Y., 2011. RNA secondary structure in mutually exclusive splicing. *Nat. Struct. Mol. Biol.* 18, 159–168.
- Yap, K., Makeyev, E.V., 2013. Regulation of gene expression in mammalian nervous system through alternative pre-mRNA splicing coupled with RNA quality control mechanisms. *Mol. Cell. Neurosci.* <http://dx.doi.org/10.1016/j.mcn.2013.01.003> (Epub ahead of print).
- Zavolan, M., van Nimwegen, E., 2006. The types and prevalence of alternative splice forms. *Curr. Opin. Struct. Biol.* 16, 362–367.
- Zhang, Z., Xin, D., Wang, P., Zhou, L., Hu, L., Kong, X., Hurst, L.D., 2009. Noisy splicing, more than expression regulation, explains why some exons are subject to nonsense-mediated mRNA decay. *BMC Biol.* 7, 23.
- Zheng, C.L., Kwon, Y.-S., Li, H.-R., Zhang, K., Coutinho-Mansfield, G., Yang, C., Nair, T.M., Gribskov, M., Fu, X.-D., 2005. MAASE: an alternative splicing database designed for supporting splicing microarray applications. *RNA* 11, 1767–1776.
- Zhou, A., Breese, M.R., Hao, Y., Edenberg, H.J., Li, L., Skaar, T.C., Liu, Y., 2012. Alt event finder: a tool for extracting alternative splicing events from RNA-seq data. *BMC Genomics* 13 (Suppl. 8), S10.

## 2. PUBLICATIONS

---

## 3

# General Discussion

### 3.1 Amino acid side chain combinations restrained to a fraction of their potential

Our study on combinatorics in AA side chains showed that there is a vast potential from the combination of chemical elements into AAs. Even though we restricted the calculation to aliphatic side chains without cycles and double bonds, the number of possible AAs grows exponentially in the number of used carbon atoms<sup>118</sup>. The question is, why nature uses exactly the present 20 AAs for the assembly of proteins. The first intuitive reason is that the combinations must be tractable, the cells must be able to manage them. The diverse proteins are not only assembled once in the cells. They are synthesized many times again. Also, a multitude of descendants of a cell must inherit the trait of utilizing proteins. For these reasons, information of the synthesis of proteins is stored and not only the proteins themselves. Hence, cells possess machineries to store, read and write information. How big the protein set of a cell may ever be, it is embedded in an information processing system. One step of this processing is translation of the information (DNA/RNA) into a functioning, physical representation (proteins). The mapping is implemented in the genetic code, and could be considered a translation from one symbolic representation to another one. It is reasonable that this code is small. This is because during translation the right elementary translation units, the tRNAs, must be chosen again and again. Though this takes place in aqueous solution wherein molecules can move fast, the bigger the code is, the more time is needed for a single translation step.

Despite these reasons for a rather small code, still, the questions are, why the number of used AAs are 20, and why the genetic code has its given structure with the properties redundancy and robustness, in exactly the implementation as we see it. In principle, a bigger set of AAs, would allow for higher diversity of proteins, and thus, the exploration of a greater search space from

### 3. GENERAL DISCUSSION

---

which species could have been selected during evolution. In such a larger space, potentially more and better solutions for fitness optimization could be constructed, which could have evolutionary benefits against a less diverse system with fewer variation. An argument against this could be that even with only 20 AAs, variety is big enough to construct any needed protein function. Moreover, finding optimal solutions in a larger search space would take more time. Thus, the bottleneck of protein evolution is perhaps not the amount of variety from combinations, but rather the time needed to explore it during many generations.

The first significant restriction of the genetic code comes from having codons with three positions using four different symbols each<sup>123,124</sup>. This limits the code to have at most 64 translational units, and hence, at most 64 building blocks for proteins and stop codons.

A simple reason for the current partition of the 64 codons was given in a study showing that 19 or 20 is the most probable number of partitions when having 0 to 3 stop codons. Assuming that an early code had random variations, this lead to the most probable number of 20 AAs<sup>125</sup>. In 1968, Crick<sup>126</sup> coined the notion “frozen accident” shortly after the discovery that three bases code for one AA in 1961<sup>46</sup>, and the deciphering of the genetic code in 1965<sup>47</sup>. It is an attempt to explain the unusual partitioning of the code. According to Crick, the genetic code is evolutionary old. Once it was used by higher level mechanisms, as translation, it could not be changed because minute alterations would have had a huge impact on macro-structures, and likely be lethal<sup>126</sup>. This is why the code is relatively universal in nature and only two extensions are known so far, which are the incorporation of selenocysteine and pyrrolysine by the repurpose of stop codons<sup>127</sup>. Crick’s theory poses too few constraints on the code’s evolution in order to explain its exact organization and the properties of error-tolerance and redundancy<sup>128</sup>.

Newer research gives evidence that the code’s structure did not arise randomly (by “accident”), but that the current code’s properties were favored by three forces: diversity, error-tolerance and cost<sup>129</sup>. I have discussed the need for diversity already. The more diverse AAs there are, the more unusual niches of the fitness landscape can be explored, which entails a fitness advantage. Indeed, there seems to be a selection for the use of versatile AAs<sup>130</sup>. The other two forces make the code evolve to have a rather small set of AAs. According to the rate-distortion model, whose basis is borrowed from information theory<sup>131</sup>, translation is considered a transfer of a message (RNA sequence) via a noisy channel (tRNA), which models the misreading of triplets<sup>132</sup>. It was argued that the current code structure is a topological feature of this noisy information transmission process<sup>129</sup>. Other error sources are replication and transcription. The question is how to make a code in a way that the message has the highest chance to stay the same or very similar after translational misreading, and when also mistakes during transcription happen. First, the code degeneracy is built in a way that codons of a given AA are very similar, i.e., they differ in only one, and rarely in two positions. Thus, it is more likely that the translation

### 3.1 Amino acid side chain combinations restrained to a fraction of their potential

is the same when one position is misread or mutated. Misreadings happen more often in the first codon position. Accordingly, the code had evolved in a way that AAs differing in the first position of their codon are more similar<sup>128</sup>. Transitions, i.e., conversions from pyrimidine (Y) to Y and purine (R) to R, happen more often than transversions (Y to R, or R to Y). The code has evolved in a way that all AAs with a Y or R in the second position are similar to AAs of the same group. Thus, the chance of having a similar AA after a mutation at the second position is maximized, and hence there is a smaller negative effect. This property also groups the codons into encoding hydrophobic and hydrophilic AAs<sup>128</sup>. Not only is the number of 20 AAs the most likely one in random partitions (see above<sup>125</sup>), simulations showed that the number is stable in a certain range of mutational and selective parameters, even if taken errors in translation and replication into account. This range is called encoding plateau. These simulations also showed that the exact partition of codons are determined when taking the errors into account and having a model of code and message co-evolution<sup>128</sup>.

Except for glycine and the AAs with six codons, the production pathways of AAs are the same for codons with the same initial letter<sup>133,134,135</sup>. In the beginning there was likely a simple genetic code that was sufficient to grasp the few extant AAs, some of which may have originated by abiogenesis, as supposed by experiments of Miller and Urey<sup>136</sup>. Accordingly, less discriminative codes based on triplets were proposed, e.g., RRY<sup>137</sup>, RNY<sup>138,139</sup>, and G-nonG-N<sup>140</sup> (N stands for any AA). The code developed gradually to a more information rich one<sup>133</sup>, which was paralleled by gradual sharing of codons as AA synthetic pathways developed<sup>135</sup>. Interestingly, the simplest AAs from the “primordial soup”<sup>136</sup> are encoded as GNN. With the assimilation of the bigger AAs into the code<sup>141</sup>, the last codon position became more refined, and e.g., glutamate and aspartate could be discriminated<sup>133</sup>. The third mentioned force during the evolution of the genetic code are the costs to store, replicate, and correct damage of the stored information. These are further necessities that favored the evolution of a simple code with few AAs<sup>129</sup>.

In “thawing” of the “frozen accident” theory<sup>142</sup>, it is clear that the code is indeed very rigid (“frozen”). It has not changed for millions of years, probably because of its high optimization and the many macro-level processes depending on it. However, all the mentioned, markedly involved, properties of the code make an assignment of codons to AAs by chance (“accident”) unlikely<sup>133</sup>. Although, there is great potential in a protein “construction kit” with more aliphatic and other AAs with special chemical properties, the given reasons make a further expansion unlikely. Thus, we only see a small subset of the vast possibilities of protein building blocks.

One outcome of our study was that despite their combinatorial potential, AA side chains with more carbon atoms are rarely used. In fact, the largest proteinogenic aliphatic AAs have only four carbon atoms (leucine, isoleucine), or five in case of a less strict definition (lysine). One argument for this is that bigger AAs would render proteins more cumbersome, and thus, less

### 3. GENERAL DISCUSSION

---

able to execute their function. Another reason are higher costs in production, both in terms of energy and production time<sup>143</sup>. A third reason against use of large aliphatic AAs is their decreased solubility in aqueous solution. Their hydrophobicity would make them well-suited as inner protein parts, but harder to manage in processes as the production and transport in cytosol. The last reason we have nearly neglected so far is steric instability. Multiple branching pushes side chains closer together. More highly branched molecules experience higher strain on their bonds, which render them thermodynamically more unstable. From a certain complexity on, compounds will spontaneously fragment<sup>144</sup>. Even if this reaction is endothermic, entropic movements will push molecules above this hurdle. C16, the nearly fully saturated, and thus, branched alkane made of 16 C atoms, is the smallest structure that can not be made due to thermodynamical instability<sup>144</sup>. Note, C17 is the according fully branched alkane. It was found that disallowed structures outnumber allowed ones, already when only excluding C16 and C17 substructures. Still, the number grows exponentially in the number of incorporated C atoms<sup>144</sup>. Interestingly, a randomly constructed alkane of above 53kDa (the average yeast protein mass<sup>8</sup>) likely does not exist for its high chance of thermodynamical instability. This puts a firm upper bound on possible AA structures. It was argued, that there are further thermodynamically disallowed structures in practice. “Molecular crowding” can take place with less branching, i.e., even in molecules without C16 and C17 sub-structures<sup>144</sup>.

The rigorous restrictions on the first level of protein combinatorics, where elements are assembled into only few used AAs, do not hold for non-proteinogenic AAs. These are not bound to be incorporated into a firm code, and thus, their diversity is much higher. Non-proteinogenic AAs are found in all species and comprise signal transducers, secondary metabolites, fungal toxins and many more compounds of yet unknown function<sup>45,145,146</sup>. The majority of those AAs are unknown to researchers and new ones are discovered consistently, as e.g. AAs contained in non-ribosomally synthesized peptides in fungi of the *Trichoderma* genus<sup>147</sup>. Though their synthetic pathway and the pathway of incorporation into macro-molecular structures needs additional genes, energetic and temporal costs are not that much of a factor as for proteinogenic AAs. Non-proteinogenic AAs can have longer synthetic pathways and reach greater molecular mass, because they rather take specific functions and are thus not as abundantly present as proteinogenic AAs. One example from our publication is 2-amino-9,13-dimethyl-heptadecanoic acid from *Streptomyces* sp. 1010 ( $\text{CO}_2\text{-C}_{18}\text{H}_{36}\text{-NH}_3^+$ , mol. mass 313g/mol, Trp mol. mass 204g/mol)<sup>45</sup>. However, the restrictions on branching complexity<sup>144</sup> and solubility also hold for non-proteinogenic AA.

2-amino butanoic acid, also known as  $\alpha$ -aminobutyric acid (ABU), is an AA that is not incorporated into the genetic code, despite its simplicity. We argued that this is for the low gain of diversity, which means the number of isomeric variants (one) per synthetically “invested” car-

bon atoms (four) (Fig. 4 in Grützmann et al.<sup>118</sup>). To recapitulate, one driving force during the origin of the genetic code was diversity<sup>129</sup>. ABU is physico-chemically very similar to the other small AAs alanine and valine, which are already in the code. Therefore, it was likely disfavored as a proteinogenic “building block”. However, it was suggested that ABU or similar small AAs (norvaline, norleucine, and ornithine) were present in an earlier version of the code<sup>133</sup>. The other AAs may have also been selected against because of their similarity. Nevertheless, for its low absolute synthetic costs and flexibility, ABU is used apart from incorporation into proteins<sup>43</sup>.

## 3.2 Alternative splicing in fungi

The idea of this project was to extend the research of AS to the fungal kingdom. Fungi are eukaryotes and have genes harboring introns, which have to be spliced out from the mRNA during gene expression. Opposed to the other not so well investigated eukaryotic domains, Chromista and Protista, many fungi are multi-cellular, and have sophisticated lifestyles<sup>102</sup>, which necessitate a diverse gene inventory. As in animals and plants, one could expect further variation of the genetic information via AS. Indeed, many instances of AS associated fungal genes have been discovered in the last decade<sup>81,148,149,150</sup>. To our knowledge, there are only few systematic studies of AS capacities in the fungal kingdom<sup>78,79</sup>. Also, the question of how AS may contribute to complexity of the fungal lifestyle and cellular structures has not been approached yet in a more global multi-species study. As new transcriptome and genome sequences are uploaded to public databases on a regular basis<sup>151,152</sup>, a bioinformatics analysis of those allows for further clearance of the relevance of fungal AS.

In the beginning of our study, only few NGS transcriptome data sets were available for fungi. In principle, NGS data is preferable for its high sampling depth, good accuracy and low cost. However, a special requirement of our study was also the availability of a well assembled and well annotated genome. This narrowed the selection of data sets, and excluded most NGS studies. Thus, we preferred to base our study on one kind of sequencing technology, with respect to the transcriptome, namely, Sanger sequencing (classical ESTs) and added one species (*Arthroderma benhamiae*) with NGS data for testing.

Our study shows that analysis of NGS reads of greater length (here, *A. benhamiae*, Roche 454, average length 297 bases) is feasible with the same workflow. We indeed find more AS events for this species, which illustrates a higher sensitivity due to the higher sampling depth. Thus, we reason that our method is ready to be applied to NGS data. However, due to different properties, results from short-read NGS data might well have a bias compared to results from classical ESTs. Conclusions on species comparisons would be refutable. To my knowledge, no one has tried to use both, classical ESTs and short-read NGS in a meta-comparison of



### 3. GENERAL DISCUSSION

---

transcriptomes. Furthermore, the wealth of reads from technologies as RNA-Seq should be mapped against the genome sequence using faster methods as, e.g. TopHat<sup>153</sup>. Now-a-days transcriptome sets consist of several millions of reads<sup>154,155</sup> as opposed to ca. one million reads of our test data set. This also makes parallelization of some of the other analysis steps by exploiting common multi-core/CPU architectures reasonable.

In our first study<sup>118</sup> we already found interesting results, which show that it is worth analyzing fungal AS and its significance in more depth. The rate of AS is a first obvious proxy for understanding its significance for a species. A central issue was how to estimate AS rates in a way that they are comparable between species. The attempt in this study was to normalize the rates for the number of sampled introns, that is, dividing the number of AS events of a species by the number of introns detected by ESTs. In view of the wide range of data amounts as well as varying detection and filtering approaches known to the research community, it is questionable whether all so far found AS capacities are representative and comparable. An extreme example are the rates of 53% in mouse, and 24% in rat of one study<sup>73</sup>. For the close evolutionary relation of both species, their strong divergence in reported AS rates are probably an artifact. The abundance and distribution of available transcript samples of a species have to be considered. In cases in which few transcript data are available and cover only small fractions of genes, AS capacities are typically underestimated, whereas vigorous sampling of transcripts enables the detection of rare splicing events that may be consequence of stochastic noise of the splicing machinery. Fox-Walsh and Hertel argued that every multi-exon gene is alternatively spliced with a certain frequency<sup>156</sup>. Thus, the detection of an alternative isoform is only a matter of sensitivity of the method applied, and considering all detectable isoforms can lead to an over-appreciation of a species' AS capacity.

Hence, in our second study on fungal AS<sup>120</sup>, we use the normalization strategy of randomly sampling ESTs to estimate AS rates. Also, we extended the analysis, investigated more features and put them into the greater context of multi-cellular complexity. Random sampling produced AS rates that were independent of the transcript amounts, and are thus more comparable. We found that the AS rates increased with sampling more ESTs per locus (supplemental Figure S2 in Grützmann et al. 2013<sup>120</sup>), presumably because AS events with rarer isoforms were found. Rates quickly reached a plateau, since most species had a low coverage (EST density), and no additional events were found. Note, sampling depth of 10 meant to sample at most 10 ESTs, and still accounting for loci with lower depths. We found that, except for the species with higher EST coverages, the ratios between the AS rates of the species remain nearly the same at different sampling depths. That is, the curves progress with proportional distance. Hence, it seems that if sampling depth is in or below the order of all species' EST coverages, the AS rates are highly comparable. Furthermore, for species with higher EST coverage (*Neurospora*



*crassa* and *Arthroderma benhamiae*), AS rates kept rising with higher random sampling depths, although their ESTs were from different platforms (Sanger sequencing, and Roche 454). Thus, it seems that sequencing technology does not have such a great impact as do read amounts. This holds as long as EST lengths are clearly above exon and intron sizes, so that AS events can be detected unambiguously, as was the case here.

An interesting question is, how far the estimates are away from the real AS capacities. This question also falls back to how frequent a minor splice isoform may occur, that is, the one whose expression is in general lower than that of the so-called major isoform. At a sampling depth of 100 reads per SS, feasible with NGS data, the minor isoform can be present as rarely as in 1% of the cases. It is hard to give a universal lower boundary for a meaningful frequency. How meaningful a boundary is, should also be evaluated with a statistical test. In principle, one mRNA molecule could be enough for translation into several proteins, because most eukaryotic mRNAs have half-lives of several hours<sup>8</sup>. Further, for some processes the presence of a few proteins is enough to have an effect in a cell. This is, because enzymes, for example, are re-used for multiple reactions and their half-lives reach over several orders of magnitude<sup>36</sup>. However, it is doubtful whether a function can be reliably executed when having only a single or few molecules of an mRNA isoform. This is, because small fluctuations could switch off the function completely.

Moreover, it is important to distinguish between functional and spurious or so-called noisy splicing in order to characterize AS capacities. It is estimated that the raw error rate of the spliceosome is in the magnitude of  $10^{-4}$ <sup>156</sup>. Another error source for AS could be transcription. Fox-Walsh and Hertel state that "It is expected that 1 of every 25,000 exons transcribed contains a single splice-site mutation that induces alternative splicing"<sup>156</sup>. This could lead to false positive detected AS events, especially for genes with multiple introns and high EST coverage. Though further steps in an analysis could sort out these events later, it is reasonable to disregard very rare isoforms in first place. In our study, this is achieved by the random sampling step. At a maximal sampling depth of 10, only 10 ESTs from one locus are regarded and assessed for AS. This yields AS events with RNA isoforms of not rarer than 10% after repeating the random selection for 20 times and averaging the observations in terms of found AS events. Though a minor isoform expression cutoff of 10% seems very strict, this number was also chosen because of the low EST coverage in our data, which made deeper sampling unreasonable.

A great part of the fungal AS events are RIs. There are specific error sources that lead to detection of false positive RIs. An mRNA isoform with a RI could mean that the intron was not spliced out but retained by accident. Note, for all other AS types, all isoforms were produced by the spliceosome and the AS event could not be the result of absent spliceosomal processing. Another error source of false positive RIs can be the contamination with not yet processed pre-mRNA.

### 3. GENERAL DISCUSSION

---

With current methods in molecular biology it is difficult to extract only cytosolic mRNA. Hence, EST libraries are made from whole cell mRNA extracts, which could in principle contain unprocessed mRNA from the nucleus (Karol Szafranski, personal communication, December 2012). However, a poly(A) tail capture was made for most of the used libraries. Thus, the detected mRNAs were completely transcribed because polyadenylation is the last step in transcription. Then, because intron splicing occurs predominantly co-transcriptionally<sup>53</sup>, it had likely taken place in these RNAs also. Further, it was found that several spliceosomes can assemble onto one mRNA<sup>54</sup>, both of which means that splicing can occur concurrently and, in principle, be finished fast. This reduces the chance of having false positive IRs due to unprocessed pre-mRNA.

Besides all these hypotheses, we tested the RI-supporting ESTs for evidence of splicing in a similar manner as it was done in another study<sup>157</sup>. For nearly all IR events of our study, there was at least one EST supportive of the RI isoform, that was spliced at another position, that is, it harbored another processed intron elsewhere. Thus, it is no pre-mRNA that we observe, and the spliceosome could have spliced the intron that was actually retained. Still, in these cases, IR could have happened accidentally. It is sensible to limit analyses to regulated, functional AS events. Only in these cases, the cell controls the event and can exploit it for its cellular function. However, it is hard to proof regulation and function with bioinformatics analysis only. Even if retention is not regulated, it will have an effect and the cell has to deal with it. Thus, we did not put further restrictions on the investigated IR cases.

There are other potential roles for mRNA isoforms with RIs that are not encoded into proteins. We showed that only one third of the IR events preserve the reading frame. Stop codons have a very high chance to occur in random sequences (e.g.  $\approx 50\%$  in a 15 triplet long sequence), and thus, only few of the observed RI containing mRNA isoforms can code for alternative proteins. However, they could have an impact on cellular function via coupling of AS with NMD. In such a scenario, splicing (co-)factors are recruited in dependence on external conditions, such as signals transduced from outside the cell, and lead to IR. Then, a premature stop codon in the RI could trigger the NMD pathway, eventually leading to downregulation of this gene's expression. We found that, at least for the fungi present in HomoloGene database, most of the core components of NMD machinery are conserved. Other studies showed already that NMD could have a functional role in fungi. Most RIs of the yeast *Yarrowia lipolytica* contain PTCs, and there is evidence that corresponding transcripts are degraded by NMD<sup>92</sup>. Evidence for functional NMD was seen for *N. crassa*<sup>158</sup>.

Transcripts with RIs could also have a function in evolutionary perspective. Species may test out new alternative gene material and evolve new functions<sup>159,160</sup>. If an isoform does not harm the cell, its expression is unlikely inhibited.

With our study we expanded the knowledge of AS capacities in the fungal kingdom. In their

study, McGuire et al. found varying numbers of AS events in 14 fungi<sup>79</sup>. The AS event numbers for species common to our study are roughly the same, while for *N. crassa* we found many more AS events (860 vs. 20). Although we found similar “raw” numbers for *C. neoformans*<sup>79</sup>, our AS rate estimation is much higher than previously expected (ca. 18% vs. 4%<sup>78,81</sup>). A few studies of fungal AS appeared that are based on NGS transcriptome data. By trend, more AS events were found using this technology compared to our study. For instance, 433 AS events were found for *S. pombe* by Rhind et al.<sup>161</sup>, while we discovered three events. 231 events were found for *Fusarium graminearum*<sup>162</sup>, and we found 42 events in *Fusarium oxysporum*. Though the latter two species are not the same, the numbers may still give a rough orientation because the species are from the same genus. And last, 1375 AS events were found for *A. oryzae*<sup>80</sup>, while we discovered 89 events.

More significant than the mere estimated AS rates is the effect an AS event has on a species’ cellular function, and ultimately, its life. This is out of scope for a sole bioinformatics analysis, and only can be proven with biological observations. However, further evidence can be found beforehand by computational means. One hint can be a protein-coding potential of an alternative isoform<sup>163</sup>. A further strong evidence for functionality is conservation of an isoform in other species, reasoning that a beneficial function would be evolutionary conserved. This also holds for isoforms that are not translated into proteins, because these can have other functions as argued earlier. Another hint at functionality of an AS event would be a condition dependent use of isoforms. An isoform that is observed to be exclusively expressed in a certain growth condition, tissue type, or developmental stage, is likely regulated and has its dedicated function. An example for a conserved AS event in fungi whose isoforms are translated into proteins with different functions, is the gene SKI7/HBS1. Interestingly, a whole genome duplication in *S. cerevisiae* resulted in a loss of this AS event, but both functions were conserved in form of the two sub-functionalized genes SKI7 and HBS1<sup>164</sup>.

The question for this thesis is, if there is complexity via combinations in fungi by the means of AS? One kind of complexity we see in fungi are the diverse types of growth structures, especially sexual structures. We found a coincidence of higher AS rates with more complex structures of the diverse investigated fungal taxa. Especially the simple ascomycetous yeasts (*P. stipitis*, *S. cerevisiae*, *S. pombe*) have lower AS rates than the other fungi. Another coincidence we found is that pathogenic fungi have a higher AS rate than non-pathogenic fungi. Virulence is a complex trait because it involves increased abilities to adapt to changing environmental conditions as during host invasion. For example, the supply with nutrients changes when proceeding through diverse host tissues, as is the case for e.g. *C. albicans* or *A. fumigatus*<sup>165</sup>. Additionally, fungi have to react to the host defense mechanisms for a successful infection<sup>166,167</sup>. Furthermore, the dimorphic switch, the change from yeast to hyphal growth or the other way around, can be seen

### 3. GENERAL DISCUSSION

---

as a complex trait as well, though some non-pathogenic fungi, as *Yarrowia lipolytica*<sup>168</sup>, are also capable to do this.

*UmRrm75* gene of *Ustilago maydis* putatively encoding an RNA binding protein is upregulated during dimorphic switch to filamentous growth induced by acid pH.  $\Delta UmRrm75$  mutants show reduced virulence, mating and post-mating filamentous growth. Interestingly, *UmRrm75* has an alternatively spliced 3' SS, affecting one of its three RNA recognition domains. Although nothing is known about the mode of action of the protein for dimorphism, it is tempting to speculate that the dimorphic switch could be triggered by a switch of the gene's alternative isoform expressions. Each isoform could be involved in determining one of the two fungi's growth forms. In this respect, the highly pathogenic opportunistic fungus *Candida albicans* would be an interesting study object. As a Saccharomycotina it is closely related to the budding yeast and harbours more, yet few, introns<sup>169</sup> than its relative. So far, only few AS events have been discovered, which are not closely involved in pathogenicity<sup>169,170,171</sup>. The hypothesis of an AS-switch for dimorphism could be one of its intricate mechanism for its opportunistic virulence in contrast to budding yeast.

The coincidences of high AS rates with complexity in multicellularity and behavior lead us to speculate that AS could contribute to the phenomenon of multi-cellular complexity. To achieve this, on the one hand, new proteins could be necessary to encode further functions. Which could be done by elevated gene numbers or by additional protein isoforms encoded by the genes. On the other hand, higher complexity can be achieved by more intricate regulation of gene expression, for example, by coupling of AS and NMD as has been speculated already<sup>92</sup>.

In terms of combinations, AS can be seen as a layer of diversification of genetic material between transcription and translation. In this study, we approached the question to which extent AS occurs in fungi. We put special emphasis on comparability of the propensities, which allowed observations of coincidences with functional and cell structural complexity. Overall, the AS capacities even of the higher fungi (Basidiomycetes) seem to be clearly lower than that of higher animals. Still, apart from the ascomycetous yeasts, a considerable fraction of the fungi's genes is involved in AS, and a combination of differentially expressed isoforms is possible. The lower AS capacities also reflect in less intricate AS patterns from single genes. Human genes, e.g., often harbor several alternatively spliced parts, which can be composed to a multitude of different isoforms<sup>172</sup>. We rarely found this kind of combination in fungi. The majority of AS affected genes show only one AS event. Thus, there is few combinatorial complexity form AS in the individual affected fungal genes. We speculate that up to two thirds of fungal IR events could be coupled to NMD because they introduce frame-shifts. This could be a way to regulate gene expression. Thus, there could be multiplicity from combinatorial gene expression regulated or mediated by splicing (co-)factors. However, the coupling of AS and NMD in fungi remains

### 3.3 Mutually exclusive exons - A interdependent type of combinatorial splicing

---

to be investigated, and it can hardly be estimated from current knowledge to which extent it eventually contributes to organismic complexity.

Finally, the lack of a coding potential of many IR events may be a hint on a rather evolutionary role of fungal AS. It was hypothesized that AS may accelerate evolution of new functions<sup>173</sup>. AS could enable a species to test new sequences in alternative isoforms, thereby, leaving the established function of the original isoform unaffected<sup>159,174</sup>. In this respect, we found that two of the over-represented Pfam domains in the AS affected genes are fungi-specific. These may be hotspots of accelerated evolution.

### 3.3 Mutually exclusive exons - A interdependent type of combinatorial splicing

In contrast to cassette exons and retained introns, MXEs have not been investigated as extensively in genome-wide fashion. They mostly occur either as side results in AS databases<sup>175,176</sup>, or in single gene studies<sup>177,178</sup>. Thus, we decided to investigate the extent, structure and features of MXEs in human and mouse. In the years during our study, MXEs have received some more attention with the appreciation of complex splicing patterns in the human transcriptome<sup>172,179</sup>. Also, recently, a genome-wide study of MXEs in *D. melanogaster* was published<sup>180</sup>.

In the classical definition of MXE splicing, exactly one exon out of several appears in the mature mRNA after splicing. In a less strict definition, also all of the involved exons may be spliced out in the mature transcript<sup>172</sup>. The exclusiveness of these exons strongly depends on the considered number of tissues and/or conditions the analyzed transcripts are extracted from. For example, the human gene *TCL6* shows MXE properties when comparing only few tissues. However, the pattern is lost when mRNA from all known tissues is considered<sup>62</sup>. In the end, detection of a MXE splicing event always depends on finding a, potentially very rare, condition for which exclusiveness of the exons is rejected. To approach such an uncertain classification, we suggested that MXEs are categorized as such, as long as the majority of reported transcripts support the exclusive pattern<sup>122</sup>. A similar dichotomy is the rare case of a hand with six fingers, which is still called a hand despite the deviation from the usual case<sup>122</sup>. After all, the question is, if the isoforms have a physiological role that can only be executed with the exclusive occurrences, and if the presence of non exclusive patterns in other tissues and conditions interferes with the function.

MXE is the only AS type where the resultant protein isoforms can have the same length, when only considering AS affected translated mRNA regions. This opens the opportunity to exchange protein parts while maintaining the overall protein structure. Thus, various protein isoforms can easily be encoded that have the same overall function but differ in specificity, e.g., an

### 3. GENERAL DISCUSSION

---

RNA binding protein with sequence specific isoforms. An example are MXEs in chicken beta-tropomyosins that are alternatively incorporated into smooth or skeletal muscle cell mRNA<sup>178</sup>. Another example is a neuronal ion channel, whose activation speed is altered upon exchange of MXEs<sup>177</sup>. Also, MXEs in *D. melanogaster* are enriched in transmembrane transporter and ion-channel activity<sup>180</sup>, corroborating the idea that MXEs contribute to specificity of isoforms. MXEs are thought to have originated from exon duplication<sup>181</sup>. For this reason, one expects MXEs to lie adjacently on the pre-mRNA. Thus, usually, AS analysis pipelines are designed in a way that only discovers adjacent MXEs. Another reason for the design is probably the ease of implementation. When allowing more structural exceptions, the implementation is more difficult, including testing and assuring reliability of the predictions. However, in our analysis we did not restrict the MXEs to lie in close vicinity. In doing so, we found that most of them are non-adjacent. That is, there are constitutively spliced exons between them. We found that non-adjacent MXEs have a higher tendency to cause a frame shift, which makes them more likely to trigger NMD. We also found that non-adjacent MXEs have a bigger length difference and a lower sequence similarity (not published). All this suggests, that they unlikely originate from exon duplication, and an alternative explanation for their origin should be sought.

Another difficulty in implementation is the question of how well-defined the bordering introns and exons must be. That is, if exact genomic positions of adjacent introns and exons are found in all transcripts supporting an MXE isoform. The more strictly defined these borders are, the more reliable are the predicted MXEs. However, implementation of strict borders is difficult because there is significant variation in the exons and introns next to an MXE. This is due to the high splice variation of the human transcriptome, which often yields multiple AS events for one gene<sup>76,172,179</sup>. Adherence to strictly defined MXE borders can lead to exclusion of potential MXEs, merely, because other AS events happen adjacently. This in turn is a significant issue in terms of number of predicted MXEs and comparability to other studies predicting MXEs. Last but not least, the length of sequenced ESTs restricts how well-defined the border of an MXE can be.

We only found MXEs involving two exclusive partners in our study on mouse and human. This is interesting because in less complex organisms as *D. melanogaster* the rather involved MXE patterns with multiple exons were found. However, this does not mean that mammalian splicing patterns are less complex. The average human gene has 7.8 introns<sup>95</sup>, and the average *D. melanogaster* gene has less than four introns per gene<sup>1</sup>. AS patterns in fruit fly should be less intricate in general than in human. There are between 1841 and 4275 AS affected genes reported for *D. melanogaster* in the databases DEDB, FlyBase, ASAP II, ECgenec, which are

---

<sup>1</sup>3.6 introns per gene, calculated from 13,379 reported genes and 48,039 splice junctions, based on Misra et al.<sup>182</sup>



### 3.3 Mutually exclusive exons - A interdependent type of combinatorial splicing

---

about 11-30% of all genes<sup>175</sup>. By contrast, more than 90% of the human genes were found to be involved in AS<sup>75,76</sup>, with an estimated 88,000-132,000 AS events, and at least seven AS events per multiexon gene on average<sup>76</sup>. Even when considering these results obtained from NGS data as overestimation, AS was found to affect more than half of the human genes in previous studies<sup>74,183</sup>. On top of that, both, the absolute number and the proportion of complex AS events (others than IR, CE, A3'SS and A5'SS) are generally higher in vertebrates than in invertebrates<sup>62</sup>. Finally, we found cluster spliced MXEs in human, which are complex forms of MXEs, in which a group of exons is spliced in mutually exclusive manner to another exon or exon group. On top of that, most of these cluster spliced MXEs are non-adjacent, all of which allows for a complex restructuring of coding material.

Opposed to other basic AS types, the way MXE splicing is implemented by the cell is not obvious, and has not been elucidated to full extent. Several regulatory mechanisms have been proposed, some of which were already proven to be utilized in different species. These are spliceosome incompatibility<sup>184</sup>, steric hindrance<sup>185,186</sup>, coordinated splicing control coupled to NMD<sup>187</sup>, and last, a whole class of regulatory mechanisms in which RNA secondary structure elements are involved in exon choice. The latter ones are docker selector pairing<sup>188,189</sup>, induced steric hindrance<sup>190</sup>, and forming of a splicing activation complex directed by RNA pairing and recruitment of *cis*-control elements<sup>189</sup>. Most of the proposed regulatory mechanisms are inappropriate for non-adjacent MXEs. The reasons are given already in our review<sup>122</sup>.

Only the mechanism of forming a splicing activation complex led by *cis*-control elements, is feasible for non-adjacent MXEs, as long as no further details are discovered that may pose restrictions on close exons and introns. Further mechanisms for non-adjacent MXE splicing regulation are conceivable that make use of regulatory microRNAs and epigenetic factors involving chromatin restructuring<sup>65</sup>. Likely, as in adjacent MXE splicing, there are different modes of action, which are used in different species. However, as non-adjacent and clustered MXEs are an only recently appreciated AS type, substantial research on their regulation still has to be undertaken.

It becomes clear that it is not enough to only consider the basic established AS types to understand the meaning of AS of higher eukaryotes. For instance, considering the exons of an MXE event separately, they appear to be skipped exons. One could wrongly conclude that a part of the putatively resultant protein is excised and a certain function is switched off, or a spliced-out localization sequence yields the transport of the protein to another place in the cell. However, only when appreciating the whole MXE event, it is clear that the protein part is actually exchanged with another one. Now this is a completely new property that can not be realized by two ordinarily skipped exons. Moreover, it is possible that the disallowed isoforms of some MXEs are deleterious. Likewise, one can imagine emergence of new functions by

### 3. GENERAL DISCUSSION

---

other dependent or even more complex AS patterns. While considering only local patterns and analyzing the basic AS types allows for a rough scan of the AS capacities, only the analysis of their combinations, if there are any, enables the proper understanding of the involved biological phenomena. Therefore, the goal must be to delineate the complete AS patterns of all genes as well as their conditions and locations of expression. As mentioned earlier, some scientists speak of an *AS code*, with reference to the notion *genetic code*. It is a set of rules that determines the outcome of AS on transcript level<sup>191,192</sup>. Clearly, the code can not be a direct translation of symbols as in case of the genetic code. First, the current knowledge points at such a code to consist of many components that act in cumulative manner, meaning the outcome is determined by the sum (or difference) of their influence<sup>65</sup>. Second, there will be code components that are rather soft determiners of the AS outcome. Examples are the concentration of *trans*-factors, or the similarity of found sequences to known *cis*-splicing motifs. This also can necessitate the modelling of probabilities rather than fixed rules, because the mere presence of a factor in the cell is no guarantee that it is effective, e.g. by binding. And third, AS is tightly involved in other processes of gene expression, for example transcription<sup>67,68</sup>. Thus, an AS code can not be a simple mapping, or the application of a few simple rules. In sum, I propose to use the notion “code” only with caution. One step of success was the tissue specific prediction of exon inclusions with high accuracy involving around 200 features derived from transcript sequences<sup>193</sup>. While the actual regulatory relation is hidden and only indirectly represented in the algorithmic predictor, feature maps can be generated that can guide mechanistic studies<sup>193</sup>. Still, an immense investigative task remains: the incorporation of further tissue types, the influence of external and internal stimuli and conditions, and especially, the prediction of more involved AS patterns into such an algorithm.

An interesting question is, if dependencies as in MXE splicing could be predicted by more sophisticated computer programs. Simply, if the dependencies are caused by or correlated with features that are learned by or in other ways implemented in the algorithm, such a prediction is feasible. Probably, as extension of the study of Barash et al.<sup>193</sup>, a fully accurate predictor will involve features of *trans*-acting splice factors, too. Once, the putatively black box-like predictor is trained, the most significant learned features could be extracted. Using these features, hypotheses for MXE regulation mechanisms could be generated that are testable in biological experiments.



## 3.4 Combinatorial complexity - final discussion

### 3.4.1 Complexity through combination - a universal principle

The central theme of this thesis is complexity arising from combination. Two areas were investigated in which this reflects: aliphatic AA side chains, and alternative mRNA splicing. Combination is a universal principle in biology. Many biological systems comprise distinct parts or building blocks, which are combined to form a whole. One great benefit of this principle is, that novelties in nature emerge by combination in a fast and inexpensive way. Here, inexpensiveness coins, e.g., low metabolic costs and few effort in genetic coding. To corroborate the hypothesis of the universality of the principle, I will give some examples of implementations in nature.

Nucleotides are combined in sequential order to code genes in the DNA of chromosomes. Their order strictly determines the order of AAs, which are combined into proteins<sup>8</sup>. In case of functional non-coding RNA, ribonucleotides are combined to RNA molecules that form secondary structures to execute their function<sup>194</sup>. Spatial structure and physico-chemical properties of proteins and functional RNAs are determined by the order and properties of their building blocks. Their sequential combination is influenced by different mechanisms taking place on different time scales. One way of combination is the genetic recombination of gene material during meiosis. During the process of crossing-over, homologous chromosomes of different mating partners are aligned and DNA sequences are exchanged by the chromosomes, necessitating DNA strand breaks and ligations<sup>8</sup>. In another example, on the scale of populations, new nucleotide combinations can emerge by horizontal gene transfer. It was shown that, e.g., bacteria and fungi interchange coding DNA sequences and thus novel gene material can spread in a population of species<sup>195,196</sup>. This material is not only tolerated but also used by its new host, in that, for instance, resistances against antibiotics via special encoded enzymes can quickly propagate in bacteria<sup>197</sup>.

Another example on cellular level, cell types can be seen as combinations of organelles. Following the endosymbiont theory, for instance, mitochondria arose by incorporation of prokaryotes into the bigger eukaryotic cells million years ago<sup>198</sup>. Another example of organelles contributing to cell type identity in chordates are the number of nuclei, which ranges from zero in red blood cells to several hundreds in muscle cells<sup>8</sup>.

A relatively new research field is the influence of histone marks on gene expression. Histones can be post-translationally modified at diverse AA positions and with several types of modifications, for example phosphorylation, methylation, and acetylation<sup>16</sup>. These marks are not static but change during a cell's life and determine expression of a nearby gene in a combinatorial way<sup>199</sup>. Interestingly, these modifications can be inherited and constitute an alternative way of carrying

### 3. GENERAL DISCUSSION

---

over information to a succeeding generation<sup>200</sup>.

The multitude of immunoglobins, which are needed by B cells of the immune system to defend against a yet unseen antigen of a new pathogen, are constructed from a small set of antibody genes by different genetic mechanisms. One of them is the V(D)J recombination of the variable domains of antibodies, during which a random combination of various genomic segments is selected during B cell maturation in the bone marrow<sup>201</sup>.

The last example is the conjoint expression of genes in cells. Though (almost) all somatic cells of a species carry the same genomic blueprint, only distinct expressed combinations of them determine cell and development type identity in metazoan. On a more abstract level on which genes encode proteins that take part in metabolic pathways, a cell can also be seen as a combination of a subset of all possible such pathways. For example, in contrast to skeletal muscle cells, red blood cells do not produce energy via the Krebs cycle, but rather use the Embden-Meyerhof pathway<sup>36</sup>. Another example are amino acid (AA) synthetic pathways. There are groups of AAs that share the same pathway. While the beginning reactions are the same for similar AAs, the later parts are branched, and other enzymes complete the AAs synthesis<sup>133</sup>. In the following, I will show how combinations underlie the two major topics of my thesis, alternative splicing and side chains of aliphatic AAs. Thereby, it will become evident that conceiving the combinatorial structure of these systems is often critical for understanding their entire nature.

#### 3.4.2 Combinatorial complexity beyond the genetic code

Only few amino acids are used for the assembly of proteins, despite the vast potential from the combinatorial nature of their side chain structure. Even though we restricted the study to aliphatic AAs without rings and double bonds the calculation of the combinations gets rather complicated. AAs with aromatic rings, which appear in natural proteins, could only be considered if double bonds and rings were allowed. Furthermore, if one wanted to also regard the proteinogenic AA proline, rings without double bonds and bonds to the AAs' amine nitrogen would have to be considered as well. Then, AAs derived from ring shaped azacycloalkanes could also be regarded. These are another class of AAs with relevance in nature. For example, azetidine-2-carboxylic acid is a toxin in *Liliaceae* plants<sup>202</sup>, and piperidine-2-carboxylic acid occurs in the lysine metabolism of the murine brain<sup>203</sup>. However, in this case, one may have to set a threshold for the allowed size of rings, because such an AA with a three-membered ring is under strain, very reactive, and thus toxic<sup>204</sup>. It was not found in living organisms so far<sup>204</sup>. In the end, despite a better appreciation of occurrences in nature, these kind of exceptions and extensions would inflate the calculation formula considerably. At some point, calculation would be too complicated and simulations of the chemical structures would be necessary.

### 3.4 Combinatorial complexity - final discussion

---

The main finding was that the number of theoretical possibilities given the restrictions grows exponentially. This is the essence of the potential of combinations. A small set of micro-level rules, yields an inconceivable number of combinations on the macro-level. For the assembly of proteins, there are two micro-levels involved in combinations. On the first one, chemical elements are combined to molecules, in this case AAs with diverse side chains. On the second level, the AAs are united to chains by peptide bonds. On both levels, combinations have exponential potential in terms of their compounds: there are approximately  $2.8^n$  aliphatic side chains from  $n$  carbon atoms, given the restrictions of our study, and there are  $20^m$  possible proteins of length  $m$  when having a “construction kit” of 20 different AAs. Interestingly, nature seems to exploit a bigger part of the potential for the construction of proteins only on the second level. In principle, a cell could choose from  $(2.8^n)^m$  possibilities to construct its proteins. However, for the reasons I gave already above that involve an efficient genetic code (3.1, page 87), there are exactly 20 AAs with the exact partitioning of the code as we see it. Opposed to this small “construction kit”, the amount of stored information need not to be finite or tractable. A cell’s genome comprises thousands to billions of nucleic acid bases of data harboring genetic information<sup>8,205</sup>. For these reasons, on the first micro-level of combinatorial protein diversity, only few of the many possible combinations are realized in form of proteinogenic AAs. However, many more of the possibilities of the second level of combination are realized: several million of AA sequences have been found and characterized so far<sup>206</sup>.

Still, the amount of constructable, meaningful proteins with specific functions is not nearly grasped by research. This reflects in the task of predicting protein structures from AA sequences that keeps challenging scientists for decades<sup>207</sup>. Not less difficult are the reliable prediction of protein function<sup>208,209</sup> and protein interaction<sup>210,211</sup>, which are particularly difficult in cases where no homologs of the underlying sequences are available. Despite the great variety and universality of proteins constructable via the genetic code, nature seems to have the need for further expansion. This happens on many levels, one of which is the expansion of the genetic code in form of the encoded AAs selenocysteine (Sel) and pyrrolysine (Pyl)<sup>39,40,41</sup>. As of the year 2003, there were 25 selenoproteins known in human, most of which have Sel in their enzymatic active site<sup>212</sup>. These proteins could probably not execute their function as efficiently without Sel, because selenium makes it more reactive than sulfur in the analogous cysteine<sup>213</sup>. Similar to Sel, Pyl is situated in the active site of enzymes, for example in methyltransferases. Its pyrroline ring is assumed to take a specific role in catalyzation<sup>214</sup>. Again, this AA probably contributes to a more efficient execution of catalytic functions than alternative AAs would do.

Standard codons are re-purposed by the incorporation of specific RNA secondary structures to realize the incorporation of Pyl and Sel. Besides encoding further AAs, many other codons were found to have dual functions. Examples are the utilization of the leucine codon CUG for

### 3. GENERAL DISCUSSION

---

serine in the yeast *Candida cylindracea*, the read through of stop codons enhanced by *cis*- and *trans*-acting factors, and the coding of the codon UGA for tryptophan instead of a translation stop in mitochondria of vertebrates<sup>42</sup>.

Another expansion of the protein variety is realized in form of post-translational modifications (PTMs), as e.g. phosphorylation, methylation, glycosylation, hydroxylation, and acetylation<sup>16</sup>. Prohaska et al. showed that complexity in chromatin regulation indeed increased in step-wise fashion during evolution<sup>215</sup>. With respect to the extent of PTMs, it was found that 16% of the Swiss-Prot annotated proteins are modified as observed experimentally, and around 44% are putatively modified<sup>216</sup>. Some types of PTMs add chemical groups that can not be found in proteins right after translation (e.g. phosphate), and thus, result in specific properties that can not be encoded in AA sequences directly. Some changes result in addition of small chemical groups (e.g. alkyl groups). During others, whole proteins or peptides are added (e.g. SUMOylation<sup>217</sup>). Again others result in structural changes (e.g. disulfide bonds). In sum, PTMs offer a very high degree of diversification of protein structures and functions. Many features introduced by PTMs are not achievable by direct encoding into AA sequences using the current genetic code.

In respect of these extensive modifications, the genetic code can be regarded as an “accident”, i.e. being suboptimal, in the form in which it became nearly “frozen” a long time ago. However, I hypothesize that the later addition of diversity through PTMs was the optimal solution in the overall perspective. A first small and tractable code was advantageous for the primordial species of the early history of life. Diversifying expansion in form of, e.g., PTMs is advantageous for the exploration of and specification into niches, presumably not until less relative costs of a species were spent for mere survival. Setting up a more extended, flexible code first would have had a negative effect on the fitness of the simple early organisms.

PTMs have the further benefit of adding specificity fast. Some PTMs, as phosphorylation and acetylation, are only temporal. They are, for example, used for signal transduction<sup>8</sup>. They have a switch-like, alternating nature: while kinases add phosphate groups, phosphatases remove these<sup>8</sup>. It is rather difficult to encode this functionality into the primary sequence.

The need for diverse “building blocks” can also be seen on the level of nucleic acid sequences. The four standard nucleotides are sufficient for secondary structure formation in single-stranded nucleic acids. However, the alphabet of tRNA is expanded for several unusual nucleotides derived by, for example, methylated standard nucleotides<sup>36</sup>. These allow for interactions with synthetases and ribosomes by better accessibility or higher hydrophobicity<sup>36</sup>.

#### 3.4.3 Combinatorial complexity in alternative splicing

In general, AS is understood to be a major contributor of the complex phenotype of higher eukaryotes. This is by encoding protein isoforms with distinct cellular destinations, switching

### 3.4 Combinatorial complexity - final discussion

---

off binding capacities of transcription factors, altering enzyme catalytic sites and many more different properties<sup>65</sup>. AS can be seen as the recombination of parts of pre-mRNA. An AS event gives rise to at least two mRNA isoforms that can distinguish in as few as one nucleotide, up to as many as several hundred nucleotides. Complexity arises by combination, especially when an mRNA is affected by multiple AS events. On average, a human gene has nearly eight introns<sup>95</sup>. More than 90% of all human genes are affected by AS, and there are at least seven AS events per multi-exon gene on average<sup>76</sup>. An example for the huge combinatorial potential of AS are the neurexin genes with more than 2000 putative isoforms<sup>218</sup>. And again, the extreme example in *D. melanogaster* is the DSCAM gene. It contains four clusters of 12, 48, 33 and 2 exons. Out of each cluster, one exon is selected in a mutually exclusive way, respectively. If the selection happens independently,  $12 \cdot 48 \cdot 33 \cdot 2 = 38,016$  isoforms could be produced from this gene<sup>63</sup>.

A further diversification by AS is achieved by differential expression of isoforms from different genes that act together to have an outcome. The combinatorial effect of splice isoforms can be seen during chaperone synthesis during proteotoxic stress response. The synthesis is mainly regulated by the transcription factors (TFs) Heat Shock Factor 1 and 2 (HSF1, HSF2), which have two alternative isoforms. It was found that the HSF2 $\beta$  isoform inhibits HSF1 $\beta$  activity and that the isoform expression ratio of both factors determines the cellular level of stress response in a quantitative way<sup>219</sup>.

On a more systemic level, AS contributes to organ development and disease. For example, the heart consists of several tissues and cell types with specific functions that allow to execute the central role of this organ in a concerted way. AS crucially adds diversity to the transcriptome that is needed for the complex structures and functions of a heart. For example, cardiac troponin T, and Ca<sup>2+</sup>/calmodulin-dependent protein kinase mRNAs are alternatively spliced. In addition, many SR proteins, hnRNP proteins and other splicing *trans*-factors were found to regulate heart development<sup>220</sup>.

In the beginning of AS research, AS may have only been appreciated as a process to add protein isoforms with a slightly different function or, which even may have been doubted to have any biological relevance. With nowadays perspective, many cellular processes can only be fully understood when taking the role of AS into account. Another intriguing example next to the above mentioned heat shock factor genes, is a tri-geminal ganglion-specific splice isoform of a cation channel in vampire bats. It enables the bats to sense warm-blooded animals in contrast to fruit-feeding bats whose orthologous receptor gene does not produce the necessary isoform<sup>221</sup>. Some AS events do not only influence a single function of a terminal protein with a restricted involvement, but affect many downstream processes or even play a central regulatory role as controlling transcriptional networks in animal development<sup>222</sup>. An example is a switching AS event specific to embryonic stem cells that changes the DNA binding affinity of the FOXP1

### 3. GENERAL DISCUSSION

---

TF. This in turn, induces the expression of central TFs of pluripotency and inhibits expression of genes for differentiation<sup>223</sup>. In another example, AS affects many conserved protein-coding regions during the transition from mitotic to meiotic growth of murine sperm cells. Moreover, many key regulators of splicing were differentially regulated<sup>224</sup>. AS may significantly affect global transcriptome expression, and contribute to the fundamental change to meiotic growth. Splicing is often involved in such complex phenomena as diseases. It is estimated that more than one third of single nucleotide polymorphisms causing diseases affect splicing<sup>225</sup>. These effects are mainly mediated via disruption of core and *cis*-splicing elements<sup>225</sup>. An alteration in splicing can not only cause a single form of a disease, but AS can contribute to diversity of complex diseases as, e.g., prostate cancer. This impairs standardization of therapies and makes individualized treatment necessary<sup>226</sup>. Furthermore, also different mutations of splicing factors can yield subtypes of diseases, as was shown in a recent study on myelodysplastic syndrome<sup>227</sup>.

#### 3.4.4 Costs of the combinatorial principle

Undoubtedly, there is a vast potential behind creating biological complexity through combinations. The main benefit is that new things as macromolecular structures or biological functions can be created from using already existing parts. However, there are significant costs involved. Often there has to be a kind of machinery that assembles, recombines or interprets the combinations. Examples are the spliceosome with its several 100 compounds<sup>50</sup>, the molecules involved in setting and reading histone marks, and the enzymes responsible for V(D)J recombination of antibodies<sup>201</sup>. Those combination-managing systems require to be produced and assembled themselves, as well as be modified and transported to their right cell locations. This all involves costs in terms of time and energy, the latter one often in form of ATP. For AS, further costs involve the degradation of the excised introns (lariat)<sup>36</sup>, and the buildup and management of the compounds involved in NMD, if this pathway is used. The mutual dependencies of the investigated MXEs necessitate additional controlling factors as proteins, which go not without expenses. As discussed earlier, large AA side chains have a higher potential in structural diversity. The downside here is that large side chains also involve higher costs in production in terms of energy consumption and time needed before they can be used<sup>143</sup>. Significant costs are involved in the implementation of the genetic code. Transfer RNA have to be synthesized as well as their corresponding aminoacyl synthetases, which connect anticodons with AAs. Similar to the spliceosome, yet not that large, is the ribosome - a ribonucleoprotein whose heterogeneous subunits have to be produced and assembled. Eventually, all the information has to be stored in a huge genome, consisting of nucleic acids and histones, all of which have to be produced too. DNA has to be replicated, restructured and DNA damage must be repaired. This consumes large amounts of energy<sup>36</sup>.

Another downside of managing combinatorial complexity is the evolutionary time needed to evolve such machineries. In short term perspective, there is an advantage to encode and build complex structures directly. For example, having a gene expression regulator that binds the DNA and inhibits a specific gene's transcription seems more practicable than adding splice regulators and a whole NMD machinery for the downregulation, as introduced earlier. However, once a combination-managing apparatus is developed, it can be used universally with slight modifications. As examples, the NMD pathway is used for many genes<sup>91</sup>, and the recombination system is built for many meiotic divisions to come<sup>36</sup>. Then, the gained evolutionary benefit probably made up for the put about effort during evolution. In the light of the universality of the combinatorial principle, this conclusion can be drawn for many aspects of biological complexity.

#### 3.4.5 Restriction of combinations

Another phenomenon of the principle of complexity by combination is the restriction of the huge combinatorial potential. One inherent restriction lies in the fact that the time is not sufficient to test every of the often exponentially increasing number of possibilities. One example is that the immune system usually needs some hours up to days to find a matching antibody to an antigen<sup>201</sup>. Here, finding an appropriate combination fast, decides about nothing short of the mere survival of the whole organism. In another example, there are  $20^{361} \approx 4.7 \cdot 10^{469}$  possible eukaryotic proteins when calculating with the median protein length of 361 AAs<sup>228</sup>. Many of those are far from having a structural relevance (e.g. 361 times alanine). However, even with a heuristic search and only testing of, at first glance not meaningless proteins, likely, only a marginal fraction has ever been tested by nature. Thus, there is space for evolutionary improvement even if we froze the current environmental conditions so that species only had to optimally adapt to a static surrounding. A practical reason for restriction is that novelties remain controllable. If for instance, new proteins were tested at high throughput and in parallel, the chance of having harmful outcomes with uncontrollable effects for a species would be high. From an evolutionary perspective, populations would have a high risk of extinction. This also reflects in the example of extending the set of proteogenic AAs by ones with new side chains. Indeed, an evolutionary very slow expansion took place. Only selenocysteine and pyrrolysine have been discovered so far<sup>39,40,41</sup>. The reason is, that the genetic code is a very central system on which many cell processes depend in a fundamental way. For both new AAs, the reuse of existing stop codons via additional signals, the secis<sup>229</sup> and pylis<sup>230</sup> RNA structure elements, was likely the easiest way of incorporation. A reprogramming of parts of the genetic code, that is, the re-assignment of anticodons to different AAs, would certainly have had great negative impact on the affected species, or would have taken a very long time for evolution.



### 3. GENERAL DISCUSSION

---

For these reasons, the exploration of the combinatorial potential often takes place in a piece by piece fashion, whereupon only few novelties are tested at a time. This is, for instance achieved, by very low mutation rates of species genomes<sup>231,232</sup>. Though meiotic recombinations yield huge numbers of new combinations, the exchanged parts (alleles) are very similar, both in mere sequence and actual function. There are combinations that show clearly altered phenotypic outcomes, for example, for aging<sup>233</sup> and disease<sup>234</sup>. However, new allele combinations have no frequent, dramatically different outcomes - a claim that is underlined by billions of successful recombinations during mankind's history. With evolutionary selection, combinations that lead to detrimental results usually have a lower fitness resulting in lower chance to maintain in a population. Thus, the effect with respect to populations is even smaller. An example for piece by piece exploration in AS is, that cassette exons that arose newly from exonization of *Alu* elements have low inclusion rates. That is, the expression of the isoform harboring the new exon is lower than that of the original isoform<sup>52</sup>. This way, a potential malicious effect of the minor isoform can be held down more easily.

A piece by piece discovery of combinations is unreasonable for the vast amount of recombinations that need to be screened during antibody construction. The immune system has developed special mechanisms to avoid deleterious effects resultant from recognition of antigens of the own body. Lymphocytes are tested for the reactivity to these self antigens during maturation, and undergo apoptosis if such a reaction is found<sup>201</sup>.

MXEs harbor an interesting type of restriction. An MXE of two exons could be regarded as two cassette exons for which two of the possible splice combinations are inhibited: the exclusion and inclusion of both exons. While at first glance, this restricts the potential of combinations, it implements a specific function, as explained earlier (section 3.3, page 97). Such a function is not easily achievable without the avoidance of certain splice combinations in MXEs.

Summarizing, through the various types of restrictions of combinations, species are able to better control the vast potential of combinatorial complexity. This enables them to keep the chance for deleterious effects low, and to develop and implement new cellular functions at the same time.



# References

1. J. C. VENTER, M. D. ADAMS, E. W. MYERS, ET AL. **The sequence of the human genome.** *Science*, **291**(5507):1304–1351, Feb 2001. 1
2. J. HARROW, A. FRANKISH, J. M. GONZALEZ, ET AL. **GENCODE: the reference human genome annotation for The ENCODE Project.** *Genome Res*, **22**(9):1760–1774, Sep 2012. 1, 2
3. M. B. GERSTEIN, Z. J. LU, E. L. VAN NOSTRAND, ET AL. **Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project.** *Science*, **330**(6012):1775–1787, Dec 2010. 1
4. R. BRENCHELEY, M. SPANNAGL, M. PFEIFER, ET AL. **Analysis of the bread wheat genome using whole-genome shotgun sequencing.** *Nature*, **491**(7426):705–710, Nov 2012. 1
5. T. CAVALIER-SMITH. **Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox.** *J Cell Sci*, **34**:247–278, Dec 1978. 1
6. J. M. CLAVERIE. **Gene number. What if there are only 30,000 human genes?** *Science*, **291**(5507):1255–1257, Feb 2001. 1
7. F. CRICK. **Central dogma of molecular biology.** *Nature*, **227**(5258):561–563, Aug 1970. 1
8. H. LODISH, A. BERK, C. A. KAISER, ET AL. *Molecular Cell Biology (Lodish, Molecular Cell Biology)*. W. H. Freeman, New York, USA, 6th edition, June 2007. 1, 2, 4, 5, 90, 93, 101, 103, 104
9. H. KOGA, S. KAUSHIK, AND A. M. CUERVO. **Protein homeostasis and aging: The importance of exquisite quality control.** *Ageing Res Rev*, **10**(2):205–215, Apr 2011. 2
10. J. ZHANG, G. VEMURI, AND J. NIELSEN. **Systems biology of energy homeostasis in yeast.** *Curr Opin Microbiol*, **13**(3):382–388, Jun 2010. 2
11. D. E. KOSHLAND, JR. **Special essay. The seven pillars of life.** *Science*, **295**(5563):2215–2216, Mar 2002. 2
12. M. F. ROJAS-DURAN AND W. V. GILBERT. **Alternative transcription start site selection leads to large differences in translation activity in yeast.** *RNA*, **18**(12):2299–2305, Dec 2012. 2
13. D. MCSHEA. **Functional complexity in organisms: parts as proxies.** *Biology and Philosophy*, **15**:641–668, 2000. 2, 3
14. M. W. WRIGHT AND E. A. BRUFORD. **Naming 'junk': human non-protein coding RNA (ncRNA) gene nomenclature.** *Hum Genomics*, **5**(2):90–98, Jan 2011. 2
15. A. MINSKY. **Information content and complexity in the high-order organization of DNA.** *Annu Rev Biophys Biomol Struct*, **33**:317–342, 2004. 2
16. C. KANNICHT, editor. *Posttranslational Modification of Proteins: Tools for Functional Proteomics (Methods in Molecular Biology)*, **194**. Humana Press, Berlin, Germany, 2002. 2, 101, 104
17. J.-F. COUTURE AND R. C. TRIEVEL. **Histone-modifying enzymes: encrypting an enigmatic epigenetic code.** *Curr Opin Struct Biol*, **16**(6):753–760, Dec 2006. 2
18. C. E. SHANNON. **A Mathematical Theory of Communication.** *Bell System Technical Journal*, **27**(3):379–423, 1948. 2
19. Y. ZHANG, H. LIU, J. LV, ET AL. **QDMR: a quantitative method for identification of differentially methylated regions by entropy.** *Nucleic Acids Res*, **39**(9):e58, May 2011. 2
20. J. LEE, D. MCMANUS, AND K. CHON. **Atrial Fibrillation detection using time-varying coherence function and Shannon Entropy.** *Conf Proc IEEE Eng Med Biol Soc*, **2011**:4685–4688, 2011. 2
21. F. M. REZA. *An Introduction to Information Theory*. Dover Publications, New York, USA, 2010. 2
22. D. SOSA, P. MIRAMONTES, W. LI, ET AL. **Periodic distribution of a putative nucleosome positioning motif in human, nonhuman primates, and archaea: mutual information analysis.** *Int J Genomics*, **2013**:963956, 2013. 2
23. I. GROSSE, H. HERZEL, S. V. BULDYREV, ET AL. **Species independence of mutual information in coding and noncoding DNA.** *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, **61**(5 Pt B):5624–5629, May 2000. 2
24. A. N. KOLMOGOROV. **On tables of random numbers.** *Sankhya Ser. A*, **25**:369–376, 1963. 2
25. T. BROWN. *Genomes*. Wiley-Liss, Manchester, UK, 2002. 2, 6, 7
26. E. N. TRIFONOV. **Making Sense of the Human Genome.** In R. H. SARMA AND M. H. SARMA, editors, *Structure and Methods: Human Genome Initiative and DNA Recombination*, **1**, pages 69–77, Schenectady, USA, 1990. Adenine Press. 3
27. E. N. TRIFONOV. **The multiple codes of nucleotide sequences.** *Bull Math Biol*, **51**(4):417–432, 1989. 3
28. Y. CUI, G. KANG, K. SUN, ET AL. **Gene-centric genomewide association study via entropy.** *Genetics*, **179**(1):637–650, May 2008. 3
29. W. RITCHIE, S. GRANJEAUD, D. PUTHIER, ET AL. **Entropy measures quantify global splicing disorders in cancer.** *PLoS Comput Biol*, **4**(3):e1000011, Mar 2008. 3
30. L. SHAMIR, C. A. WOLKOW, AND I. G. GOLDBERG. **Quantitative measurement of aging using image texture entropy.** *Bioinformatics*, **25**(23):3060–3063, Dec 2009. 3

## REFERENCES

31. J. T. BONNER. *The Evolution of Complexity by Means of Natural Selection*. Princeton University Press, Princeton, USA, 1988. 3
32. K. KRIPPENDORFF. **Combinatorial Explosion - Web Dictionary of Cybernetics and Systems**, 2013. <http://pespmc1.vub.ac.be/ASC/COMBIN.EXPLO.html>, accessed 17 June 2013. 3
33. L. VAUQUELIN AND P. ROBIQUET. **The discovery of a new plant principle in *Asparagus sativus***. *Ann.Chim.*, **572**(1):88–93, 1806. 4
34. W. WOLLASTON. **On cystic oxide: a new species of urinary calculus**. *Trans. R. Soc. London*, **100**:223–230, 1810. 4
35. J. S. FRUTON. *Contrasts in Scientific Style: Research Groups in the Chemical and Biochemical Sciences (Memoirs of the American Philosophical Society)*. Amer. Philosophical Society, Philadelphia, USA, 1990. 4
36. J. M. BERG, J. L. TYMOCZKO, AND L. STRYER. *Biochemistry*. W. H. Freeman, New York, USA, 2002. 4, 5, 93, 102, 104, 106, 107
37. A. LESK. *Introduction to Bioinformatics*. Oxford University Press, New York, USA, 3 edition, 2008. 4
38. C. LEVINthal. **How to Fold Graciously**. In J. T. P. DEBRUNNEN AND E. MUNCK, editors, *Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House*, pages 22–24, Monticello, USA, 1969. University of Illinois Press. 4
39. B. J. LEE, P. J. WORLAND, J. N. DAVIS, ET AL. **Identification of a selenocysteyl-tRNA(Ser) in mammalian cells that recognizes the nonsense codon, UGA**. *J Biol Chem*, **264**(17):9724–9727, Jun 1989. 5, 103, 107
40. W. LEINFELDER, T. C. STADTMAN, AND A. BÖCK. **Occurrence in vivo of selenocysteyl-tRNA(SERUCA) in *Escherichia coli*. Effect of sel mutations**. *J Biol Chem*, **264**(17):9720–9723, Jun 1989. 5, 103, 107
41. G. SRINIVASAN, C. M. JAMES, AND J. A. KRZYCKI. **Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA**. *Science*, **296**(5572):1459–1462, May 2002. 5, 103, 107
42. A. V. LOBANOV, A. A. TURANOV, D. L. HATFIELD, ET AL. **Dual functions of codons in the genetic code**. *Crit Rev Biochem Mol Biol*, **45**(4):257–265, Aug 2010. 5, 104
43. E. RIGA, D. R. HOLDSWORTH, R. N. PERRY, ET AL. **Electrophysiological analysis of the response of males of the potato cyst nematode, *Globodera rostochiensis*, to fractions of their homospecific sex pheromone**. *Parasitology*, **115** (Pt 3):311–316, Sep 1997. 5, 91
44. X.-F. MING, A. G. RAJAPAKSE, J. M. CARVAS, ET AL. **Inhibition of S6K1 accounts partially for the anti-inflammatory effects of the arginase inhibitor L-norvaline**. *BMC Cardiovasc Disord*, **9**:12, 2009. 5
45. V. IVANOVA, M. ORIOL, M. J. MONTES, ET AL. **Secondary metabolites from a *Streptomyces* strain isolated from Livingston Island, Antarctica**. *Z Naturforsch C*, **56**(1-2):1–5, 2001. 5, 90
46. F. H. CRICK, L. BARNETT, S. BRENNER, ET AL. **General nature of the genetic code for proteins**. *Nature*, **192**:1227–1232, Dec 1961. 5, 88
47. M. NIRENBERG, P. LEDER, M. BERNFIELD, ET AL. **RNA codewords and protein synthesis, VII. On the general nature of the RNA code**. *Proc Natl Acad Sci USA*, **53**(5):1161–1168, May 1965. 5, 88
48. M. YARUS. *Life from an RNA World: The Ancestor Within*. Harvard University Press, Cambridge, USA, April 2010. 5
49. A. ELZANOWSKI AND J. OSTELL. **The Genetic Codes**, April 2013. <http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi>, accessed 26 June 2013. 5
50. A. A. HOSKINS, L. J. FRIEDMAN, S. S. GALLAGHER, ET AL. **Ordered and dynamic assembly of single spliceosomes**. *Science*, **331**(6022):1289–1295, Mar 2011. 6, 106
51. T. W. NILSEN. **The spliceosome: the most complex macromolecular machine in the cell?** *Bioessays*, **25**(12):1147–1149, Dec 2003. 6
52. E. KIM, A. GOREN, AND G. AST. **Alternative splicing: current perspectives**. *Bioessays*, **30**(1):38–47, Jan 2008. 6, 7, 9, 108
53. G. DUJARDIN, C. LAFAILLE, E. PETRILLO, ET AL. **Transcriptional elongation and alternative splicing**. *Biochim Biophys Acta*, **1829**(1):134–140, Jan 2013. 6, 7, 94
54. Y. BRODY, N. NEUFELD, N. BIEBERSTEIN, ET AL. **The in vivo kinetics of RNA polymerase II elongation during co-transcriptional splicing**. *PLoS Biol*, **9**(1):e1000573, 2011. 6, 94
55. A. A. PATEL AND J. A. STEITZ. **Splicing double: insights from the second spliceosome**. *Nat Rev Mol Cell Biol*, **4**(12):960–970, Dec 2003. 6
56. S. M. BERGET, C. MOORE, AND P. A. SHARP. **Spliced segments at the 5' terminus of adenovirus 2 late mRNA**. *Proc Natl Acad Sci*, **74**(8):3171–3175, Aug 1977. 6
57. M. G. ROSENFELD, S. G. AMARA, B. A. ROOS, ET AL. **Altered expression of the calcitonin gene associated with RNA polymorphism**. *Nature*, **290**(5801):63–65, Mar 1981. 6
58. J. MERKIN, C. RUSSELL, P. CHEN, ET AL. **Evolutionary dynamics of gene and isoform regulation in Mammalian tissues**. *Science*, **338**(6114):1593–1599, Dec 2012. 7
59. H. BAO, E. LI, S. D. MANSFIELD, ET AL. **The developing xylem transcriptome and genome-wide analysis of alternative splicing in *Populus trichocarpa* (black cottonwood) populations**. *BMC Genomics*, **14**:359, 2013. 7
60. F. KEMPKEN. **Alternative splicing in ascomycetes**. *Appl Microbiol Biotechnol*, **97**(10):4235–4241, May 2013. 7
61. L. LA VIA, D. BONINI, I. RUSSO, ET AL. **Modulation of dendritic AMPA receptor mRNA trafficking by RNA splicing and editing**. *Nucleic Acids Res*, **41**(1):617–631, Jan 2013. 7
62. M. SAMMETH, S. FOISSAC, AND R. GUIGÓ. **A general definition and nomenclature for alternative splicing events**. *PLoS Comput Biol*, **4**(8):e1000147, Aug 2008. 7, 97, 99
63. B. R. GRAVELEY. **Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures**. *Cell*, **123**(1):65–73, Oct 2005. 7, 105

## REFERENCES

64. G. AST. **How did alternative splicing evolve?** *Nat Rev Genet*, **5**(10):773–782, Oct 2004. 7
65. O. KELEMEN, P. CONVERTINI, Z. ZHANG, ET AL. **Function of alternative splicing.** *Gene*, **514**(1):1–30, Feb 2013. 7, 8, 99, 100, 105
66. J. T. WITTEN AND J. ULE. **Understanding splicing regulation through RNA splicing maps.** *Trends Genet*, **27**(3):89–97, Mar 2011. 7
67. M. J. MUÑOZ, M. DE LA MATA, AND A. R. KORNBLIHTT. **The carboxy terminal domain of RNA polymerase II and alternative splicing.** *Trends Biochem Sci*, **35**(9):497–504, Sep 2010. 7, 100
68. L. I. GÓMEZ ACUÑA, A. FISZBEIN, M. ALLÓ, ET AL. **Connections between chromatin signatures and splicing.** *Wiley Interdiscip Rev RNA*, **4**(1):77–91, 2013. 7, 8, 100
69. D. AUBOEUF, D. H. DOWHAN, Y. K. KANG, ET AL. **Differential recruitment of nuclear receptor coactivators may determine alternative RNA splice site choice in target genes.** *Proc Natl Acad Sci USA*, **101**(8):2270–2274, Feb 2004. 7
70. M. DE LA MATA, C. R. ALONSO, S. KADENER, ET AL. **A slow RNA polymerase II affects alternative splicing in vivo.** *Mol Cell*, **12**(2):525–532, Aug 2003. 7
71. R. F. LUCO, Q. PAN, K. TOMINAGA, ET AL. **Regulation of alternative splicing by histone modifications.** *Science*, **327**(5968):996–1000, Feb 2010. 8
72. H.-L. ZHOU, M. N. HINMAN, V. A. BARRON, ET AL. **Hu proteins regulate alternative splicing by inducing localized histone hyperacetylation in an RNA-dependent manner.** *Proc Natl Acad Sci USA*, **108**(36):E627–E635, Sep 2011. 8
73. N. KIM, A. V. ALEKSEYENKO, M. ROY, ET AL. **The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species.** *Nucleic Acids Res*, **35**(Database issue):D93–D98, Jan 2007. 8, 92
74. J. M. JOHNSON, J. CASTLE, P. GARRETT-ENGELE, ET AL. **Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays.** *Science*, **302**(5653):2141–2144, Dec 2003. 8, 99
75. E. T. WANG, R. SANDBERG, S. LUO, ET AL. **Alternative isoform regulation in human tissue transcriptomes.** *Nature*, **456**:470–476, Nov 2008. 8, 99
76. Q. PAN, O. SHAI, L. J. LEE, ET AL. **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.** *Nat Genet*, **40**(12):1413–1415, Dec 2008. 8, 9, 98, 99, 105
77. M. A. CAMPBELL, B. J. HAAS, J. P. HAMILTON, ET AL. **Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis.** *BMC Genomics*, **7**:327, 2006. 8
78. M. IRIMIA, J. L. RUKOV, D. PENNY, ET AL. **Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing.** *BMC Evol Biol*, **7**:188, 2007. 8, 91, 95
79. A. M. MCGUIRE, M. D. PEARSON, D. E. NEAFSEY, ET AL. **Cross-kingdom patterns of alternative splicing and splice recognition.** *Genome Biol*, **9**(3):R50, 2008. 8, 91, 95
80. B. WANG, G. GUO, C. WANG, ET AL. **Survey of the transcriptome of *Aspergillus oryzae* via massively parallel mRNA sequencing.** *Nucleic Acids Res*, **38**(15):5075–5087, Aug 2010. 8, 95
81. B. J. LOFTUS, E. FUNG, P. RONCAGLIA, ET AL. **The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*.** *Science*, **307**(5713):1321–1324, Feb 2005. 8, 91, 95
82. P. GRABOWSKI. **Alternative splicing takes shape during neuronal development.** *Curr Opin Genet Dev*, **21**(4):388–394, Aug 2011. 8
83. D. L. BLACK. **Mechanisms of alternative pre-messenger RNA splicing.** *Annu Rev Biochem*, **72**:291–336, 2003. 8
84. L. BALVAY AND M. Y. FISZMAN. **Analysis of the diversity of tropomyosin isoforms.** *C R Seances Soc Biol Fil*, **188**(5-6):527–540, 1994. 8
85. K. KEMPER, M. J. P. M. TOL, AND J. P. MEDEMA. **Mouse tissues express multiple splice variants of prominin-1.** *PLoS One*, **5**(8):e12325, 2010. 8
86. R. SINHA, T. LENSER, N. JAHN, ET AL. **TasDB2 - A comprehensive database of subtle alternative splicing events.** *BMC Bioinformatics*, **11**:216, 2010. 8
87. T.-M. CHERN, E. VAN NIMWEGEN, C. KAI, ET AL. **A simple physical model predicts small exon length variations.** *PLoS Genet*, **2**(4):e45, Apr 2006. 8
88. M. HILLER, K. HUSE, K. SZAFRANSKI, ET AL. **Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity.** *Nat Genet*, **36**(12):1255–1257, Dec 2004. 8
89. A. BUSCH AND K. J. HERTEL. **Extensive regulation of NAGNAG alternative splicing: new tricks for the spliceosome?** *Genome Biol*, **13**(2):143, Feb 2012. 8
90. S. RAYSON, L. ARCIGA-REYES, L. WOOTTON, ET AL. **A role for nonsense-mediated mRNA decay in plants: pathogen responses are induced in *Arabidopsis thaliana* NMD mutants.** *PLoS One*, **7**(2):e31917, 2012. 9
91. J. HWANG AND L. E. MAQUAT. **Nonsense-mediated mRNA decay (NMD) in animal embryogenesis: to die or not to die, that is the question.** *Curr Opin Genet Dev*, **21**(4):422–430, Aug 2011. 9, 107
92. M. MEKOUAR, I. BLANC-LENFLE, C. OZANNE, ET AL. **Detection and analysis of alternative splicing in *Yarrowia lipolytica* reveal structural constraints facilitating nonsense-mediated decay of intron-retaining transcripts.** *Genome Biol*, **11**(6):R65, 2010. 9, 94, 96
93. L. HUANG AND M. F. WILKINSON. **Regulation of nonsense-mediated mRNA decay.** *Wiley Interdiscip Rev RNA*, **3**(6):807–828, 2012. 9
94. L. F. LAREAU, M. INADA, R. E. GREEN, ET AL. **Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements.** *Nature*, **446**(7138):926–929, Apr 2007. 9
95. M. K. SAKHARKAR, V. T. K. CHOW, AND P. KANGUEANE. **Distributions of exons and introns in the human genome.** *In Silico Biol*, **4**(4):387–393, 2004. 9, 98, 105

## REFERENCES

---

96. R. H. WHITTAKER. **New concepts of kingdoms or organisms. Evolutionary relations are better represented by new classifications than by the traditional two kingdoms.** *Science*, **163**(3863):150–160, Jan 1969. 9
97. A. G. B. SIMPSON AND A. J. ROGER. **The real 'kingdoms' of eukaryotes.** *Curr Biol*, **14**(17):R693–R696, Sep 2004. 9
98. S. M. ADL, A. G. B. SIMPSON, M. A. FARMER, ET AL. **The new higher level classification of eukaryotes with emphasis on the taxonomy of protists.** *J Eukaryot Microbiol*, **52**(5):399–451, 2005. 9
99. A. J. ROGER AND A. G. B. SIMPSON. **Evolution: revisiting the root of the eukaryote tree.** *Curr Biol*, **19**(4):R165–R167, Feb 2009. 9
100. T. CAVALIER-SMITH. **A revised six-kingdom system of life.** *Biol Rev Camb Philos Soc*, **73**(3):203–266, Aug 1998. 9
101. T. BRUNS. **Evolutionary biology: a kingdom revised.** *Nature*, **443**(7113):758–761, Oct 2006. 9
102. M. J. CARLILE, S. C. WATKINSON, AND G. W. GOODAY. *The Fungi*. Academic Press, London, UK, 2 edition, 2001. 9, 10, 11, 91
103. J. BARNARD. **Oregon's monster mushroom is world's biggest living thing.** *The Independent*, 06 August 2000. 9
104. M. BLACKWELL. **The fungi: 1, 2, 3 ... 5.1 million species?** *Am J Bot*, **98**(3):426–438, Mar 2011. 9
105. D. S. HIBBETT, M. BINDER, J. F. BISCHOFF, ET AL. **A higher-level phylogenetic classification of the Fungi.** *Mycol Res*, **111**(Pt 5):509–547, May 2007. 10
106. S. A. KARPOV, K. V. MIKHAILOV, G. S. MIRZAEVA, ET AL. **Obligately phagotrophic aphelids turned out to branch with the earliest-diverging fungi.** *Protist*, **164**(2):195–205, Mar 2013. 10
107. K. HOFFMANN AND P. VOIGT, K. AND KIRK. ***Mortierellomycotina* subphyl. nov., based on multi-gene genealogies.** *Mycotaxon*, **115**:353–363, 2011. 10
108. C. ALEXOPOULOS, C. MIMS, AND M. BLACKWELL. *Introductory Mycology*. Wiley, New York, USA, 1996. 10
109. N. JAIN, F. HASAN, AND B. C. FRIES. **Phenotypic Switching in Fungi.** *Curr Fungal Infect Rep*, **2**(3):180–188, Sep 2008. 10
110. D. LANGOR AND J. SWEENEY, editors. *Ecological Impacts of Non-Native Invertebrates and Fungi on Terrestrial Ecosystems*. Springer, Canada, 2009. 11
111. A. N. B. ELLEPOLA AND C. J. MORRISON. **Laboratory diagnosis of invasive candidiasis.** *J Microbiol*, **43 Spec No**:65–84, Feb 2005. 11
112. A. MCCORMICK, J. LOEFFLER, AND F. EBEL. ***Aspergillus fumigatus*: contours of an opportunistic human pathogen.** *Cell Microbiol*, **12**(11):1535–1543, Nov 2010. 11
113. J. R. PERFECT. **The impact of the host on fungal infections.** *Am J Med*, **125**(1 Suppl):S39–S51, Jan 2012. 11
114. C. S. KUROKAWA, M. F. SUGIZAKI, AND M. T. PERAÇOLI. **Virulence factors in fungi of systemic mycoses.** *Rev Inst Med Trop Sao Paulo*, **40**(3):125–135, 1998. 11
115. J. KARKOWSKA-KULETA, M. RAPALA-KOZIK, AND A. KOZIK. **Fungi pathogenic to humans: molecular bases of virulence of *Candida albicans*, *Cryptococcus neoformans* and *Aspergillus fumigatus*.** *Acta Biochim Pol*, **56**(2):211–224, 2009. 11
116. J. W. BENNETT. ***Aspergillus*: a primer for the novice.** *Med Mycol*, **47 Suppl 1**:S5–12, 2009. 11
117. G. HU, S. JI, Y. YU, ET AL. **Organisms for Biofuel Production: Natural Bioresources and Methodologies for Improving Their Biosynthetic Potentials.** *Adv Biochem Eng Biotechnol*, Sep 2013. [Epub ahead of print]. 11
118. K. GRÜTZMANN, S. BÖCKER, AND S. SCHUSTER. **Combinatorics of alphatic amino acids.** *Naturwissenschaften*, **98**(1):79–86, Jan 2011. 14, 87, 91, 92
119. K. GRÜTZMANN, K. SZAFRANSKI, M. POHL, ET AL. **The alternative messages of fungal genomes.** In D. SCHOMBURG AND A. GROTE, editors, *Short Papers and Poster Abstracts, German Conference on Bioinformatics*, pages 35–39, Braunschweig, 2010. 23
120. K. GRÜTZMANN, K. SZAFRANSKI, M. POHL, ET AL. **Fungal alternative splicing is associated with multicellular complexity and virulence - A genome-wide multi-species study.** *DNA Research*, **accepted**, 2013. 28, 92
121. M. POHL, D. HOLSTE, R. BORTFELDT, ET AL. **Mutually exclusive spliced exons show non-adjacent and grouped patterns.** In I. GROSSE, S. NEUMANN, S. POSCH, ET AL., editors, *Short Papers and Poster Abstracts, German Conference on Bioinformatics*, pages 19–24, Halle, 2009. 70
122. M. POHL, R. H. BORTFELDT, K. GRÜTZMANN, ET AL. **Alternative splicing of mutually exclusive exons-A review.** *Biosystems*, **114**(1):31–38, Oct 2013. 77, 97, 99
123. E. SZATHMÁRY. **Why are there four letters in the genetic alphabet?** *Nat Rev Genet*, **4**(12):995–1001, Dec 2003. 88
124. P. V. BARANOV, M. VENIN, AND G. PROVAN. **Codon size reduction as the origin of the triplet genetic code.** *PLoS One*, **4**(5):e5708, 2009. 88
125. D. G. SALINAS, M. O. GALLARDO, AND M. I. OSORIO. **The most probable number of blocks for the partitions of the set of codons could have determined the number of standard amino acids.** *Biosystems*, **109**(2):133–136, Aug 2012. 88, 89
126. F. H. CRICK. **The origin of the genetic code.** *J Mol Biol*, **38**(3):367–379, Dec 1968. 88
127. N. P. LUKASHENKO. **Expanding genetic code: amino acids 21 and 22-selenocysteine and pyrrolysine.** *Genetika*, **46**(8):1013–1032, Aug 2010. 88
128. G. SELLA AND D. H. ARDELL. **The coevolution of genes and genetic codes: Crick's frozen accident revisited.** *J Mol Evol*, **63**(3):297–313, Sep 2006. 88, 89
129. T. TLUSTY. **A colorful origin for the genetic code: information theory, statistical mechanics and the emergence of molecular codes.** *Phys Life Rev*, **7**(3):362–376, Sep 2010. 88, 89, 91
130. D. H. ARDELL AND G. SELLA. **On the evolution of redundancy in genetic codes.** *J Mol Evol*, **53**(4-5):269–281, 2001. 88
131. T. BERGER. *Rate Distortion Theory: Mathematical Basis for Data Compression*. Prentice Hall, Englewood Cliffs, USA, 1971. 88

132. T. TLUSTY. **Rate-distortion scenario for the emergence and evolution of noisy molecular codes.** *Phys Rev Lett*, **100**(4):048101, Feb 2008. 88
133. F. J. TAYLOR AND D. COATES. **The code within the codons.** *Biosystems*, **22**(3):177–187, 1989. 89, 91, 102
134. S. R. PELC AND M. G. WELTON. **Stereochemical relationship between coding triplets and amino-acids.** *Nature*, **209**(5026):868–870, Feb 1966. 89
135. J. T. WONG. **A co-evolution theory of the genetic code.** *Proc Natl Acad Sci USA*, **72**(5):1909–1912, May 1975. 89
136. H. C. UREY. **On the Early Chemical History of the Earth and the Origin of Life.** *Proc Natl Acad Sci U S A*, **38**(4):351–363, Apr 1952. 89
137. F. H. CRICK, S. BRENNER, A. KLUG, ET AL. **A speculation on the origin of protein synthesis.** *Orig Life*, **7**(4):389–397, Dec 1976. 89
138. M. EIGEN AND P. SCHUSTER. **The Hypercycle. A Principle of Natural Self-Organization. Part C: The Realistic Hypercycle.** *Die Naturwissenschaften*, **65**(7):341–369, July 1978. 89
139. M. EIGEN. **The Hypercycle: A Principle of Natural Self-Organization. Part A: Emergence of the Hypercycle.** *Die Naturwissenschaften*, **64**(11):541–565, November 1977. 89
140. E. N. TRIFONOV. **Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16 S rRNA nucleotide sequences.** *J Mol Biol*, **194**(4):643–652, Apr 1987. 89
141. A. L. WEBER AND S. L. MILLER. **Reasons for the occurrence of the twenty coded protein amino acids.** *J Mol Evol*, **17**(5):273–284, 1981. 89
142. D. SÖLL AND U. L. RAJBHANDARY. **The genetic code - thawing the 'frozen accident'.** *J Biosci*, **31**(4):459–463, Oct 2006. 89
143. H. SELIGMANN. **Cost-minimization of amino acid usage.** *J Mol Evol*, **56**(2):151–161, Feb 2003. 90, 106
144. R. S. PATON AND J. M. GOODMAN. **Exploration of the accessible chemical space of acyclic alkanes.** *J Chem Inf Model*, **47**(6):2124–2132, 2007. 90
145. Y. LU AND S. FREELAND. **On the evolution of the standard amino-acid alphabet.** *Genome Biol*, **7**(1):102, 2006. 90
146. L. FOWDEN. **Plant amino acid research in retrospect: from Chinball to Singh.** *Amino Acids*, **20**(3):217–224, 2001. 90
147. P. K. MUKHERJEE, B. A. HORWITZ, AND C. M. KENERLEY. **Secondary metabolism in Trichoderma—a genomic perspective.** *Microbiology*, **158**(Pt 1):35–45, Jan 2012. 90
148. D. J. EBBOLE, Y. JIN, M. THON, ET AL. **Gene discovery and gene expression in the rice blast fungus, *Magnaporthe grisea*: analysis of expressed sequence tags.** *Mol Plant Microbe Interact*, **17**(12):1337–1347, Dec 2004. 91
149. L. WANG, Y. XI, J. YU, ET AL. **A statistical method for the detection of alternative splicing using RNA-seq.** *PLoS One*, **5**(1):e8529, 2010. 91
150. E. C. H. HO, M. J. CAHILL, AND B. J. SAVILLE. **Gene discovery and transcript analyses in the corn smut pathogen *Ustilago maydis*: expressed sequence tag and genome sequence comparison.** *BMC Genomics*, **8**:334, 2007. 91
151. D. A. BENSON, M. CAVANAUGH, K. CLARK, ET AL. **GenBank.** *Nucleic Acids Res*, **41**(Database issue):D36–D42, Jan 2013. 91
152. T. BARRETT, S. E. WILHITE, P. LEDOUX, ET AL. **NCBI GEO: archive for functional genomics data sets—update.** *Nucleic Acids Res*, **41**(Database issue):D991–D995, Jan 2013. 91
153. C. TRAPNELL, L. PACTHER, AND S. L. SALZBERG. **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics*, **25**(9):1105–1111, May 2009. 92
154. Z. WANG, M. GERSTEIN, AND M. SNYDER. **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet*, **10**:57–63, Nov 2008. 92
155. C. TRAPNELL, B. A. WILLIAMS, G. PERTEA, ET AL. **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol*, **28**(5):511–515, May 2010. 92
156. K. L. FOX-WALSH AND K. J. HERTEL. **Splice-site pairing is an intrinsically high fidelity process.** *Proc Natl Acad Sci USA*, **106**(6):1766–1771, Feb 2009. 92, 93
157. A. C. ENGLISH, K. S. PATEL, AND A. E. LORRAINE. **Prevalence of alternative splicing choices in *Arabidopsis thaliana*.** *BMC Plant Biol*, **10**:102, 2010. 94
158. H. M. HOOD, D. E. NEAFSEY, J. GALAGAN, ET AL. **Evolutionary roles of upstream open reading frames in mediating gene regulation in fungi.** *Annu Rev Microbiol*, **63**:385–409, 2009. 94
159. S. BOUE, I. LETUNIC, AND P. BORK. **Alternative splicing and evolution.** *Bioessays*, **25**(11):1031–1034, Nov 2003. 94, 97
160. F.-C. CHEN, S.-S. WANG, C.-J. CHEN, ET AL. **Alternatively and constitutively spliced exons are subject to different evolutionary forces.** *Mol Biol Evol*, **23**(3):675–682, Mar 2006. 94
161. N. RHIND, Z. CHEN, M. YASSOUR, ET AL. **Comparative functional genomics of the fission yeasts.** *Science*, **332**(6032):930–936, May 2011. 95
162. C. ZHAO, C. WAALWIJK, P. J. G. M. DE WIT, ET AL. **RNA-Seq analysis reveals new gene models and alternative splicing in the fungal pathogen *Fusarium graminearum*.** *BMC Genomics*, **14**:21, 2013. 95
163. F. MIGNONE, G. GRILLO, S. LIUNI, ET AL. **Computational identification of protein coding potential of conserved sequence tags through cross-species evolutionary analysis.** *Nucleic Acids Res*, **31**(15):4639–4645, Aug 2003. 95
164. A. N. MARSHALL, M. C. MONTEALEGRE, C. JIMÉNEZ-LÓPEZ, ET AL. **Alternative splicing and subfunctionalization generates functional diversity in fungal proteomes.** *PLoS Genet*, **9**(3):e1003376, 2013. 95
165. C. B. FLECK, F. SCHÖBEL, AND M. BROCK. **Nutrient acquisition by pathogenic fungi: nutrient availability, pathway regulation, and differences in substrate utilization.** *Int J Med Microbiol*, **301**(5):400–407, Jun 2011. 95



## REFERENCES

---

166. A. G. SORGO, C. J. HEILMANN, S. BRUL, ET AL. **Beyond the wall: *Candida albicans* secret(e)s to survive.** *FEMS Microbiol Lett*, **338**(1):10–17, Jan 2013. 95
167. A. ABAD, J. V. FERNÁNDEZ-MOLINA, J. BIKANDI, ET AL. **What makes *Aspergillus fumigatus* a successful pathogen? Genes and molecules involved in invasive aspergillosis.** *Rev Iberoam Micol*, **27**(4):155–182, 2010. 95
168. A. T. MORALES-VARGAS, A. DOMÍNGUEZ, AND J. RUIZ-HERRERA. **Identification of dimorphism-involved genes of *Yarrowia lipolytica* by means of microarray analysis.** *Res Microbiol*, **163**(5):378–387, Jun 2012. 96
169. Q. M. MITROVICH, B. B. TUCH, C. GUTHRIE, ET AL. **Computational and experimental approaches double the number of known introns in the pathogenic yeast *Candida albicans*.** *Genome Res*, **17**(4):492–502, Apr 2007. 96
170. A. RICCOMBENI, G. VIDANES, E. PROUX-WÉRA, ET AL. **Sequence and analysis of the genome of the pathogenic yeast *Candida orthopsilosis*.** *PLoS One*, **7**(4):e35750, 2012. 96
171. K. STRUIJBS, J. VAN DEN BURG, W. F. VISSER, ET AL. **Alternative splicing directs dual localization of *Candida albicans* 6-phosphogluconate dehydrogenase to cytosol and peroxisomes.** *FEMS Yeast Res*, **12**(1):61–68, Feb 2012. 96
172. M. SAMMETH. **Complete alternative splicing events are bubbles in splicing graphs.** *J Comput Biol*, **16**(8):1117–1140, Aug 2009. 96, 97, 98
173. Y. XING AND C. LEE. **Alternative splicing and RNA selection pressure—evolutionary consequences for eukaryotic genomes.** *Nat Rev Genet*, **7**(7):499–509, Jul 2006. 97
174. F.-C. CHEN AND T.-J. CHUANG. **The effects of multiple features of alternatively spliced exons on the K(A)/K(S) ratio test.** *BMC Bioinformatics*, **7**:259, 2006. 97
175. Y. LEE, Y. LEE, B. KIM, ET AL. **ECgene: an alternative splicing database update.** *Nucleic Acids Res*, **35**(Database issue):D99–103, Jan 2007. 97, 99
176. D. HOLSTE, G. HUO, V. TUNG, ET AL. **HOLLYWOOD: a comparative relational database of alternative splicing.** *Nucleic Acids Res*, **34**(Database issue):D56–D62, Jan 2006. 97
177. M. SOOM, G. GESSNER, H. HEUER, ET AL. **A mutually exclusive alternative exon of *slol* codes for a neuronal BK channel with altered function.** *Channels (Austin)*, **2**(4):278–282, 2008. 97, 98
178. M. E. GALLEGU, L. BALVAY, AND E. BRODY. **cis-acting sequences involved in exon selection in the chicken beta-tropomyosin gene.** *Mol Cell Biol*, **12**(12):5415–5425, Dec 1992. 97, 98
179. J. E. KRULL, P. A. F. GALANTE, D. T. OHARA, ET AL. **SPLOOCE: a new portal for the analysis of human splicing variants.** *RNA Biol*, **9**(11):1339–1343, Nov 2012. 97, 98
180. K. HATJE AND M. KOLLMAR. **Expansion of the mutually exclusive spliced exome in *Drosophila*.** *Nat Commun*, **4**:2460, Sep 2013. 97, 98
181. R. R. COPLEY. **Evolutionary convergence of alternative splicing in ion channels.** *Trends Genet*, **20**(4):171–176, Apr 2004. 98
182. S. MISRA, M. A. CROSBY, C. J. MUNGALL, ET AL. **Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review.** *Genome Biol*, **3**(12):RESEARCH0083, 2002. 98
183. E. S. LANDER, L. M. LINTON, B. BIRREN, ET AL. **Initial sequencing and analysis of the human genome.** *Nature*, **409**(6822):860–921, Feb 2001. 99
184. I. LETUNIC, R. R. COPLEY, AND P. BORK. **Common exon duplication in animals and its role in alternative splicing.** *Hum Mol Genet*, **11**(13):1561–1567, Jun 2002. 99
185. C. F. KENNEDY AND S. M. BERGET. **Pyrimidine tracts between the 5' splice site and branch point facilitate splicing and recognition of a small *Drosophila* intron.** *Mol Cell Biol*, **17**(5):2774–2780, May 1997. 99
186. M. P. MULLEN, C. W. SMITH, J. G. PATTON, ET AL. **Alpha-tropomyosin mutually exclusive exon selection: competition between branchpoint/polypyrimidine tracts determines default exon choice.** *Genes Dev*, **5**(4):642–655, Apr 1991. 99
187. R. SPELLMAN, A. RIDEAU, A. MATLIN, ET AL. **Regulation of alternative splicing by PTB and associated factors.** *Biochem Soc Trans*, **33**(Pt 3):457–460, Jun 2005. 99
188. D. ANASTASSIOU, H. LIU, AND V. VARADAN. **Variable window binding for mutually exclusive alternative splicing.** *Genome Biol*, **7**(1):R2, 2006. 99
189. Y. YANG, L. ZHAN, W. ZHANG, ET AL. **RNA secondary structure in mutually exclusive splicing.** *Nat Struct Mol Biol*, **18**(2):159–168, Feb 2011. 99
190. Y. JIN, Y. YANG, AND P. ZHANG. **New insights into RNA secondary structure in the alternative splicing of pre-mRNAs.** *RNA Biol*, **8**(3):450–457, 2011. 99
191. A. CHOUDHARY AND K. KRITHIVASAN. **Network of evolutionary processors with splicing rules and permitting context.** *Biosystems*, **87**(2-3):111–116, Feb 2007. 100
192. Z. WANG AND C. B. BURGE. **Splicing regulation: from a parts list of regulatory elements to an integrated splicing code.** *RNA*, **14**(5):802–813, May 2008. 100
193. Y. BARASH, J. A. CALARCO, W. GAO, ET AL. **Deciphering the splicing code.** *Nature*, **465**(7294):53–59, May 2010. 100
194. A. MANZOUROLAJDAD, Y. WANG, T. I. SHAW, ET AL. **Information-theoretic uncertainty of SCFG-modeled folding space of the non-coding RNA.** *J Theor Biol*, **318**:140–163, Feb 2013. 101
195. M. SYVANEN. **Evolutionary implications of horizontal gene transfer.** *Annu Rev Genet*, **46**:341–358, 2012. 101
196. D. A. FITZPATRICK. **Horizontal gene transfer in fungi.** *FEMS Microbiol Lett*, **329**(1):1–8, Apr 2012. 101
197. C. P. ANDAM, G. P. FOURNIER, AND J. P. GOGARTEN. **Multilevel populations and the evolution of antibiotic resistance through horizontal gene transfer.** *FEMS Microbiol Rev*, **35**(5):756–767, Sep 2011. 101
198. M. W. GRAY. **Mitochondrial evolution.** *Cold Spring Harb Perspect Biol*, **4**(9):a011403, Sep 2012. 101

199. O. J. RANDO. **Combinatorial complexity in chromatin structure and function: revisiting the histone code.** *Curr Opin Genet Dev*, **22**(2):148–155, Apr 2012. 101
200. C. HUANG, M. XU, AND B. ZHU. **Epigenetic inheritance mediated by histone lysine methylation: maintaining transcriptional states without the precise restoration of marks?** *Philos Trans R Soc Lond B Biol Sci*, **368**(1609):201110332, Jan 2013. 102
201. C. JANEWAY, P. TRAVERS, M. WALPOT, ET AL. *Immunobiology*. Garland Science, New York, USA, 5 edition, 2001. 102, 106, 107, 108
202. L. FOWDEN. **Azetidine-2-carboxylic acid: a new cyclic imino acid occurring in plants.** *Biochem J*, **64**(2):323–332, Oct 1956. 102
203. H. INOUE, Y. SAKATA, H. NISHIO, ET AL. **A simple and highly sensitive HPLC method with fluorescent detection for determination of pipecolic acid in mouse brain areas.** *Biol Pharm Bull*, **34**(2):287–289, 2011. 102
204. J. BEHRE, R. VOIGT, I. ALTHÖFER, ET AL. **On the evolutionary significance of the size and planarity of the proline ring.** *Naturwissenschaften*, **99**(10):789–799, Oct 2012. 102
205. A. NAKABACHI, A. YAMASHITA, H. TOH, ET AL. **The 160-kilobase genome of the bacterial endosymbiont *Carsonella*.** *Science*, **314**(5797):267, Oct 2006. 103
206. U. CONSORTIUM. **Update on activities at the Universal Protein Resource (UniProt) in 2013.** *Nucleic Acids Res*, **41**(Database issue):D43–D47, Jan 2013. 103
207. K. A. DILL AND J. L. MACCALLUM. **The protein-folding problem, 50 years on.** *Science*, **338**(6110):1042–1046, Nov 2012. 103
208. S. ERDIN, A. M. LISEWSKI, AND O. LICHTARGE. **Protein function prediction: towards integration of similarity metrics.** *Curr Opin Struct Biol*, **21**(2):180–188, Apr 2011. 103
209. R. D. SLEATOR AND P. WALSH. **An overview of in silico protein function prediction.** *Arch Microbiol*, **192**(3):151–155, Mar 2010. 103
210. J. REIMAND, S. HUI, S. JAIN, ET AL. **Domain-mediated protein interaction prediction: From genome to network.** *FEBS Lett*, **586**(17):2751–2763, Aug 2012. 103
211. H. X. TA AND L. HOLM. **Evaluation of different domain-based methods in protein interaction prediction.** *Biochem Biophys Res Commun*, **390**(3):357–362, Dec 2009. 103
212. G. V. KRYUKOV, S. CASTELLANO, S. V. NOVOSELOV, ET AL. **Characterization of mammalian selenoproteomes.** *Science*, **300**(5624):1439–1443, May 2003. 103
213. B. J. BYUN AND Y. K. KANG. **Conformational preferences and pK(a) value of selenocysteine residue.** *Biopolymers*, **95**(5):345–353, May 2011. 103
214. B. HAO, W. GONG, T. K. FERGUSON, ET AL. **A new UAG-encoded residue in the structure of a methanogen methyltransferase.** *Science*, **296**(5572):1462–1466, May 2002. 103
215. S. J. PROHASKA, P. F. STADLER, AND D. C. KRAKAUER. **Innovation in gene regulation: the case of chromatin computation.** *J Theor Biol*, **265**(1):27–44, Jul 2010. 104
216. G. A. KHOURY, R. C. BALIBAN, AND C. A. FLOUDAS. **Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database.** *Sci Rep*, **1**:90, Sep 2011. 104
217. V. WILSON, editor. *Sumoylation: Molecular Biology and Biochemistry (Horizonbioscience)*. Taylor & Francis, Wymondham, UK, 2004. 104
218. K. TABUCHI AND T. C. SÜDHOF. **Structure and evolution of neurexin genes: insight into the mechanism of alternative splicing.** *Genomics*, **79**(6):849–859, Jun 2002. 105
219. S. LECOMTE, L. REVERDY, C. LE QUÉMENT, ET AL. **Unraveling complex interplay between heat shock factor 1 and 2 splicing isoforms.** *PLoS One*, **8**(2):e56085, 2013. 105
220. C. GAO AND Y. WANG. **Global impact of RNA splicing on transcriptome remodeling in the heart.** *J Zhejiang Univ Sci B*, **13**(8):603–608, Aug 2012. 105
221. E. O. GRACHEVA, J. F. CORDERO-MORALES, J. A. GONZÁLEZ-CARCACÍA, ET AL. **Ganglion-specific splicing of TRPV1 underlies infrared sensation in vampire bats.** *Nature*, **476**(7358):88–91, Aug 2011. 105
222. M. IRIMIA AND B. J. BLENCOWE. **Alternative splicing: decoding an expansive regulatory layer.** *Curr Opin Cell Biol*, **24**(3):323–332, Jun 2012. 105
223. M. GABUT, P. SAMAVARCHI-TEHRANI, X. WANG, ET AL. **An alternative splicing switch regulates embryonic stem cell pluripotency and reprogramming.** *Cell*, **147**(1):132–146, Sep 2011. 106
224. R. SCHMID, S. N. GRELLSCHEID, I. EHRMANN, ET AL. **The splicing landscape is globally reprogrammed during male meiosis.** *Nucleic Acids Res*, Sep 2013. [Epub ahead of print]. 106
225. K. H. LIM, L. FERRARIS, M. E. FILLoux, ET AL. **Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes.** *Proc Natl Acad Sci USA*, **108**(27):11093–11098, Jul 2011. 106
226. P. RAJAN, D. J. ELLIOTT, C. N. ROBSON, ET AL. **Alternative splicing and biological heterogeneity in prostate cancer.** *Nat Rev Urol*, **6**(8):454–460, Aug 2009. 106
227. F. DAMM, O. KOSMIDER, V. GELSI-BOYER, ET AL. **Mutations affecting mRNA splicing define distinct clinical phenotypes and correlate with patient outcome in myelodysplastic syndromes.** *Blood*, **119**(14):3211–3218, Apr 2012. 106
228. L. BROCCIERI AND S. KARLIN. **Protein length in eukaryotic and prokaryotic proteomes.** *Nucleic Acids Res*, **33**(10):3390–3400, 2005. 107
229. R. WALCZAK, E. WESTHOF, P. CARBON, ET AL. **A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs.** *RNA*, **2**(4):367–379, Apr 1996. 107
230. A. THÉOBALD-DIETRICH, R. GIEGÉ, AND J. RUDINGER-THIRION. **Evidence for the existence in mRNAs of a hairpin element responsible for ribosome dependent pyrrolysine insertion into proteins.** *Biochimie*, **87**(9-10):813–817, 2005. 107
231. C. D. CAMPBELL, J. X. CHONG, M. MALIG, ET AL. **Estimating the human mutation rate using autozygosity in a founder population.** *Nat Genet*, **44**(11):1277–1281, Nov 2012. 108

## REFERENCES

---

232. S. WIELGOSS, J. E. BARRICK, O. TENAILLON, ET AL. **Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load.** *Proc Natl Acad Sci USA*, **110**(1):222–227, Jan 2013. 108
233. M. BARBIERI, V. BOCCARDI, A. ESPOSITO, ET AL. **A/ASP/VAL allele combination of IGF1R, IRS2, and UCP2 genes is associated with better metabolic profile, preserved energy expenditure parameters, and low mortality rate in longevity.** *Age (Dordr)*, **34**(1):235–245, Feb 2012. 108
234. J. HULKKONEN, P. LAIPPALA, AND M. HURME. **A rare allele combination of the interleukin-1 gene complex is associated with high interleukin-1 beta plasma levels in healthy individuals.** *Eur Cytokine Netw*, **11**(2):251–255, Jun 2000. 108



# Beitrag der Autoren

Title	Literaturangabe	Autoren	Arbeitsanteil
The alternative messages of fungal genomes	short papers and poster abstracts, German Conference on Bioinformatics, pp. 35-39, Hrsg. Dietmar Schomburg, Andreas Grote, Braunschweig, 2010	<b>Konrad Grützmann</b> Karol Szafranski Martin Pohl Matthias Platzer Stefan Schuster	50% 15% 20% 5% 10%
Fungal alternative splicing is associated with multicellular complexity and virulence - A genome-wide multi-species study	<i>DNA Research</i> , accepted, September 3 <sup>rd</sup> , 2013	<b>Konrad Grützmann</b>  Karol Szafranski Martin Pohl Kerstin Voigt Andreas Petzold Stefan Schuster	50%  15% 15% 10% 5% 5%
Mutually exclusive spliced exons show non-adjacent and grouped patterns	short papers and poster abstracts, German Conference on Bioinformatics, pp. 19-24, Hrsg. Ivo Grosse, Steffen Neumann, Stefan Posch, Falk Schreiber, Peter Stadler, Halle, 2009	Martin Pohl Dirk Holste Ralf Bortfeldt <b>Konrad Grützmann</b> Stefan Schuster	50% 15% 15% 15% 5%
Alternative splicing of mutually exclusive exons – A review	<i>BioSystems</i> , 2013, 114(1): 31-38	Martin Pohl Ralf H. Bortfeldt <b>Konrad Grützmann</b> Stefan Schuster	60% 10% 20% 10%
Combinatorics of aliphatic amino acids	<i>Naturwissenschaften</i> , 2011, 98(1): 79-86	<b>Konrad Grützmann</b> Stefan Schuster Sebastian Böcker	50% 25% 25%

.....  
bestätigt Prof. Dr. Stefan Schuster

# Lebenslauf

## ■ Persönliche Daten

Konrad Grützmann  
geboren 22.07.1981, Halle (Saale)  
Wohnort Burgsdorffstraße 21, 01129 Dresden

## ■ Wissenschaftlicher Werdegang

04/2015 **Promotionsverteidigung**, *Friedrich-Schiller-Universität*, Jena  
2008–2013 **Promotionsstudium**, *Friedrich-Schiller-Universität*, Jena  
2001–2008 **Diplom Bioinformatik**, *Diplomarbeit "Rekonstruktion von Genregulationsnetzwerken bei Candida albicans"*, *Friedrich-Schiller-Universität*, Jena

## ■ Beruflicher Werdegang

seit 02/2014 **wissenschaftlicher Mitarbeiter**, *Helmholtz-Zentrum für Umweltforschung UFZ*, Leipzig  
04-09/2013 **Schreiben und Abgabe der Dissertation, Elternzeit**, Dresden  
06/2008–03/2013 **wissenschaftlicher Mitarbeiter und Doktorand**, *Lehrstuhl für Bioinformatik Prof. Schuster*, Universität, Jena  
05/2008 **wissenschaftlicher Mitarbeiter**, *AG Systembiologie/Bioinformatik, Prof. Reinhard Guthke*, Hans-Knöll-Institut, Jena  
09–10/2005, 01–03/2006 **wissenschaftliche Hilfskraft**, *Lehrstuhl für Bioinformatik, Prof. Peter Stadler*, Universität Leipzig  
03–04/2004 **wissenschaftliche Hilfskraft**, *AG Biosystemanalyse, Prof. Peter Dittrich*, Universität Jena

## ■ Wehrdienst und Schulausbildung

10/2000–09/2001 **Zivildienst**, *Kindergarten der AWO*, Stadtroda  
1992–2000 **Gymnasium Stadtroda**, Abitur  
1988–1992 **Grundschule Stadtroda**

## ■ Veröffentlichungen

1. Endogenous metabolites and inflammasome activity in early childhood and links to respiratory diseases, G. Herberth, K. Offenberg, U. Rolle-Kampczyk, M. Bauer, W. Otto, S. Röder, **K. Grützmann**, U. Sack, J.C. Simon, M. Borte, M. von Bergen, I. Lehmann, *J Allergy Clin Immunol*, S0091-6749(15)00112-8, 2015
2. Categorized Counting Mediated by Blotting Membrane Systems for Particle-based Data Mining and Numerical Algorithms, T. Hinze, **K. Grützmann**, B. Höckner, P. Sauer, S. Hayat, *Proceedings of*

- the 15th International Conference on Membrane Computing*, Prague, LNCS 8961, 241-257, 2014
3. Fungal alternative splicing is associated with more complex multicellularity and virulence - A genome-wide multi-species study, **K. Grützmann**, K. Szafranski, M. Pohl, K. Voigt, A. Petzold, S. Schuster, *DNA Research*, 21(1), 2014, 27-39
  4. Alternative splicing of mutually exclusive exons, M. Pohl, R. Bortfeldt, **K. Grützmann**, S. Schuster, *BioSystems*, 114 (1), 2013, 31-38
  5. Combinatorics of Aliphatic Amino Acids, **K. Grützmann**, S. Böcker, S. Schuster, *Naturwissenschaften* 98 (1), 2011, 79-86
  6. The alternative messages of fungal genomes, **K. Grützmann**, K. Szafranski, M. Pohl, M. Platzer, S. Schuster, short papers and poster abstracts, German Conference on Bioinformatics, 35-39, Hrsg. D. Schomburg, A. Grote, Braunschweig, 2010
  7. Mutually exclusive spliced exons show non-adjacent and grouped patterns, M. Pohl, D. Holste, R. Bortfeldt, **K. Grützmann**, S. Schuster, short papers and poster abstracts, German Conference on Bioinformatics, 19-24, Hrsg. I. Grosse, S. Neumann, S. Posch, et al., Halle, 2009

## Wissenschaftliche Vorträge

- Fungal alternative splicing associates with higher cellular complexity and virulence, **K. Grützmann**, K. Szafranski, M. Pohl, A. Petzold, K. Voigt, S. Schuster, BITS, Catania 2012, Italien
- Alternative Splicing in the Fungal Kingdom, **K. Grützmann**, K. Szafranski, M. Pohl, A. Petzold, K. Voigt, S. Schuster, VAAM Tübingen, 2012, Germany
- Combinatorics of Aliphatic Amino Acids and Alternative Splicing in the Fungal Kingdom, **K. Grützmann**, K. Szafranski, S. Böcker, M. Pohl, A. Petzold, K. Voigt, S. Schuster, JCB Seminar Jena 2011
- The alternative messages of fungal genomes, **K. Grützmann**, K. Szafranski, M. Pohl, M. Platzer, S. Schuster, GCB Braunschweig 2010
- Mutually exclusive spliced exons show non-adjacent and grouped patterns, M. Pohl, D. Holste, R. Bortfeldt, **K. Grützmann**, S. Schuster, GCB Halle 2009

## Poster

- The evolution of zygomycetes, the most basal terrestrial fungi: lessons from new genome projects, V. U. Schwartz, K. Hoffmann, G. Walther, H. Vogel, M. Felder, S. Müller, K. Shelest, **K. Grützmann**, M. Pohl, S. Winter, S. Böcker, S. Schuster, A. Petzold, K. Szafranski, M. Nowrousian, M. Platzer, A. Vilcinskis, A.A. Brakhage, K. Voigt, ECFG, Marburg 2012, Germany
- Towards the extent and meaning of fungal alternative splicing, **K. Grützmann**, K. Szafranski, M. Pohl, A. Petzold, S. Schuster, ECFG, Marburg 2012, Germany
- Combinatorics of aliphatic amino acids, **K. Grützmann**, S. Böcker, S. Schuster, GCB Weihenstephan 2011, Germany
- The Extent of Alternative Splicing in the Fungal Kingdom, **K. Grützmann**, K. Szafranski, M. Pohl, S. Schuster, Special Interest Group Meeting on Alternative Splicing (at ISMB/ECCB), 2011, Vienna, Austria
- The Alternative Messages of Fungal Genomes, **K. Grützmann**, K. Szafranski, M. Pohl, S. Schuster, RNA Society meeting 2011, Kyoto, Japan
- The alternative messages of fungal genomes, **K. Grützmann**, K. Szafranski, M. Pohl, S. Schuster, JCB Workshop 2011, Jena, Germany
- The alternative messages of fungal genomes, **K. Grützmann**, K. Szafranski, M. Pohl, M. Platzer, S. Schuster, GCB Braunschweig 2010, Germany
- Alternative splicing in fungal aldo-keto reductases, **K. Grützmann**, K. Hoffmann, M. Eckart, S. Schuster, K. Voigt, Asian Mycological Congress, 15-19 November, 2009, Taiwan
- Towards a workflow of cross-species analysis of alternative splicing - A case study of the fungal domain, **K. Grützmann**, M. Pohl, K. Szafranski, S. Schuster, GCB 2009, Halle/Germany



# Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Hilfsmittel angefertigt habe. Mir ist die geltende Promotionsordnung bekannt und ich habe weder die Hilfe eines Promotionsberaters in Anspruch genommen, noch haben Dritte unmittelbare oder mittelbare geldwerte Leistungen für Arbeit erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die vorgelegte Dissertation wurde noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht. Weiterhin habe ich mich mit der vorliegenden Arbeit an keiner anderen Hochschule um den akademischen Grad *doctor rerum naturalium* (Dr. rer. nat.) beworben und weder früher noch gegenwärtig die Eröffnung eines Verfahrens zum Erwerb des o.g. akademischen Grades an einer anderen Hochschule beantragt.

Bei der Auswahl und Auswertung des Materials, sowie bei der Herstellung des Manuskripts haben mich meine Kollegen am Lehrstuhl für Bioinformatik unter der Leitung von Prof. Dr. Stefan Schuster unterstützt.

Dresden, den ...