Bioinformatische Analysen der transkriptionellen Regulation morphologischer Variationen in Cyanobakterien der Ordnung Stigonematales

Inaugural-Dissertation

zur Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät der
Christian-Albrechts-Universität zu Kiel

vorgelegt von
Robin Koch
aus Wesel
Kiel, Januar 2016

aus dem Institut für Allgemeine Mikrobiologie der Christian-Albrechts-Universität zu Kiel

Erklärung

Die vorliegende Dissertation habe ich eigenständig, ohne unerlaubte Hilfe und unter Einhaltung der Regeln guter wissenschaftlicher Praxis der deutschen Forschungsgemeinschaft angefertigt. Die Dissertation wurde weder in der vorgelegten noch in ähnlicher Form bei einer anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Kiel, 14. Januar 2016

Robin Koch

Referentin: Prof. Dr. Tal Dagan
Koreferentin: Prof. Dr. Ruth Schmitz-Streit
Tag der mündlichen Prüfung: 20.04.2016
Zum Druck genehmigt: Kiel, den 20.04.2016

Der Dekan

Im Laufe dieser Arbeit wurden mit Zustimmung des Betreuers folgende Beiträge veröffentlicht:

Rezensierte Fachzeitschriften

Tal Dagan, Mayo Roettger, Karina Stucken, Giddy Landan, **Robin Koch**, Peter Major, Sven B. Gould, Vadim V. Goremykin, Rosmarie Rippka, Nicole Tandeau de Marsac, Muriel Gugger, Peter J. Lockhart, John F. Allen, Iris Brune, Irena Maus, Alfred Pühler, William F. Martin. 2013. Genomes of Stigonematalean Cyanobacteria (Subsection V) and the Evolution of Oxygenic Photosynthesis from Prokaryotes to Plastids. *Genome Biol Evol*. 5:31-44

Karina Stucken, **Robin Koch**, Tal Dagan. 2013. Cyanobacterial defense mechanisms against foreign DNA transfer and their impact on genetic engineering. *Biol Res* 46:373–382.

Tagungsbeiträge (Poster)

Robin Koch, Karina Stucken, Anne Kupczok, Judith Ilhan, Tal Dagan. 2015. Transcriptional regulation of morphological transitions in stigonematalean cyanobacteria. Gesellschaft für Genetik (GFG), Kiel, Deutschland (Preis für bestes Poster).

Robin Koch, Karina Stucken, Anne Kupczok, Judith Ilhan, Tal Dagan. 2015. Transcriptional regulation of morphological transitions in stigonematalean cyanobacteria. Gesellschaft für Molekulare Biologie und Evolution (SMBE), Wien, Österreich.

Robin Koch, Anja Stefanski, Kai Stühler, Karina Stucken, Tal Dagan. 2014. The small peptides repertoire of *Scytonema hofmanni* PCC 7110. Europäische Tagung zur Molekularen Biologie der Cyanobakterien (EWMBC), Texel, Niederlande.

Robin Koch, Giddy Landan, Karina Stucken, Tal Dagan. Reductive Evolution of the heterocyst-forming filamentous, true-branching Cyanobacteria *Fischerella thermalis* PCC 7521 and *Fischerella* sp. JSC-11. 2014. Gesellschaft für Molekulare Biologie und Evolution (SMBE) Anschlusstagung, Kiel, Deutschland.

Robin Koch, Giddy Landan, Karina Stucken, and Tal Dagan. Reductive Evolution of the heterocyst forming filamentous, true-branching Cyanobacteria *Fischerella thermalis* PCC 7521 and *Fischerella* sp. JSC-11. 2013. Föderation europäischer mikrobieller Gesellschaften FEMS, Leipzig, Germany.

Inhaltsverzeichnis

1 Zusammenfassung

Cyanobakterien umfassen eine monophyletische Gruppe von Spezies, welche eine Vielzahl morphologischer Variationen aufweisen. Spezies der Gattungen *Chlorogloeopsis* und *Fischerella*, die der Ordnung Stigonematales zugehörig sind, entwickeln multiseriate sowie verzweigte Filamente und repräsentieren die morphologisch komplexesten Cyanobakterien. Induktionen einer Saccharose- bzw. einer NaCl-Konzentrationserhöhung erzeugten morphologische Transitionen in Spezies dieser Gattungen. Für zwei *Fischerella*-Spezies (*F. muscicola* PCC 7414, *F. thermalis* PCC 7521) zeigte sich dies durch unverzweigtes und filamentöses Wachstum. In *Chlorogloeopsis* (*C. fritschii* PCC 6912) wurden aseriate Kolonien von Zellaggregaten beobachtet. Für Cyanobakterien der Ordnung Stigonematales sind keine spezifischen Signaturgene bekannt, welche die komplexen Morphologien erklären könnten. Daher vermittelte diese Beobachtung die Vermutung, dass differentielle Transkription einen Einfluss auf morphologische Transitionen nimmt.

In dieser Arbeit wurde in *Fischerella* und *Chlorogloeopsis* eine vergleichende quantitative Transkriptomanalyse zwischen den unterschiedlichen Morphotypen und zwischen den Spezies durchgeführt. Mittels der 5'-dRNAseq Methode konnten Transkriptionsstartpunkte (TSSe) für jede Spezies genomweit detektiert werden. Die Mehrheit der TSSe wurde als TSSe komplementär zu einem CDS (aTSSe) klassifiziert oder konnte multiplen TSS-Klassen zugeordnet werden. Diese Beobachtung lässt einen hohen regulatorischen Anteil der Transkriptome und multiple Funktionen annotierter CDSe vermuten. Ein Vergleich von Einzelkopie-Proteinfamilien zwischen den *Fischerella*-Spezies ermittelte im Durchschnitt 35% konservierte TSSe. Zwischen *F. muscicola* und *C. fritschii* wurden im Durchschnitt 11% der TSSe als konserviert ermittelt. Eine Analyse korrelierender Expressionsänderungen zwischen den *Fischerella*-Spezies zeigte 198 konservierte TSSe, welche Startpunkte von Transkripten darstellen, die mit der morphologischen Transition einhergehen können. Beispielsweise wurde ein in der Saccharose hochregulierter TSS stromaufwärts eines potentiellen Operon lokalisiert, welches ein Homolog des Morphogens *bolA* beinhaltet. In *Escherichia coli* stellt BolA einen Repressor des bakteriellen Aktinhomologs *mreB* dar. Zwei weitere konservierte TSSe für *mreC* und *mreD* konnten detektiert werden, sodass angenommen wird, dass die differentielle Expression des *mre*-Operon einen Einfluss auf die beobachteten morphologischen Transitionen genommen hat.

2 Abstract

Cyanobacteria constitute a monophyletic group, yet species in this group are highly diverse in their cellular morphology, including unicellular and linear or branching filaments. Stigonematalean species from *Chlorogloeopsis* and *Fischerella* genera, forming multiseriate and branching filaments respectively, represent the most phenotypically complex cyanobacteria. Interestingly, culturing these species under increasing salt or sucrose concentrations results in different colony morphologies with the *Fischerella* (*F. muscicola* PCC 7414, *F. thermalis* PCC 7521) forming linear filaments and the *Chlorogloeopsis* (*C. fritschii* PCC 6912) forming aseriate clumped cell clusters. Since no clear signature genes can explain complex morphologies within Stigonematalean species, this observation suggests that such morphological transitions are facilitated by transcriptional regulation rather than differential gene content.

Here we quantify the transcriptome conservation among the three species and compare the transcriptional regulation between the natural and transformed morphologies. Using 5'-dRNAseq data we mapped transcription start sites (TSSs) within these cyanobacteria. Interestingly, the majority of TSSs where annotated as antisense TSSs (aTSSs) or were classified in more than one TSS-class, suggesting a huge antisense regulatory repertoire of transcripts and multiple functions for annotated genes. A comparison of single-copy orthologous genes revealed that 35% of the TSSs are conserved between the *Fischerella* species and 11% of the TSSs are conserved between *F. muscicola* and *C. fritschii.* An analysis of *Fischerella* orthologs where the conserved transcriptional changes between the branched and linear filaments were correlated, revealed 198 conserved TSSs that are putatively related to the morphological transitions. For example, one upregulated TSS is located upstream of a putative operon that includes a *bolA*-like gene. This gene is homologous to a known morphogene in *Escherichia coli* (*bolA*) that acts as a repressor of the bacterial actin homolog, *mreB*. In addition, two alternative conserved TSSs were detected for *mreC* and *mreD* in both *Fischerella* species, suggesting that differential expression of the *mre*-operon may have a role in cyanobacterial morphological transitions.

3 Einleitung

3.1 Das Phylum der Cyanobakterien

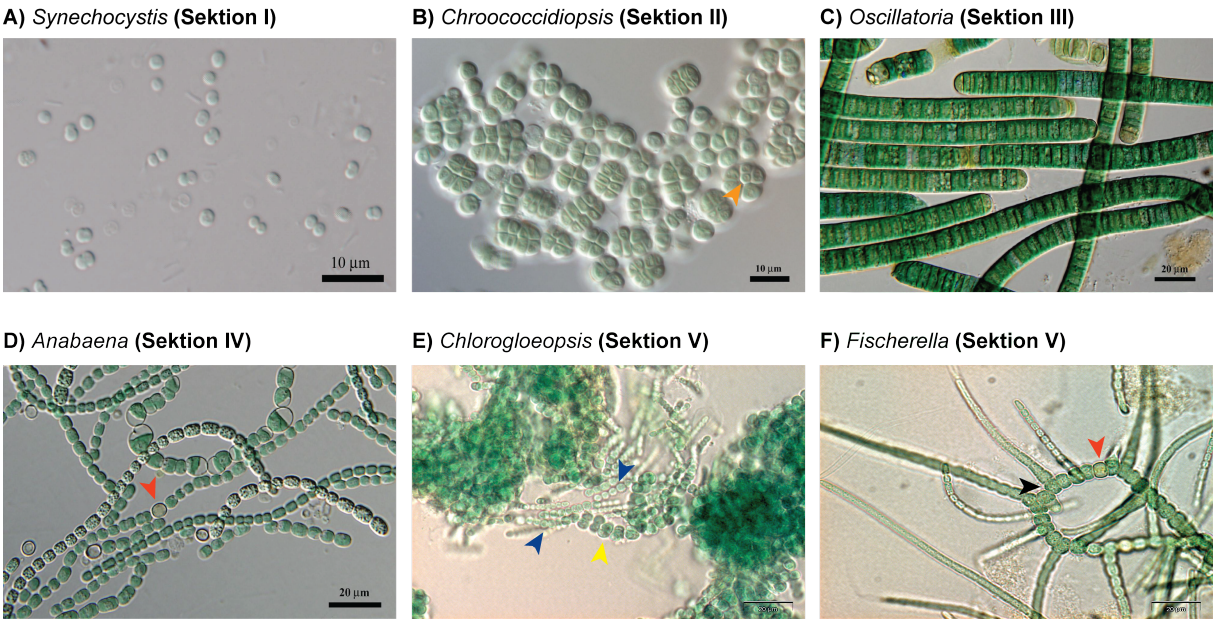


Abb. 1: Cyanobakterielle Morphologien.

B: Orangefarbener Pfeil = Baeozyte. **D, F:** Roter Pfeil = Heterozyste. **E:** Blaue Pfeile = Hormogonien, gelber Pfeil = multiseriates Filament. **F:** Schwarzer Pfeil = echtes Verzweigungswachstum. **A-D:** Vertreter der Sektionen I-IV (entnommen aus: http://ccala.butbn.cas.cz). **E-F:** Vertreter der Sektion V (**E** = *Chlorogloeopsis fritschii* PCC 6912, **F** = *Fischerella muscicola* PCC 7414).

Cyanobakterien besitzen die Fähigkeit der oxygenen Photosynthese als Hauptressource ihres Energiestoffwechsels. Sie kommen in multiplen Habitaten wie beispielsweise Thermalquellen, Brackwasser, Wüsten und Polarregionen vor (Garcia-Pichel und Pringault 2001, Garcia-Pichel et al. 2003, Pointing et al. 2009). Durch die Fähigkeit der oxygenen Photosynthese, haben Vorfahren heutiger Vertreter dieses Phylums vor 2,95-2,45 Milliarden Jahren maßgeblich an der Entstehung atmosphärischen Sauerstoffs auf der Erde beigetragen (Bekker et al. 2004, Planavsky et al. 2014). Sie werden heute als wichtiger Primärproduzent in den Ozeanen und als wichtiger Bestandteil in der Produktion atmosphärischen Sauerstoffs und des Kohlenstoffkreislaufs angesehen (Shi und Falkowski 2008). Die Stickstofffixierung ist für einige Gattungen dieses Phylums ein wichtiges Charakteristikum (Carpenter und Romans 1991, Canfield et al. 2010), welche dadurch oft in Symbiosen mit anderen Spezies gefunden werden. Im Austausch anderer Stoffe unterstützen diese Cyanobakterien ihre Symbionten wie Pflanzen, Moose, Pilze und Diatomeen mit fixiertem Stickstoff (Adams et al. 2006, Thompson et al. 2012). Während der Evolution der Eukaryoten spielten Vorfahren der heutigen

Cyanobakterien eine wichtige Rolle, da jene Vorfahren auch die Vorfahren der heutigen Plastiden darstellen und durch Gentransfer das Pflanzengenom maßgeblich mitgeformt haben (Martin et al. 1998, Dagan et al. 2013). Dieser Gentransfer zwischen den Plastiden und den Pflanzengenomen ist möglicherweise immer noch zu beobachten (Wang et al. 2012). Einige Gattungen der heutigen Cyanobakterien weisen die komplexesten Morphologien bei Bakterien auf. Daher sind sie auch Gegenstand der Analyse multizellulären Ursprungs (Claessen et al. 2014).

Basierend auf ihren unterschiedlichen Morphologien werden Cyanobakterien in fünf Sektionen eingeteilt (Rippka et al. 1979). Gattungen der Sektion I sind *Synechococcus*, *Prochlorococcus* und *Synechocystis* (Abb. 1A). Durch binäre Zellteilung und kompletter Abschnürung der Tochterzellen, weisen diese ein unizelluläres Wachstum auf. Die Sektion II umfasst *Chroococcidiopsis* (Abb. 1B) und *Stanieria*. Ein wichtiges Charakteristikum dieser Sektion ist die Entwicklung einer Baeozyte. Eine Baeozyte ist eine kleine vegetative Ausgangszelle. Diese vergrößert sich während des vegetativen Wachstums stark. Es entwickelt sich eine extrazelluläre Matrix (F-layer). Innerhalb des F-layers repliziert sich die DNA und die entstehenden Nukleoide verteilen sich auf das Zytoplasma. Anschließend erfolgt die Teilung des Zytoplasmas und innerhalb des F-layers entstehen durch diesen Vorgang neue kleine Baeozyten. Diese werden durch Aufbrechen des F-layers entlassen (Angert 2005). Der beschriebene Vorgang kann in den Sektion II Cyanobakterien unterschiedlich verlaufen. Innerhalb der Sektion III befinden sich die Gattungen *Oscillatoria* (Abb. 1C) und *Trichodesmium*. Die morphologische Erscheinungsform dieser Sektion umfasst filamentös wachsende Cyanobakterien. Arten der Gattung *Trichodesmium* sind die einzig bekannten diazotrophen Bakterien, welche Stickstoff unter aeroben Bedingungen bei Licht und ohne die Ausbildung von Heterozysten zu Ammonium reduzieren können (Bergman et al. 2012). Heterozysten können bei Cyanobakterien der Gattungen *Anabaena* (Abb. 1D) und *Rivularia* beobachtet werden, welche der Sektion IV zugehörig sind. Dies sind spezialisierte Zellen für die Stickstofffixierung, da die beteiligten Gene extrem sauerstoffempfindlich sind. Heterozysten können in sequenzieller Abfolge im Filament auftauchen. Dies ist bei *Anabaena* der Fall (Flores und Herrero 2010). Sie können aber auch polar auftauchen, was bei *Rivularia* vorkommt. Die komplexesten Morphologien sind bei Sektion V Cyanobakterien zu beobachten (Abb. 1E-F). Diese Sektion wird maßgeblich von Gattungen der Ordnung Stigonematales gebildet. Ein Charakteristikum ist das echte Verzweigungswachstum und die Zellteilung auf multiplen Teilungsebenen. Die Art des Verzweigungswachstums kann in den Gattungen der Sektion V variieren und wird als T-Typ und Y-Typ bezeichnet. Spezies der Gattung *Mastigocladus* zeigen vermehrt den Y-Typ des Verzweigungswachstums. In Spezies der Gattung *Fischerella* kann der T-Typ der echten Verzweigungsform beobachtet werden (Abb. 1F). Vertreter der Gattung *Chlorogloeopsis*

stellen eine Ausnahme dar, da sie multiseriate Zellaggregate (Abb. 1E) bilden. Alle Sektion V Cyanobakterien bilden die Heterozyste für die Stickstofffixierung und sind in der Lage Akineten (Dauerzellen) und freibewegliche Filamente (Hormogonien) zu bilden (Hoffmann und Castenholz 2015). Cyanobakterien der Sektionen III-V sind wichtige Beispiele mikrobieller Multizellularität. Die Zellen der Filamente sind durch sogenannte Septosomen miteinander verbunden und gewährleisten eine Kontinuität des Zytoplasmas (Giddings und Staehelin 1981, Wilk et al. 2011, Nürnberg et al. 2014).

Phylogenetische Analysen des cyanobakteriellen Phylums lassen einen polyphyletischen Ursprung des filamentösen Wachstums der Sektion III Cyanobakterien vermuten. Diese morphologische Erscheinungsform scheint demnach mehrfach unabhängig (konvergent) entstanden zu sein (Schirrmeister et al. 2013, Shih et al. 2013). Vertreter der Sektion IV-V sind monophyletisch (Dagan et al. 2013, Shih et al. 2013). Die Entwicklung der Heterozyste hat in evolutionären Prozessen vermutlich nur einmal stattgefunden. Des Weiteren sind Sektion V Cyanobakterien monophyletischen Ursprungs innerhalb der Sektion IV-V Klade (Dagan et al. 2013, Shih et al. 2013). Dies lässt vermuten, dass die Stigonematales einen Nostocales ähnlichen Vorfahren hatten und sich die komplexen Morphologien der heutigen Sektion V Cyanobakterien erst später in evolutionären Prozessen entwickelten.

Während die Zelldifferenzierung und die interzelluläre Kommunikation in Cyanobakterien der Nostocales intensiv analysiert wird (Flores und Herrero 2010), ist in den Stigonematales wenig über Komponenten bekannt, die einen Einfluss auf die komplexe Koloniemorphologie haben könnten. Die Sequenzierung von Genomen der Sektion V Cyanobakterien ermöglichte Analysen der An- und Abwesenheit von Genen, die mit der cyanobakteriellen Multizellularität und dem filamentösen Wachstum zusammenhängen könnten. Jedoch wurden keine spezifischen Sektion V Kandidatengene ermittelt, die mit dem echten Verzweigungswachstum oder der multiseriaten Koloniemorphologie einhergehen (Dagan et al. 2013, Shih et al. 2013). Vertreter der Sektion V Cyanobakterien weisen vermehrt Gene auf, die in signaltransduktionalen- und in transkriptionellen Kategorien eingeordnet werden (Shih et al. 2013). Allerdings konnte diesen Genen kein eindeutiger Einfluss auf die komplexen Morphotypen zugeschrieben werden. Analysen befassten sich mit der Identifizierung von Genen, deren Abwesenheit zu multizellulärem Wachstum führt. In einem Transposon Mutagenese Experiment des Sektion I Cyanobakteriums *Synechococcus elongatus* PCC 7942 konnte die Abwesenheit, bzw. der Verlust der Expression einiger Gene, zu einem filamentösen Wachstum führen (Miyagishima et al. 2005). Diesen Genen wurden daher Funktionen während der Zellteilung zugeschrieben. In allen Sektion V Cyanobakterien sind Homologe dieser Gene nachgewiesen worden (Stucken et al. 2010, Dagan et al. 2013). Daher kann der Verlust der Gene nicht die Ursache des Sektion III-V filamentösen Wachstums sein. Ob die Präsenz auch mit der Expression

zusammenhängt ist jedoch unbekannt. Proteine des Septosoms könnten ebenfalls einen Einfluss auf die Morphologie der Kolonie haben. Ein Experiment, in dem in *Nostoc punctiforme* ATCC 29133 (Sektion IV) das Gen *amiC2* ausgeschaltet wurde, zeigte Kolonien von Zellaggregaten im Gegensatz zu den linearen Filamenten des Wildtyps (Lehner et al. 2011). Das Gen *amiC2* hat Homologe in allen cyanobakteriellen Sektionen. Daher hängt die Abwesenheit dieses Gens wahrscheinlich nicht mit den beispielsweise multiseriaten Zellaggregaten von *Chlorogloeopsis* (Sektion V) zusammen. Die Überexpression des konservierten Tubulinhomologs FtsZ kann in *S. elongatus* PCC 7942 (Sektion I) ein filamentöses Wachstum induzieren (Mori und Johnson 2001). Daher könnten komplexe Morphologien der Sektion V Cyanobakterien transkriptionellen Ursprungs sein. Gestützt wird diese Vermutung von Studien an *Chlorogloeopsis fritschii* (Sektion V), in denen die Spezies unterschiedlichen Umweltfaktoren in Form von hohen Saccharosekonzentrationen ausgesetzt wurde. In den Experimenten konnte eine morphologische Transition vom multiseriaten zu einem aseriaten Morphotyp beobachtet werden (Evans et al. 1976). Diese Analyse demonstriert, dass komplexe Phänotypen in *C. fritschii* plastisch sind. Es unterstützt die Vorstellung, dass differentielle Expressionsmuster einen Einfluss auf morphologische Transitionen haben und nicht ausschließlich differentielle Genpräsenz ursächlich sein muss.

In der vorliegenden Arbeit wurde eine komparative Transkriptomanalyse der Spezies *F. muscicola*, *F. thermalis* und *C. fritschii* vorgenommen. Diese gehören den Gattungen *Fischerella* und *Chlorogloeopsis* der Sektion V Cyanobakterien an. Die vergleichende Transkriptomanalyse sollte Aufschluss darüber geben, ob eine differentielle Transkription bei Änderung der Wachstumsbedingungen einen Einfluss auf beobachteten morphologischen Transitionen hat. Zu diesem Zweck wurden die Primärtranskriptome der drei Spezies von jeweils zwei unterschiedlichen Wachstumsbedingungen sequenziert und primäre Transkriptionsstartpunkte (TSSe) detektiert. Dies sollte eine Analyse der Transkriptionsaktivität in Abhängigkeit der beobachteten morphologischen Transitionen ermöglichen.

3.2 Die Transkription in Eubakterien

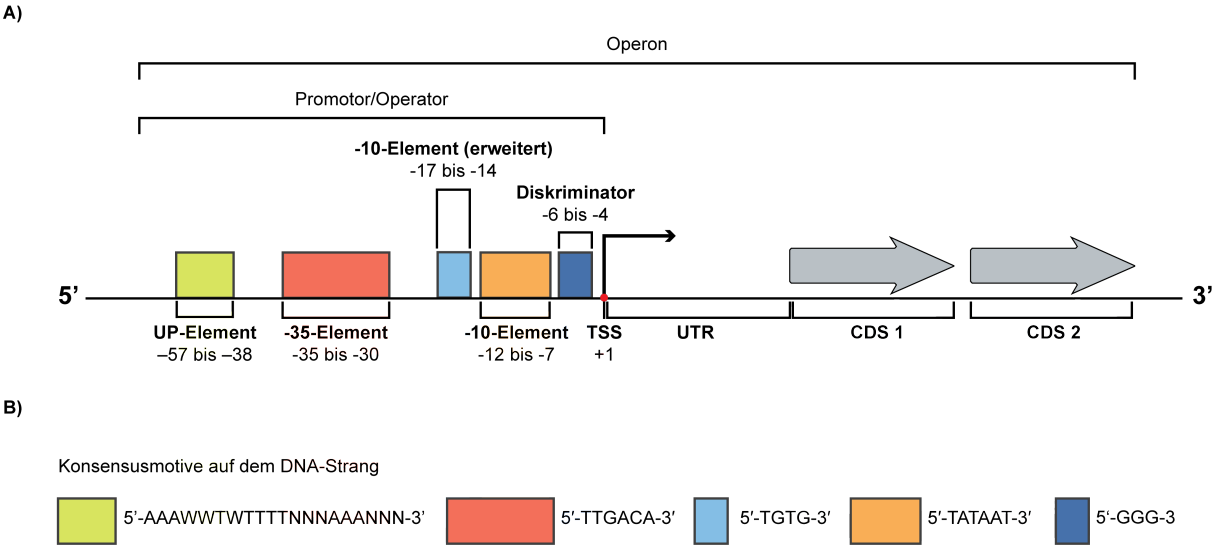


Abb. 2: Schematische Darstellung eines Operon.

A) Von rechts nach links: CDS = kodierender Sequenzabschnitt, UTR = untranslatierter Bereich, TSS = Transkriptionsstartpunkt. Innerhalb des Promotors/Operators sind Bindestellen für die RNA-Polymerase vorhanden. **B)** Konsensusmotive der Bindestellen. A = Adenin, T = Thymin, C = Cytosin, G = Guanin, W = A oder T, N = beliebiges Nukleotid.

Kodierende Sequenzabschnitte (CDSe) sind in Eubakterien in Form eines Operon arrangiert und weisen stromaufwärts einen operativen Sequenzbereich auf, der essentiell für die Transkription ist (Abb. 2). Erstmals beschrieben wurde dieses Modell in der Spezies *E. coli* (Jacob und Monod 1961). Die Transkription in Eubakterien verläuft über drei definierte Schritte. Im ersten Schritt (Transkriptionsinitiation) werden Sequenzmotive des abzulesenden DNA-Strangs (Abb. 2, Promotor/Operator) von Proteindomänen des Holoenzyms RNA-Polymerase (RNAP) erkannt (Campbell et al. 2008). Die meisten Sequenzabschnitte werden von einer modularen Komponente, dem sogenannten σ-Faktor, gebunden. Der σ-Faktor stellt eine eubakterielle Eigenschaft dar (Haugen et al. 2008). Eubakterien besitzen mindestens einen σ-Faktor (Campbell et al. 2008, Güell et al. 2011). Das -10-Element und das -35-Element werden von Proteindomänen der σ-Faktoren erkannt. Die meisten σ-Faktoren haben diese Domänen gemeinsam (Campbell et al. 2008). Diese Elemente auf dem DNA-Strang sind daher durch Konsensusmotive charakterisiert (Abb. 2B) (Harvey und Reynolds 1987). Nicht alle Promotoren, die vom selben σ-Faktor gebunden werden, weisen das gleiche Motiv auf. Es können jedoch Gene stärker transkribiert werden, je ähnlicher die Promotoren den Konsensusmotiven sind. Die σ^{70}-Faktoren stellen die größte Familie der σ-Faktoren dar (Campbell et al. 2008). Auch in Cyanobakterien wurden

σ^{70}-Faktoren identifiziert (Imamura und Asayama 2009). Die Familie wird in vier unterschiedliche Gruppen eingeteilt (Kumar et al. 2010). Die Gruppen Eins und Zwei sind hauptsächlich an der Transkription von lebensnotwendigen Komponenten, den sogenannten Haushaltsgenen, einer eubakteriellen Zelle beteiligt (Kumar et al. 2010). Die Gruppe Drei ist an der transkriptionellen Aktivität von Genen beteiligt, die für eine Stressantwort (Hitzestress) essentiell sind. Die vierte Gruppe beinhaltet σ^{70}-Faktoren, welche die Transkription von Genen unterschiedlicher Funktionen regulieren (Kumar et al. 2010). Der Unterschied der Gruppen liegt in der Kombination der Sequenzmotive an die die σ^{70}-Faktoren binden (Imamura und Asayama 2009). Die klassische Kombination ist die Präsenz des -35-Elements und des -10-Elements (Abb. 2) (Campbell et al. 2008). Es kann ein zusätzliches Motiv (UP-Element) stromaufwärts des -35-Elements vorhanden sein (Abb. 2). Diese Region weist ebenfalls ein Konsensusmotiv auf und interagiert direkt mit der C-terminalen Domäne der α-Untereinheit einer RNAP (Haugen et al. 2008). Des Weiteren kann das -35-Element abwesend sein und nur das -10-Element vorliegen (Imamura und Asayama 2009). Neuere Analysen zeigen die Kombination eines -35-Elements, eines -10-Elements und stromabwärts die Präsenz einer Region, welche als Diskriminator bezeichnet wird (Abb. 2) und ebenfalls ein Konsensusmotiv aufweist (Haugen et al. 2006). In *E. coli* wurde ein weiteres Konsensusmotiv entdeckt, das ein erweitertes -10-Element darstellt (Abb. 2) (Mitchell et al. 2003).

Nach der Initiation erfolgt der Übergang vom geschlossenen Komplex in den offenen Komplex (Transkriptionselongation). Die RNAP löst sich vom Promotor. Das erste von der RNAP transkribierte Nukleotid stellt den TSS dar. Stromabwärts eines TSS liegt die untranslatierte Region (Abb. 2, UTR). Dieser Bereich kodiert für ribosomale Bindestellen (Shine-Dalgarno-Sequenz) oder für Elemente, die während der Transkription, bzw. posttranskriptional, Einfluss auf die Translationsregulation nehmen können. Riboswitches besitzen eine solche Funktion (Barrick et al. 2004, Güell et al. 2011). Stromabwärts der UTR liegen die kodierenden Sequenzabschnitte (Abb. 2, CDS 1 u. CDS 2). Die Expression mehrerer CDSe kann durch einen Promotor erfolgen. Alle CDSe können eine UTR aufweisen. Durch das Vorhandensein mehrerer CDSe wird das entstehende Transkript auch als polycistronisches Transkript bezeichnet. Die Transkription endet mit der Termination, die ebenfalls von einem Sequenzabschnitt auf dem DNA-Strang vermittelt wird. Dieser Sequenzabschnitt vermittelt in Eubakterien eine rho-abhängige bzw. eine rho-unabhängige Termination (Güell et al. 2011). Im ersten Fall ist das Protein Rho an der Termination beteiligt. Im zweiten Fall kann durch das Vorliegen eines speziellen Sequenzabschnittes eine Haarnadelschleife am Ende des RNA-Transkripts entstehen, welche die Transkription beendet (Santangelo und Artsimovitch 2011). Die Transkription ist in Eubakterien ein stark regulierter Prozess. Die σ-Faktoren können durch anti-σ-Faktoren reguliert werden (Campbell et al. 2008).

3.3 Transkriptionelle Regulation

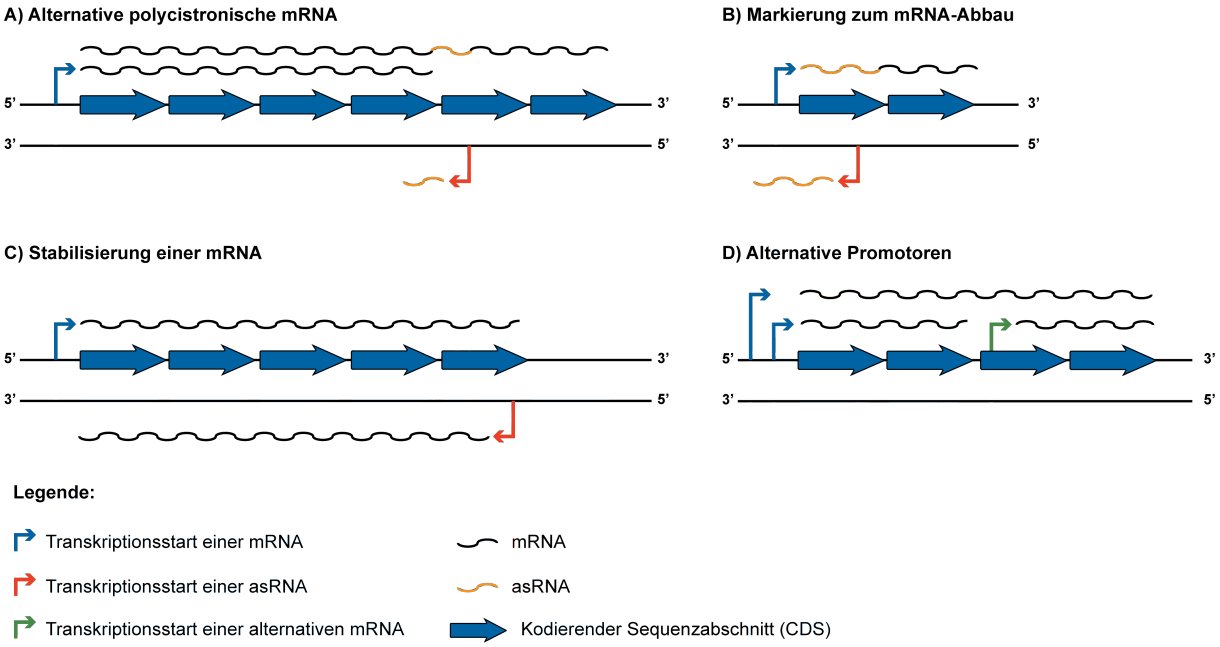


Abb. 3: Beispiele transkriptioneller Regulation.

Die transkriptionelle Regulation in Eubakterien verläuft nicht ausschließlich über unterschiedliche σ-Faktoren. Weitere Transkriptionsfaktoren wie NusA, NusG und GreA können Einfluss auf die Elongationsphase der Transkription nehmen (Güell et al. 2011). Viele regulatorische Funktionen können über kleine RNAs (sRNAs), lange nicht-kodierende RNAs (ncRNAs) und durch komplementäre RNAs (asRNAs) erfolgen (Sesto et al. 2013). Abb. **3** zeigt vier Beispiele wie eine solche Regulation vermittelt werden kann. Im ersten Beispiel (Abb. 3A) wird von einer polycistronischen mRNA ausgegangen. In einer Operonregion auf dem komplementären Strang kann sich ein TSS für eine asRNA befinden (Abb. 3A, roter Pfeil). Wird diese transkribiert ist eine differentielle Expression des Operons möglich. Die asRNA weist Komplementarität mit der UTR eines CDS der polycistronischen mRNA auf. Durch Basenpaarung kann sich eine Haarnadelschleife bilden, welche die Transkription vorzeitig terminiert. Zwei alternative polycistronische mRNA-Varianten sind möglich und durch die direkte Translation einer entstehenden mRNA ist die Expression der Gene differentiell. In der Spezies *Vibrio anguillarum* wurde eine solche Regulation entdeckt. Durch die Existenz einer asRNA wird das *fad*-Operon differentiell exprimiert (Stork et al. 2007). Die kleinere Variante des polycistronischen Transkripts lag über 17 mal höher exprimiert vor als die längere Variante (Stork et al. 2007). Im zweiten Beispiel (Abb. 3B) wird eine Markierungsfunktion zum mRNA-Abbau illustriert. Ein Operon mit zwei kodierenden Sequenzabschnitten wird transkribiert. Ein Transkriptionsfaktor kann die Transkription einer asRNA aktivieren, welche

komplementär zum ersten CDS ist. Es entsteht ein mRNA/asRNA Duplex, der abgebaut wird. Im Cyanobakterium *Synechocystis* sp. PCC 6803 ist eine solche Funktionsweise beschrieben worden. Unter limitiertem Eisengehalt in der Umgebung wird die Transkription einer asRNA für *isiA* aktiv (Dühring et al. 2006). Der Duplex der *isiA*-mRNA und der asRNA markiert das Transkript zum Abbau (Dühring et al. 2006). Ein Gegenbeispiel stellt die transkriptionelle Regulation über eine lange asRNA dar (Abb. 3C). Eine polycistronische mRNA wird transkribiert. Auf dem komplementären Strang wird eine lange asRNA gebildet. Dieser Duplex kann Erkennungsmotive in der mRNA maskieren und stabilisiert so das Transkript. In einer Analyse des Cyanobakteriums *Prochlorococcus* sp. MED4 wurden asRNAs entdeckt, welche diese Funktion aufweisen. Entdeckt wurden diese asRNAs durch den Expressionsvergleich zwischen einem mit einem Phagen infizierten Stamm und einem nicht infizierten Stamm. Der infizierte Stamm zeigte einen großen Abfall der mRNA-Expression. Einige Regionen zeigten jedoch Überexpression. Diese Regionen waren durch lange asRNA-Präsenz gekennzeichnet (Stazic et al. 2011). Die stabilisierende Funktion wurde durch die Maskierung von Erkennungssequenzen für das RNase E Enzym vermittelt (Stazic et al. 2011). Das vierte Beispiel stellt die Existenz alternativer Promotoren (Abb. 3D) dar. Von einem Operon wird eine polycistronische mRNA synthetisiert. Stromabwärts des primären Promotors liegen zusätzliche alternative Promotoren. Unter variierenden Bedingungen werden die alternativen Promotoren aktiv und führen zu unterschiedlichen Expressionsraten der kodierenden Sequenzabschnitte. Solche Suboperonstrukturen können durch bedingungsabhängige Faktoren aktiviert werden. In der Spezies *Mycoplasma pneumoniae* wurden potentielle interne Promotoren in einigen Operonregionen detektiert. Die den Operon assoziierten CDSe zeigten Heterogenität in deren Expressionsprofil (Güell et al. 2009).

Die transkriptionelle Regulation von Genen mittels sRNAs und alternativen Promotoren ist in vielen bakteriellen Spezies beschrieben. In Cyanobakterien wurden in Vertretern der Sektion I (Mitschke, Georg, et al. 2011, Kopf et al. 2014, Voigt et al. 2014), der Sektion III (Pfreundt et al. 2014) und der Sektion IV (Mitschke, Vioque, et al. 2011, Kopf et al. 2015) genomweite Transkriptomanalysen durchgeführt. Alle Analysen dokumentieren die Präsenz vieler asRNAs sowie potentielle alternative Promotoren in Operonregionen. Die transkriptionelle Regulation und differentielle Expression von Genen scheint daher in Cyanobakterien ein generelles Charakteristikum. Für Vertreter der Sektion II und V Cyanobakterien fehlt eine solche Transkriptomanalyse bisher. Seit der Einführung der differentiellen RNA-Sequenzierung (dRNAseq) ist es jedoch möglich genomweite Transkriptionsstartpunkte mit geringem Aufwand zu detektieren.

3.4 Die differentielle RNA-Sequenzierung (dRNAseq)

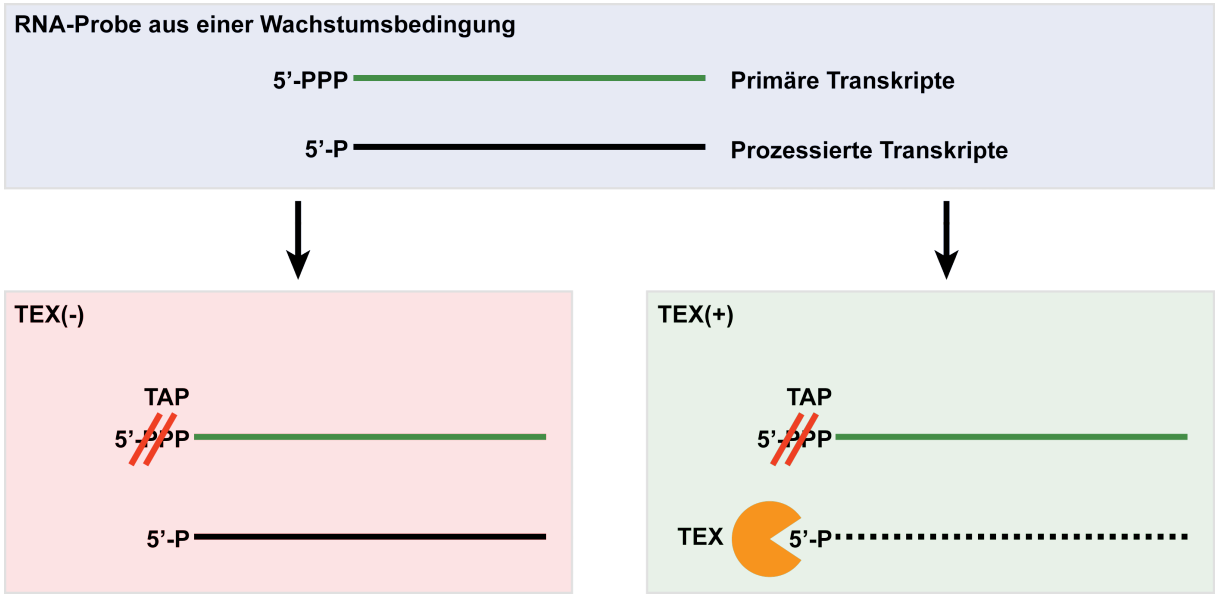


Abb. 4: Die dRNAseq Methode (modifiziert aus (Sharma und Vogel 2014)).

5'-PPP: Triphosphatrest, 5'-P: Monophosphatrest, TEX: Terminator Exonuklease, TAP: Tobacco acid pyrophosphatase.

Die Identifizierung von Transkriptionsstartpunkten (TSSe) mittels der differentiellen RNA Sequenzierung (dRNAseq) hat seine Anfänge in der Transkriptomanalyse des humanpathogenen Bakteriums *Helicobacter pylori* (Sharma et al. 2010). Sie wurde ursprünglich entwickelt, um in dieser Spezies den Mangel an ncRNAs zu erforschen. Doch das Gegenteil wurde beobachtet (Sharma und Vogel 2014). Für viele andere Spezies ermöglichte diese Methode erst eine Transkriptomanalyse. In den folgenden Jahren führten die Ergebnisse vieler Transkriptomanalysen zu einem Paradigmenwechsel. In bakteriellen Genomen wird starke pervasive transkriptionelle Aktivität angenommen und unabhängig von der Genomgröße nehmen sRNAs viele regulatorische Funktionen ein (Wade und Grainger 2014). Die differentielle RNA-Sequenzierung hat zu dieser neuen Perspektive beigetragen (Sharma und Vogel 2014). Seitdem sind mit dieser Methode zahlreiche Transkriptomprojekte in Eubakterien (*Schewanella* (Shao et al. 2014)), Archeen (*Methanosarcina* (Jäger et al. 2009)) und Eukaryoten (*Hordeum* (Zhelyazkova et al. 2012)) durchgeführt worden. Die Vorgehensweise der dRNAseq wird anschaulich von Sharma und Vogel (Sharma und Vogel 2014) beschrieben und wird zusammenfassend für die Analyse eingeleitet.

Das Transkriptom einer eubakteriellen Zelle besteht aus Primärtranskripten, die ein Triphosphatrest am 5'-Ende tragen. Dies stellt ein Analogon der 5'-Kappe eukaryotischer RNA

dar (Sorek und Cossart 2010). Zusätzlich entstehen prozessierte Transkripte, welche ein Monophosphatrest am 5'-Ende tragen. Prozessierte Transkripte sind ribosomale RNAs (rRNAs) sowie transfer RNAs (tRNAs) (Sorek und Cossart 2010). Es handelt sich aber nicht um Transkripte, die während der Transkriptionsinitiation entstehen. Die differentielle RNA-Sequenzierung ist eine Methode der Sequenzierung, die über einen selektiven Einsatz einer Terminator Exonuklease (TEX) eine der aufgeteilten Proben für Primärtranskripte im Allgemeinen anreichern soll (Sharma et al. 2010). Das bedeutet, dass der Transkriptionsstartpunkt (TSS) eines Transkripts auf Nukleotidebene ermittelt werden kann, da nur ein Primärtranskript ein 5'-Triphosphatrest aufweist (Sharma et al. 2010).

Die extrahierte RNA wird unter jeder Bedingung in eine TEX(-)- und eine TEX(+)-Probe aufgeteilt (Abb. 4). Die TEX(+)-Probe wird mit einer Terminator Exonuklease (TEX) behandelt. Beide Proben werden mit einer Tobacco acid pyrophosphatase (TAP) behandelt. In der TEX(+)-Probe wird kein rRNA-Degradierungskit verwendet. In der TEX(-)-Probe wird ein rRNA-Verdau vorgenommen, da quantitativ betrachtet annähernd 95-99% des gesamten Transkriptoms einer eubakteriellen Zelle ausschließlich aus tRNA- und rRNA-Transkripten besteht (Sorek und Cossart 2010). Die resultierende TEX(+)-Probe weist im Idealfall keine prozessierten Transkripte mehr auf. In der TEX(-)-Probe sind primäre und prozessierte Transkripte, die von dem rRNA-Degradierungskit nicht abgebaut wurden, vorhanden. Die TEX(-)-Proben dienen bei der Bestimmung des Transkriptionsstartpunktes als Vergleichsgröße, um die exakte Position des ersten transkribierten Nukleotids zu ermitteln. Der Einsatz von TAP dient dem Abspalten des 5'-Diphosphatrest vom 5'-Triphosphatrest der Primärtranskripte. Dieser Schritt wird für die Ligation eines RNA-Adapters durchgeführt. Alle Transkripte werden zusätzlich am 3'-Ende polyadenyliert. Eine reverse Transkriptase wird verwendet, um die RNA in cDNA umzuschreiben. Die cDNA wird im Anschluss sequenziert. In dieser Arbeit wird die dRNAseq Methode für jeweils zwei Wachstumsbedingungen in den Spezies *F. muscicola*, *F. thermalis* und *C. fritschii* angewendet.

4 Zielsetzung

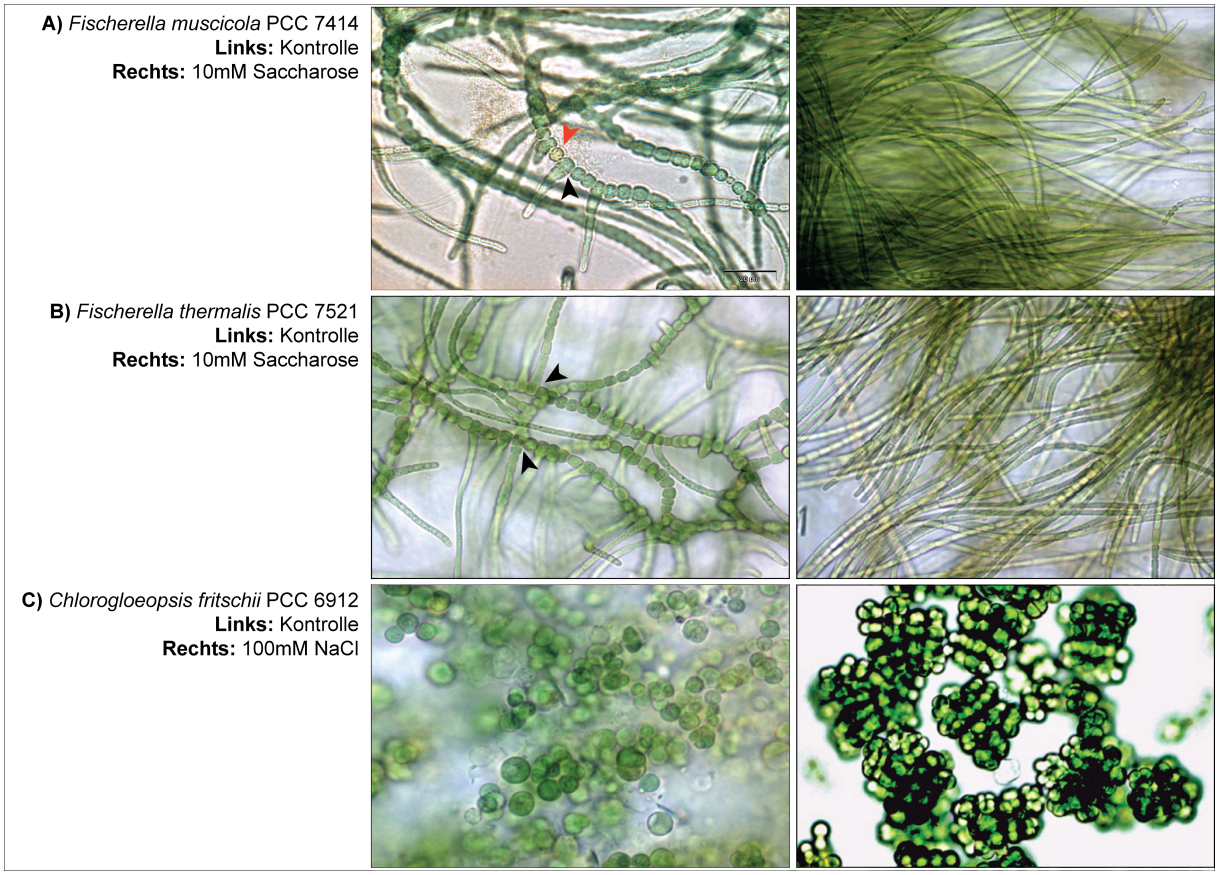


Abb. 5: Morphologische Transitionen in *Fischerella* und *Chlorogloeopsis*.

A, B (links): Roter Pfeil = Heterozyste. Schwarze Pfeile = echtes Verzweigungswachstum. **A)** und **B)** zeigen Morphologien der *Fischerella* Spezies in der Kontrollbedingung (**links**) und bei Induktion einer Konzentrationserhöhung von Saccharose (**rechts**). Bei 10mM wurde ein synchronisierter Morphotyp beobachtet, der weder Heterozysten noch echtes Verzweigungswachstum aufwies. In **C)** wird *Chlorogloeopsis* in der Kontrollbedingung (**links**) und bei Induktion einer Konzentrationserhöhung von NaCl (**rechts**) dargestellt. Bei 100 mM NaCL wurde ein synchronisierter Morphotyp von aseriaten Zellaggregaten ohne Heterozysten beobachtet.

Morphologische Transitionen innerhalb der Sektion V Cyanobakterien könnten nicht mit einem differentiellen Gengehalt einhergehen, sondern mit unterschiedlichen transkriptionellen Expressions- und Regulationsmustern. Um diese Hypothese zu testen, wurden die Spezies *F. muscicola* PCC 7414, *F. thermalis* PCC 7521 und C. *fritschii* PCC 6912 in jeweils zwei unterschiedlichen Wachstumsbedingungen kultiviert und die Transkriptome sequenziert. Bei der ersten Bedingung handelte es sich um die generalisierten Morphologien (Kontrolle) der drei Spezies (Abb. 5, links). Die Spezies der Gattung *Fischerella* wuchsen in Form von filamentösen und echt verzweigten Trichomen (Abb. 5AB, links). Bei *C. fritschii* wurden unizelluläre und multiseriate Zellaggregate beobachtet (Abb. 5C, links). Unterschiedliche Zelldifferenzierungen (Heterozysten) und Wachstumsformen (Hormogonien) konnten in allen

drei Spezies beobachtet werden. Die unterschiedlichen Zelltypen und Wachstumsformen weisen mit hoher Wahrscheinlichkeit andere transkriptionelle Aktivität auf. Zwischen Heterozysten und vegetativen Zellen ist dies in *Anabaena* sp. PCC 7120 dokumentiert (Flores und Herrero 2010). Die Koloniemorphologie der drei Spezies wurde synchronisiert, um die transkriptionellen Signale solcher Zelltypen und Wachstumsformen zu minimieren. Die Induktion einer steigenden Saccharosekonzentration konnte in den *Fischerella* Spezies einen synchronisierten Morphotypen erzeugen (Abb. 5AB, rechts). Bei einer 10 mM Saccharoseinduktion stellten *F. muscicola* und *F. thermalis* das Verzweigungswachstum ein und bildeten Sektion III ähnliche Filamente. Bei *C. fritschii* wurden nach einer 100 mM NaCl-Induktion aseriate Zellaggregate beobachtet. Dieser Morphotyp ähnelte den Sektion II.

Um die transkriptionellen Änderungen der synchronisierten Morphotypen zu analysieren, wurde die RNA der drei Spezies aus jeweils beiden Bedingungen extrahiert und sequenziert. Eine Detektion der Transkriptionsstartpunkte wurde in allen Bedingungen durchgeführt. Vergleichende Transkriptomanalysen zwischen den Bedingungen und zwischen den Spezies sollten Gene identifizieren, die auf die Induktion mit differentieller Expression reagieren. Im Vergleich zu den Analysen in *S. elongatus* PCC 7942 (Miyagishima et al. 2005) und *N. punctiforme* ATCC 29133 (Lehner et al. 2011), stellte diese Analyse einen reversen Ansatz der genomweiten Detektion genetischer Komponenten des Divisom- und Elongationskomplexes dar.

5 Material und Methoden

5.1 Überblick

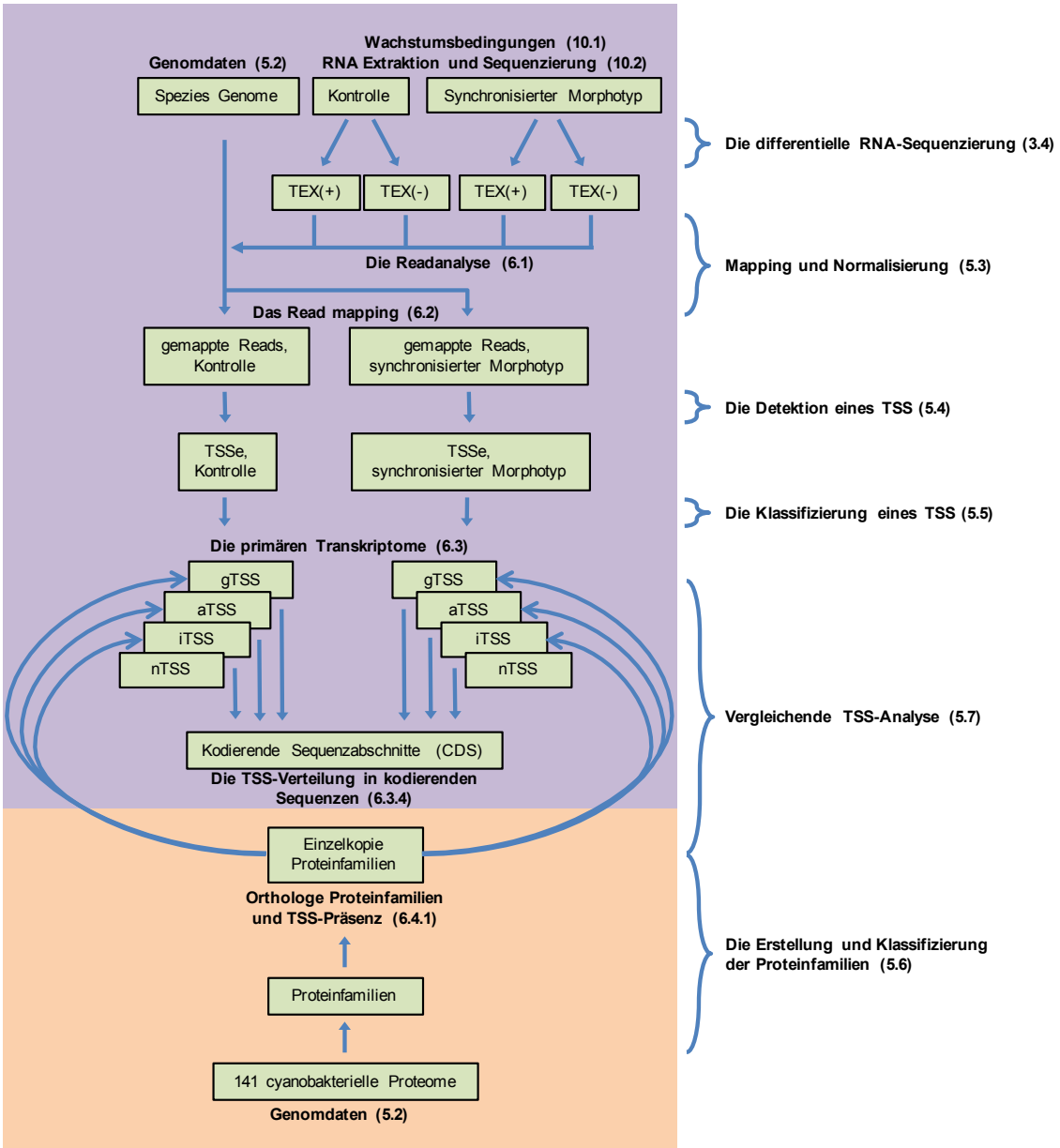


Abb. 6: Arbeitsablauf der Analyse.

Die purpurfarbene Box stellt die Transkriptomebene in dieser Arbeit dar und umfasst die Analyse aller Bedingungen der drei Spezies. Die orangefarbene Box beschreibt die Proteomebene und stellt die Identifizierung der Einzelkopie-Proteinfamilien dar. Die blauen Pfeile stehen für Methoden dieser Arbeit. Die grünen Kästchen bezeichnen Ergebnisse. Die Nummern in den Klammen zeigen die jeweiligen Abschnitte des Methoden- und Ergebnisteils an, in denen der entsprechende Sachverhalt erläutert wird.

Für die Analyse auf Transkriptomebene wurden Genom- und Transkriptomdaten der Spezies *F. muscicola* PCC 7414, *F. thermalis* PCC 7521 und *C. fritschii* PCC 6912 verwendet. Abschnitt 5.2 erläutert die verwendeten Genomdaten. Die Abschnitte 3.4 und 10.2 beschreiben die Erstellung der Transkriptomdaten aus den jeweils unterschiedlichen Wachstumsbedingungen der drei Spezies. Das Ziel war Transkriptionsmuster zu detektieren, die einen möglichen Einfluss auf die synchronisierten Morphotypen haben können. Die Transkriptomebene wurde für jede der drei Spezies separat durchgeführt und deckt damit eine allgemeine TSS-Analyse auf Speziesebene ab.

Auf Proteomebene wurden Proteomdaten von 141 Cyanobakterien (Tab. 11) verwendet. Bei Beginn der Analyse war dies die höchste Anzahl an Spezies für die ein sequenziertes Genom und Genannotationsdaten vorlagen. Für die entsprechenden Speziesproteome wurden in der Analyse Homologien für jedes Protein ermittelt, um Einzelkopie-Proteinfamilien zu identifizieren (Abschnitt 5.6). Der Zweck dieser Vorgehensweise lag im vergleichenden Aufbau der Analyse. Die Einzelkopie-Proteinfamilien ermöglichten eine direkte Verbindung von TSS und CDS zwischen den Spezies. Des Weiteren wurde ein differentieller Gengehalt als Ursache für die morphologischen Transitionen nicht angenommen. Daher wurden die zu identifizierenden Transkriptionsmuster in Einzelkopie-Proteinfamilien vermutet. Deren Funktion und transkriptionelle Regulation eröffnet zusätzlich eine mögliche phylumweite Analyse.

Insgesamt wurden zwei Interspeziesanalysen durchgeführt. Das Bindeglied der Ebenen sind identifizierte Einzelkopie-Proteinfamilien aus der Proteomebene und assoziierte TSSe aus der Transkriptomebene (Abb. 6). Durch die Einzelkopie-Proteinfamilien war ein direkter Speziesvergleich der TSSe ohne Genomalignment möglich (Abschnitt 5.7). Zusätzlich eröffnete dies die Definition orthologer TSSe (Abschnitt 5.7). Die Sequenzidentitäten von TSS-Regionen und orthologen Genpaaren wurden verglichen (Abschnitt 6.4.3). Des Weiteren wurde eine vergleichende Analyse der Änderung der Transkriptabundanz orthologer TSSe (Abschnitt 6.4.4) durchgeführt. Dieser Aspekt der Analyse ermöglichte die Detektion einer Korrelation der Änderung der Transkriptabundanz zwischen *F. muscicola*, *F. thermalis*. Im Falle des *F. muscicola* / *F. thermalis* Vergleichs wurde zusätzlich durch die Saccharose induzierten und sehr ähnlichen Morphotypen (Abb. 5A u. B) eine Identifizierung von Kanditatengenen möglich. Diese Kandidaten stellen bekannte und potentiell neue Komponenten des Elongations- und Divisomkomplex dar (Abschnitt 6.4.4).

5.2 Genomdaten

Die in dieser Analyse verwendeten Sequenzdaten wurden vom Joint Genome Institute (JGI)[1] heruntergeladen. Das JGI bietet Dienstleistungen im Bereich der Sequenzierung an. Darunter fallen die Sequenzierung von Genomen, Metagenomen, Transkriptomen und deren automatisierte Annotation. Die assemblierten Genome, Proteome und Annotationsdaten mikrobieller Spezies können über die Integrated Microbial Genomes Datenbank (IMG) analysiert werden.

Von den 141 cyanobakteriellen Sequenzdaten wurden 135 im April 2013 von der IMG mit einem FTP-Zugang[2] heruntergeladen. Die Struktur der IMG-Sequenzdaten ist der Struktur der Daten des National Center of Biotechnology and Information (NCBI)[3] sehr ähnlich. Jede Spezies besitzt eine IMG-Taxon ID[4] (Tab. 11). Mit dieser ID können in der IMG zusätzliche Informationen über eine Spezies erhalten werden. Alle Gene besitzen eine IMG-Gen ID. Diese sind deckungsgleich zu den IDs der entsprechenden Proteinsequenzen, sofern ein Gen ein Protein kodiert. Zu jedem IMG-Spezieseintrag gibt es Metadaten, wie beispielsweise die NCBI-Taxon ID. Diese stellt eine direkte Verbindung zu abgeschlossenen und bereitgestellten Projekten des NCBI dar. Die Metadaten sind über den IMG-Webauftritt zu erhalten. Dieser erfordert eine Registrierung. Das JGI umfasst publiziertes genomisches Sequenzdatenmaterial, laufende Projekte und fragmentierte Genomsequenzen (Drafts).

Weitere sechs Genomprojekte wurden vom Zentrum für Biotechnologie (CeBiTec) der Universität Bielefeld durchgeführt[5]. Diese Projekte wurden vom Institut für Molekulare Evolution der Heinrich-Heine-Universität Düsseldorf in Auftrag gegeben. Es handelt sich um die Genome folgender Spezies:

- *Chlorogloeopsis fritschii* PCC 6912 (NCBI Accession: AJLN00000000)
- *Chlorogloeopsis fritschii* PCC 9212 (NCBI Accession: AJLM00000000)
- *Fischerella muscicola* PCC 73103 (NCBI Accession: AJLJ00000000)
- *Fischerella thermalis* PCC 7521 (NCBI Accession: AJLL00000000)
- *Fischerella muscicola* PCC 7414 (NCBI Accession: AJLK00000000)
- *Scytonema hofmanni* PCC 7110 (NCBI Accession: ANNX00000000)

[1] http://jgi.doe.gov
[2] von engl. file transfer protocol = Protokoll zum Herunterladen von Dateien eines entfernten Computers.
[3] http://www.ncbi.nlm.nih.gov
[4] von engl. Identifier = Eine Zahlen- und/oder Buchstabenkombination zum Identifizieren von Datenbankeinträgen.
[5] http://www.cebitec.uni-bielefeld.de

Zu Beginn der Analyse waren die Sequenzdaten bereits publiziert (Dagan et al. 2013), aber vom NCBI noch nicht verarbeitet worden. Entsprechend wurden die Sequenzdaten vom CeBiTec benutzt. Diese wurden von der Annotationsdatenbank GenDB[6] heruntergeladen. Die verwendeten Genome liegen in sogenannten Scaffolds[7] vor. Beim NCBI liegen die Genome als Contigs[8] vor. Das Vorliegen von Scaffolds hat den Vorteil, dass für einige Contigs eine Richtung und Sortierung innerhalb des Genoms ermittelt werden konnte. Die Genomkarte ist feiner aufgelöst. Zwischen den Contigs und zwischen den Scaffolds liegen unsequenzierte Bereiche. Unsequenzierte Bereiche sind durch Linker[9] gekennzeichnet. Bei diesen Linkern handelt es sich um eine repetitive Sequenzabfolge `CTAGCTAGCTAG`, die an den Start- und Endpunkten der Contigs lokalisiert ist. Zwischen den Contigs befinden sich die unsequenzierten Bereiche. Diese sind durch eine Abfolge des Buchstabens `N` gekennzeichnet. In einigen Fällen ist die Lückengröße bekannt und wird mit einer entsprechenden Anzahl von `N` aufgefüllt. Dadurch wird die tatsächliche Genomlänge genauer aufgelöst. Bis auf *Chlorogloeopsis fritschii* weisen alle verwendeten Genome diese Eigenschaften auf. Die Scaffolds wurden aufgrund der ermittelten Orientierung der Contigs verwendet. Durch diese Richtungsinformation war eine genauere Zuordnung zwischen TSSe und CDSe möglich. Das Genom der Spezies *Fischerella thermalis* konnte während der Analyse geschlossen werden. Die automatisierten Annotationen wurden jedoch nicht überprüft. Deshalb wurde auch bei *F. thermalis* auf die Scaffold-Version zurückgegriffen. Insgesamt hatte die Analyse einen Genomsequenz-, Gensequenz- und Proteinsequenzumfang von 141 cyanobakteriellen Spezies. In Tab**. 11** sind die von der GenDB verwendeten Genome hervorgehoben.

Die Annotationsdaten und die Anzahl kodierender Sequenzen hängen vom verwendeten automatisierten Annotationsablauf ab (Kisand und Lettieri 2013, Bakke et al. 2009). Bei der GenDB werden die Programme `GLIMMER` (Delcher et al. 1999) und `CRITICA` (Badger und Olsen 1999) zur Detektion offener Leseraster (ORFs) verwendet (Meyer et al. 2003). Beim NCBI kommt eine modifizierte Version des Programms `GeneMarkS` (Besemer et al. 2001) zum Einsatz. Daher ist es möglich, dass die Annotationsdaten des NCBI und des GenDB qualitative und quantitative Unterschiede für die sechs Cyanobakterien aufweisen. Die Anzahl der vorhergesagten ORFs ist bei den GenDB Annotationen höher als beim NCBI. Deshalb haben nicht alle identifizierten Kandidatengene einen NCBI-Locus Tag erhalten (Abschnitt 10.8).

[6] https://gendb.cebitec.uni-bielefeld.de/cgi-bin/login.cgi
[7] http://genome.jgi.doe.gov/help/scaffolds.html
[8] von engl. contiguous = angrenzend, zusammenhängend.
[9] zu deutsch = Verknüpfer. Hier eine künstlich erzeugte, repetitive Sequenzabfolge.

5.3 Mapping und Normalisierung

Während der Erstellung der cDNA-Bibliotheken oder der Sequenzierung, konnten technische Artefakte entstehen. Daher wurden die Reads[10] der Sequenzierungen einer Qualitätskontrolle unterzogen. Die erhaltenen Reads der TEX(-)- und TEX(+)-Proben wurden für jede Spezies und Bedingung separat analysiert. Eine allgemeine Nukleotidkompositionsanalyse pro Position sollte Aufschluss über mögliche Anreicherungen von Nukleotiden geben (Abschnitt 6.1). Im Anschluss erfolgte eine Homopolymeranalyse, die das sequenzielle Auftreten vieler gleicher Nukleotide analysieren sollte. Bei dieser Analyse wurde jeder Read auf Polymere eines Nukleotids hin untersucht. Dabei wurde immer das größtmögliche Polymer der Nukleotide Adenin, Thymin, Guanin und Cytosin gezählt. Demzufolge waren Monomere und kleine Polymere oft zu finden. Hat man einen Read mit einer Länge von acht Nukleotiden und der Sequenz `ATGCATGC`, wird die Homopolymeranalyse zwei Monomere jedes Nukleotids finden. Hat man einen Read mit einer Länge von acht Nukleotiden und der Sequenz `ATGCCATG`, werden jeweils zwei Monomere für `A`, `T` und `G`, aber ein Polymer des Nukleotids `C` gezählt. Die Länge eines Homopolymers durfte die Länge eines Homopolymers auf dem Genom nicht überschreiten. Waren Anreicherungen längerer Polymere in einer TEX-Probe vorhanden, wurden die Reads am Startpunkt des Polymers bis zum Ende des Reads gekürzt. Die Reads durften nach Kürzung nicht kleiner als 20 Nukleotide sein. War dies für einen Read der Fall, wurde dieser verworfen.

Das Mapping der analysierten Reads auf die jeweiligen Genome wurde mit dem Programm `blastall` (v. 2.2.17) mit der Option `blastn` (Altschul et al. 1997) durchgeführt. Dieses Programm ist in vorherigen Transkriptomanalysen eingesetzt worden (Wurtzel et al. 2009, Wurtzel et al. 2012). Die Ergebnisse der Mappings sind mit denen anderer Programme wie `Maq` (Li et al. 2008) oder `Novoalign` [11] vergleichbar (Wurtzel et al. 2009). Ein Vorteil ergibt sich aus der einfachen Analyse der tabellarischen Ausgabe. Diese wird durch den Parameter `-m8` erzielt. Die Nachteile von `blastall` liegen in einer langen Laufzeit und einem erhöhten Speicherplatzbedarf durch unkomprimierte Ausgabe. Der Parameter des Eingabesequenzfilters `-F` wurde auf `FALSE` gesetzt. Alle anderen Parameter besaßen Standardeinstellungen.

Um einen Read als gemappt zu definieren, mussten folgende Kriterien erfüllt sein: Erstens, ein Read durfte nicht auf einer ribosomalen Operonregion aligniert werden. Zweitens, der von `blastall` berechnete `expectation value (E)` durfte den Wert von 0,0001 nicht

[10] Reads = kurze Abschnitte eines Transkripts, hier 50 Nukleotide lang.
[11] http://www.novocraft.com/products/novoalign

überschreiten. Drittens, höchstens vier positionelle Fehltreffer und keine Lücken durften innerhalb des Alignments existieren. Viertens, das Alignment musste 80% der Eingabesequenzlänge umfassen und fünftens, es durfte nur ein Alignment mit maximalen Bitscore vorkommen. Die tabellarischen Ausgaben wurden anschließend in das `wiggle` Dateiformat konvertiert.

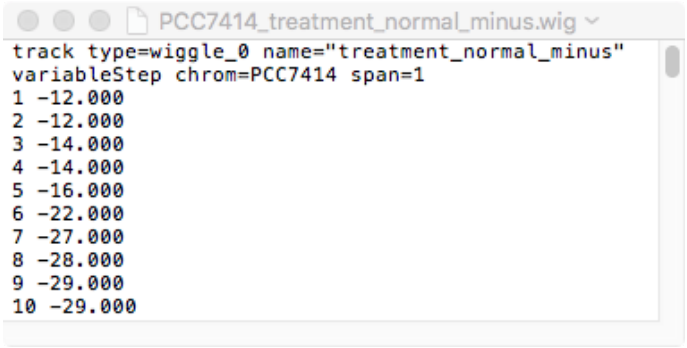


Abb. 7: Beispiel des `wiggle` (.wig) Dateiformats.

Die ersten zwei Zeilen bezeichnen die Probe (hier: TEX(-) auf dem (-)-Strang der Saccharosebedingung von *F. muscicola*). Spalte 1 ab Zeile 3 sind die Genompositionen. Spalte 2 ab Zeile 3 sind die Anzahl der gemappten Reads pro Genomposition. Das (-)-Zeichen soll auf das Mapping auf dem (-)-Strang hinweisen.

Das Abspeichern gemappter Readinformationen kann über das sogenannte `wiggle` (.wig) Dateiformat erzielt werden (Abb. 7). Für jede Genomposition wird die Anzahl der gemappten Readnukleotide eines DNA-Stranges gezählt. Die Informationen des DNA-Strangs für ein Mapping wurde aus den Alignments entnommen. Das `wiggle` Dateiformat kann im Integrated Genome Browser (IGB) (Nicol et al. 2009) als Mappinggraph dargestellt werden. Für jede TEX-Probe und jede Bedingung wurden jeweils zwei `wiggle` Dateien erzeugt. Insgesamt wurden 24 `wiggle` Dateien erstellt.

Die verschiedenen Entnahmeproben und deren Sequenzierung wiesen eine unterschiedliche Anzahl von Reads auf. Jede gemappte Position eines DNA-Stranges und einer TEX-Probe ($\mathrm{TEX_{Pos}}$) wurde daher speziesspezifisch mit einem Faktor normalisiert, der zu einer normalisierten gemappten Position ($\mathrm{TEX_{Norm}}$) führte:

Formel 1: Berechnung einer normalisierten Mappingposition.

$$\mathrm{TEX_{Norm}} = \frac{\mathrm{TEX_{Pos}}}{\left(\frac{\mathrm{TEX}}{\mathrm{Min_{Tex}}}\right)}$$

Wobei TEX für die Mappings der zu normalisierenden TEX-Probe steht und $\mathrm{Min_{Tex}}$ die kleineste TEX-Probe einer Spezies mit Mappings bezeichnet.

5.4 Die Detektion eines TSS

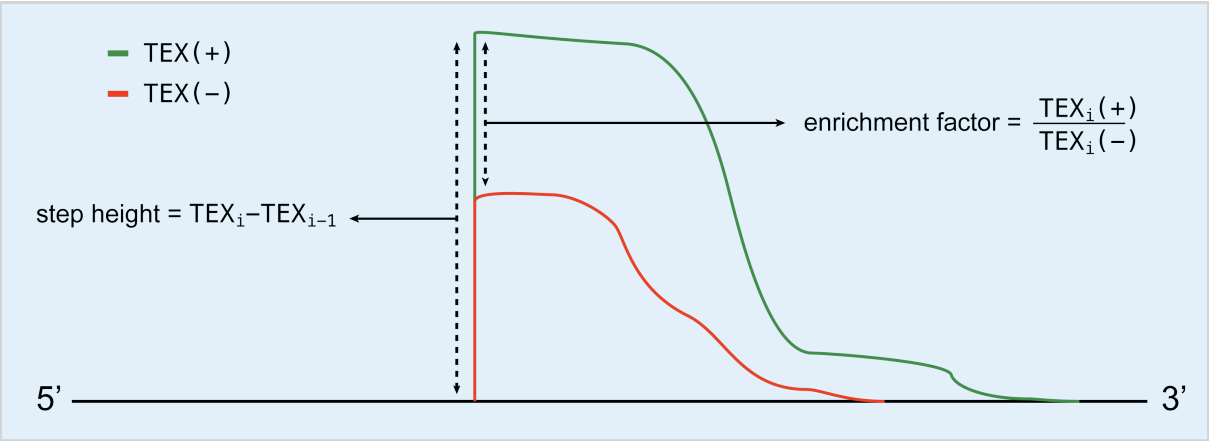


Abb. 8: Modell eines Peaks[12] im Mappinggraphen (modifiziert aus (Dugar et al. 2013)).

Auf einer Region des (+)-Strangs haben mehrere Reads aus der TEX(+)-Probe (grün) und der TEX(-)-Probe (rot) gemappt. Die Schritthöhe (`step height`) wird für jede TEX-Probe separat ermittelt und errechnet sich aus der Differenz der Reads an der betrachteten Position (i) und der vorhergegangenen Position (i-1). Der Anreicherungsfaktor (`enrichment factor`) ist der Quotientenwert aus der Schritthöhe der TEX_i(+)-Position als Dividend und der Schritthöhe der TEX_i(-)-Position als Divisor.

Abb. 8 stellt schematisch die Mappinggraphen zweier TEX-Proben dar. Innerhalb einer Region auf einem Genom wurden Reads aus der TEX(-)- und der TEX(+)-Probe einer Wachstumsbedingung gemappt und normalisiert. In beiden Proben sind die 5’-Enden von Primärtranskripten enthalten. Der Anteil prozessierter Transkripte ist in der TEX(+)-Probe weitaus geringer. Ein Anreicherungsfaktor kann ermittelt werden, um innerhalb eines Peaks[12] eine potentielle TSS zu bestimmen. Dabei spielt die Anzahl der gemappten und normalisierten Reads pro Position in der TEX(-)- und in der TEX(+)-Probe eine große Rolle. Die Anzahl der gemappten und normalisierten Reads pro Position ist die Ausgangsbasis der Berechnung von Schritthöhe und Anreicherungsfaktor. Die Schritthöhe einer Position (i) errechnet sich aus der Differenz der Reads an der normalisierten Position (i) und der vorhergegangenen Position (i-1). Die berechneten Schritthöhen beider TEX-Proben an der Position (i) werden für den Anreicherungsfaktor benutzt. Dieser errechnet sich aus der Schritthöhe der TEX_i(+)-Position als Dividend und aus der TEX_i(-)-Position als Divisor. Für jede Position (i) wird der Anreicherungsfaktor ermittelt. Ein Schwellenwert für den Anreicherungsfaktor bestimmt über die Speicherung des errechneten Wertes an Position (i). Der Schwellenwert entscheidet über Sensitivität bzw. Spezifität der TSS-Detektion. In dieser Analyse lag der Schwellenwert bei

[12] zu deutsch = Gipfel, Spitze, Scheitelwert.

größer oder gleich 2. Generell gilt ein Wert des Anreicherungsfaktors unterhalb von 1 als Anreicherung der TEX(-)-Probe. Es handelt sich dann um eine prozessierte Position.

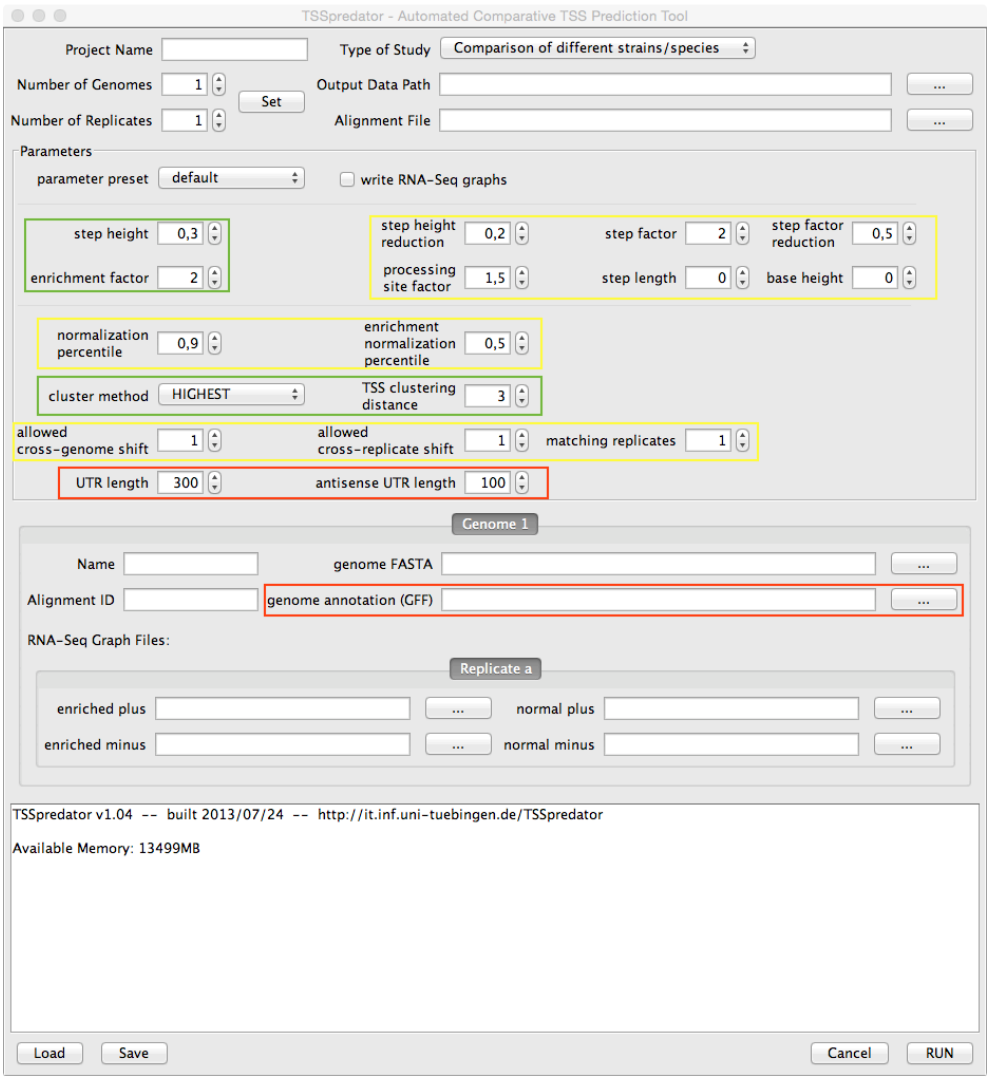


Abb. 9: Aufruf des `TSSpredator` (v. 1.04).

Für die Detektion von Transkriptionsstartpunkten erfordert das Programm Parametereinstellungen. Die wichtigsten Parameter sind grün markiert. Die gelben Kästchen stellen Parameter dar für die die Standardeinstellungen beibehalten worden sind. Die roten Kästchen stellen Parameter dar, die bei der TSS-Detektion nicht angegeben worden sind. Das Programm kann im Bedingungsvergleich oder im Speziesvergleich gestartet werden.

Die Methode der automatisierten Detektion von Transkriptionsstartpunkten ist jünger als die dRNAseq Methode. Früher wurden die Mappinggraphen manuell inspiziert, um TSSe zu bestimmen (Sharma et al. 2010). Dieses Vorgehen kann fehlerbehaftet sein (Sharma und Vogel 2014). Mittlerweile sind automatisierte Lösungen der TSS-Detektion vorhanden. Beispiele hierfür sind die Programme `TSSpredator` (Dugar et al. 2013), `TSSer` (Jorjani und Zavolan 2014) und `TSSAR` (Amman et al. 2014). Das Programm `TSSpredator` wurde in dieser Analyse für die TSS-Detektion benutzt. Die oben beschriebenen Variablen

(Schritthöhe und Anreicherungsfaktor) werden im Programm `step height` und `enrichment factor` genannt (Abb. 9, grün).

Es soll auf zwei weitere Parameter eingegangen werden, die im `TSSpredator` `TSS clustering distance` und `cluster method` genannt werden (Abb. 9, grün). Rauschen innerhalb einer Sequenzierung zeigt sich durch unscharfe Readmappingkurven in den Mappinggraphen. Dadurch werden zu viele Positionen als angereichert detektiert. Eine automatisierte Detektion muss in der Lage sein, ein solches Rauschen mit einer Routine zu erkennen und zu interpretieren. Für eine Analyse von Transkriptionsstartpunkten zwischen zwei Bedingungen gibt es zwei Ebenen, in der sich dieses Rauschen manifestieren kann. Erstens, durch unscharfe und/oder verschobene Peaks innerhalb einer Wachstumsbedingung. Die Reads der TEX(-)- oder der TEX(+)-Probe können, kontinuierlich oder verschoben, abrupt gemappt sein. Multiple Positionen in einem Peak erfüllen dann den Schwellenwert des Anreicherungsfaktors. Dies erschwert jedoch eine exakte Positionsdetektion. Zweitens, eine detektierte Position in der ersten Bedingung kann in der zweiten Bedingung verschoben detektiert vorliegen. Solche Positionen müssen von alternativen Transkriptionsstartpunkten unterschiedlicher Bedingungen unterschieden werden. Um in einer Genomregion mit einer Weite von zwei Nukleotiden nicht einen unterschiedlichen TSS für je eine Bedingung zu erhalten, werden die detektierten Positionen in allen TEX-Proben und Bedingungen gruppiert. Dies übernimmt eine Gruppierungsroutine. Im `TSSpredator` wird diese Routine `TSS clustering distance` genannt (Abb. 9, grün). Dieser Parameter definiert ein Fenster, in dem solche Ereignisse behandelt werden. Das Genom mit den gemappten normalisierten Reads wird mit diesem Fenster abgesucht und multiple angereicherte Positionen innerhalb des Fensters auf eine Position zusammengefasst. Der `TSSpredator` wurde in dieser Analyse im Bedingungsvergleich benutzt. Daher wurden innerhalb einer Bedingung und anschließend zwischen den Bedingungen die Positionen zusammengefasst. An welcher Position die potentiellen TSSe zusammengefasst werden, entscheidet die Methode der Gruppierung, die im Programm unter `cluster method` ausgewählt werden kann (Abb. 9, grün). Es gibt zwei Methoden der Gruppierung. Erstens, die Gruppierung an der ersten detektierten Position eines Fensters und zweitens, die Gruppierung an der detektierten Position mit höchster Schritthöhe eines Fensters. Diese Methoden werden `first` und `highest` im `TSSpredator` genannt. In dieser Arbeit wurde die Methode `highest` gewählt.

Zusätzlich zum `TSS clustering distance` Parameter wurde ein eigener Gruppierungsalgorithmus mit einer Fensterweite von 35 Nukleotiden nach der Detektion implementiert. Die Gruppierung wurde für jede Bedingung und jeden DNA-Strang separat durchgeführt. Die vom `TSSpredator` ermittelten Positionen und normalisierten `step heights` (Abb. 10) wurden in Gruppen eingeteilt. Die Entfernungen der ermittelten Positionen zueinander entschieden über die Gruppengrößen. Von einer detektierten Position

ausgehend, wurde 35 Nukleotide stromabwärts nach einer weiteren detektierten Position gesucht. War eine weitere Position vorhanden, wurde diese Position der Gruppe zugeordnet. Sobald eine Position außerhalb eines Fensters gefunden wurde, definierte diese Position eine neue Gruppe. Für jede Gruppe wurde die Position mit der höchsten `step height` bestimmt. Anschließend wurden die jeweils gleichen Gruppen beider Bedingungen miteinander verglichen. Die Position mit der höchsten `step height` zwischen den Bedingungen definierte dann die Position des TSS. Dieser Position wurden die ermittelten höchsten `step heights` beider Bedingungen und entsprechender Gruppen zugeordnet.

TSSpredator_output.txt

Pos	Strand	Condition	det	enr	stepHeight	enrichmentFactor
209	-	control	1	1	13.79	2.35
209	-	treatment	1	1	10.25	8.5
521	-	control	1	1	25.14	3.8
521	-	treatment	1	0	26.23	1.97
539	+	control	1	1	4.06	2.37
539	+	treatment	1	1	14.34	4.37
605	-	control	1	1	9.73	2.1
605	-	treatment	1	0	13.11	0.73
1124	+	control	1	1	802.2	4.84
1124	+	treatment	1	1	2126.23	3.64
1402	+	control	1	1	6.89	>100
1402	+	treatment	1	1	23.77	2.86
1802	+	control	0	0	NA	NA
1802	+	treatment	1	1	25.0	2.03
3011	+	control	1	0	8.52	1.52
3011	+	treatment	1	1	19.67	2.77
3065	-	control	1	1	17.44	2.73
3065	-	treatment	1	0	17.62	1.79
3825	-	control	1	0	4.46	1.42
3825	-	treatment	1	1	12.71	5.17

Abb. 10: Ergebnis einer TSS-Detektion mit dem `TSSpredator` (vereinfacht).

Spalten von links nach rechts: Pos = Genomposition eines detektierten TSS; Strand = DNA-Strang, auf dem der detektierte TSS liegt; Condition = Die Wachstumsbedingung, in welcher der TSS gefunden wurde; det = Logischer Wert für TSS-Detektion an entsprechender Genomposition (1 = ja, 0 = nein); enr = Logischer Wert für Anreicherung der detektierten Position bei einem Schwellenwert von ≥ 2 (1 = ja, 0 = nein); stepHeight = Schritthöhe an entsprechender Genomposition nach Anwenden der Gruppierung; enrichmentFactor = Anreicherungsfaktor.

Für die Analyse wurde das Programm mit den Standardeinstellungen gestartet. Lediglich der `clustering distance` Parameter wurde auf 30 gesetzt und nach erfolgreicher Detektion die zusätzliche Gruppierungsroutine implementiert. Das Programm wurde im Bedingungsvergleich benutzt. Genomalignments wurden nicht erstellt. Der Vergleich zwischen den Spezies wurde mit einer selbst entwickelten Vorgehensweise durchgeführt. Für *F. muscicola*, *F. thermalis* und *C. fritschii* existierten keine standardisierten Annotationstabellen (Abb. 9, rot). Daher wurden die detektierten TSSe mit einer eigenen Vorgehensweise klassifiziert.

5.5 Die Klassifizierung eines TSS

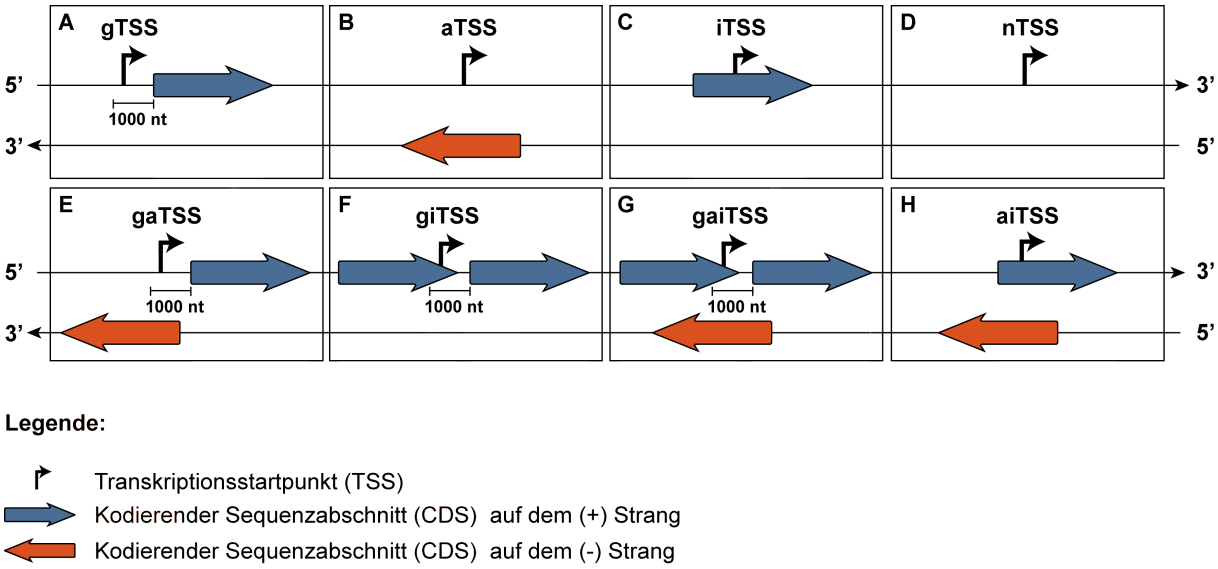


Abb. 11: Schema zur Definition der TSS-Klassen.

Basierend auf der Lokalisierung relativ zu einem CDS wurden die TSSe in folgende Klassen eingeteilt. **A:** Genische TSSe (gTSSe) resultieren nach Transkriptionsinitiation und Elongation in putative mRNA-Transkripte. Der Schwellenwert für die Klassifizierung eines gTSS wurde technisch bestimmt (siehe Text). **B:** Zu einem CDS komplementäre TSSe (aTSSe). Solche Transkripte können regulatorische Funktionen aufweisen. Diese können in *cis* und/oder in *trans* agieren. **C:** Interne TSSe (iTSSe) können zu verkürzten Transkripten führen, wenn alternative Startkodonen stromabwärts vorliegen. Zusätzlich können solche TSSe ein Indikator für falsch annotierte ORFs sein oder für sRNAs stehen. **D:** TSSe unbekannter Funktion (nTSSe). TSSe solcher Transkripte können zu sRNAs oder lncRNAs führen. Sie können auch für kleine, nicht detektierte CDSe kodieren. **E-H**: Intermediäre TSS-Klassen. Diese Klassen können multiple Funktionen aufweisen. Zusätzlich könnten sie auf Suboperonstrukturen hinweisen.

Nach der Detektion wurden die Transkriptionsstartpunkte den TSS-Klassen zugeordnet. Für die Klassifizierung wurden Start- und Stopkoordinaten der kodierenden Sequenzabschnitte (CDSe) verwendet (Abb. 11). Durch die Detektion waren die Genompositionen der TSSe bekannt und konnten mit den CDS-Koordinaten verknüpft werden. Alle TSSe innerhalb eines CDS wurden als interne TSSe (Abb. 11C) oder als TSSe komplementär zu einem CDS klassifiziert (Abb. 11B). Wenn sich CDSe auf den beiden DNA-Strängen überlappen, wurden die TSSe multiplen Klassen zugeordnet (Abb. 11H). Für die gTSS-Klasse wurde ein UTR-Schwellenwert bestimmt, da die Längen der UTRs in den drei Cyanobakterien unbekannt waren. Intergenische Längen zweier direkt benachbarter CDSe, die auf dem gleichen Strang kodiert vorliegen, wurden für die Schwellenwertberechnung verwendet. In *F. muscicola* beträgt eine intergenische Länge im Durchschnitt 1.168 Nukleotide (Median: 251 nt). Bei *F. thermalis* beträgt eine intergenische Länge im Durchschnitt 1.243 Nukleotide (Median: 290 nt). In

C. fritschii existieren intergenische Längen mit einer durchschnittlichen Länge von 1.162 Nukleotiden (Median: 245 nt). Mittelwert und Median unterschieden sich stark voneinander. Bei einem Schwellenwert ähnlich des Median wären entfernte, aber putativ alternative gTSSe als TSSe unbekannter Funktion (Abb. 11D) klassifiziert worden. Ein Ziel der Analyse ist die Detektion alternativer gTSSe für CDSe. Daher sollte ein größtmöglicher Anteil von Transkriptionsstartpunkten in den CDS-Vergleich fließen. Entsprechend diente der Mittelwert als Orientierung für einen Schwellenwert. Dieser wurde auf 1.000 Nukleotide eingestellt und führt zu vier weiteren TSS-Klassen (Abb. 11A, E, F, G).

Die Verwendung von CDS-Koordinaten und UTR-Schwellenwert führt bei der TSS-Klassifizierung zu drei eindeutigen (puren) TSS-Klassen (Abb. 11A-C). Durch die Lokalisierung der CDSe auf dem (+)-Strang und dem (-)-Strang, sind für TSSe auch multiple Zuordnungen möglich (Abb. 11E-H). Im Extremfall kann ein TSS drei TSS-Klassen zugeordnet werden (Abb. 11G). Nach der TSS-Klassifizierung verbleiben nicht klassifizierte TSSe in der nTSS-Klasse (Abb. 11D). Für die nTSS-Klasse wurden keine Analysen auf Interspeziesebene vorgenommen, da die Assoziation mit einem kodierenden Genabschnitt ein essentieller Orientierungspunkt in der Analyse darstellt.

5.6 Die Erstellung der Proteinfamilien

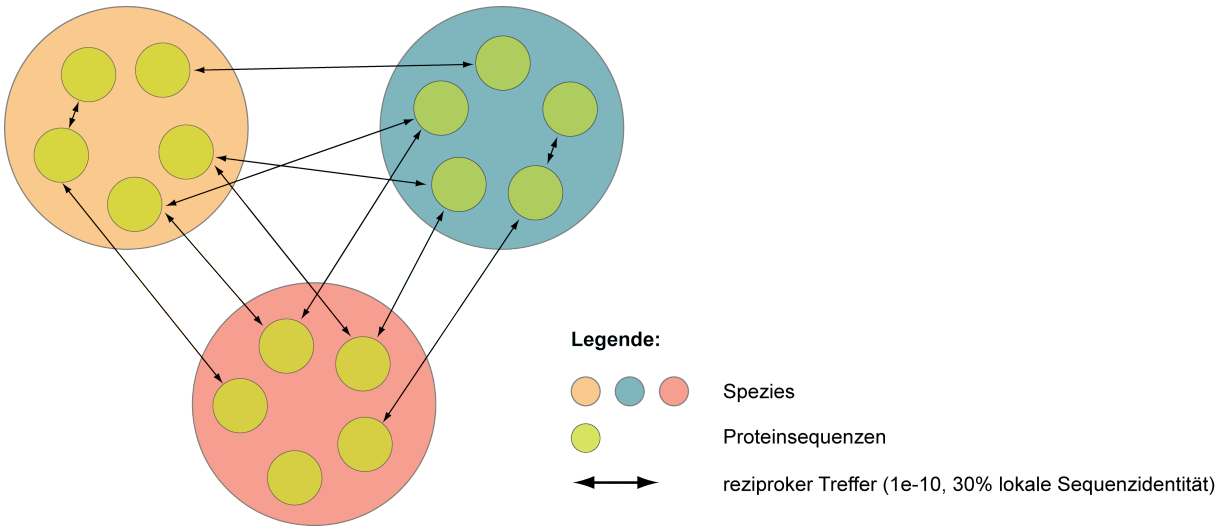


Abb. 12: Bidirektionaler BLAST Graph am Beispiel von drei Spezies.

Für die Erstellung der Proteinfamilien wurde eine bidirektionale Suche (Tatusov et al. 1997) mit allen Proteinsequenzen der 141 cyanobakteriellen Spezies durchgeführt. Dabei wurde nach signifikanten Treffern für jedes Protein aus jedem cyanobakteriellen Proteom in allen Proteomen sequenzierter Cyanobakterien gesucht (Abb. 12). Die Suche wurde mit dem Programm `blastall` (v.2.2.17) mit der Option `blastp` (Altschul et al. 1997) durchgeführt. Nur bidirektionale Paare wurden gespeichert (Abb. 12). Die Proteinsequenzen mussten sich gegenseitig mit mindestens 30% lokaler Sequenzidentität und mit mindestens einem `expectation value (E)` von 1×10^{-10} finden. Der Datensatz wies insgesamt 652.598 Proteinsequenzen auf. Von diesen Proteinsequenzen erzeugten 589.692 Sequenzen Knotenpaare, die mit 126.137.698 signifikanten Kanten verbunden waren. Davon erfüllten 587.089 Knotenpaare mit 45.492.676 Kanten das bidirektionale Kriterium. Die Knotenpaare mit den entsprechenden Kanten wurden global aligniert und nur Knotenpaare deren Alignment mindestens 30% globale Sequenzidentität aufwiesen wurden gespeichert. Für 533.724 Knotenpaare, die mit 28.785.928 Kanten verbunden waren, traf dieses Kriterium zu. Diese Knotenpaare und die globalen Identitätswerte wurden für das Programm MCL (Enright et al. 2002) benutzt. Das Programm erzeugte 33.701 Cluster. Jede Spezies hatte nach der Clustererzeugung eine individuelle Proteinfamiliensignatur (Tab. 11). Die vergleichende TSS-Analyse wurde mit Einzelkopie-Proteinfamilien zwischen den Spezies *F. muscicola* und *F. thermalis* sowie zwischen *F. muscicola* und *C. fritschii* durchgeführt.

5.7 Vergleichende TSS-Analyse

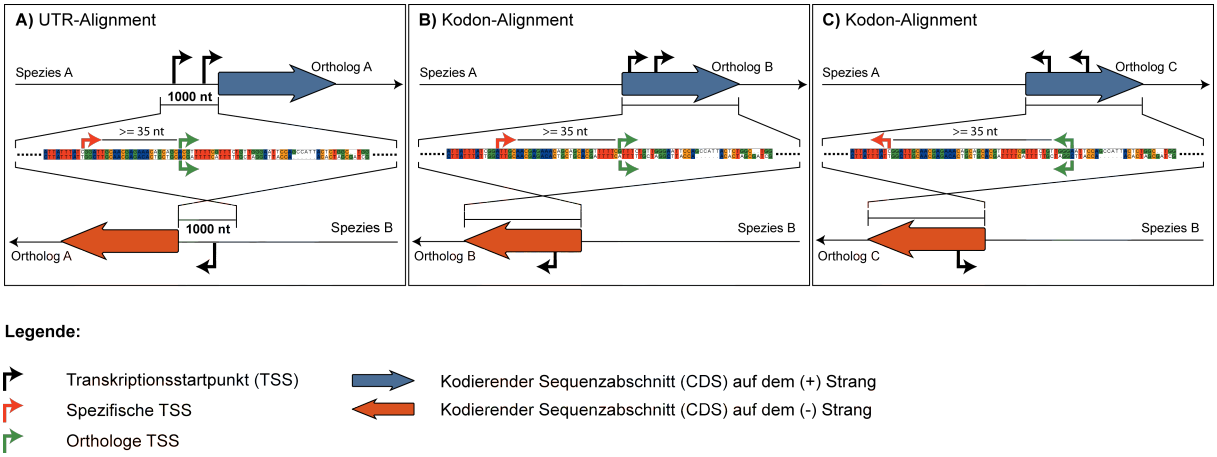


Abb. 13: Die Definition orthologer und speziesspezifischer TSSe.

Die transkriptionellen Gemeinsamkeiten zweier Spezies wurden mit der Erstellung von globalen Alignments detektiert (Abb. 13). Es wurden zwei Arten von Alignments erstellt. Erstens, ausgehend vom UTR-Schwellenwert (Abschnitt 5.5) wurden von den identifizierten Einzelkopie-Proteinsequenzpaaren zwischen *F. muscicola* und *F. thermalis* und zwischen *F. muscicola* und *C. fritschii* die Sequenzbereiche stromaufwärts der CDSe miteinander aligniert (Abb. 13A). Zweitens, die globalen paarweisen Alignments der Einzelkopie-Proteinsequenzpaare wurden in CDS-Alignments transformiert. Die Transformationen der Alignments erfolgte über die entsprechenden CDSe (Abb. 13B, C). Nach der Erstellung der Alignments erfolgte die Lokalisierung der Transkriptionsstartpunkte. Die TSS-Klasse entschied über die Art des Alignments, in der der TSS einer Spezies lokalisiert wurde. Ein TSS mit multiplen TSS-Klassen wurde entsprechend in beiden Arten von Alignments lokalisiert, wenn beide CDSe in Einzelkopie-Proteinfamilien zwischen *F. muscicola* und *F. thermalis* bzw. zwischen *F. muscicola* und *C. fritschii* vorlagen. Wurden alle TSSe in den entsprechenden Alignments lokalisiert, ist die Orthologie eines TSS-Paares bestimmt worden. Ein TSS-Paar wurde als ortholog definiert, wenn sich die TSSe nicht weiter als 35 Alignmentpositionen voneinander entfernt befanden (Abb. 13). Traf dies nicht zu, wurde ein TSS als speziesspezifisch definiert (Abb. 13). Die Wahl der Distanz von 35 Nukleotiden wurde analog zum Gruppierungsalgorithmus nach der TSS-Detektion gewählt (Abschnitt 5.4).

In der Analyse der TSS-Regionen (Abschnitt 6.4.3) wurde von jedem orthologen und speziesspezifischen TSS die Sequenzidentität in einem maximal 70 Positionen weiten Fenster bestimmt. Befand sich ein TSS weniger als 35 Alignmentpositionen vom Ende eines Alignments entfernt, konnte das Fenster entsprechend kleiner sein. Ein orthologes Genpaar

konnte mehrere orthologe TSSe und speziesspezifische TSSe aufweisen. Daher wurde der Vergleich der TSS-Regionen mit den Proteinsequenzidentitäten in zwei Varianten durchgeführt. Erstens, jede TSS-Regionidentität wurde einzeln mit der Identität des entsprechenden Einzelkopie-Proteinsequenzpaares verglichen. Zweitens, von allen orthologen und speziesspezifischen TSS-Regionidentitäten eines CDS-Paares wurde der Mittelwert bestimmt. Dieser Mittelwert wurde mit der Identität des Proteinsequenzpaares verglichen. Die entsprechenden Varianten werden in dieser Arbeit als “pro TSS” und “pro CDS-Paar” bezeichnet (Abschnitt 6.4.3, Tab. 8 u. Tab. 9, Spalte Analysebereich).

Die Identifizierung der Kandidatengene mit möglichen Einfluss auf die beobachtete morphologische Transition wurde über einen Transkriptabundanzvergleich erzielt. Hierfür wurde der Logarithmus der Schritthöhe zur Basis zwei für orthologe TSSe ($\mathrm{oTSS_e}$) berechnet:

Formel 2: Berechnung der Veränderung der Transkriptabundanz zwischen den Bedingungen.

$$\mathrm{oTSS_e} = log_2\left(\frac{\mathrm{oTSS_{treat}}}{\mathrm{oTSS_{norm}}}\right)$$

Wobei $\mathrm{oTSS_{treat}}$ der Schritthöhe eines orthologen TSS aus der Saccharose- bzw. der NaCl-Bedingung entspricht und $\mathrm{oTSS_{norm}}$ die Schritthöhe eines orthologen TSS aus der Kontrollbedingung bezeichnet. Das Logarithmieren wird vorgenommen, um extreme Änderungen der Transkriptabundanz miteinander vergleichen zu können. Die Änderung der Abundanz ist negativ, wenn der orthologe TSS in der Kontrollbedingung eine höhere Anzahl an Transkripten aufweist und ist positiv, wenn der TSS in der Saccharosebedingung bzw. in der NaCl-Bedingung eine höhere Anzahl an Transkripten aufweist. Da jeweils zwei Spezies miteinander verglichen werden, liegen für einen orthologen TSS jeweils zwei Änderungswerte der Transkriptabundanz vor. Zwischen *F. muscicola* und *F. thermalis* und zwischen *F. muscicola* und *C. fritschii* wurden diese Werte in einem Streudiagram dargestellt (Abschnitt 6.4.4, Abb. 28) und der Pearson Korrelationskoeffizient berechnet. Für jede TSS-Klasse wurden 5% der orthologen TSSe markiert. Diese zeigten die höchste gemeinsame Änderung der Transkriptabundanz im Speziesvergleich. Die der markierten orthologen TSSe zugeordneten Einzelkopie-Proteinfamilien stellen potentielle Kandidaten der morphologischen Transition dar.

6 Ergebnisse

6.1 Die Readanalyse

Die Sequenzierungen der Kontrollbedingungen und der Saccharosebedingungen bzw. der NaCl-Bedingung wurden von der Firma Vertis Biotechnologie AG[13] durchgeführt und sind in Abschnitt 10.2 näher beschrieben. Insgesamt wurden zwölf Proben sequenziert, von denen jeweils vier Proben immer einer Spezies zugeordnet sind. Pro Spezies wurden zwei Bedingungen zu je zwei Proben, eine TEX(-)-Probe und eine TEX(+)-Probe, sequenziert. Die erhaltenen Reads hatten eine uniforme Länge von 50 nt und wiesen über die komplette Länge eines Reads gute Qualitätswerte (Ewing et al. 1998) auf. Daher wurde keine Qualitätssteigerung vorgenommen.

Abb**. 14** stellt einen Überblick über alle Nukleotide (nt) der Reads aus den TEX(-)-Proben dar. Die TEX(-)-Reads aus der Kontrollbedingung (links) sind jeweils der Saccharosebedingung (rechts) im Falle von *F. muscicola* und *F. thermalis* bzw. der NaCl-Bedingung von *C. fritschii* gegenübergestellt. In den TEX(-)-Proben der Kontrollbedingungen sind von Position eins bis neun Adenin/Guanin Anteile von 30-40% zu erkennen. Ab Position zehn beginnt jedoch ein starker Anstieg des Anteils von Adenin mit Maxima an Position 28, welche in einem Bereich von 60-65% liegen. Diese Anteile sind nach Position 28 kontinuierlich bis unter 40% abfallend. Die Anteile von Adenin und Guanin in den Reads der Saccharosebedingungen bzw. der NaCl-Bedingung variieren stärker im Vergleich zu den Kontrollbedingungen. Von Position eins bis neun sind Anteile von Guanin zu erkennen, die zwischen 40-50% liegen (Abb. 14). Auch in den Reads dieser Bedingungen sind starke Anstiege des Anteils von Adenin ab Position zehn zu erkennen. Die Maxima liegen bei Position 28. Die Anteile sind bis zum Ende der Reads nur schwach abnehmend.

Diese Betrachtungsweise der Reads gab nur unzureichende Informationen über ein potentielles Artefakt. Eine Polymeranalyse wurde durchgeführt, um Informationen auf Homopolymere in den Sequenzierungen zu erhalten. Abb**. 15** zeigt die Polymerlängenverteilungen der Reads aus den TEX(-)-Proben der Kontrollbedingungen (links), sowie aus den Saccharosebedingungen bzw. der NaCl-Bedingung (rechts) für *F. muscicola*, *F. thermalis* und *C. fritschii*.

[13] http://www.vertis-biotech.com

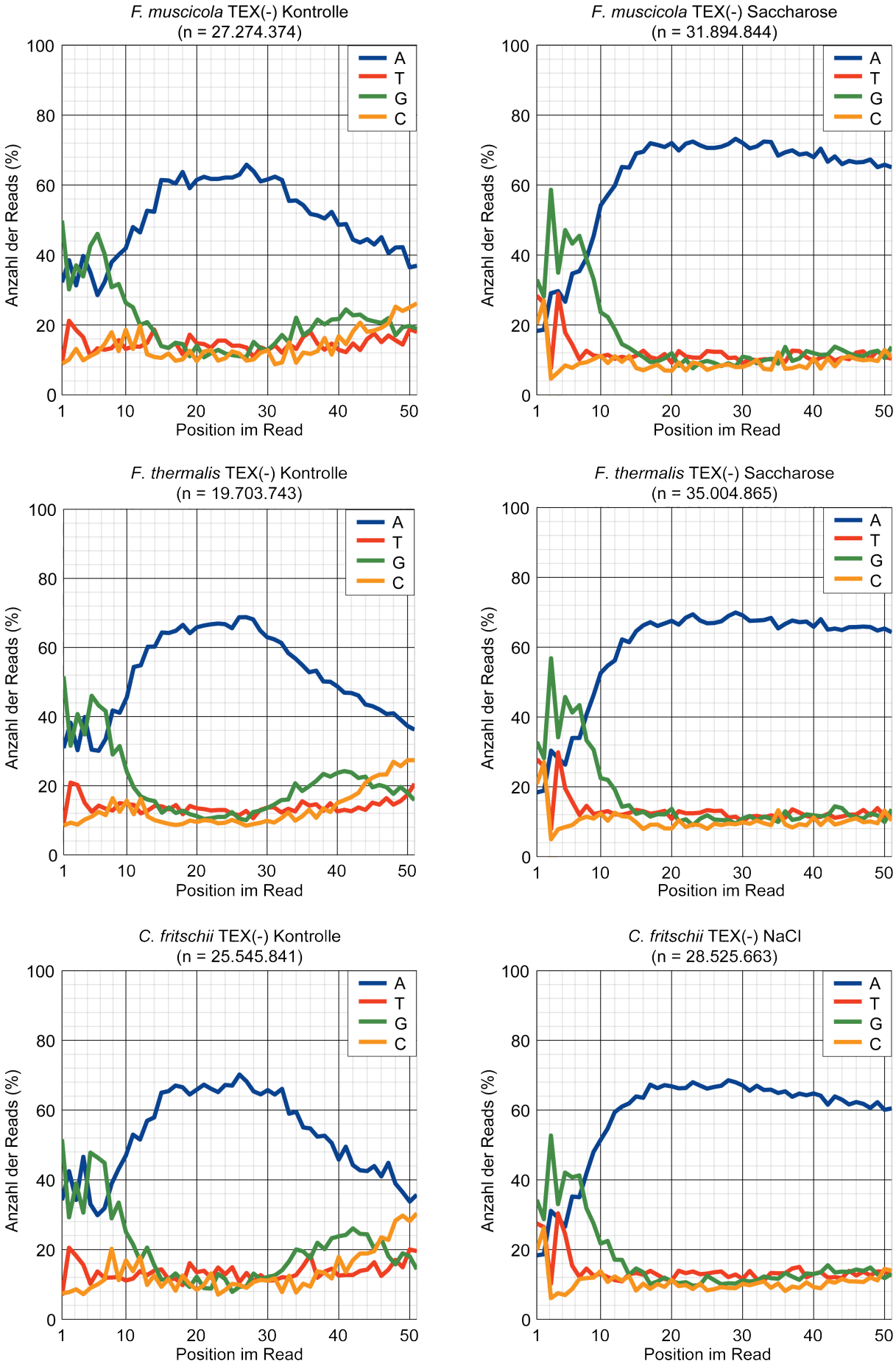


Abb. 14: Nukleotidanteile pro Readposition in den TEX(-)-Reads.

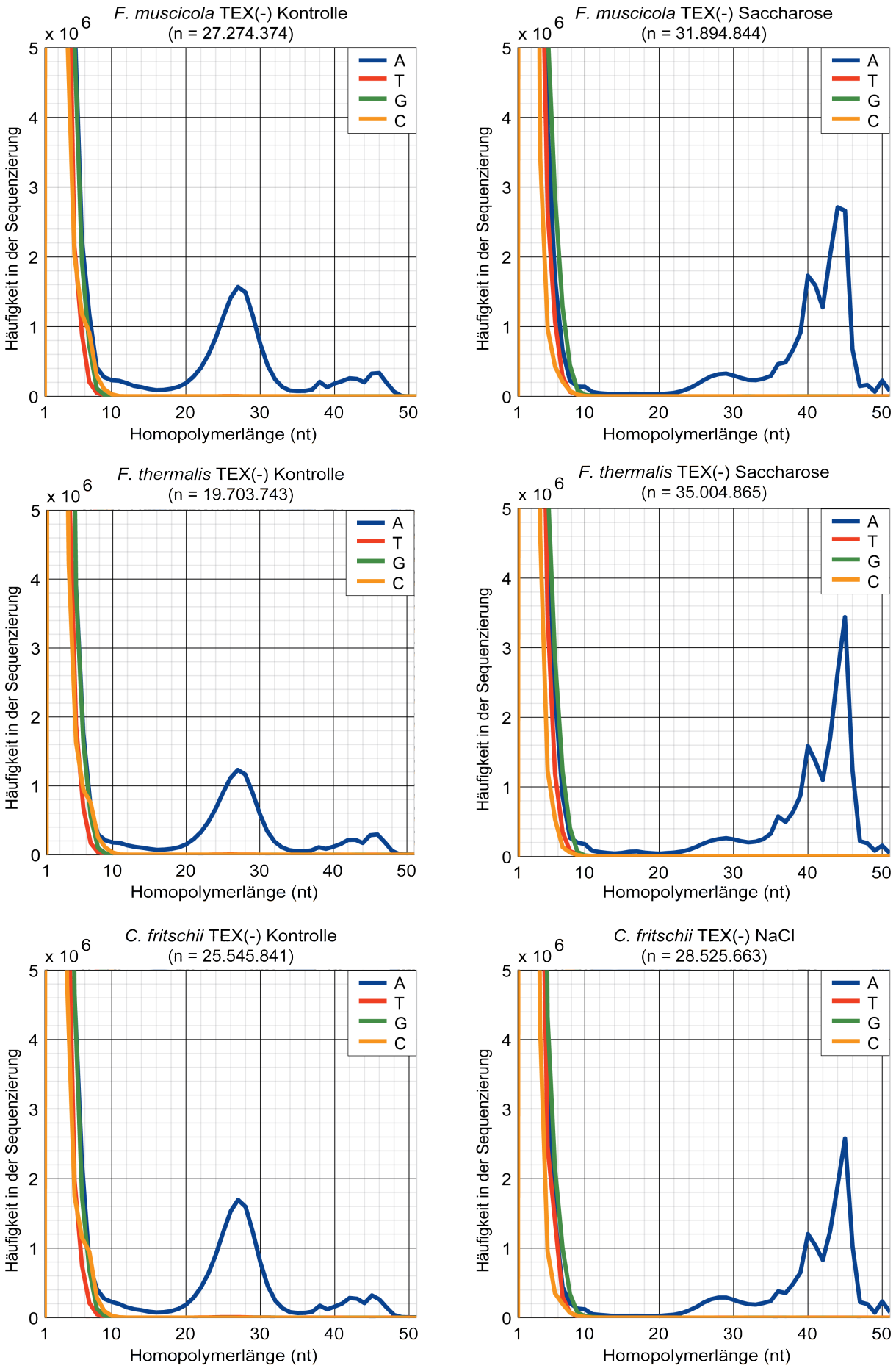


Abb. 15: Homopolymerhäufigkeiten in der TEX(-)-Sequenzierung.

Deutlich zu erkennen ist ein hoher Anteil von Adenin-Homopolymeren mit Längen von 20 nt bis 35 nt. Die TEX(-)-Reads der Saccharosebedingungen bzw. der NaCl-Bedingung weisen diese nicht auf. Hier hat die Mehrheit aller Polymere des Adenin Längen von 35 nt bis 45 nt (Abb. 15). Die Maxima liegen speziesübergreifend bei ca. 45 nt. Ab einer Polymerlänge von 25 nt ist die Hälfte einer Readlänge erreicht. Es kann davon ausgegangen werden, dass die Häufigkeit eines Polymers dieser Länge oder größeren Längen mit der Anzahl der Reads deckungsgleich ist.

Abb**. 16** und Abb**. 17** zeigen die Ergebnisse derselben analytischen Vorgehensweise für die TEX(+)-Proben. Auch hier werden die Kontrollbedingungen der Saccharosebedingungen bzw. der NaCl-Bedingung gegenübergestellt. Über die Wachstumsbedingungen und den Spezies hinweg sind keine auffälligen Anteile des Adenins erkennbar (Abb. 16). Die Anteile des Adenins sind pro Position höher als die Anteile der anderen Nukleotide. Im Vergleich zu den TEX(-)-Reads sind diese Anteile stark unterschiedlich (vgl. Abb. 14 u. Abb. 16). Auch für die TEX(+)-Reads wurde eine Homopolymeranalyse durchgeführt.

Die Polymerlängenverteilung in Abb**. 17** zeigt in allen TEX(+)-Proben Homopolymere des Adenin. Im Vergleich zu den TEX(-)-Reads sind die Häufigkeiten der Homopolymere mindestens um den Faktor 10 geringer (vgl. Abb. 15 u. Abb. 17). Die Häufigkeiten von Homopolymerlängen zwischen 20 nt und 35 nt fallen deutlich geringer aus. In der Kontrollbedingung von *C. fritschii* sind diese Längen nicht erkennbar. Mit Ausnahme von *F. muscicola* befinden sich die Maxima im Bereich der Readlänge. In *F. muscicola* liegt das Maximum bei einer Länge von 36 nt.

Für die TEX(+)-Proben wurde nur ein kleines Signal von Homopolymeren beobachtet. In den TEX(-)-Proben wurde ein eindeutiges Artefakt von Adenin-Polymeren detektiert. Bei der Berechnung des `expectation value (E)` des Programms `blastall` ist die Readlänge ein Berechnungselement. Viele Reads könnten daher einen schlechten E-Wert für Alignments aufweisen. Um ein Mapping auf die Genome zu gewährleisten, wurden die Polymere von den Reads entfernt. Für die TEX(-)- und TEX(+)-Reads wurden zwei Filterparamter definiert. Für alle Reads wurde ein Schwellenwert für Adenin-Homopolymere und für Thymin-Homopolymere als Kontrolle mit der Länge 10 nt gesetzt. Sobald mindestens eines dieser Polymere innerhalb eines Reads vorkam, wurde der Read ab Beginn des ersten Polymers bis zum Ende des Reads abgetrennt. Die Gesamtlänge eines Reads durfte anschließend die Länge von 20 nt nicht unterschreiten.

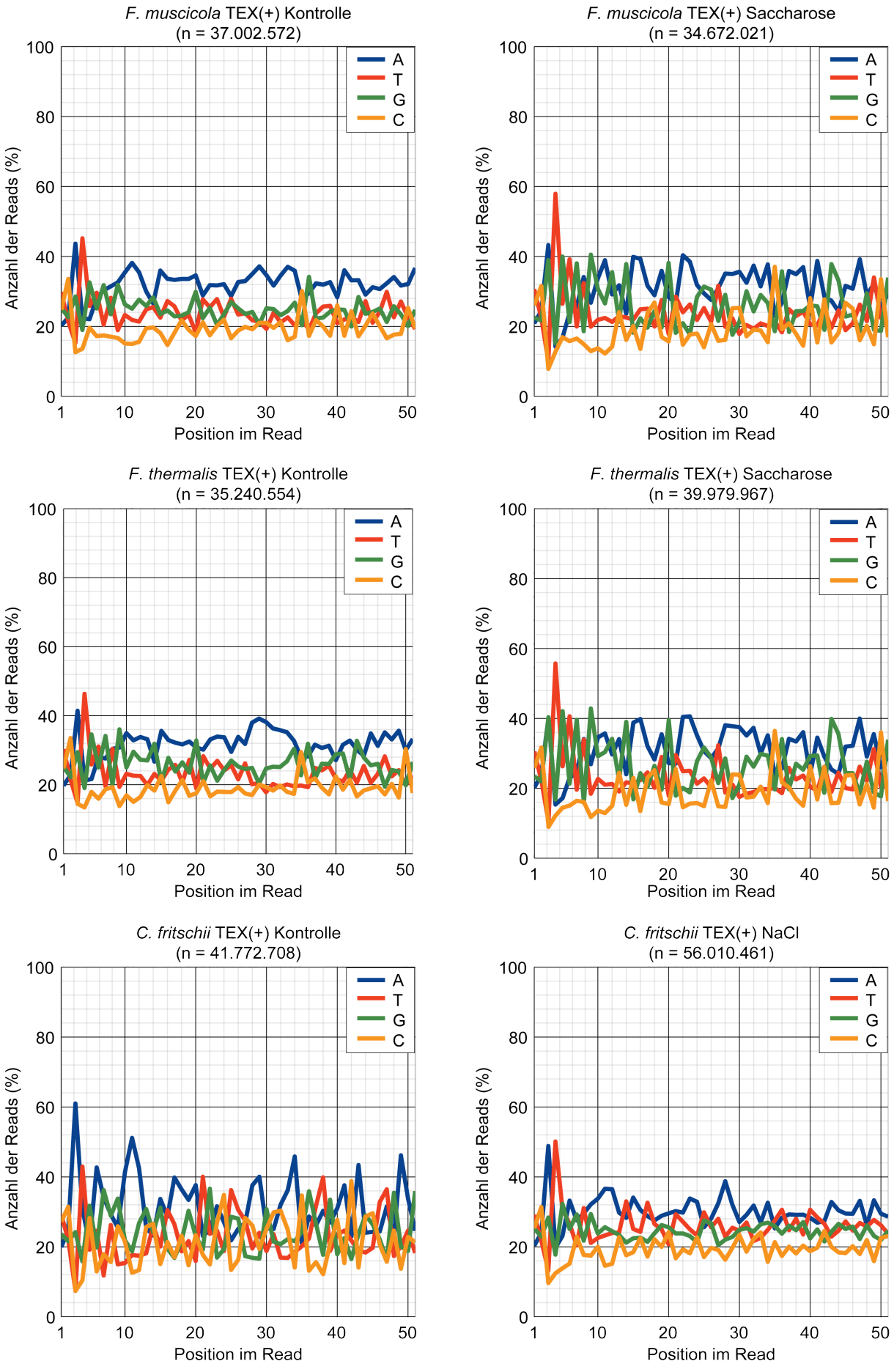


Abb. 16: Nukleotidanteile pro Readposition in den TEX(+) Reads.

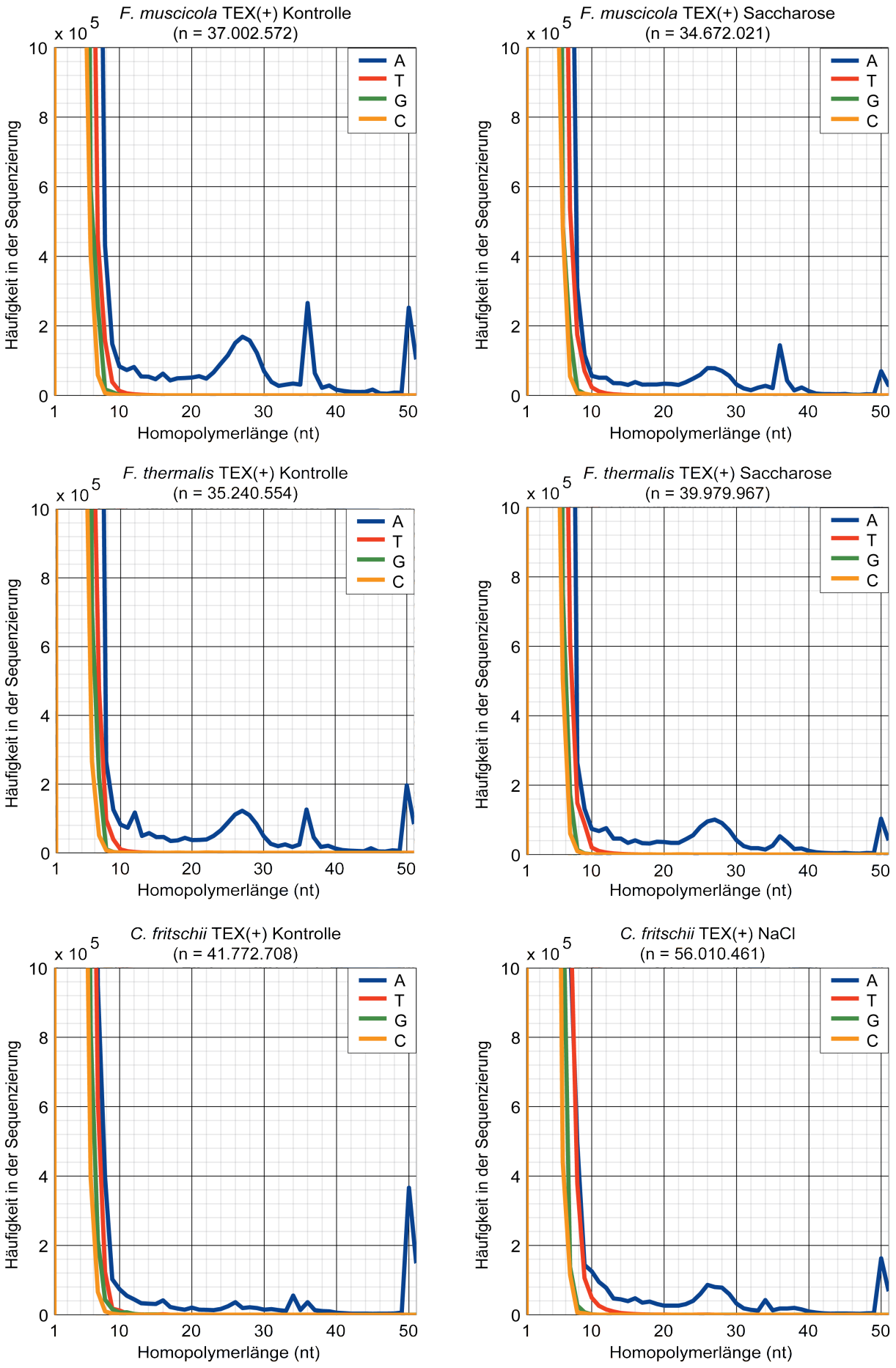


Abb. 17: Homopolymerhäufigkeiten in der TEX(+) Sequenzierung.

6.2 Das Read mapping

Tab. 1: Mappingverlauf in *F. muscicola*, *F. thermalis* und *C. fritschii*.

Für jede Spezies ist der Verlust von Reads bei schrittweiser Anwendung der Mappingkriterien dokumentiert. Jede Zeile einer Tabelle bezeichnet den Schritt der Anwendung. Nach Anwenden eines Schrittes wird die Anzahl der in der Analyse verbleibenden Reads dokumentiert. Die Ausgangsanzahl der Reads und die Anzahl der gemappten Reads ist in grün hervorgehoben. Der relative Readverlust in jedem Schritt ist im Bereich „**Kumulativer Readverlust**" dargestellt. Die Schritte mit besonders hohem Readverlust sind in rot hervorgehoben. Nach dem Mapping wurde der Normalisierungsfaktor berechnet und ist in **schwarz** hervorgehoben.

***F. muscicola* PCC 7414**

	Verbleibende Reads				Kumulativer Readverlust (%)			
	Kontrolle		Saccharose		Kontrolle		Saccharose	
	TEX(+)	TEX(-)	TEX(+)	TEX(-)	TEX(+)	TEX(-)	TEX(+)	TEX(-)
Sequenzierte Reads	**37.002.572**	**27.274.374**	**34.672.021**	**31.894.844**	0	0	0	0
Nach Poly-A Abtrennung	35.310.945	14.491.163	33.901.430	13.054.811	**5**	**47**	**2**	**59**
Nach Poly-T Abtrennung	35.309.659	14.486.971	33.900.573	13.051.944	5	47	2	59
BLAST Treffer	34.278.834	13.526.837	32.456.833	12.109.068	7	50	6	62
Gemappt	**23.235.040**	**11.766.369**	**26.904.765**	**11.005.486**	**37**	**57**	**22**	**65**
Normalisierungsfaktor	**2,11**	**1,07**	**2,44**	**1,00**				

***F. thermalis* PCC 7521**

	Verbleibende Reads				Kumulativer Readverlust (%)			
	Kontrolle		Saccharose		Kontrolle		Saccharose	
	TEX(+)	TEX(-)	TEX(+)	TEX(-)	TEX(+)	TEX(-)	TEX(+)	TEX(-)
Sequenzierte Reads	**35.240.554**	**19.703.743**	**39.979.967**	**35.004.865**	0	0	0	0
Nach Poly-A Abtrennung	34.136.007	9.423.281	39.175.119	16.243.222	**3**	**52**	**2**	**54**
Nach Poly-T Abtrennung	34.134.954	9.417.046	39.174.181	16.240.405	3	52	2	54
BLAST Treffer	34.018.464	9.096.945	39.018.821	15.899.801	3	54	2	55
Gemappt	**22.929.996**	**8.580.024**	**32.588.644**	**15.392.917**	**35**	**56**	**18**	**56**
Normalisierungsfaktor	**2,67**	**1,00**	**3,80**	**1,79**				

***C. fritschii* PCC 6912**

	Verbleibende Reads				Kumulativer Readverlust (%)			
	Kontrolle		NaCl		Kontrolle		NaCl	
	TEX(+)	TEX(-)	TEX(+)	TEX(-)	TEX(+)	TEX(-)	TEX(+)	TEX(-)
Sequenzierte Reads	**41.772.708**	**25.545.841**	**56.010.461**	**28.525.663**	0	0	0	0
Nach Poly-A Abtrennung	40.929.302	11.902.386	55.269.988	134.133.10	**2**	**53**	**1**	**53**
Nach Poly-T Abtrennung	40.928.282	11.893.915	55.268.130	134.109.97	2	53	1	53
BLAST Treffer	40.196.602	11.012.755	54.425.209	12.740.949	4	57	3	55
Gemappt	**33.892.486**	**10.370.776**	**42.279.566**	**11.995.068**	**19**	**59**	**25**	**58**
Normalisierungsfaktor	**3,27**	**1,00**	**4,08**	**1,16**				

Zu Beginn waren im Durchschnitt ca. 34 Million Reads pro Probe und Spezies in der Analyse vorhanden (Tab. 1). Nach Anwenden der ersten Filterprozedur wurden in allen TEX(-)-Proben im Durchschnitt 53% der Reads verworfen (Tab. 1). In diesen Proben betrug das Minimum und das Maximum des relativen Verlusts von Reads bei *F. muscicola* in der Kontrollbedingung 47% und in der Saccharosebedingung 59%. Für die TEX(+)-Proben wurden keine hohen Verluste beobachtet. Diese wiesen im Schnitt einen Verlust von 2,5% auf. Das Minimum war

in der NaCl-Bedingung von *C. fritschii* mit 1% zu beobachten. Das Maximum war in der Kontrollbedingung von *F. muscicola* mit 5% zu beobachten. Die zweite Filterprozedur befasste sich mit Thymin-Homopolymeren. Da die Abb**. 15** und Abb**. 17** keine Homopolymere des Thymin auf der benutzten Y-Skala zeigten, war hier in allen Proben kein hoher Verlust zu erwarten. Geringe Verluste von Reads, welche sich in den relativen Anteilen nicht manifestierten, waren zu beobachten. Die nach Anwenden der ersten beiden Filterprozeduren verbleibenden Reads wurden mit dem Programm `blastall` mit den jeweiligen Genomen aligniert. Unabhängig von der Probe waren überall leichte Verluste von Reads zu beobachten. Im Durchschnitt haben 2% der Reads aller Proben nach der zweiten Filterprozedur keine signifikanten Alignments mit den jeweiligen Genomen erzeugt (Tab. 1). Nach Anwenden der letzten Filterprozedur wurde ein Read als gemappt definiert (Abschnitt 5.3). Einem Read wurde eine eindeutige Position auf dem Genom zugeordnet. Im Vergleich zu den TEX(-)-Proben wurden starke Verluste von Reads in den TEX(+)-Proben beobachtet. In den TEX(+)-Proben gingen im letzten Schritt im Durchschnitt 22% der Reads verloren. In den TEX(-)-Proben waren es im Durchschnitt nur 3%. Das Maximum der Verluste in den TEX(+)-Proben wurde in *F. thermalis* mit 32% beobachtet. Das Minimum der Verluste war in der TEX(+)-Probe der Kontrollbedingung von *C. fritschii* mit 19% zu beobachten.

Zusammenfassend wiesen die TEX(-)-Proben einen relativen Verlustbereich zwischen 56%-65% auf. Die TEX(+)-Proben verloren im Verlauf des Mappings zwischen 18%-37% der Reads (Tab. 1). Die Proben jeder Spezies wurden speziesspezifisch normalisiert und in `wiggle` Dateiformaten abgespeichert (Abschnitt 5.3). Anschließend erfolgte die Detektion der TSSe mithilfe des `TSSpredator` (Abschnitt 5.4) sowie die daran anknüpfende TSS-Klassifizierung (Abschnitt 5.5).

6.3 Die primären Transkriptome

6.3.1 Die TSS-Verteilung auf die TSS-Klassen

Der `TSSpredator` detektierte für *F. muscicola* 15.351 TSSe, für *F. thermalis* 12.092 TSSe und für *C. fritschii* 19.818 TSSe. Nach Implementierung der eigenen Gruppierungsroutine (Abschnitt 5.4) wurden für *F. muscicola* 15.137 TSSe, für *F. thermalis* 11.968 TSSe und für *C. fritschii* 19.419 TSSe erhalten.

Basierend auf der Klassendefinition (Abschnitt 5.5) wurden den Transkriptionsstartpunkten TSS-Klassen zugeordnet. In allen Spezies war ein ähnliches Bild der TSS-Verteilung auf die Klassen zu beobachten (Abb. 18). Die gTSS-Klasse wies im Durchschnitt einen relativen Anteil von 22% auf (Abb. 18, blau). In der aTSS-Klasse war durchschnittlich mit 27% der höchste Anteil zu beobachten (Abb. 18, rot). Die iTSS-Klasse wies im Durchschnitt einen relativen Anteil von 16% auf (Abb. 18, grün) und die nTSS-Klasse war im Durchschnitt mit einem relativen Anteil von 9% vertreten (Abb. 18, grau). Ungefähr 26% der TSSe wurden multiplen TSS-Klassen zugeordnet. Die Anteile in den intermediären Klassen (giTSS, gaTSS, gaiTSS) waren in erster Linie auf den gewählten UTR-Schwellenwert eines CDS zurückzuführen (Abschnitt 5.5).

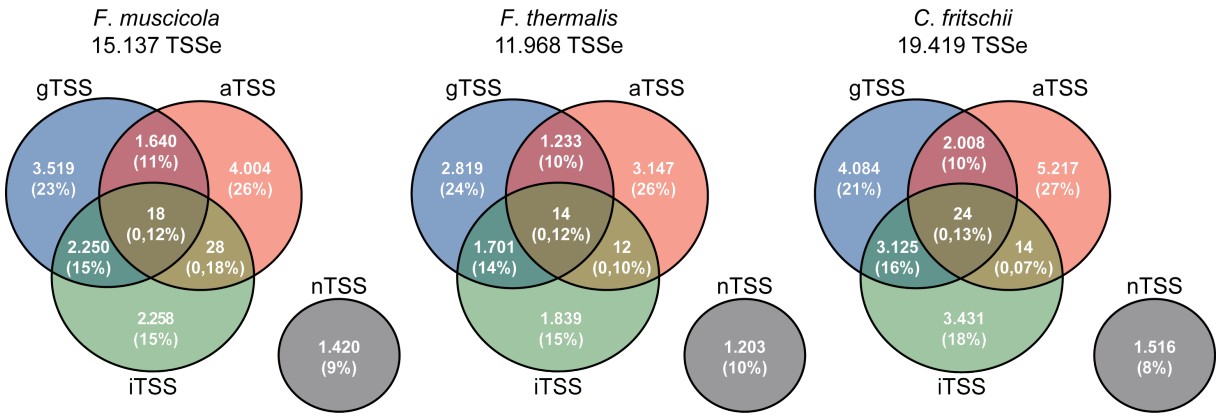


Abb. 18: TSS-Verteilung auf die TSS-Klassen.

Die Venn-Diagramme stellen die TSS-Zuordnungen für *F. muscicola* (links), *F. thermalis* (mitte) und *C. fritschii* (rechts) in den jeweiligen TSS-Klassen dar. Die überlappenden Bereiche stellen TSSe dar, welche multiplen TSS-Klassen zugeordnet worden sind. Unterhalb jeder absoluten TSS-Anzahl ist der relative Anteil an der Gesamtverteilung innerhalb der entsprechenden Spezies dargestellt.

6.3.2 Eigenschaften der Klassen

6.3.2.1 UTR-Längen

Die Entfernungen der TSSe zum Startkodon eines assoziierten CDS wurden analysiert. Als Orientierungspunkt diente das erste Nukleotid des Startkodon. Für Positionen intermediärer Klassen waren mehrere Orientierungspunkte möglich, da solche TSSe mehreren CDS zugeordnet sind. Die intermediären Klassen gaiTSS und aiTSS wurden in dieser Analyse nicht berücksichtigt, da deren relativer Anteil an der Gesamtverteilung weniger als 1% beträgt (Abb. 18). In Abb**. 19** ist die Entfernungsverteilung der gTSSe, gaTSSe und giTSSe für *F. muscicola* (links), *F. thermalis* (mitte) und *C. fritschii* (rechts) dargestellt. Ungefähr 24% aller gTSSe wurden in einem Entfernungsbereich zwischen 25 nt und 75 nt zu einem CDS gefunden (Abb. 19, blau). Im Vergleich zu Transkriptionsstartpunkten intermediärer Klassen sind gTSSe signifikant näher zum assoziierten CDS (Abb. 19, Kolmogorov-Smirnov Test). Die gaTSSe und giTSSe zeigen keine eindeutigen Maxima. Die giTSSe liegen signifikant näher zu ihrem assoziierten CDS als gaTSSe (Abb. 19, Kolmogorov-Smirnov-Test).

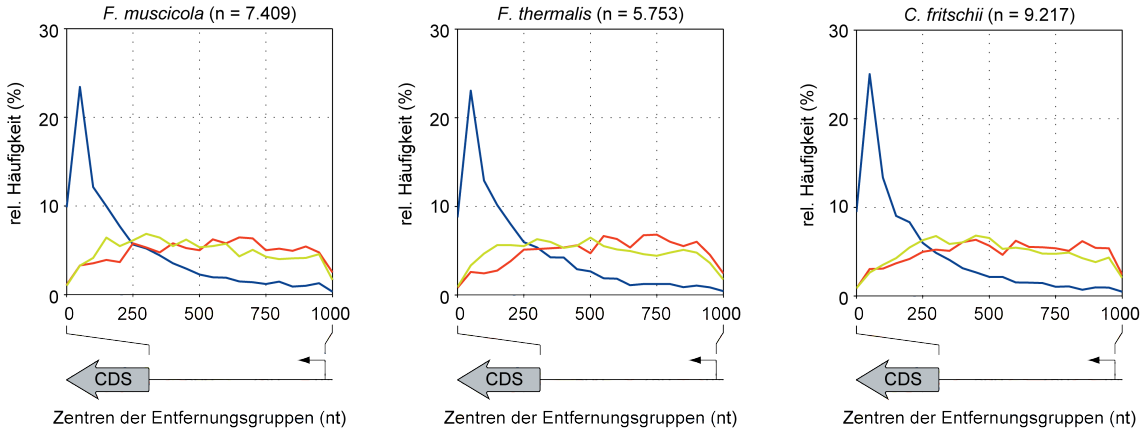


Abb. 19: TSS zu CDS Entfernung für gTSSe, gaTSSe und giTSSe.

Es handelt sich bei den Diagrammen um Histogramme in Liniendarstellung für *F. muscicola* (links), *F. thermalis* (mitte) und *C. fritschii* (rechts). Bei den relativen Häufigkeiten auf der Y-Achse ist die TSS-Anzahl in Entfernungsgruppen, dividiert mit der Gesamtanzahl der TSSe innerhalb einer TSS-Klasse und multipliziert mit 100 dargestellt. Die UTR von 1.000 nt eines CDS wurde in 21 nicht überlappende Entfernungsgruppen aufgeteilt. Ein Wert auf der X-Achse markiert das Zentrum einer Entfernungsgruppe. Eine Entfernungsgruppe ist definiert als ein nicht überlappender Entfernungsbereich von +/- 25 nt.

6.3.2.2 TSS-Positionen in kodierenden Sequenzabschnitten

In Abb. **20** ist die Entfernungsverteilung der iTSSe und giTSSe für *F. muscicola* (links), *F. thermalis* (mitte) und *C. fritschii* (rechts) dargestellt. Die Entfernungsgruppen sind relative Entfernungsbereiche zum ersten Nukleotid des Startkodon eines assoziierten CDS. Die relativen Entfernungen der TSSe ($\mathrm{TSS_{rel}}$) wurden wie folgt berechnet:

Formel 3: Berechnung relativer Entfernungen in einem CDS.

$$\mathrm{TSS_{rel}} = \left(\frac{|(\mathrm{ATG_{abs}} - \mathrm{TSS_{abs}})|}{\mathrm{ORF}}\right) * 100$$

Wobei $\mathrm{ATG_{abs}}$ der genomischen Koordinate des ersten Nukleotids des Startkodon eines CDS, $\mathrm{TSS_{abs}}$ der genomischen Koordinate eines TSS und ORF der Länge des assoziierten CDS entspricht. TSSe der iTSS-Klasse liegen signifikant häufiger am 5'-Ende eines CDS als giTSSe (Abb. 20, Kolmogorov-Smirnov-Test). Die Maxima für iTSSe variieren zwischen den Spezies, liegen aber alle in den ersten 25% eines CDS (Abb. 20, blau). Die Maxima der giTSSe zeigen eine Lokalisierung in den letzten 25% eines CDS (Abb. 20, grün).

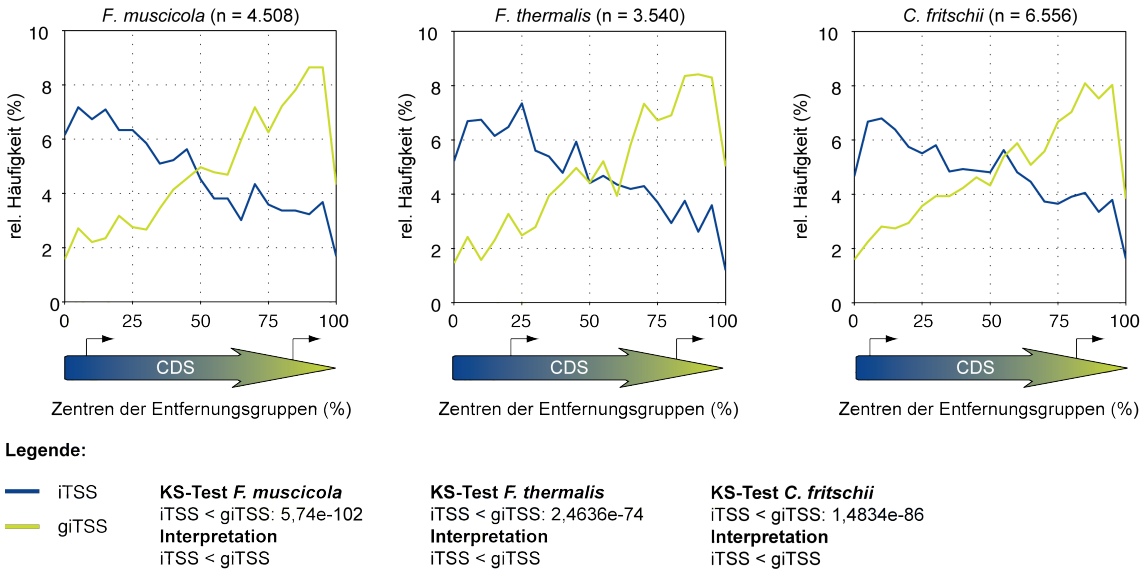
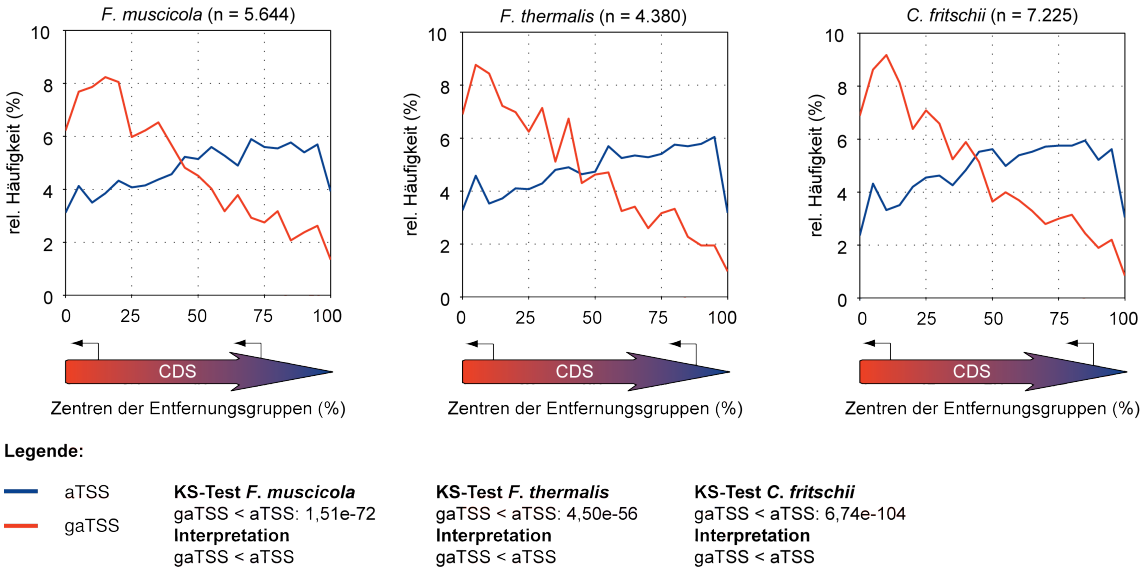


Abb. 20: TSS-Verteilung in kodierenden Sequenzabschnitten (CDS) für iTSSe und giTSSe.

Es handelt sich bei den Diagrammen um Histogramme in Liniendarstellung für *F. muscicola* (links), *F. thermalis* (mitte) und *C. fritschii* (rechts). Bei den relativen Häufigkeiten auf der Y-Achse ist die TSS-Anzahl in Entfernungsgruppen, dividiert durch die Gesamtanzahl der TSSe einer TSS-Klasse und multipliziert mit 100 dargestellt. Die relative Entfernung zum ersten Nukleotid des Startkodon eines assoziierten CDS wurde in 21 nicht überlappende Entfernungsgruppen aufgeteilt. Ein Wert auf der X-Achse markiert das Zentrum einer Entfernungsgruppe. Eine Entfernungsgruppe ist definiert als ein nicht überlappender relativer Entfernungsbereich von +/- 2,5%.

In Abb**. 21** ist die Entfernungsverteilung der aTSSe und gaTSSe für *F. muscicola* (links), *F. thermalis* (mitte) und *C. fritschii* (rechts) dargestellt. Die Entfernungsgruppen sind relative Entfernungsbereiche zum ersten Nukleotid des Startkodon eines assoziierten komplementären CDS. Die relativen Entfernungen der TSSe (TSS_{rel}) wurden wie für die internen TSSe berechnet (siehe Formel 3). Die aTSSe tendieren dazu vermehrt am 3'-Ende eines CDS lokalisiert zu sein (Abb. 21, blau). Die Tendenz zum 3'-Ende ist jedoch nicht so stark wie für giTSSe (vgl. Abb. 20 u. Abb. 21). Mit Ausnahme von *F. muscicola* liegen die Maxima der aTSSe in den letzten 25% eines komplementären CDS. Für *F. muscicola* liegt das Maximum im relativen Entfernungsbereich von 67,5% bis 72,5% (Abb. 21, Entfernungsgruppe 70). Die gaTSSe sind signifikant häufiger am 5'-Ende eines komplementären CDS als aTSSe (Abb. 21, Kolmogorov-Smirnov-Test). Für die gaTSSe liegen die Maxima für alle Spezies in den ersten 25% eines komplementären CDS.



Legende:

aTSS
gaTSS

KS-Test *F. muscicola*
gaTSS < aTSS: 1,51e-72
Interpretation
gaTSS < aTSS

KS-Test *F. thermalis*
gaTSS < aTSS: 4,50e-56
Interpretation
gaTSS < aTSS

KS-Test *C. fritschii*
gaTSS < aTSS: 6,74e-104
Interpretation
gaTSS < aTSS

Abb. 21: TSS-Verteilung im komplementären CDS für aTSSe und gaTSSe.

Es handelt sich bei den Diagrammen um Histogramme in Liniendarstellung für *F. muscicola* (links), *F. thermalis* (mitte) und *C. fritschii* (rechts). Bei den relativen Häufigkeiten auf der Y-Achse ist die TSS-Anzahl in Entfernungsgruppen, dividiert durch die Gesamtanzahl der TSSe einer TSS-Klasse und multipliziert mit 100 dargestellt. Die relative Entfernung zum Nukleotid des Startkodon eines komplementären CDS wurde in 21 nicht überlappende Entfernungsgruppen aufgeteilt. Ein Wert auf der X-Achse markiert das Zentrum einer Entfernungsgruppe. Eine Entfernungsgruppe ist definiert als ein nicht überlappender relativer Entfernungsbereich von +/- 2,5%.

6.3.2.3 Intergenische Entfernungen

Bei den Längen der intergenischen Regionen wurde getestet, ob bei der Präsenz eines intermediären TSS, die assoziierten CDSe eine andere Entfernung zueinander aufweisen als CDSe, die keinen intermediären TSS aufweisen. CDSe auf demselben DNA-Strang und ohne intermediäre TSS wurden mit Entfernungen von kodierenden Sequenzabschnitten verglichen, die eine intermediäre TSS aufweisen. Abb. 22 stellt eine kumulative Verteilung der intergenischen Entfernungen dar. CDSe mit einem giTSS sind signifikant näher zueinander als CDSe mit einer gaTSS oder als CDSe ohne intermediäre TSSe (Abb. 22, Kolmogorov-Smirnov-Test). CDSe mit einem gaTSS weisen signifikant kleinere Entfernungen zueinander auf, als CDSe ohne eine intermediäre TSS-Zuweisung (Abb. 22, Kolmogorov-Smirnov-Test). Des Weiteren ist dies am 80. Perzentil zu erkennen (Abb. 22). Während es für CDSe mit giTSS-Assoziation für alle Spezies bei ca. 250 nt liegt, liegt es für CDSe mit gaTSS-Assoziation bei ca. 500 nt. Für CDSe ohne intermediäre TSSe ist es größer als 1.000 nt (Abb. 22). Das 80. Perzentil bedeutet, dass 80% der CDS zu CDS Entfernungen nicht weiter als den entsprechenden Wert auf der X-Achse voneinander entfernt liegen.

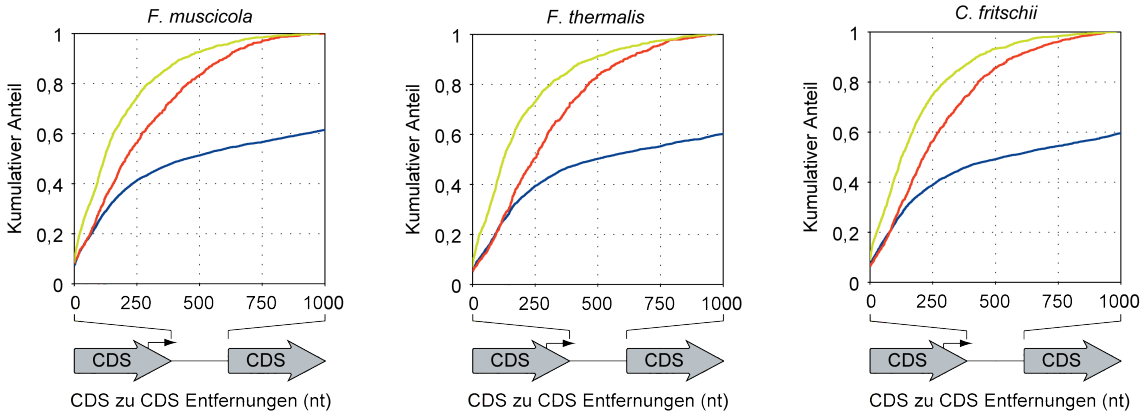


Abb. 22: Intergenische Entfernungen für CDSe mit und ohne giTSSe oder gaTSSe.

In den kumulativen Verteilungsdiagrammen sind die Entfernungen zweier benachbarter CDSe für *F. muscicola* (links), *F. thermalis* (mitte) und *C. fritschii* (rechts) dargestellt. Mit Ausnahme benachbarter CDSe, welche eine gaTSS aufweisen, sind nur Entfernungen benachbarter CDSe auf dem selben DNA-Strang dargestellt.

6.3.3 Differentielle TSS-Aktivität

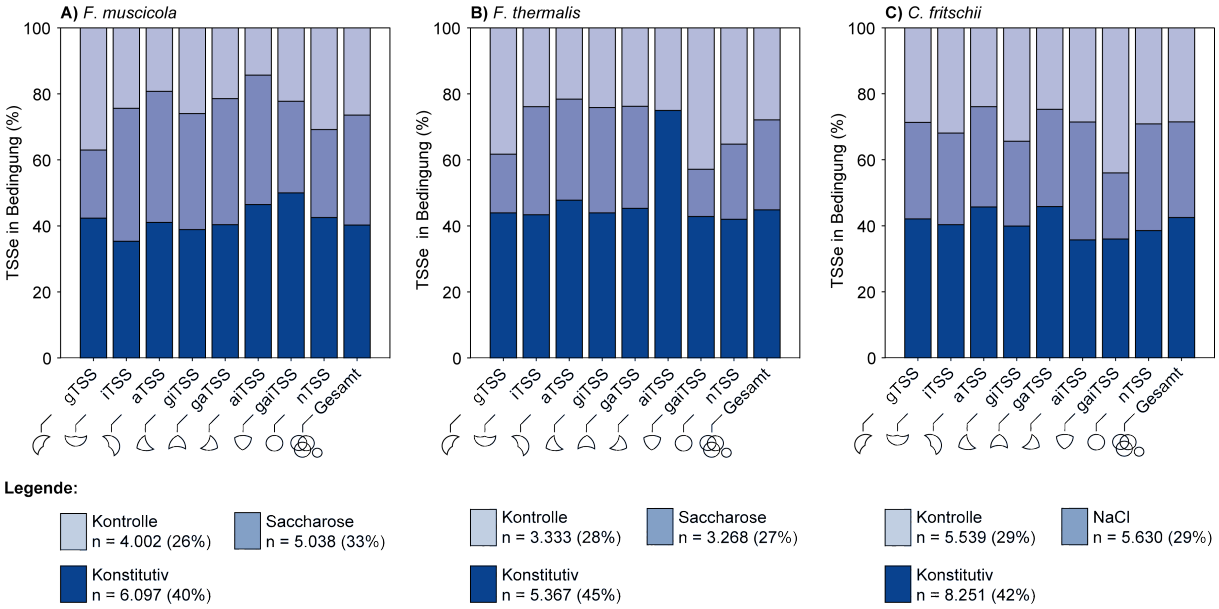


Abb. 23: TSS-Verteilung auf die Wachstumsbedingungen.

Die gestapelten Säulendiagramme stellen die differentielle TSS-Aktivität in den jeweiligen TSS-Klassen für die Spezies *F. muscicola* (**A**), *F. thermalis* (**B**) und *C. fritschii* (**C**) dar. In dunkelblau sind TSSe für beide Bedingungen in der TEX(+)-Probe angereichert gewesen. In blau sind TSSe nur für die Saccharose- bzw. die NaCl-Bedingungen in der TEX(+)-Probe als angereichert detektiert worden. In hellblau waren TSSe nur für die Kontrollbedingungen in der TEX(+)-Probe angereichert. Die jeweiligen Stichprobengrößen (n) entsprechen der Summe aller differentiell- bzw. konstitutiv aktiven TSSe einer Spezies.

TSSe, welche nur in einer der beiden Wachstumsbedingungen gefunden wurden, können Kandidaten einer differenziellen Transkript-Aktivität sein. Daher bat sich die Vorgehensweise einer Anreichungsfaktoranalyse zwischen den Bedingungen an. Abb**. 23** stellt die differentiell detektierten TSSe dar. Speziesübergreifend wurden ähnliche TSS-Verteilungen in den Wachstumsbedingungen von *F. muscicola* (Abb. 23A), *F. thermalis* (Abb. 23B) und *C. fritschii* (Abb. 23C) beobachtet. Konstitutive TSS-Aktivität war für 40%-45% der TSSe in allen drei Spezies vorhanden (Abb. 23, dunkelblau). Hierbei handelt es sich möglicherweise um grundlegende transkriptionelle Systeme, die unabhängig von der Wachstumsbedingung aktiv waren. Diese Anteile können feiner aufgelöst werden, wenn die Anzahl der Wachstumsbedingungen erhöht wird. Zwischen 27%-33% der TSSe waren ausschließlich im synchronisierten Morphotypen aktiv (Abb. 23, blau). Für *F. muscicola* und *F. thermalis* ist ein gemeinsamer konservierter transkriptioneller Einfluss der Saccharose möglich. Des Weiteren können in diesen Anteilen konservierte transkriptionelle Komponenten vorhanden sein, welche einen Einfluss auf die *Fischerella*-Morphologien haben. Beide Parameter in diesen Spezies (Konzentration der Saccharose bei RNA-Extraktion und beobachteter Morphotyp) waren identisch bzw. sehr ähnlich. In *C. fritschii* waren beide Parameter unterschiedlich. In den

Kontrollbedingungen waren zwischen 26%-29% der TSSe aktiv (Abb. 23, hellblau). Das Ausgangsmedium war bei Start des Experiments in allen drei Spezies identisch. TSS-Aktivitäten in der Kontrolle könnten daher auf die unterschiedlichen Morphologien und Wachstumsformen zurückzuführen sein. In der gTSS-Klasse wird ein Unterschied von *F. muscicola* und *F. thermalis* im Vergleich zu *C. fritschii* erkennbar. Während in beiden *Fischerella*-Spezies zwischen 37% und 38% der gTSSe in den Kontrollbedingungen aktiv waren (Abb. 23AB, gTSS), war bei *C. fritschii* ein geringerer Anteil von 29% aktiv (Abb. 23C, gTSS). In der NaCl-Bedingung war ein ähnlicher Anteil detektiert worden. Ob es sich bei *C. fritschii* um die gleichen Gene mit alternativen Transkriptionsstartpunkten handelt, wurde nicht untersucht. In der iTSS-Klasse wurden zwischen den *Fischerella*-Spezies ebenfalls mehr Ähnlichkeiten beobachtet. Während in beiden *Fischerella*-Spezies 24% der iTSSe in den Kontrollbedingungen aktiv waren (Abb. 23AB, iTSS), waren es in *C. fritschii* 32% (Abb. 23C, iTSS). Es wurden jedoch auch Unterschiede zwischen *F. muscicola* und *F. thermalis* beobachtet. Für aTSSe war ein Unterschied der konstitutiven TSS-Aktivität und der TSS-Aktivität in der Saccharose am höchsten. In *F. muscicola* waren 41% der aTSSe konstitutiv aktiv und 40% ausschließlich in der Saccharose (Abb. 23A, aTSS). Bei *F. thermalis* waren 48% aller aTSSe konstitutiv aktiv und 31% in der Saccharose (Abb. 23B, aTSS).

Es zeigten sich also bereits auf dieser Ebene Gemeinsamkeiten einer differentiellen TSS-Aktivität in den *Fischerella*-Spezies. Für eine komparative Transkriptomanalyse ohne Genomalignments fehlte jedoch ein gemeinsamer Vergleichspunkt auf Interspeziesebene. Deshalb wurde eine weiterführende Analyse auf dieser Ebene nicht vorgenommen. Zusätzlich erschwerte die hohe TSS-Anzahl und die multiplen Funktionen eines potentiellen Transkripts eine zielgerichtete Analyse auf dieser Ebene. Daher sind die Vergleichspunkte zwischen den Spezies sind in dieser Arbeit Einzelkopie-Proteinfamilien.

6.3.4 TSS-Verteilung in den kodierenden Sequenzen

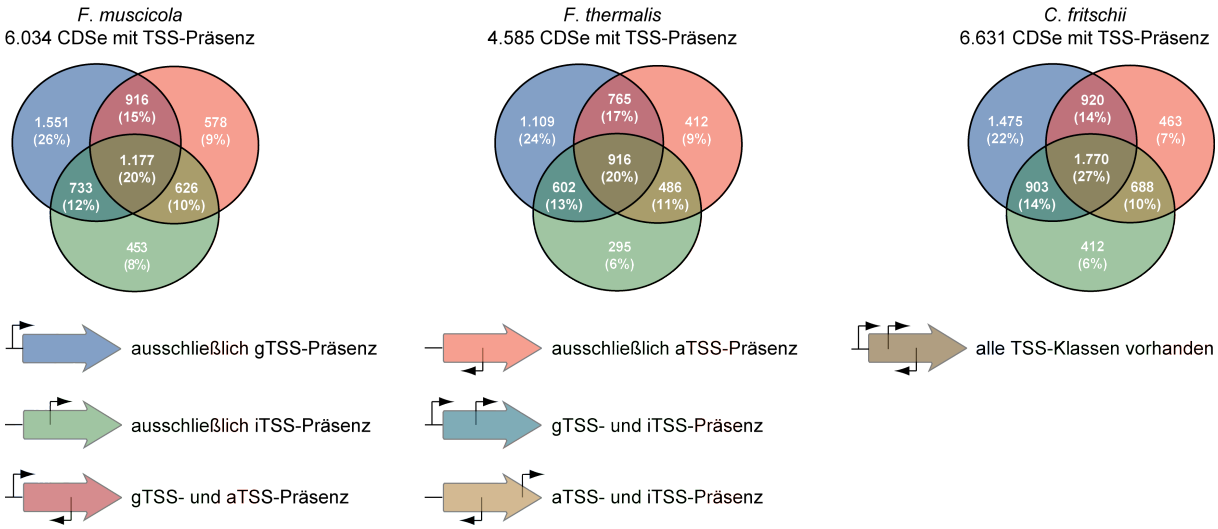


Abb. 24: TSS-Präsenz in den kodierenden Sequenzabschnitten (CDS).

Die dargestellten Venn-Diagramme illustrieren die Art der TSS-Präsenz in den kodierenden Sequenzabschnitten (CDSe) der Spezies *F. muscicola* (links), *F. thermalis* (mitte) und *C. fritschii* (rechts). Die überlappenden Bereiche der Venn-Diagramme stellen CDSe dar, die mehr als eine TSS-Klasse aufweisen. Solche CDSe weisen entsprechend immer mehr als einen TSS auf. Unter der absoluten CDS-Anzahl ist jeweils der relative CDS-Anteil an der Gesamtanzahl der CDSe mit TSS-Präsenz dargestellt.

Auf der Ebene der kodierenden Sequenzabschnitte (CDSe) konnte die Analyse für 84% aller CDSe in *F. muscicola* (Abb. 24, links), für 86% aller CDSe in *F. thermalis* (Abb. 24, mitte) und für 89% aller CDSe in *C. fritschii* (Abb. 24, rechts) TSSe ermitteln. Die TSS-Verteilungen pro Klasse und CDS sind in *F. muscicola*, *F. thermalis* und *C. fritschii* nicht signifikant unterschiedlich voneinander (χ^2: $p = 0.31$).

Im Durchschnitt wurden für 25% der CDSe in *F. muscicola* und *F. thermalis* ausschließlich gTSSe detektiert (Abb. 24). Für 20% der CDSe wurden TSSe aller Klassen gefunden. Dies bedeutet, dass solche CDSe mindestens drei TSSe aufgewiesen haben (Abb. 24). In ca. 16% der Fälle zeigten CDSe das Vorhandensein von mindestens einem gTSS und einem aTSS (Abb. 24). Für *C. fritschii* wurden für 27% der CDSe Transkriptionsstartpunkte aller Klassen gefunden (Abb. 24). Für 22% wurden ausschließlich gTSSe detektiert und 14% der CDSe zeigten mindestens einen gTSS und einen aTSS (Abb. 24). Weitere 14% zeigten mindestens einen gTSS und einen iTSS (Abb. 24). Diese Beobachtungen lassen auf multiple Genfunktionen bzw. auf einen hohen regulatorischen Anteil des Transkriptoms schließen, da im Speziesdurchschnitt 61% der CDSe mehr als eine TSS-Klasse aufwiesen.

Tab. 2: TSS-Anzahl pro CDS und TSS-Klasse.

Deskriptive Statistiken zur TSS-Anzahl pro CDS und Klasse für die Spezies *F. muscicola* (**A**), *F. thermalis* (**B**) und *C. fritschii* (**C**). CAI (Sharp und Li 1987) = Codon Adaption Index. Ein bioinformatisches Maß für die theoretische Expressionsstärke eines CDS. Je näher der Wert an eins ist, desto besser entsprechen die Kodonen des CDS dem Kodonrepertoire der Spezies.

A) *F. muscicola*							CDS-Länge		CAI	
Klasse	**TSSe**	**CDSe**	**Min TSS**	**Max TSS**	**Median**	**Mittelwert**	**Spearmans rho**	**p-val**	**Spearmans rho**	**p-val**
gTSS	7.427	4.377	1	8	1	1,70	**3,90E-02**	**9,80E-03**	1,19E-02	4,31E-01
aTSS	5.690	3.297	1	8	1	1,73	**4,79E-01**	**1,88E-182**	2,06E-02	2,37E-01
iTSS	4.554	2.989	1	8	1	1,52	**3,66E-01**	**2,93E-95**	-2,02E-02	2,70E-01
nTSS	1.420	-	-	-	-	-	-	-	-	-

B) *F. thermalis*							CDS-Länge		CAI	
Klasse	**TSSe**	**CDSe**	**Min TSS**	**Max TSS**	**Median**	**Mittelwert**	**Spearmans rho**	**p-val**	**Spearmans rho**	**p-val**
gTSS	5.767	3.392	1	7	1	1,70	2,22E-02	1,95E-01	2,30E-03	8,96E-01
aTSS	4.406	2.579	1	9	1	1,71	**4,77E-01**	**5,81E-147**	-1,14E-02	5,62E-01
iTSS	3.566	2.299	1	10	1	1,55	**3,28E-01**	**6,00E-59**	1,03E-02	6,20E-01
nTSS	1.203	-	-	-	-	-	-	-	-	-

C) *C. fritschii*							CDS-Länge		CAI	
Klasse	**TSSe**	**CDSe**	**Min TSS**	**Max TSS**	**Median**	**Mittelwert**	**Spearmans rho**	**p-val**	**Spearmans rho**	**p-val**
gTSS	9.242	5.068	1	7	2	1,82	3,19E-02	2,29E-02	-4,70E-03	7,36E-01
aTSS	7.264	3.841	1	18	1	1,89	**5,16E-01**	**5,92E-260**	**-6,00E-02**	**2,01E-04**
iTSS	6.595	3.773	1	10	1	1,75	**4,17E-01**	**1,35E-158**	**-3,73E-02**	**2,19E-02**
nTSS	1.516	-	-	-	-	-	-	-	-	-

Für die TSS-Anzahl pro CDS und TSS-Klasse konnten für *F. muscicola* und *F. thermalis* ähnliche Werte beobachtet werden (Tab. 2AB). Generell wurden starke Unterschiede zwischen Median und Mittelwert in der TSS-Anzahl pro CDS und TSS-Klasse ermittelt. In *F. muscicola* und *F. thermalis* wurde im Median ein TSS pro CDS und Klasse gefunden (Tab. 2AB). Für *C. fritschii* wurde mit Ausnahme der gTSS-Klasse das gleiche beobachtet. In der gTSS-Klasse wurde ein Median von zwei TSS pro CDS gefunden. (Tab. 2C)

Die Abweichungen von Mittelwert und Median erklären sich durch CDSe mit hoher TSS-Anzahl einer Klasse. In *F. muscicola* wurden für eine Succinyl-CoA-Synthetase insgesamt acht gTSSe detektiert (Tab. 15, Anhang). Zwei weitere CDSe, dessen Produkte als hypothetische Proteine annotiert sind, wiesen acht aTSSe bzw. acht iTSSe auf (Tab. 16, Tab. 17, Anhang). In *F. thermalis* war ähnliches zu beobachten. Für drei als hypothetisch annotierte CDSe wurden sieben gTSSe (Tab. 15), neun aTSSe und zehn iTSSe (Tab. 16, Tab. 17, Anhang) detektiert. In *C. fritschii* wies eine Histidinkinase sieben gTSSe (Tab. 15, Anhang) auf. Ein weiterer CDSs wies zehn iTSSe auf (Tab. 17, Anhang) und das Maximum wurde von einem CDS mit einer Länge von 15.336 Nukleotiden erreicht. Dieser wies insgesamt 18 aTSSe auf (Tab. 16, Anhang).

Hohe multiple TSS-Zuweisungen pro CDS und Klasse wurden selten beobachtet (Abb. 25). In den meisten Fällen wurden nicht mehr als zwei TSSe gleicher Klassenzuweisung

detektiert (Abb. 25, 80. Perzentil). Besonders im letzten CDS-Beispiel in *C. fritschii* fällt die hohe TSS-Anzahl und die Länge des CDS auf. Für die aTSS- und iTSS-Klasse konnte ein Zusammenhang zwischen CDS-Länge und TSS-Anzahl ermittelt werden (Tab. 2). Kein Zusammenhang wurde zwischen der TSS-Anzahl und dem CAI eines CDS ermittelt (Tab. 2).

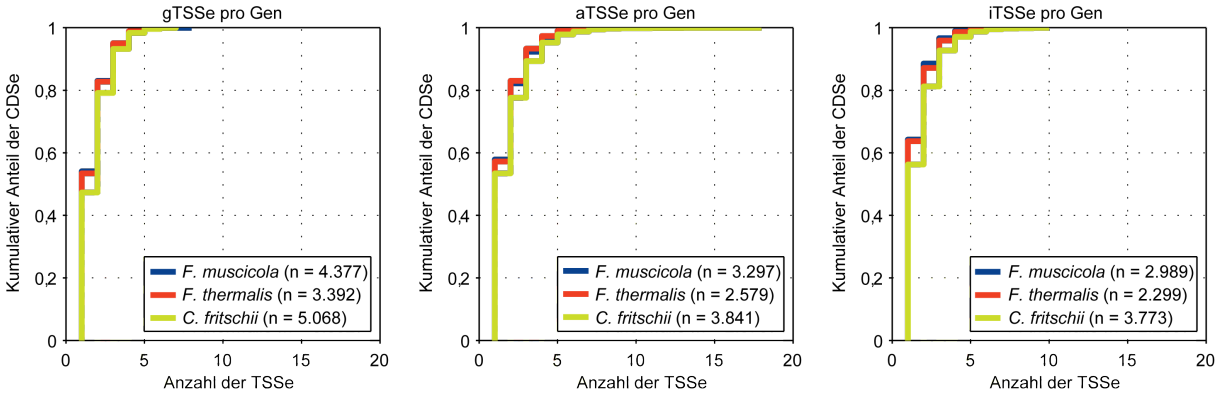


Abb. 25: Kumulative Verteilung der CDSe in Abhängigkeit zur TSS-Anzahl.

Die kumulativen Verteilungsdiagramme stellen die CDS-Verteilung in Abhängigkeit zur TSS-Anzahl in *F. muscicola* (blau), *F. thermalis* (rot) und *C. fritschii* (grün) für die TSS-Klassen dar. In den meisten Fällen wird ein TSS pro CDS und Klasse gefunden. 80% der CDSe mit einer TSS-Detektion zeigen nicht mehr als einen alternativen TSS, unabhängig von der TSS-Klasse.

6.3.5 Eigenschaften der Promotoren

Für einige CDSe konnten extrem viele Transkriptionsstartpunkte detektiert werden. Diese Eigenschaft eines Transkriptoms wurde ebenfalls in Analysen anderer Cyanobakterien beobachtet (Mitschke, Georg, et al. 2011, Mitschke, Vioque, et al. 2011). Im Nachfolgenden wird auf einen Genbereich eingegangen, der stromaufwärts eines TSS liegt, dem Promotor. Es wurde untersucht, ob die Mehrheit der detektierten TSSe auch essentielle Bindestellen der RNA-Polymerase aufweisen.

Abb**. 26** stellt in den drei Spezies die Nukleotidhäufigkeiten aller TSS-Regionen dar, welche 50 nt stromaufwärts eines TSS liegen. Stromaufwärts eines CDS (Abb. 26A), komplementär zu einem CDS (Abb. 26B), innerhalb eines CDS (Abb. 26C) oder weit entfernt von einem CDS (Abb. 26D) sind daher die Nukleotidhäufigkeiten pro Position berechenbar. In den 5'-Regionen der TSSe wurden speziesübergreifend und klassenübergreifend sehr ähnliche Nukleotidanteile pro Position beobachtet.

Bis ca. 12 Nukleotide vor dem Transkriptionsstart wurden keine eindeutigen Präferenzen für bestimmte Nukleotide beobachtet (Abb. 26). Die Anteile von Adenin und Thymin nahmen in diesen Bereichen jedoch immer einen höheren Anteil ein als die Anteile von Guanin und Cytosin. Besonders in der gTSS- und nTSS-Klasse war dies zu beobachten (Abb. 26AD, Bereiche -50 bis -12). Das -35-Element war mit dieser Vorgehensweise nicht nachweisbar. Daher wurde das Programm `MEME` (Bailey et al. 2009) für eine *de novo* Motivsuche benutzt. Das -35-Element konnte in dieser Region jedoch auch nicht mit diesem Programm eindeutig nachgewiesen werden. Die Schwierigkeit der Detektion eines -35-Elements ist auch für das Cyanobakterium *Prochlorococcus* MED4 in früheren Analysen beschrieben worden (Voigt et al. 2014, Vogel 2003). Ab Position -12 bzw. -10 wurden eindeutige positionelle Präferenzen für die Nukleotide Adenin und Thymin beobachtet. Bis zum Transkriptionsstart (Abb. 26, Position +1) entsprachen diese Präferenzen in etwa dem charakteristischen `TATAAT` Konsensusmotiv des -10-Elements. Das Programm `MEME` konnte ebenfalls ein Konsensusmotiv in allen TSS-Klassen, Spezies und in annähernd allen Regionen ermitteln (Abb. 26, Sequenzlogos). Für die Intermediären TSS-Klassen ist die Analyse der Promotorregionen in Abb**. 32** im Anhang dargestellt und zeigt die gleiche Präferenz für ein potentielles -10-Element.

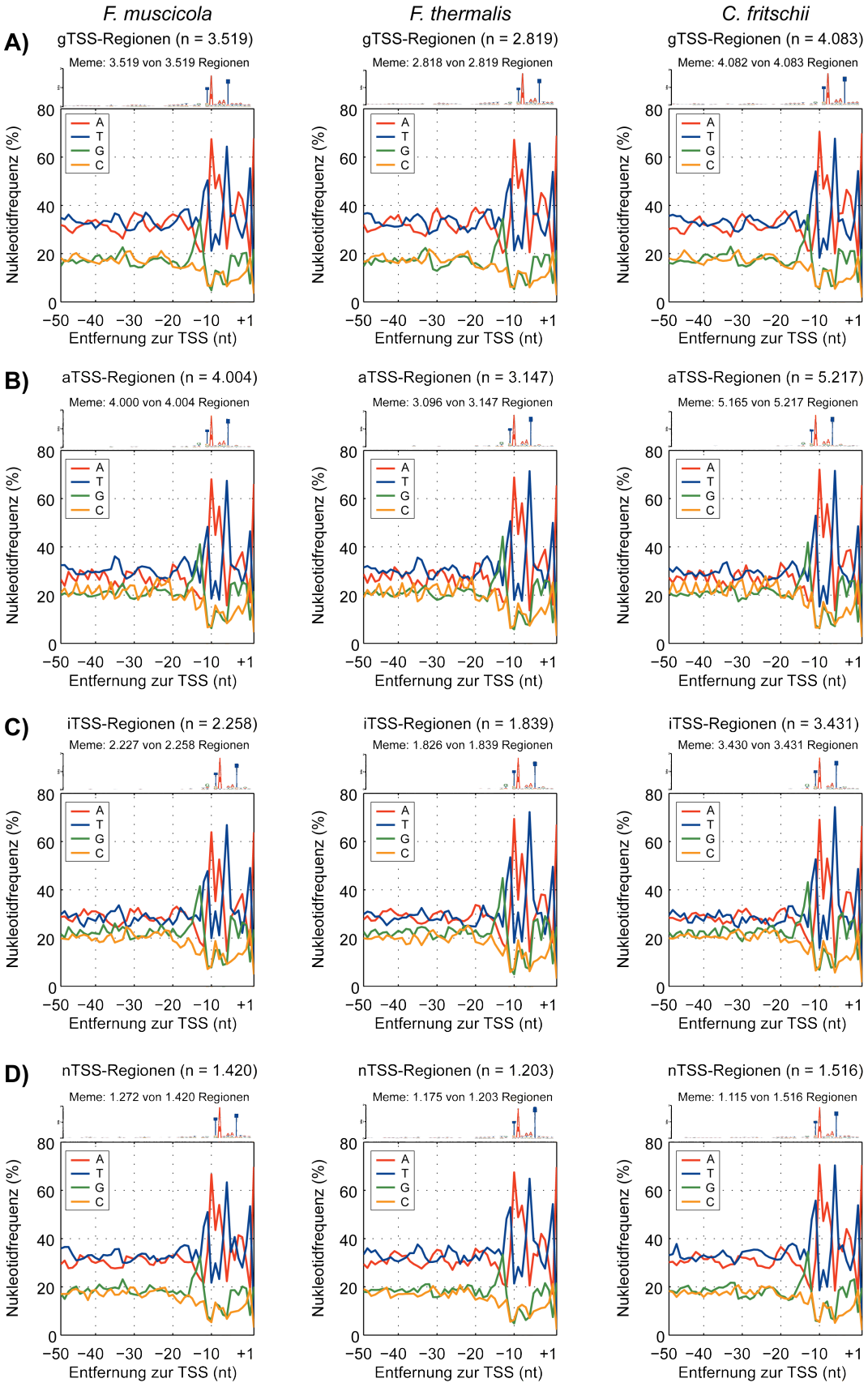


Abb. 26: Promotorregionen - pure TSS-Klassen.

6.4 Die Interspezies TSS-Analyse

6.4.1 Orthologe Proteinfamilien und TSS-Präsenz

Basierend auf der Vorgehensweise in Abschnitt 5.6 wurden in der Analyse insgesamt 33.563 Proteinfamilien, in denen sich mindestens zwei unterschiedliche Spezies befanden, erzeugt. 112 Proteinfamilien zeigten jeweils exakt eine Proteinkopie jeder Spezies. Um die Evolution transkriptioneller Regulation in *F. muscicola*, *F. thermalis* und *C. fritschii* zu studieren, fokussierte sich die Analyse auf homologe Proteine zwischen *F. muscicola* und *F. thermalis* sowie zwischen *F. muscicola* und *C. fritschii*. Aufgrund inhärenter Probleme Paralogien zu detektieren, wurden nur homologe Proteine miteinander verglichen, die als Einzelkopie vorlagen.

76% der *F. muscicola* Proteinsequenzen, 82% der *F. thermalis* Proteinsequenzen und 89% der *C. fritschii* Proteinsequenzen konnten Proteinfamilien zugeordnet werden (Tab. 11). 3.363 Proteinfamilien enthielten Proteine von *F. muscicola* und *F. thermalis*, von denen 2.805 Familien jeweils nur eine Genkopie in beiden Spezies aufwiesen (Tab. 4). Für *F. muscicola* und *C. fritschii* wurden 3.410 gemeinsame Proteinfamilien beobachtet. Von diesen enthielten 2.697 für beide Spezies genau eine Genkopie (Tab. 3). Insgesamt enthielten 2.321 Proteinfamilien genau eine Genkopie für alle drei Cyanobakterien.

Durch die Identifizierung der Einzelkopie-Proteinfamilien wurde das Vorliegen von gemeinsamen Transkriptionstartpunkten aller TSS-Klassen analysierbar. Nur bei TSS-Präsenz in jeweils beiden kodierenden Sequenzabschnitten, konnte die Methode aus Abschnitt 5.7 zu einem orthologen TSS im Alignment führen. Im *F. muscicola* und *F. thermalis* Vergleich wurden für 55% der Einzelkopie-Proteinfamilien gTSSe in jeweils beiden Spezies detektiert (Tab. 5A). Für 45% der Einzelkopie-CDSe wurden aTSSe in beiden Spezies beobachtet und 38% der Einzelkopie-CDSe zeigten iTSS-Präsenz in beiden Spezies (Tab. 5A). Im *F. muscicola* und *C. fritschii* Vergleich wurde eine ähnliche Anzahl an TSSe für Einzelkopie-CDSe gefunden. In 54% der Einzelkopie-CDSe wurden gTSSe in beiden Spezies beobachtet. Für 41% der Einzelkopie-CDSe konnten aTSSe in beiden Spezies detektiert werden und 37% der Einzelkopie-CDSe zeigten iTSSe in beiden Cyanobakterien (Tab. 5B).

Tab. 4: Proteinfamilien zwischen *F. muscicola* und *F. thermalis*.

Proteinfamilien in Abhängigkeit der Genkopien von *F. muscicola* und *F. thermalis*. Pro Zeile und Spalte sind jeweils die Anzahl an Proteinfamilien dargestellt, in der eine bestimmte Anzahl an Genkopien der beiden Spezies gefunden wurden. In **schwarz** ist die Gesamtanzahl an Einzelkopie-Proteinfamilien zwischen *F. muscicola* und *F. thermalis* dargestellt.

		F. thermalis																	
	Proteinkopien	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Proteinfamilien *F. muscicola*	CDSe *F. muscicola*
F. muscicola	1	**2.805**	34	2	1	0	0	0	0	0	0	0	0	0	0	0	0	2.842	2.842
	2	124	189	7	2	0	0	0	0	0	0	0	0	0	0	0	0	322	644
	3	22	40	36	2	1	0	0	0	0	0	0	0	0	0	0	0	101	303
	4	6	14	12	7	0	0	0	0	0	0	0	0	0	0	0	0	39	156
	5	1	1	7	6	6	0	1	0	0	0	0	0	0	0	0	0	22	110
	6	1	3	0	4	7	4	0	0	0	0	0	0	0	0	0	0	19	114
	7	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2	14
	8	0	0	0	2	0	4	1	0	0	0	0	0	0	0	0	0	7	56
	9	0	0	1	0	0	1	2	0	0	0	0	0	0	0	0	0	4	36
	10	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	10
	11	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	2	22
	12	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	12
	13	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	13
	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Proteinfamilien *F. thermalis*	2.959	281	65	26	14	9	5	1	1	2	0	0	0	0	0	0	3.363	4.332
	CDSe *F. thermalis*	2.959	562	195	104	70	54	35	8	9	20	0	0	0	0	0	0	4.016	

Tab. 3: Proteinfamilien zwischen *F. muscicola* und *C. fritschii*.

Proteinfamilien in Abhängigkeit der Genkopien von *F. muscicola* und *C. fritschii*. Pro Zeile und Spalte sind jeweils die Anzahl an Proteinfamilien dargestellt, in der eine bestimmte Anzahl an Genkopien der beiden Spezies gefunden wurden. In **schwarz** ist die Gesamtanzahl an Einzelkopie-Proteinfamilien zwischen *F. muscicola* und *C. fritschii* dargestellt.

		C. fritschii																	
	Proteinkopien	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Proteinfamilien *F. muscicola*	CDSe *F. muscicola*
F. muscicola	1	**2.697**	162	19	4	1	0	1	1	0	0	0	1	1	0	0	0	2.887	2.887
	2	107	174	34	13	1	1	0	0	0	0	0	0	0	0	0	0	330	660
	3	16	23	34	8	8	3	1	2	0	1	0	0	0	0	0	0	96	288
	4	4	6	10	9	4	0	0	1	1	0	0	0	0	0	0	0	35	140
	5	0	2	3	8	5	1	2	0	1	1	0	0	0	0	0	0	23	115
	6	1	0	2	7	2	4	2	0	1	1	0	0	0	0	0	0	20	120
	7	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	3	21
	8	0	0	0	0	1	2	1	0	0	2	0	0	1	0	0	0	7	56
	9	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0	0	4	36
	10	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	10
	11	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	2	22
	12	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	12
	13	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	13
	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Proteinfamilien *C. fritschii*	2.826	368	102	49	22	11	8	6	3	7	0	3	3	1	0	1	3.410	4.380
	CDSe *C. fritschii*	2.826	736	306	196	110	66	56	48	27	70	0	36	39	14	0	16	4.546	

Tab. 5: TSS-Präsenz in Einzelkopie-CDS Paaren.

A) TSS-Präsenz in Einzelkopie-CDS Paaren zwischen *F. muscicola* und *F. thermalis*. Pro Zeile sind die Anzahl an Proteinfamilien dargestellt, in denen für CDSe beider Spezies, nur für *F. muscicola*, nur für *F. thermalis* oder für keinen CDS einer Spezies TSSe einer bestimmten TSS-Klasse gefunden wurden.
B) TSS-Präsenz in Einzelkopie-CDS Paaren zwischen *F. muscicola* und *C. fritschii*. Pro Zeile sind die Anzahl an Proteinfamilien dargestellt, in denen für CDSe beider Spezies, nur für *F. muscicola*, nur für *C. fritschii* oder für keinen CDS einer Spezies TSSe einer bestimmten TSS-Klasse gefunden wurden.

A)

Klasse	Beide Spezies		Nur *F. muscicola*		Nur *F. thermalis*		Keine TSS		Gesamt
gTSS	1.552	55%	358	13%	354	13%	541	19%	2.805
aTSS	1.254	45%	334	12%	298	11%	919	33%	2.805
iTSS	1.075	38%	399	14%	306	11%	1.025	37%	2.805

B)

Klasse	Beide Spezies		Nur *F. muscicola*		Nur *C. fritschii*		Keine TSS		Gesamt
gTSS	1.449	54%	372	14%	556	21%	320	12%	2.697
aTSS	1.117	41%	393	15%	547	20%	641	24%	2.697
iTSS	1.004	37%	387	14%	619	23%	687	25%	2.697

6.4.2 Orthologe Transkriptionsstartpunkte

Die Identifizierung orthologer TSSe verlief über zwei unterschiedliche Typen von Alignments. In beiden Typen wurde als Vergleichsebene ein Einzelkopie-CDS Paar verwendet, welches aus der Proteinfamilienanalyse aus Abschnitt 6.4.1 resultierte. Die Typen der Alignments und die Kriterien für die Detektion eines orthologen TSS sind im Abschnitt 5.7 beschrieben. Bei gemeinsamer gTSS-Präsenz für ein Einzelkopie-CDS Paar (Tab. 5) sind Alignments mit Sequenzen erzeugt worden, die dem gewählten UTR-Schwellenwert entsprachen. Bei einem Einzelkopie-CDS Paar mit gemeinsamer aTSS- oder iTSS-Präsenz wurden die paarweisen globalen Alignments der Proteinsequenzen aus Abschnitt 6.4.1 in CDS-Alignments transformiert und die TSSe innerhalb der Alignments bestimmt (siehe Abschnitt 5.7).

Tab. 6: Orthologe- und speziesspezifische TSSe für Einzelkopie-CDS Paare.

A) Orthologe- und speziesspezifische TSSe für Einzelkopie-CDS Paare zwischen *F. muscicola* und *F. thermalis*. Die Anzahl an Proteinfamilien ist deckungsgleich mit der Summe an Proteinfamilien, für die mindestens ein TSS in einer Spezies gefunden wurde (Tab. 5A).

B) Orthologe- und speziesspezifische TSSe für Einzelkopie-CDS Paare zwischen *F. muscicola* und *C. fritschii*. Die Anzahl an Proteinfamilien ist deckungsgleich mit der Summe an Proteinfamilien, für die mindestens ein TSS in einer Spezies gefunden wurde (Tab. 5B).

A)

Klasse	TSSe	Proteinfamilien	TSSe/Proteinfamilie	Orthologe TSSe	%	Spezifische TSSe	%
gTSS	5.112	2.264	2,26	1.622	32%	3.490	68%
aTSS	4.082	1.886	2,16	1.556	38%	2.526	62%
iTSS	3.382	1.780	1,90	1.213	36%	2.169	64%

B)

Klasse	TSSe	Proteinfamilien	TSSe/Proteinfamilie	Orthologe TSSe	%	Spezifische TSSe	%
gTSS	6.323	2.377	2,66	824	13%	5.499	87%
aTSS	5.391	2.057	2,62	553	10%	4.838	90%
iTSS	4.695	2.011	2,33	438	9%	4.257	91%

Tab. 6 zeigt die Verteilung orthologer- und speziesspezifischer TSSe im intragenerischen Vergleich (A) und im intergenerischen Vergleich (B). Die Anzahl der TSSe pro Familie unterscheidet sich von der Anzahl an TSSe pro CDS (vgl. Tab. 2, Abschnitt 6.3.4). Dies ist in beiden Speziesvergleichen und allen TSS-Klassen zu beobachten. Im intragenerischen Vergleich sind orthologe aTSSe am häufigsten, dann orthologe iTSSe und orthologe gTSSe (Tab. 6A). Im intergenerischen Vergleich sind orthologe gTSSe am häufigsten, dann orthologe aTSSe und orthologe iTSSe (Tab. 6B). Die Häufigkeit orthologer TSSe pro Klasse ist geringer

als im intragenerischen Vergleich (vgl. Tab. 6A u. B). Mehrfachzählungen eines TSS können diese Analyse verzerren, da intermediäre TSS-Klassen mitgezählt werden. Ein Beispiel soll dies erläutern. Ein Einzelkopie-CDS Paar mit einem gTSS für jede Spezies, in einer Spezies jedoch als gaTSS klassifiziert, kann eine Mehrfachzählung hervorrufen, wenn der komplementär zur gaTSS liegende CDS ein weiteres Einzelkopie-CDS Paar darstellt. Für die gTSS-Klasse kann der TSS ortholog sein. Für die aTSS-Klasse kann der TSS jedoch als speziesspezifisch gezählt werden. Der Grund ist ein konservierter, aber nicht identischer genomischer Kontext zwischen den Spezies. Daher wurde die Analyse mit puren TSS-Klassen für beide Spezies wiederholt. Für jeweils beide Spezies mussten die TSSe von Einzelkopie-CDS Paaren in puren TSS-Klassen vorliegen, um gezählt zu werden. Ein sehr ähnlicher genomischer Kontext war daher Bedingung. Bei dieser strikteren Vorgehensweise werden im intragenerischen Vergleich orthologe aTSSe signifikant häufiger als orthologe gTSSe oder orthologe iTSSe gefunden (Tab. 18). Im intergenerischen Vergleich sind signifikant häufiger orthologe gTSSe als orthologe aTSSe zu beobachten. Des Weiteren sind orthologe aTSSe signifikant häufiger als orthologe iTSSe (Tab. 19).

In der nächsten Analyse sollte getestet werden, ob Einzelkopie-CDS Paare mit steigender globaler Proteinsequenzidentität auch mehr orthologe TSSe aufweisen. Hierfür wurden die CDSe in sieben Gruppen von Sequenzidentitäten eingeteilt und die Anzahl orthologer TSSe mit der Anzahl speziesspezifischer TSSe in einem gestapelten Säulendiagram verglichen (Abb. 27A,B). Auch hier wurden nur die puren TSS-Klassen berücksichtigt. Im intragenerischen Vergleich wurde in allen TSS-Klassen ein Zusammenhang zwischen orthologer TSS-Anzahl und Proteinsequenzidentität gefunden (Abb. 27A). Einzelkopie-Proteinpaare mit ausschließlich orthologer TSS-Präsenz tendieren ähnlicher zueinander zu sein (Tab. 7, Tukey's Test). Diese Beobachtungen wurden nur im intragenerischen Vergleich gemacht. Im intergenerischen Vergleich war zwischen orthologer TSS-Anzahl und Proteinsequenzidentität kein Zusammenhang zu beobachten. Nur für die iTSS-Klasse konnte eine Tendenz beobachtet werden (Abb. 27B). Für diese TSS-Klasse wurde ebenfalls bei sehr ähnlichen Proteinsequenzen das Vorhandensein von ausschließlich orthologen TSS beobachtet. (Tab. 7, Tukey's Test).

Auf intragenerischer Ebene zeigt sich ein Zusammenhang zwischen der CDS-Sequenzidentität und der Anzahl orthologer TSSe. Mit Ausnahme der iTSS-Klasse konnte dies auf der intergenerischen Ebene nicht beobachtet werden.

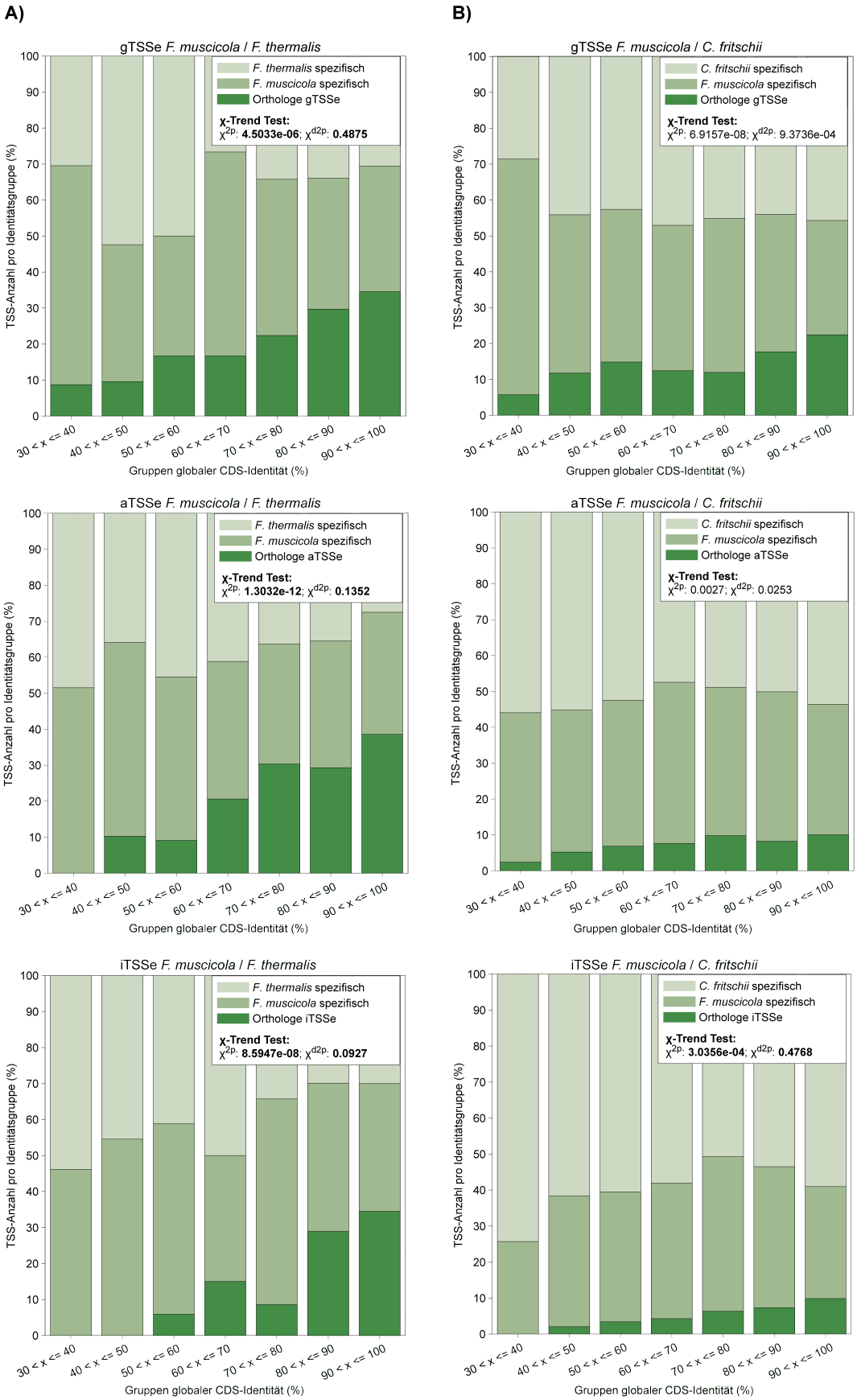


Abb. 27: Orthologe TSS-Anzahl in Abhängigkeit zur CDS-Identität.

A) Intragenerischer Vergleich, **B)** Intergenerischer Vergleich. Der, linear positive Anteilstrend für orthologe TSSe ist in **schwarz** hervorgehoben (chi-Trend-Test).

Tab. 7: CDS-Identität und Art der TSS-Präsenz.

Diese Tabelle stellt den Vergleich dreier Gruppen von CDS-Paaren und deren Sequenzidentität dar. Die Gruppe (I) beinhaltet Proteinsequenzidentitäten von CDS-Paaren, die nur orthologe TSS-Präsenz aufweisen. Die Gruppe (II) stellt Proteinsequenzidentitäten von CDS-Paaren dar, die orthologe und speziesspezifische TSS-Präsenz aufweisen. In der Gruppe (III) sind ausschließlich Proteinsequenzidentitäten von CDS-Paaren vorhanden, die nur speziesspezifische TSSe aufweisen. Die Mittelwerte der drei Gruppen wurden miteinander verglichen.

			CDS-Gruppen			Kruskal-Wallis Einseitiger ANOVA		Tukey's Test
Spezies 1	**Spezies 2**	**Klasse**	**I**	**II**	**III**	**χ2**	**p**	**Vergleich der Rangmittelwerte**
F. muscicola	*F. thermalis*	gTSS	412	809	1043	71,1	3,63E-16	I > III; II > III
F. muscicola	*F. thermalis*	aTSS	400	677	809	19,16	6,89E-05	I > II,III
F. muscicola	*F. thermalis*	iTSS	373	535	872	12,94	1,60E-03	I > III
F. muscicola	*C. fritschii*	gTSS	94	630	1653	40,29	1,79E-09	II > I, III
F. muscicola	*C. fritschii*	aTSS	68	404	1585	13,28	1,30E-03	keine Gruppenunterschiede
F. muscicola	*C. fritschii*	iTSS	65	323	1622	15,58	4,00E-04	I > III; II > III

Legende:

I = CDS-Paare, mit ausschließlich orthologer TSS-Präsenz.

II = CDS-Paare, mit orthologer und speziesspezifischer TSS-Präsenz.

III = CDS-Paare, mit ausschließlich speziesspezifischer TSS-Präsenz.

6.4.3 TSS-Regionen und Konservierung

Durch das Vorliegen von Alignments der Stromaufwärtsregionen der Einzelkopie-CDS Paare, sowie durch das Vorhandensein von CDS-Alignments, war es möglich zwei unterschiedliche TSS-Regionen miteinander zu vergleichen. TSS-Regionen mit der Eigenschaft der Transkriptionsinitiation (gTSS) und TSS-Regionen mit zusätzlicher kodierender Eigenschaft (iTSS oder aTSS) konnten miteinander verglichen werden. Durch das Vorhandensein zweier Eigenschaften wurde angenommen, dass aTSS- oder iTSS-Regionen unter stärkerem Selektionsdruck stehen als gTSS-Regionen. Für die Analyse wurden Sequenzidentitäten der TSS-Regionen in einem Bereich von 70 nt um einen TSS herum ermittelt und mit den entsprechenden Proteinidentitäten verglichen (Abschnitt 5.7). Da für eine Einzelkopie-Proteinfamilie mehrere TSSe pro Klasse vorlagen (Tab. 6), wurden die Sequenzidentitäten der TSS-Regionen zum einen “pro TSS” und zum anderen “pro CDS-Paar” analysiert (Tab. 8 u. Tab. 9,). Der Median in Tab**. 8** und Tab**. 9** im Wilcoxon Rangsummentest ist der Median aus den Differenzen zwischen den Proteinsequenzidentitäten und den Mittelwerten der TSS-Regionidentitäten. Der Median der Differenzen ist positiv, wenn in der Mehrheit die Proteinsequenzidentitäten höher waren als die entsprechenden TSS-Regionidentitäten. Der Median der Differenzen ist negativ, wenn die Mehrheit der Proteinsequenzidentitäten niedriger waren als die entsprechenden TSS-Regionidentitäten. Orthologe TSSe und spezifische TSSe wurden separat voneinander analysiert.

Im intragenerischen Vergleich wiesen orthologe gTSS-Regionen signifikant geringere Ähnlichkeiten als die assoziierten Einzelkopie-Proteinsequenzen zueinander auf (Tab. 8). Speziesspezifische gTSS-Regionen zeigten ebenfalls signifikant geringere Sequenzähnlichkeiten als die Einzelkopie-Proteinsequenzen. Der Median wies jedoch einen stärkeren Unterschied auf (Tab. 8). Die TSS-Regionen der aTSS- und iTSS-Klasse wiesen für orthologe TSSe eine höhere Sequenzidentität auf als die jeweiligen assoziierten Einzelkopie-Proteinsequenzen. Für speziesspezifische TSS-Regionen wurde dies nicht beobachtet. Im intragenerischen Vergleich sind aTSS-Regionen und iTSS-Regionen ähnlicher zueinander als die Einzelkopie-Proteinsequenzen. Dies wurde jedoch nur für orthologe TSS-Regionen beobachtet.

Im intergenerischen Vergleich fiel die für den *Fischerella*-Vergleich gemachte Beobachtung der gTSS-Klasse stärker aus. Hier unterschieden sich die Identitäten der orthologen- und speziesspezifischen TSS-Regionen sehr stark von den Proteinsequenzidentitäten (Tab. 9). In der aTSS- und iTSS-Klasse wurde kein signifikanter Unterschied zwischen den orthologen TSS-Regionen und den Einzelkopie-Proteinpaaren beobachtet (Tab. 9).

Tab. 8: TSS-Regionen im Vergleich zur Proteinsequenzidentität – intragenerischer Vergleich.

In dieser Tabelle ist der Vergleich der TSS-Regionidentitäten zwischen den assoziierten CDS-Paaren dargestellt. In jeder Analyse werden orthologe und speziesspezifische TSS-Region getrennt voneinander analysiert. In den Zeilen „pro TSS“ wird jede TSS-Region mit der Sequenzidentität des assoziierten Einzelkopie-Proteinpaares verglichen. Dabei wird die Differenz aus Proteinsequenzidentität und TSS-Regionidentität berechnet und aus den Ergebnissen der Median ermittelt. In den Zeilen „pro CDS-Paar“ wird zunächst der Mittelwert von allen TSS-Regionidentitäten ermittelt. Erst dann erfolgt die Berechnung der Differenz zwischen der Mittelwerte der TSS-Regionidentitäten und der entsprechenden Einzelkopie-Proteinsequenzidentitäten und dann die Ermittlung des Median.

			Spearman's Korrelaionstest		Wilcoxon Rangsummentest				
Klasse	**Analysebereich**	**Gruppe**	**r**	**p-Wert**	**p-Wert**	**H**	**Median**	**Interpretation**	**Median TSS-Region (%)**
gTSS	pro TSS	ortholog	0,146	7,35E-17	1,33E-17	1	0,46	Proteine ähnlicher als TSS-Regionen.	95,71
		spezifisch	0,187	5,61E-29	5,91E-278	1	6,97	Proteine ähnlicher als TSS-Regionen.	85,71
	pro CDS-Paar	ortholog	0,152	9,55E-08	3,84E-10	1	0,71	Proteine ähnlicher als TSS-Regionen.	95
		spezifisch	0,215	1,01E-20	1,19E-153	1	7,67	Proteine ähnlicher als TSS-Regionen.	86,43
aTSS	pro TSS	ortholog	0,277	4,30E-56	5,13E-19	1	-0,59	TSS-Regionen ähnlicher als Proteine.	95,71
		spezifisch	0,488	1,29E-151	3,66E-07	1	0,49	Proteine ähnlicher als TSS-Regionen.	94,29
	pro CDS-Paar	ortholog	0,301	4,65E-24	1,15E-06	1	-0,46	TSS-Regionen ähnlicher als Proteine.	95,71
		spezifisch	0,571	1,80E-129	5,80E-14	1	0,86	Proteine ähnlicher als TSS-Regionen.	94,29
iTSS	pro TSS	ortholog	0,297	1,24E-50	6,92E-18	1	-0,67	TSS-Regionen ähnlicher als Proteine.	95,71
		spezifisch	0,465	4,42E-117	3,96E-03	1	0,26	Proteine ähnlicher als TSS-Regionen.	94,29
	pro CDS-Paar	ortholog	0,345	1,00E-26	2,39E-08	1	-0,67	TSS-Regionen ähnlicher als Proteine.	95,71
		spezifisch	0,529	2,14E-102	1,27E-07	1	0,63	Proteine ähnlicher als TSS-Regionen.	94,29

Tab. 9: TSS-Regionen im Vergleich zur Proteinsequenzidentität – intergenerischer Vergleich.

In dieser Tabelle ist der Vergleich der TSS-Regionidentitäten zwischen den assoziierten CDS-Paaren dargestellt. In jeder Analyse werden orthologe und speziesspezifische TSS-Region getrennt voneinander analysiert. In den Zeilen „pro TSS“ wird jede TSS-Region mit der Sequenzidentität des assoziierten Einzelkopie-Proteinpaares verglichen. Dabei wird die Differenz aus Proteinsequenzidentität und TSS-Regionidentität berechnet und aus den Ergebnissen der Median ermittelt. In den Zeilen „pro CDS-Paar“ wird zunächst der Mittelwert von allen TSS-Regionidentitäten ermittelt. Erst dann erfolgt die Berechnung der Differenz zwischen der Mittelwerte der TSS-Regionidentitäten und der entsprechenden Einzelkopie-Proteinsequenzidentitäten und dann die Ermittlung des Median.

			Spearman's Korrelaionstest		Wilcoxon Rangsummentest				
Klasse	**Analysebereich**	**Gruppe**	**r**	**p-Wert**	**p-Wert**	**H**	**Median**	**Interpretation**	**Median TSS-Region (%)**
gTSS	pro TSS	ortholog	0,199	2,82E-16	7,39E-93	1	8,34	Proteine ähnlicher als TSS-Regionen.	72,86
		spezifisch	0,152	6,06E-30	0,00E+00	1	21,80	Proteine ähnlicher als TSS-Regionen.	51,43
	pro CDS-Paar	ortholog	0,207	1,94E-08	7,62E-41	1	8,49	Proteine ähnlicher als TSS-Regionen.	72,86
		spezifisch	0,197	2,61E-21	2,08E-272	1	21,10	Proteine ähnlicher als TSS-Regionen.	51,91
aTSS	pro TSS	ortholog	0,551	6,83E-89	6,35E-02	0	0,11	Kein Unterschied der Ähnlichkeit.	81,43
		spezifisch	0,57	0,00E+00	2,47E-13	1	1,43	Proteine ähnlicher als TSS-Regionen.	78,57
	pro CDS-Paar	ortholog	0,581	7,10E-44	1,43E-01	0	-0,29	Kein Unterschied der Ähnlichkeit.	81,43
		spezifisch	0,691	1,16E-282	1,43E-14	1	2,03	Proteine ähnlicher als TSS-Regionen.	78,1
iTSS	pro TSS	ortholog	0,487	2,14E-53	8,83E-01	0	0,24	Kein Unterschied der Ähnlichkeit.	81,43
		spezifisch	0,564	0,00E+00	7,63E-04	1	0,83	Proteine ähnlicher als TSS-Regionen.	78,57
	pro CDS-Paar	ortholog	0,518	5,27E-28	9,23E-01	0	0,46	Kein Unterschied der Ähnlichkeit.	81,43
		spezifisch	0,671	2,87E-254	1,66E-06	1	1,28	Proteine ähnlicher als TSS-Regionen.	78,57

6.4.4 Änderungen der Transkriptabundanz und Kandidatengene

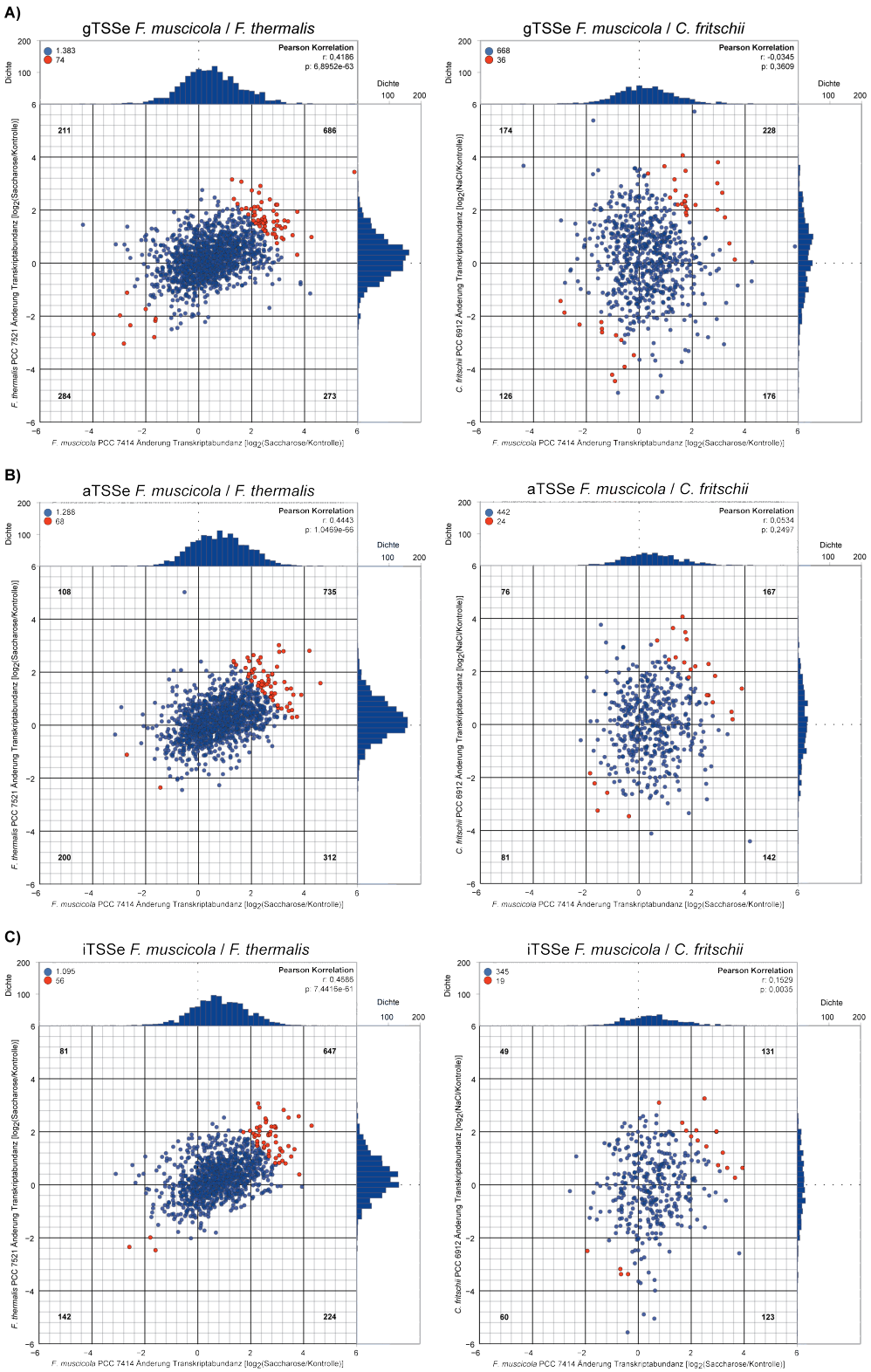


Abb. 28: Änderung der Transkriptabundanz orthologer TSSe.

Rote Punkte = 5% der orthologen TSSe, die die stärkste Änderung zwischen den Spezies aufweisen. Blaue Punkte = Orthologe TSSe und deren Änderungen zwischen den Spezies.

Tab. 10: Identifizierung genomischer Regionen durch Änderung der Transkriptabundanz.

Spalte 1 = Einzelkopie-Proteinfamilie, deren CDSe orthologe TSSe aufweisen. **Spalte 2** = Spezies in der der TSS gefunden wurde. **Spalte 3** = Änderung der Transkriptabundanz von der TSS ausgehend. **Spalte 4** = TSS-Klasse des detektierten orthologen TSS. **Spalte 5** = Entfernung zum Stromabwärts gelegenen CDS. **Spalte 6** = NCBI-Locus Tag des assoziierten CDS. **Spalte 7** = NCBI-Annotation des CDS. Die Farbe einer Zeile entspricht der Farbe eines Punktes aus Abb. 28. Eine vollständige Tabelle aller in Abb. 28 dargestellten TSSe befindet sich auf der beigelegten CD (Abschnitt 10.8).

Pfam	Spezies	Exp. Änderung	TSS-Typ	Entfernung (nt) zum 3' CDS	Locus Tag des CDS	Annotation des CDS
1610	*F. muscicola*	-0,3526	ga	43	UYI_RS0105420	hypothetisches protein
1610	*F. thermalis*	0,0163	g	43	UYK_RS0108990	
1610	*C. fritschii*	2,581	g	43	UYC_RS0116560	
1610	*F. muscicola*	2,4783	ga	356	UYI_RS0105420	
1610	*F. thermalis*	1,5054	ga	356	UYK_RS0108990	
1610	*C. fritschii*	-	-	-	UYC_RS0116560	
561	*F. muscicola*	-1,1644	g	105	UYI_RS0105415	*bolA*-ähnlich
561	*F. thermalis*	-0,6397	g	105	UYK_RS0108995	
561	*C. fritschii*	-1,1475	g	138	UYC_RS0116565	
1093	*F. muscicola*	3,1315	gi	598	UYI_RS0102400	*mreC*
1093	*F. thermalis*	1,3122	gi	388	UYK_RS0105275	
1093	*C. fritschii*	-	-	-	UYC_RS0128730	
1691	*F. muscicola*	1,1482	gi	0	UYI_RS0102395	*mreD*
1691	*F. thermalis*	0,7123	gi	0	UYK_RS0105270	
1691	*C. fritschii*	-	-	-	UYC_RS0128735	
3113	*F. muscicola*	5,879	g	303	UYI_RS0111905	*fraC*
3113	*F. thermalis*	3,4387	g	230	UYK_RS0113410	
3113	*C. fritschii*	0,6284	g	26	UYC_RS0103260	

Der Vergleich der Wachstumsbedingungen und der Spezies eröffnet die Möglichkeit einer CDS-Suche, welche speziesübergreifend eine gemeinsame Änderung der Transkriptabundanz aufweisen (Abb. 28 u. Tab. 10). Zu diesem Zweck wurden zunächst die Änderungen der Transkriptabundanz innerhalb einer Spezies ermittelt. Anschließend wurden die Änderungen für orthologe TSSe zwischen den Spezies verglichen. Zwischen den Spezies wird eine ähnliche Änderung der Transkriptabundanz für bekannte und unbekannte orthologe Komponenten angenommen.

Abb**. 28** stellt den intragenerischen- (links) und den intergenerischen Vergleich (rechts) orthologer TSSe dar. Im intragenerischen Vergleich wurde in allen TSS-Klassen eine Korrelation zwischen den Änderungen der Transkriptabundanz ermittelt (Abb. 28, Pearson-Test). Im intergenerischen Vergleich wurde keine Korrelation der Änderungen beobachtet. Beim gTSS- (Abb. 28A), aTSS- (Abb. 28B) und iTSS-Vergleich (Abb. 28C) variierten die ermittelten Änderungswerte, die die 5% stärksten Änderungen der

Transkriptabundanz in mindestens einer der Spezies bestimmen (Abb. 28, rote Punkte). Dies liegt in erster Linie an der unterschiedlichen Anzahl orthologer TSSe pro TSS-Klasse. Im intragenerischen Vergleich lag die untere Grenze des Änderungswertes für orthologe gTSSe und aTSSe bei 2,58 und für orthologe iTSSe bei 2,62. Im intergenerischen Vergleich lagen die unteren Grenzen der Änderungswerte für orthologe gTSSe und aTSSe bei 2,53 und 2,52. Für orthologe iTSSe lag die Grenze bei 2,65. Orthologe Transkriptionsstartpunkte, die nur in einer Spezies für beide Bedingungen Transkripte aufwiesen, werden nicht dargestellt. Daher unterscheidet sich die Anzahl der dargestellten TSSe von der Gesamtanzahl ermittelter orthologer Transkriptionsstartpunkte pro TSS-Klasse (vgl. Tab. 6 u. Abb. 28). Ähnliche Änderungen der Transkriptabundanz zwischen den Spezies sind im ersten und dritten Quadranten (jeweils oben rechts und unten links) in den Streudiagrammen lokalisiert. Insgesamt wurden im intragenerischen Speziesvergleich 198 orthologe TSSe und im intergenerischen Speziesvergleich 74 orthologe TSSe mit dieser Vorgehensweise detektiert (Abb. 28). Von diesen Transkriptionsstartpunkten zeigten insgesamt 35 in der Kontrolle höhere Transkriptabundanz. Im Nachfolgenden wird auf drei Beispiele genomischer Regionen eingegangen, die mit dieser Vorgehensweise identifiziert werden konnten (Tab. 10).

6.4.4.1 Genomische Region der Einzelkopie-Proteinfamilie 1610

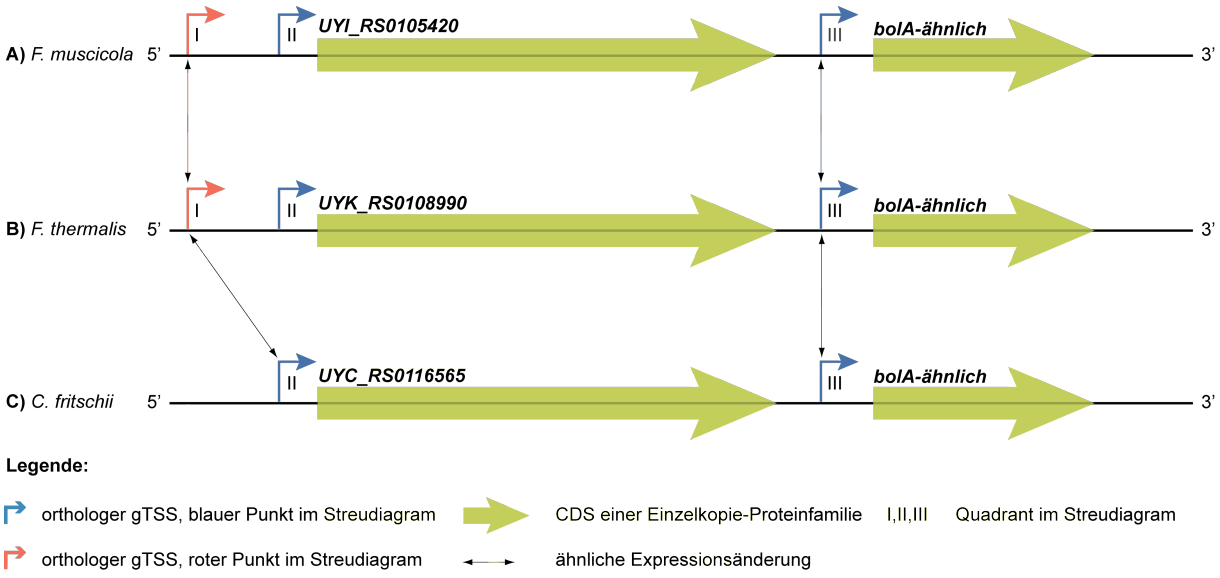


Abb. 29: Genomische Region der Einzelkopie-Proteinfamilie 1610.

A-C: Innerhalb der dargestellten Regionen befinden sich zwei CDSe. Für den hypothetischen CDS wurde ein orthologer TSS mit gemeinsamer Änderung der Transkriptabundanz zwischen *F. muscicola* (**A**) und *F. thermalis* (**B**) gefunden. Stromabwärts befindet sich der CDS *bolA*. Die römische Ziffer unter einem TSS soll die Zugehörigkeit zu einem Quadranten in Abb. 28A verdeutlichen.

Die Vorgehensweise der Analyse einer gemeinsamen Änderung der Transkriptabundanz zwischen den Spezies identifizierte eine Proteinfamilie mit einem als hypothetisch annotierten CDS (Abb. 29). Bei der dargestellten genomischen Region handelt es sich um ein potentielles bicistronisches Operon bestehend aus dem hypothetischen CDS (Tab. 10 u. Abb. 29) und dem stromabwärts liegenden CDS des Morphogens *bolA* (Aldea et al. 1988, Aldea et al. 1989). Für den hypothetischen CDS wurden in beiden *Fischerella* Spezies zwei orthologe gTSSe detektiert von denen einer in den Bedingungen des synchronisierten Morphotypen hochreguliert vorlag. Für die Spezies *C. fritschii* wurde ebenfalls ein in der NaCl-Bedingung hochregulierter orthologer gTSS identifiziert. Dieser liegt aber näher zum CDS als der hochregulierte TSS in den *Fischerella* Spezies. Das Gen *bolA* zeigte einen orthologen TSS für alle drei Spezies, welcher eine höhere Transkriptabundanz in den Kontrollbedingungen zeigte. Für die Proteinsequenz der hypothetischen CDS wurden keine konservierten Proteindomänen gefunden. Die Proteinsequenz des CDS ist in einer Proteinfamilie lokalisiert, in der Cyanobakterien aller Sektionen vorkommen. Für die Spezies *Calothrix* sp. PCC 6303 (Sektion IV) wird die orthologe Proteinsequenz als ein mögliches Membranprotein annotiert. Eine Vorhersage von Transmembrandomänen und Signalpeptiden mittels der Programme `TMHMM` (v. 2.0) (Krogh et al. 2001) und `SignalP` (v. 4.1) (Petersen et al. 2011) zeigt am N-Terminus des Proteins ein Signal für eine Transmembrandomäne.

6.4.4.2 Genomische Region der Einzelkopie-Proteinfamilien 1093 und 1691

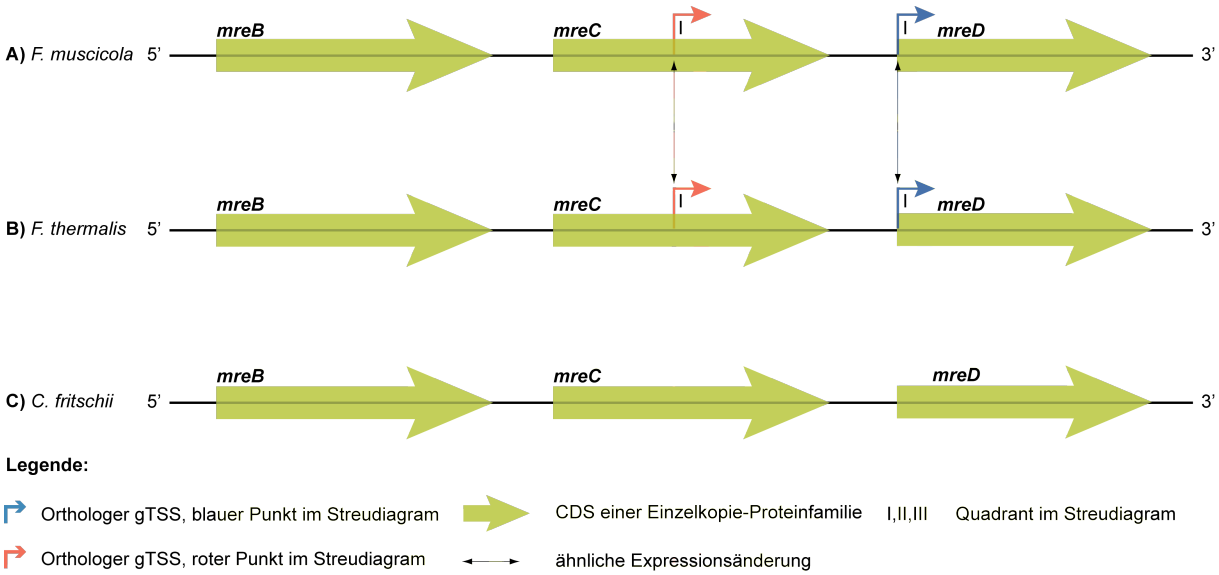


Abb. 30: Genomische Region der Einzelkopie-Proteinfamilien 1093 und 1691.

A-C: Innerhalb der dargestellten Regionen befinden sich drei CDSe. Bei *F. muscicola* (**A**) und *F. thermalis* (**B**) wurde im CDS *mreC* eine giTSS detektiert die im synchronisierten Morphotypen hochreguliert war. Stromabwärts befindet sich der CDS *mreD*. Für diesen CDS wurde ein orthologer gTSS im *Fischerella*-Vergleich detektiert, welcher direkt auf dem Adenin des Startkodon liegt. Die römische Ziffer unter einem TSS soll die Zugehörigkeit zu einem Quadranten in Abb. 28C verdeutlichen.

Bei der dargestellten genomischen Region handelt es sich um das *mre*-Operon bestehend aus *mreB*, *mreC* und *mreD* (Tab. 10 u. Abb. 30). Dem Operon werden essentielle Funktionen in der Zellmorphologie und Zellelongation zugeschrieben. Das Protein MreB ist ein Actinhomolog und Bestandteil des Zytoskeletts (Kruse et al. 2004, Singh and Montgomery 2011). Die Proteine MreC und MreD nehmen in *E. coli* strukturelle Funktionen in der Zellmembran während der Elongationsphase einer Zelle ein (Typas and Sourjik 2015). Für die *Fischerella* Spezies wurden für *mreC* und *mreD* orthologe TSSe gefunden. Der orthologe giTSS für *mreC* ermöglichte die Identifikation dieser Region und ist in Abb**. 28**C im *Fischerella* Vergleich ein roter Punkt. Die TSS für *mreD* zeigte in beiden *Fischerella* Spezies eine leicht höhere Transkriptabundanz im synchronisierten Morphotypen (Tab. 10). Diese TSS liegt direkt auf dem Adenin des Startkodon von *mreD*. Alternative TSSe sind in *C. fritschii* für dieses Operon nicht detektiert worden. Alle CDSe sind in Proteinfamilien lokalisiert, die in Cyanobakterien aller Sektionen vorkommen.

6.4.4.3 Genomische Region der Einzelkopie-Proteinfamilie 3113

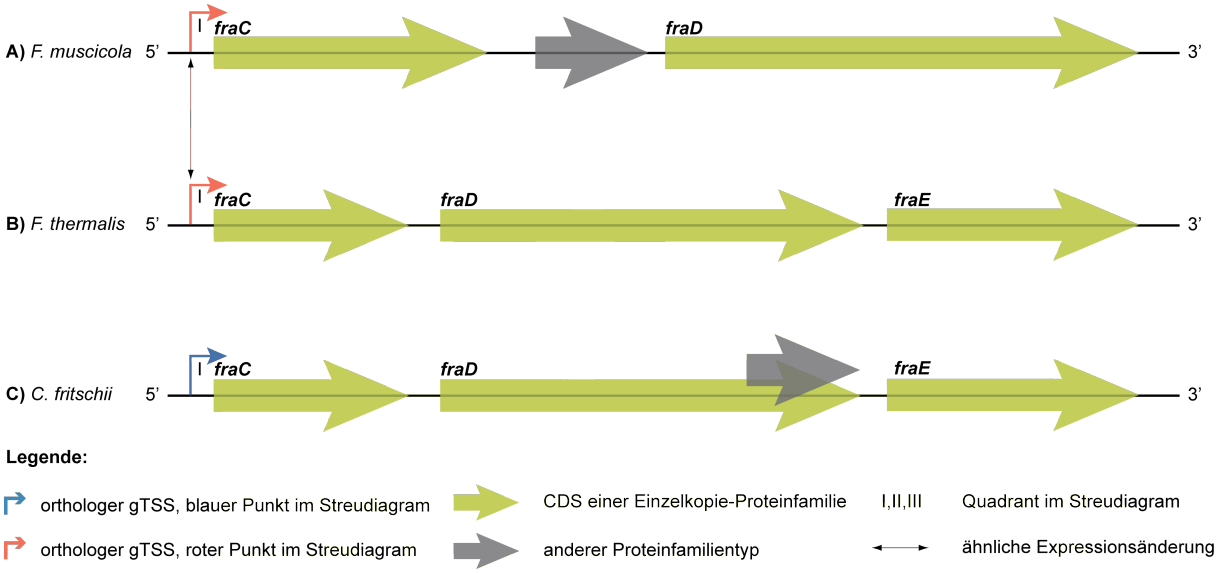


Abb. 31: Genomische Region der Einzelkopie-Proteinfamilie 3113.

A-C: Innerhalb der dargestellten Regionen befinden sich drei CDSe. Für den CDS *fraC* wurden in allen drei Spezies orthologe TSSe detektiert. Im *Fischerella*-Vergleich stellt der gTSS von *fraC* die höchste ähnliche Expressionsänderung der Analyse dar. Dieser TSS zeigte eine sehr hohe Transkriptabundanz im synchronisierten Morphotypen. In *C. fritschii* konnte keine klare Änderung der Transkriptabundanz beobachtet werden (Tab. 10). Mit Ausnahme von *F. muscicola* befinden sich stromabwärts die CDSe *fraD* und *fraE*. In *F. muscicola* wurde *fraE* in diesem Operon nicht gefunden und liegt in einer anderen Region. Die römische Ziffer unter einem TSS soll die Zugehörigkeit zu einem Quadranten in Abb. 28A verdeutlichen.

Die im *Fischerella* Vergleich stärkste Änderung der Transkriptabundanz eines orthologen TSS konnte in der genomischen Region des *fra*-Operon detektiert werden. Bei dem ersten CDS des Operon handelt es sich um *fraC*. Dieser CDS ist in *Anabaena* sp. PCC 7120 in einem Operon mit *fraD* und *fraE* beschrieben worden (Bauer et al. 1995, Merino-Puerto et al. 2010). In *F. thermalis* und *C. fritschii* wurde dieses Operon ebenfalls gefunden. In *F. muscicola* konnte *fraE* nicht mit *fraC* und *fraD* in einem Operon gefunden werden und liegt in einer anderen genomischen Region. Für *Anabaena* sp. PCC 7120 hat das Protein FraC einen Einfluss auf die Entwicklung von Zell-Zell Verbindungen zwischen zwei vegetativen Zellen und zwischen einer vegetativen Zelle und einer Heterozyste (Omairi-Nasser et al. 2015). In dieser Analyse wurde *fraC* und möglicherweise auch *fraD* von einem gTSS transkribiert, welcher im synchronisierten Morphotypen stark hochreguliert war. In *C. fritschii* wurde ebenfalls ein orthologer TSS gefunden. Für diesen TSS wurde aber keine Expressionsänderung in der NaCl-Bedingung beobachtet. Die Proteinsequenzen von FraC sind einer Proteinfamilie zugeordnet, die ausschließlich Mitglieder filamentöser Cyanobakterien (Sektion III-V) aufweist. Lediglich eine Ausnahme wurde beobachtet. *Synechococcus* sp. PCC 7335 (Sektion I) ist ebenfalls mit einer Proteinsequenz dieser Proteinfamilie zugeordnet. Es befinden sich aber

alle Sektion IV und Sektion V Cyanobakterien mit je einer Proteinsequenz in dieser Familie. Die Proteinsequenzen von FraD sind einer Proteinfamilie zugeordnet, die ausschließlich filamentösen Cyanobakterien zugeordnet werden kann. Auch in dieser Familie sind alle Sektion IV und Sektion V Cyanobakterien vorhanden. Für FraE sind die Mitglieder der Proteinfamilie wiederum allen Sektionen zugehörig. Alle drei Proteinfamilien stellen jedoch keine universellen Einzelkopie-Proteinfamilien dar, da mindestens eine Spezies einer Sektion immer abwesend ist.

Zusammenfassend wird festgehalten, dass im *Fischerella* Vergleich korrelierende Änderungen der Transkriptabundanz beobachtet wurden. Auf intergenerischer Ebene war dies nicht der Fall. Die Ursachen sind wahrscheinlich sowohl auf die unterschiedlichen Bedingungen der morphologischen Transitionen, als auch auf den generellen Unterschied der Transition in *C. fritschii* (aseriate Zellaggregate) zurückzuführen.

7 Diskussion

7.1 Sequenzierung und Read mapping

Mit dieser Arbeit konnten für *F. muscicola*, *F. thermalis* und *C. fritschii* genomweit primäre Transkriptionsstartpunkte in jeweils zwei unterschiedlichen Wachstumsbedingungen ermittelt werden. Die Anzahl und die Klassifizierung der Transkriptionsstartpunkte liegen nicht ausschließlich biologischen Ursachen zugrunde. Durch die automatisierte Vorgehensweise ist die Detektion stark von vielen einzelnen Parametern abhängig. Die richtige Parametereinstellung zugunsten von Spezifität oder Sensitivität kann sich daher als schwierig erweisen (Thomason et al. 2014). In Abschnitt 5.4 werden einige Parameter beschrieben, welche einen Einfluss auf die Anzahl der detektierten TSSe haben können. Die cDNA-Bibliothekserstellung kann in den ersten Schritten bereits Einfluss auf die TSS-Detektion nehmen (Thomason et al. 2014, Hoen et al. 2013).

Die cDNA-Bibliotheken der TEX(-)-Proben wurden im Vergleich zu den TEX(+)-Proben mit unterschiedlichen Protokollen erstellt (Abschnitt 10.2). In den TEX(-)-Proben sind ribosomale rRNA-Transkripte über das RiboZero rRNA Removal Kit (Epicentre)[14] degradiert worden. Der Einsatz eines solchen war in den TEX(+)-Proben nicht vorhanden. Hier wurde TEX eingesetzt, um Transkripte mit 5'-Monophosphatrest abzubauen. Ein weiterer Unterschied war die Reihenfolge der Polyadenylierung des 3'-Endes. Die Funktion der Polyadenylierung durch eine Poly(A)-Polymerase ist die Erzeugung eines Primer, der einer reversen Transkriptase die Transkription der RNA in eine cDNA ermöglicht. In den TEX(-)-Proben geschah dies vor Einsatz der Tobacco acid pyrophosphatase (TAP). In den TEX(+)-Proben wurden die 3'-Enden erst nach Anwenden von TAP polyadenyliert. Vermutlich ist dies sowohl der Grund für den hohen Anteil des Adenin in den TEX(-)-Reads als auch der Grund für den hohen Anteil an Homopolymeren in der Sequenzierung. Durch den hohen Homopolymeranteil innerhalb der Reads mussten diese entsprechend gekürzt werden, um signifikante Treffer mit `blastall` zu gewährleisten (Abschnitt 5.3). Eine Homopolymerlänge von 10 nt wurde gewählt, da eine Analyse von Homopolymeren innerhalb der Genome kein Auftreten längerer Polymere aufwies. Berücksichtigt wurden im Mapping nur Reads mit einer Länge von 20 nt oder länger (Abschnitt 5.3).

Während des Mappingverlaufs (Abschnitt 6.2) zeigte sich eine Eigenschaft der TEX(+)-Proben. Hier wurden im letzten Schritt hohe Verluste von Reads im Vergleich zu den

14 http://www.epibio.com

TEX(-)-Proben beobachtet. Es scheint sich hier um einen Zusammenhang zwischen rRNA-Operonregionen und TEX-Behandlung zu handeln. Die Kriterien des Mappings sahen vor, dass kein Read aus einer rRNA-Operonregion hervorgegangen ist. Dieses Kriterium galt auch für die TEX(+)-Reads. Da viele Reads aus der TEX(-)-Probe aufgrund von Homopolymeren verloren gingen und die TEX(+)-Proben geringe Anteile von Homopolymeren aufwiesen, ist es denkbar, dass Reads innerhalb der TEX(+)-Probe primäre Transkripte ribosomaler RNA darstellten. Der Einsatz eines rRNA-Degradierungskits kann die Ursache für den niedrigeren Verlust in den TEX(-)-Proben im letzten Schritt sein.

Die automatisierte TSS-Detektion ist von einem Anreicherungsfaktor auf Nukleotidebene abhängig. Daher wurde eine Normalisierung der Mappings durchgeführt. Die Anzahl der Reads pro Position, die Wahl des Anreicherungsfaktors, die TSS-Gruppierungsroutine und weitere Parameter bestimmten die Sensitivität der TSS-Detektion. Die Anzahl der TSSe ist ein Resultat des Versuchs eines Kompromisses zwischen Sensitivität und Spezifität. Während die Kriterien des Mappings relativ strikt angesetzt wurden und dadurch den Raum potentiell falsch gemappter Reads einschränkten, waren die Parameter der TSS-Detektion weniger strikt. Durch die Abwesenheit eines Schwellenwertes für die Schritthöhe blieben potentielle Transkripte mit niedriger Abundanz in der Sequenzierung detektierbar. Die Gruppierungsroutine `clustering distance` wurde hingegen wieder strikter angesetzt und auf 30 nt Fensterweite gesetzt. Des Weiteren wurde eine eigene Gruppierungsroutine mit einer Fensterweite von 35 nt verwendet. Die Gründe für eine striktere Gruppierung der detektierten TSSe lagen in der oben beschriebenen Kürzung der Reads und eine beobachtete Eigenschaft von `blastall` nicht in allen Fällen mit dem ersten Nukleotid ein Alignment zu erzeugen. Des Weiteren können Fehler in der Sequenzierung mögliche Faktoren eines Rauschens innerhalb der Mappings sein (Amman et al. 2014). Zusätzlich kam die Annahme hinzu, dass von einem Promotor auch nur ein Transkriptionsstart resultieren kann. Genomweit waren in den drei Spezies die Promotoren unbekannt. Daher wurde als Fensterweite für die Gruppierungsroutine 35 nt gewählt. Das -35-Element konnte mit dem Programm `MEME` nicht gefunden werden. Das -10-Element bzw. ein erweitertes -10-Element scheint jedoch vorhanden. Es ist denkbar, dass das -35-Element in den drei Cyanobakterien existent ist aber für weitaus weniger Stromaufwärtsbereiche und mit weniger eindeutigen Konsensusmotiven. Dies könnte die Detektion erschwert haben.

7.2 Die quantitative TSS-Verteilung

Abb**. 18** in Abschnitt 6.3.1 zeigt die Klasseneinteilung der detektierten TSSe bei einem UTR-Schwellenwert von 1.000 Nukleotiden Länge. Dieser Schwellenwert wurde hoch angesetzt. Die Analyse intergenischer Entfernungen zeigte große Unterschiede zwischen Median und Mittelwert (Abschnitt 5.5). Lange UTRs sind für CDSe jedoch nicht unmöglich und in der Spezies *Anabaena* sp. PCC 7120 sind vier TSSe für den CDS von HetR (heterocyst differentiation control protein) beschrieben. Zwei TSSe haben lange UTRs zur Folge (Mitschke, Vioque, et al. 2011).

Die Analyse soll potentielle alternative TSSe für ein CDS erfassen. Das Beispiel von HetR kann eine Ausnahme innerhalb einer Spezies darstellen. Es ist jedoch unbekannt welche transkriptionellen Muster den beobachteten Morphotypen erklären könnten. Daher soll die Analyse möglichst viele TSSe einem CDS zuordnen, um diese auf CDS-Ebene analysieren zu können. Die unterschiedlichen Alignmentmodi in der vergleichenden TSS-Analyse (Abschnitt 5.7) kompensieren mögliche Fehlklassifizierungen, da alle Perspektiven eines TSS abgedeckt werden konnten. Dieser hohe Schwellenwert führte jedoch zu hohen TSS-Anteilen intermediärer Klassen (Abb. 18). Tab**. 14** zeigt die TSS-Klassifizierungen bei variierenden Schwellenwerten. Zu erkennen ist, dass bei einem Schwellenwert ähnlich des Median der intergenischen Entfernungen eine Verteilung von 15% gTSS, 35% aTSS, 28% iTSS und 16% nTSS zu beobachten ist. 20% der intermediär klassifizierten TSSe lösen sich zu puren Klassen der aTSS, iTSS und nTSS auf. Die restlichen 6% verbleiben in den intermediären Klassen. Eine hohe aTSS-Aktivität im Vergleich zu anderen cyanobakteriellen Transkriptomanalysen mit gleicher Sequenziermethode ist das Resultat (Mitschke, Georg, et al. 2011, Mitschke, Vioque, et al. 2011, Pfreundt et al. 2014, Voigt et al. 2014, Kopf et al. 2015). Die Ergebnisse des 1.000 nt UTR-Schwellenwertes führten zu einer ähnlichen Beobachtung, sodass generell eine hohe aTSS-Transkription in *F. muscicola*, *F. thermalis* und *C. fritschii* angenommen werden kann.

Die Funktion der aTSSe ist schwierig aufzulösen, da dieser Datensatz mit Readlängen von maximal 50 nt begrenzt war. Welche Transkriptlängen von solchen Positionen entstehen ist unbekannt. RNAs, die von einem aTSS entstehen, haben vermutlich regulierende Funktionen, da sie zur mRNA des assoziierten Gens komplementär sind. Mögliche Funktionen sind in *cis* denkbar, wobei auch Funktionen in *trans* vorliegen könnten, wenn in anderen Regionen des Genoms weitere komplementäre Bereiche vorhanden sind.

Lediglich 22% der primären Transkriptome wurden der puren gTSS-Klasse zugeordnet (Abb. 18). Die gTSSe resultieren in mRNA-Transkripte des assoziierten Gens. Es können jedoch auch gTSSe, welche hinreichend weit vom assoziierten Gen entfernt liegen, mit kleinen

kodierenden Sequenzabschnitten assoziiert sein. Diese werden von Genvorhersage Programmen nicht detektiert und könnten für kleine Proteine kodieren (Sorek und Cossart 2010, Kopf und Hess 2015). Diese Arbeit konzentrierte sich hauptsächlich auf gTSSe, da diese über eine Analyse der Transkriptabundanz zu einer Identifizierung von Kandidatengenen führte.

Interne Transkriptionsstartpunkte werden im Speziesdurchschnitt in 16% der Fälle detektiert. Transkriptionsstartpunkte der iTSS-Klasse können auf Annotationsfehler des assoziierten Gens zurückzuführen sein. In einer Transkriptomanalyse des Cyanobakteriums *Synechocystis* sp. PCC 6803 konnten dadurch 58 Gene neu annotiert werden (Mitschke, Georg, et al. 2011). Interne TSSe können zu verkürzten alternativen mRNA-Transkripten führen, wenn alternative Startkodonen im selben Leseraster stromabwärts des TSS vorhanden sind. Die iTSSe können auch in kurze nicht komplementäre RNAs mit regulatorischer Funktion in *trans* resultieren. Für den CDS *ntcA*, dass den Transkriptionsfaktor NtcA (nitrogen-responsive regulatory protein) kodiert, wurden in *Synechocystis* sp. PCC 6803 solche iTSSe detektiert (Mitschke, Georg, et al. 2011). Microarray-, sowie Northern Blot Analysen untermauern die Existenz solcher Transkripte (Mitschke, Georg, et al. 2011).

Im Speziesdurchschnitt können ca. 9% aller Transkriptionsstartpunkte keinen Genen zugeordnet werden und fallen daher in die nTSS-Klasse. Transkripte, die von solchen Positionen entstehen können wie im oben beschriebenen Fall der gTSSe auch mit kleinen unbekannten kodierenden Sequenzabschnitten assoziiert sein. Des Weiteren können sRNAs mit regulatorischen Funktionen von solchen Regionen entstehen (Kopf und Hess 2015). Hier ist jedoch ebenfalls das Problem unbekannter Transkriptlängen gegeben, was die weiterführende Interpretation der nTSSe einschränkt.

Durch den festgelegten UTR-Schwellenwert bat sich die Möglichkeit die puren TSS-Klassen mit den intermediären TSS-Klassen zu vergleichen. Bei der TSS zu Startkodon Entfernungsanalyse zeigte sich zwischen gTSS-, giTSS- und gaTSS-Positionen ein Schnittpunkt bei 250 Nukleotiden Länge (Abb. 19). Dieser Schnittpunkt entspricht in etwa dem Median intergenischer Entfernungen von *F. muscicola*, *F. thermalis* und *C. fritschii*. Vermutlich wäre für eine Detektion primärer Transkriptionsstartpunkte ein Schwellenwert für UTR-Längen von 250 Nukleotiden ausreichend. Ob entferntere TSSe in den drei Spezies generell mit dem jeweiligen CDS assoziiert sind, kann nur eine TSS-Validierung im Labor zeigen. Die TSS-Lokalisierung innerhalb der Gene zeigte ebenfalls ein speziesübergreifendes Muster (Abb. 20). Während iTSSe vermehrt am 5'-Ende eines Gens zu finden waren, waren giTSSe vermehrt am 3'-Ende zu finden. Die Lokalisierung von giTSS-Positionen vermehrt am 3'-Ende könnte damit zu tun haben, dass diese TSSe nicht dem internen assoziierten CDS, sondern dem CDS stromabwärts zugehörig sind. Eine Lokalisierung der iTSSe am 5'-Ende wurde auch

in *Synechocystis* sp. PCC 6803 beobachtet (Mitschke, Georg, et al. 2011). Die Vermutung für giTSSe kann auch für die gaTSSe angenommen werden, so dass hier die Assoziation mit dem stromabwärtsliegenden CDS eher zutreffen könnte (Abb. 21). Es ist jedoch schwierig mit Entfernungsanalysen eine Aussage über die eindeutige Klassenzuweisung zu treffen. Die Beobachtung der CDS-Entfernungen bei Vorliegen einer intermediären TSS (Abb. 22), könnte auch auf eine Eigenschaft eines Operon deuten, welches differentiell transkribiert wird. Ein Operon, welches intermediäre giTSSe aufweist, könnte Hinweise auf die differentielle Expression der CDSe liefern. Das Vorliegen einer gaTSS lässt auf eine Lokalisierung zweier sehr naheliegender Operon in komplementärer Ausrichtung deuten. Es ist möglich, dass es sich jeweils um ein CDS zweier zueinander komplementärer Operon handelt. Ist dieser intermediäre TSS aktiv, könnte das komplementäre Operon inaktiviert werden. Der Promotor des Operon auf dem gleichen Strang wäre hingegen aktiv. Dieses Regulationsschema könnte eine Form des Exkludonmodells sein. Das Exkludonmodell wurde erstmals in einer Analyse der Spezies *Listeria monocytogenes* beschrieben (Sesto et al. 2013). Intermediäre TSS-Klassen könnten pure TSS-Klassen darstellen, welche die Regulation operonischer Strukturen als Funktion besitzen. Eine Analyse der intermediären TSS-Klassen auf Ebene des Operon wäre dann vergleichbar mit der Analyse purer TSS-Klassen auf CDS-Ebene. Hierfür müssten potentielle Operonstrukturen in der zu analysierenden Spezies ermittelt werden. Bei der Identifizierung von Kandidatengenen führte die Detektion eines intermediären TSS für *mreC* und *mreD* zu einer möglichen differentiellen Transkription des *mre*-Operon.

7.3 Orthologe TSSe

Orthologe TSSe bzw. speziesspezifische TSSe werden in dieser Analyse definiert, um die Analyse der Evolution transkriptioneller Regulation in den Stigonematales zu studieren. Orthologe TSSe kommen in mindestens zwei der hier analysierten Spezies vor. Orthologie eines TSS basiert auf der gemeinsamen TSS-Präsenz innerhalb einer identifizierten Einzelkopie-Proteinfamilie. Die Distanz zweier TSSe im Alignment entscheidet über die Orthologie. Der Schwellenwert für die Distanz wurde auf 35 nt gesetzt.

Die Anteile orthologer TSSe zwischen den *Fischerella* Spezies ist höher als auf intergenerischer Ebene. Dies wird beim Vergleich der gemeinsamen TSS-Präsenz für Einzelkopie-Proteinfamilien und der Anzahl orthologer TSSe deutlich (vgl. Tab. 5 u. Tab. 6). So sind beispielsweise im *Fischerella* Vergleich in 1.552 Einzelkopie-Proteinfamilien gTSSe in beiden Spezies vorhanden (Tab. 5). Es wurden 1.622 orthologe gTSSe im *Fischerella* Vergleich gefunden (Tab. 6A). Diese sind auf 1.221 unterschiedlichen Einzelkopie-Proteinfamilien verteilt. Da die Definition eines orthologen TSS die Detektion eines TSS in beiden Spezies voraussetzt, haben 79% der Einzelkopie-Proteinfamilien mit gemeinsamer TSS-Präsenz orthologe TSSe. Für den intergenerischen Vergleich sieht dies anders aus. 1.449 Einzelkopie-Proteinfamilien haben in *F. muscicola* und in *C. fritschii* eine gemeinsame TSS-Präsenz. Es konnten 824 orthologe TSSe gefunden werden (Tab. 6B). Diese sind jedoch nur auf 724 Einzelkopie-Proteinfamilien verteilt. Daher sind nur in 50% der Einzelkopie-Proteinfamilien mit gemeinsamer TSS-Präsenz auch orthologe TSSe vorhanden. Selbst Einzelkopie-Proteinfamilien mit gemeinsamer TSS-Präsenz müssen also nicht explizit konservierte Transkription aufweisen. Wird die phylogenetische Distanz erhöht, verringert sich die Möglichkeit einen orthologen TSS zu beobachten. In Abschnitt 6.4.2 werden die Anteile orthologer TSSe auf alle Einzelkopie-Proteinfamilien mit mindestens einem TSS in einer Spezies berechnet. Daher sind die Häufigkeiten orthologer TSSe in Tab**. 6** noch geringer. Betrachtet man die Anzahl speziesspezifischer TSSe für Einzelkopie-Proteinfamilien, muss eine generell hohe TSS-Diversität auf beiden Vergleichsebenen konstatiert werden.

Tendenzen einer transkriptionellen Konservierung werden aber gefunden. Es konnte ein Zusammenhang zwischen orthologer TSS-Anzahl und Proteinsequenzidentität von Einzelkopie-Proteinfamilien beobachtet werden. Im intergenerischen Vergleich, mit Ausnahme der iTSS-Klasse, konnte diese Beobachtung nicht gemacht werden (Abschnitt 6.4.2, Abb. 27B). Des Weiteren weisen im *Fischerella* Vergleich sehr ähnliche Proteinpaare auch oft ausschließlich orthologe TSSe auf (Tab. 7). Im Gegensatz dazu zeigten weniger ähnliche Proteinpaare oft ausschließlich speziesspezifische TSSe bzw. eine Variation aus speziesspezifischen- und orthologen TSS. Für den intergenerischen Vergleich konnten

diese Beobachtungen nicht gemacht werden. Interessant ist die Häufigkeit orthologer aTSSe im *Fischerella* Vergleich. Diese sind signifikant häufiger vorhanden als orthologe gTSSe (Tab. 18). Für den intergenerischen Vergleich ist genau das Gegenteil der Fall (Tab. 19). Auf den unterschiedlichen Verwandtschaftsebenen werden also unterschiedliche Signaturen orthologer TSSe beobachtet. Regulatorische Mechanismen im *Fischerella* Vergleich sind stärker konserviert. Bei steigender phylogenetischer Distanz ist die Transkription einer mRNA (orthologe gTSSe) stärker konserviert.

Eine für die aTSS- und iTSS-Klasse interessante Beobachtung wurde in der Analyse der TSS-Region gemacht. Im *Fischerella* Vergleich weisen TSS-Regionen dieser Klassen stärkere Ähnlichkeit auf als die assoziierten Proteinpaare. Auf intergenerischer Ebene wurde durch die Nichtablehnung der Nullhypothese zumindest eine ähnliche Tendenz beobachtet. Jedoch wiesen nur orthologe TSSe diese Eigenschaft auf. Vermutlich liegt der Grund dieser Beobachtung in der Lokalisierung orthologer TSSe innerhalb eines CDS. Orthologe iTSSe oder aTSSe könnten in stark konservierten Bereichen eines CDS liegen, beispielsweise in Domänen. Diese könnten funktionell essentiell sein und akkumulieren daher weniger Mutationen. Andere Regionen der CDSe könnten variabler sein. Speziesspezifische TSSe liegen in diesen variableren Regionen, weshalb diese TSS-Regionen unähnlicher zueinander sind als die Proteinsequenzen. Eine Eigenschaft synonymer Mutationen für speziesspezifische TSSe könnte hier zu beobachten sein. Während durch eine synonyme Mutation ein CDS in seiner Integrität nicht beeinträchtigt wird, könnte die transkriptionelle Regulation durch Entstehung einer speziesspezifischen TSS verändert werden. Speziesspezifische aTSSe oder iTSSe könnten durch unterschiedlichen Kodongebrauch charakterisiert sein. Synonyme Mutationen könnten einen direkten Effekt auf die transkriptionelle Regulation eines Gens haben und manifestieren sich in Form von alternativen, speziesspezifischen TSS.

7.4 Kandidatengene und transkriptionelle Regulation

Die Identifizierung von Kandidatengenen zeigte große Unterschiede zwischen dem intragenerischen und dem intergenerischen Vergleich. Auf der intragenerischen Ebene wurde eine Korrelation der Expressionsänderungen zwischen *F. muscicola* und *F. thermalis* beobachtet. Eine korrelierende Expressionsänderung konnte für den intergenerischen Vergleich nicht gefunden werden (vgl. Abb. 28, Pearson-Test). Vermutlich ist dies auf die unterschiedlichen Stressoren zurückzuführen, die in *C. fritschii* und den *Fischerella* Spezies eingesetzt wurden, um einen synchronisierten Morphotypen zu erzeugen. In den *Fischerella* Spezies wurde ein synchronisierter Morphotyp durch Saccharose induziert. In *C. fritschii* führte eine Induktion mit NaCl zu einem synchronisierten Morphotypen. Des Weiteren unterschieden sich die beobachteten Morphotypen zwischen den Gattungen. Die *Fischerella* Spezies verloren bzw. stellten das echte Verzweigungswachstum während der Saccharoseinduktion ein und bildeten lineare Filamente (Abb. 5). In *C. fritschii* wurde im Verlauf der NaCl-Induktion ein Morphotyp von aseriaten Zellaggregaten beobachtet (Abb. 5). Dieser Morphotyp konnte in *C. fritschii* bereits in früheren Arbeiten beobachtet werden (Evans et al. 1976). Die Entwicklung einer Morphologie in Form von aseriaten Zellaggregaten könnte ein Schutzmechanismus sein, um den osmotischen Druck zu verringern. Die Zelloberfläche, die dem NaCl-Medium direkt ausgesetzt ist, wird für jede Zelle verringert (Chakravarty et al. 2007). Dass andere transkriptionelle Netzwerke oder andere Expressionsraten für diesen Morphotypen verantwortlich sind, ist nicht ausgeschlossen, vermutlich ist dies der Grund warum keine Korrelation der Transkriptabundanz im intergenerischen Vergleich gefunden werden konnte.

Im *Fischerella* Vergleich wurden insgesamt 198 orthologe TSSe mit starker Ähnlichkeit in der Änderung der Transkriptabundanz identifiziert. Durch die Vorgehensweise einer vergleichenden Analyse der Transkriptabundanz konnten genomische Regionen identifiziert werden, die einen Einfluss auf die beobachteten morphologischen Transitionen haben könnten. Beispielsweise wurde stromaufwärts des Morphogens *bolA* ein hypothetisches Gen identifiziert. Dieser hypothetische CDS weist orthologe gTSSe in den *Fischerella* Spezies und in *C. fritschii* auf. In beiden *Fischerella* Spezies wurde für den TSS, der dem CDS am nächsten ist, keine Expressionsänderung beobachtet (Tab. 10 u. Abb. 28). Für den alternativen TSS konnte eine höhere Expression im synchronisierten Morphotypen beobachtet werden (Tab. 10 u. Abb. 28). Der erste TSS in den *Fischerella* Spezies ist ortholog zu einem TSS in *C. fritschii*. Dort wurde für diesen TSS jedoch eine Änderung der Transkriptabundanz ähnlich des alternativen TSS in den *Fischerella* Spezies beobachtet. Gleiche Änderungen der Transkriptabundanz für alle drei Spezies gehen in *C. fritschii* mit einer verkürzten UTR einher.

Es könnte sich um einen Promotorwechsel handeln. Für die Proteinsequenz des CDS wurden keine konservierten Proteindomänen detektiert. Die Proteinsequenz scheint in Cyanobakterien jedoch konserviert zu sein und ist für die Spezies *Calothrix* sp. PCC 6303 (Sektion IV) als ein mögliches Membranprotein annotiert. Eine Vorhersage von Transmembrandomänen und Signalpeptiden mittels der Programme `TMHMM` (v. 2.0) (Krogh et al. 2001) und `SignalP` (v. 4.1) (Petersen et al. 2011) zeigte am N-Terminus des Proteins ein Signal für eine Transmembrandomäne. Die Detektion dieser Transmembrandomäne, die Nachbarschaft zu einem Morphogen (*bolA*) und die Zuweisung der Proteinsequenz in eine Proteinfamilie mit Cyanobakterien aller Sektionen macht diesen CDS zu einem interessanten Kandidaten mit Einfluss auf morphologische Transitionen.

Für das stromabwärtsliegende Morphogen *bolA* sind bereits mehrere Funktionen bekannt. Im Cyanobakterium *Fremyella diplosiphon* (Sektion IV) wurde BolA als Repressor des *mre*-Operon beschrieben (Singh and Montgomery 2014). Bei Wachstum unter rotem Licht wurde BolA höher exprimiert als unter grünem Licht. Die höhere Expressionsrate hatte zur Folge, dass das *mre*-Operon unter rotem Licht reprimiert wurde, was in einer sphärischen Morphologie der Zellen resultierte. Unter der Einwirkung von grünem Licht wurde die Expression von BolA reprimiert, was zu einer erhöhten Expression des *mre*-Operon führte. Dies wiederum resultierte in stäbchenförmige Zellen. Für die synchronisierten Morphotypen in den *Fischerella* Spezies wurde die Abwesenheit des echten Verzweigungswachstums beobachtet. In dieser Analyse zeigt *bolA* eine höhere Transkriptabundanz in der Kontrollbedingung. Die höhere Transkriptabundanz von *bolA* in den Kontrollbedingungen könnte mit dem echten Verzweigungswachstum zusammenhängen. Die damit verbundene Repression des *mre*-Operon könnte ein Effekt auf die Zellmorphologie besitzen. Für BolA wurde auch eine regulatorische Funktion bei der Expression von Enzymen, die an der Peptidoglycansynthese beteiligt sind, beschrieben (Batista and Freire 2011). Das Gen *bolA* wird von einer Histidinkinase RcaE reguliert (Singh and Montgomery 2014). Diese ändert bei oben beschriebener unterschiedlicher Lichteinwirkung von einem phosphorylierten Zustand in einem unphosphorylierten Zustand. Der unphosphorylierte Zustand von RcaE führt zu einer Repression der *bolA*-Transkription (Singh and Montgomery 2014). Auch in *C. fritschii* wurde ein orthologer TSS für *bolA* detektiert, welcher eine ähnliche Transkriptabundanz wie die *Fischerella* Spezies aufwies. Das Gen *bolA* könnte weitere Expressionseffektoren in Form von Kinasen aufweisen, die den phosphorylierten Zustand durch Dephosphorylierung ändern und dann als Repressoren oder Aktivatoren in Frage kommen könnten.

BolA ist ein Repressor des *mre*-Operon. In der genomischen Region des *mre*-Operon wurden ebenfalls orthologe TSSe identifiziert (Abb. 30). Für *mreC* und *mreD* wurden in den *Fischerella* Spezies alternative TSSe detektiert. Auch für *mreB* konnten in dieser Analyse gTSSe in den *Fischerella* Spezies und in *C. fritschii* detektiert werden, die aber nicht für alle

drei Spezies ortholog sind und unterschiedliche bzw. keine Änderung der Transkriptabundanz im synchronisierten Morphotypen zeigten. Diese Beobachtung könnte im direkten Zusammenhang mit der BolA Expression stehen. Die TSSe von *mreC* und *mreD* zeigten erhöhte Transkriptabundanz im synchronisierten Morphotypen. Für *mreD* wurde der TSS direkt auf dem Adenin des Startkodon detektiert. In *F. diplosiphon* wird das Vorhandensein monocistronsicher Transkripte vermutet (Singh and Montgomery 2014). Es besteht daher die Vermutung, dass die Transkription einer monocistronischen Variante des *mreD* von einer Repression des *mre*-Operon durch BolA unabhängig ist und dadurch andere Funktionen aufweist als die polycistronische Variante. Der TSS würde zu einer mRNA führen, welche keine UTR besäße. Dies könnte die Translation beeinflussen. *C. fritschii* zeigte keinen alternativen TSS für *mreD*. Dies könnte auf das generell fehlende Potential eine monocistronische Variante zu transkribieren hinweisen. Für *mreC* wurde ebenfalls ein alternativer interner TSS detektiert. Hier könnte eine monocistronische- bzw. eine mit *mreD* bicistronische Variante im synchronisierten Morphotypen akkumuliert worden sein. Auch für *mreC* wurde in *C. fritschii* kein alternativer TSS gefunden. *C. fritschii* könnte nicht über das Potential verfügen monocistronische Varianten zu transkribieren.

Die stärkste gemeinsame Änderung der Transkriptabundanz wurde im *fra*-Operon detektiert. Das Operon besteht aus *fraC*, *fraD* und *fraE*. In *F. thermalis und C. fritschii* konnte dieses Operon gefunden werden. In *F. muscicola* ist *fraE* nicht Bestandteil des *fra*-Operon und liegt in einer anderen genomischen Region. FraC hat mit FraD einen Einfluss auf Zell-Zell Kontakte. Für FraC wird eine strukturgebende Funktion der Zellverbindungen angenommen, während FraD eine mögliche stabilisierende Funktion in dieser Phase besitzt (Omairi-Nasser et al. 2015). Diese Zellkontakte sind den sogenannten gab junctions der Metazoa sehr ähnlich (Nürnberg et al. 2015). Des Weiteren werden für FraC und FraD Funktionen in der Peptidoglycanerweiterung zugeschrieben (Omairi-Nasser et al. 2015). In dieser Analyse zeigte das Operon in den *Fischerella*-Spezies eine starke Änderung der Transkriptabundanz im synchronisierten Morphotypen. Für *C. fritschii* wurde nur eine sehr schwache bzw. keine Änderung des orthologen gTSS beobachtet. Eine Änderung der Transkriptabundanz von *fraC* und *fraD* im *Fischerella* Vergleich könnte auf eine erhöhte Zell-Zell Verbindungsdichte deuten. Die erhöhte Dichte könnte erhöhten Stoffaustausch zwischen den Zellen gewährleisten.

8 Zusammenfassung und Ausblick

In den Spezies *F. muscicola*, *F. thermalis* und *C. fritschii* konnten Konzentrationserhöhungen von Saccharose bzw. eine Konzentrationserhöhung von NaCl morphologische Transitionen induzieren (Abb. 5). Das primäre transkriptionelle Repertoire wurde für jede Bedingung und Spezies quantifiziert und auf komparativem Weg analysiert. Die Analyse der primären Transkriptome offenbart eine hohe aTSS-Transkription und eine hohe Anzahl von Transkriptionsstartpunkten in multiplen TSS-Klassen für kodierende Sequenzabschnitte in allen Spezies (Abschnitt 6.3). Durch die Einbindung von Einzelkopie-Proteinfamilien wurde eine komparative TSS-Analyse zwischen den Spezies ermöglicht (Abschnitt 6.4). Der Begriff „orthologer TSS“ wurde eingeführt, um die Evolution konservierter Transkriptome zu analysieren. Auf intragenerischer Ebene sind sich primäre Transkriptome ähnlicher als auf intergenerischer Ebene. Generell muss jedoch eine starke transkriptionelle Variabilität konstatiert werden. Während die aTSS-Klasse die meisten orthologen TSSe auf intragenerischer Ebene aufweist, werden auf intergenerische Ebene die meisten orthologen TSSe in der gTSS-Klasse gefunden. Im intragenerischen Vergleich können orthologe TSSe vermehrt für sehr ähnliche Proteinpaare gefunden werden. Ein linear positiver Trend zwischen orthologer TSS-Anzahl und Proteinsequenzidentität kann auf intergenerischer Ebene nicht beobachtet werden. Speziesspezifische TSSe einer Einzelkopie-Proteinfamilie sind vermehrt in variablen Regionen eines CDS lokalisiert. Diese TSSe können ein direktes Resultat synonymer Mutationen in kodierenden Sequenzabschnitten sein. Im *Fischerella* Vergleich wurde die Beobachtung einer korrelierenden Änderung der Transkriptabundanz gemacht (Abschnitt 6.4.4). Die Vorgehensweise einer komparativen Analyse der Änderung der Transkriptabundanz für orthologe TSSe ermöglichte die Identifizierung von Transkripten, die einen Einfluss auf den beobachteten Morphotypen haben könnten. In den genomischen Regionen von *bolA*, dem *mre*-Operon und dem *fra*-Operon konnten differentielle TSS-Aktivitäten detektiert werden und lassen einen Einflusses dieser genomischen Regionen auf morphologische Transitionen vermuten.

Die in dieser Arbeit beschriebenen genomischen Regionen und deren CDSe sind nur einige Beispiele von Kandidaten, welche einen potentiellen Einfluss auf den beobachteten Morphotypen haben könnten. Eine Vielzahl weiterer orthologer TSSe der gTSS-, aTSS- und iTSS-Klasse konnten im Vergleich der Änderung der Transkriptabundanz identifiziert werden (siehe Abschnitt 10.8). Viele sind mit hypothetischen CDS assoziiert. Unbekannte Funktionen kodierender Sequenzabschnitte, die multiplen regulatorischen Funktionen potentieller asRNAs, die große Anzahl differentiell aktiver TSSe ohne komparativen Orientierungspunkt

(Abschnitte 6.3.3) und die Transkriptlängen des primären Transkriptoms aus dieser Sequenzierung erschweren eine weitere zielgerichtete bioinformatische Analyse dieses Datensatzes. Weiterführende bioinformatische Analysen könnten sich auf die Promotorregionen der detektierten TSSe konzentrieren. Es wurde ein potentielles -10-Element in annähernd allen TSS-Regionen und TSS-Klassen identifiziert (Abschnitt 6.3.5). Die TSS-Stromabwärtsregionen könnten auf ribosomale Bindestellen und Riboswitches hin analysiert werden, um die Qualität der detektierten TSSe weiter zu steigern. Eine Analyse der TSS-Stromaufwärtsregionen könnte eine Analyse von Transkriptionsfaktorbindestellen einbinden. Sind Transkriptionsfaktorbindestellen genomweit detektiert, könnten diese mit Änderungen der Transkriptabundanz verknüpft werden. Des Weiteren könnten publizierte Primätranskriptome von Cyanobakterien die komparative Analyse der Transkriptabundanz und die Evolution konservierter TSSe auf weitere Gattungen der Cyanobakterien ausweiten. Durch den Orientierungspunkt von Einzelkopie-Proteinfamilien könnte die Analyse der identifizierten TSSe und Regionen auf Gattungen ohne Transkriptomsequenzierung extrapoliert werden. Beispielsweise könnten die anderen Spezies in den Einzelkopie-Proteinfamilien nach potentieller TSS-Präsenz in der gleichen TSS-Region abgesucht werden. Transkriptionsstartpunkte der nTSS-Klasse besitzen das Potential von ncRNAs. Von diesen Regionen könnten simulierte Transkripte mit unterschiedlichen Längen erstellt und mit Datenbankeinträgen für ncRNAs verglichen werden. Eine zusätzliche Analyse nach ORFs in diesen Regionen könnte potentielle kleine CDSe für kleine Proteine identifizieren. Diese Arbeit konnte zur Identifizierung bekannter und potentiell neuer morphologischer Komponenten in Cyanobakterien der Stigonematales beitragen und eröffnet weiterführende Analysen der hier detektierten primären Transkriptome.

9 Literaturverzeichnis

1. Aldea M, Garrido T, Hernandez-Chico C, *et al*. 1989. Induction of a growth-phase-dependent promoter triggers transcription of bolA, an Escherichia coli morphogene. *EMBO J* **8**:3923-3931.

2. Aldea M, Hernandez-Chico C, la Campa de AG, *et al*. 1988. Identification, cloning, and expression of *bolA*, an *ftsZ*-dependent morphogene of Escherichia coli. *J Bacteriol* **170**:5169-5176.

3. Altschul S, Madden TL, Schäffer AA, *et al*. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**:3389-3402.

4. Amman F, Wolfinger MT, Lorenz R, *et al*. 2014. TSSAR: TSS annotation regime for dRNA-seq data. *BMC Bioinformatics* **15**:89.

5. Angert ER. 2005. Alternatives to binary fission in bacteria. *Nat Rev Microbiol* **3**:214-224.

6. Badger JH, Olsen GJ. 1999. CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol* **16**:512-524.

7. Bailey TL, Boden M, Buske FA, *et al*. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**:W202-W208.

8. Bakke P, Carney N, DeLoache W, *et al*. 2009. Evaluation of three automated genome annotations for *Halorhabdus utahensis*. *PLoS One* **4**:e6291-e6291.

9. Barrick JE, Corbino KA, Winkler WC, *et al*. 2004. New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc Natl Acad Sci USA* **101**:6421–6426.

10. Batista GI, Freire P. 2011. BolA affects cell growth, and binds to the promoters of penicillin-binding proteins 5 and 6 and regulates their expression. *J Microbiol Biotechnol* **21**:243–251.

11. Bauer CC, Buikema WJ, Black K, *et al*. 1995. A short-filament mutant of *Anabaena* sp. strain PCC 7120 that fragments in nitrogen-deficient medium. *J Bacteriol* **177**:1520–1526.

12. Bekker A, Holland HD, Wang P-L, *et al*. 2004. Dating the rise of atmospheric oxygen. *Nature* **427**:117–120.

13. Bergman B, Sandh G, Lin S, *et al*. 2012. *Trichodesmium*--a widespread marine cyanobacterium with unusual nitrogen fixation properties. *FEMS Microbiol Rev* **37**:286–302.

14. Besemer J, Lomsadze A, Borodovsky M. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* **29**:2607–2618.

15. Campbell EA, Westblade LF, Darst SA. 2008. Regulation of bacterial RNA polymerase σ factor activity: a structural perspective. *Curr Opin Microbiol* **11**:121-127.

16. Canfield DE, Glazer AN, Falkowski PG. 2010. The evolution and future of Earth's nitrogen cycle. *Science* **330**:192-196.

17. Carpenter EJ, Romans K. 1991. Major role of the cyanobacterium *Trichodesmium* in nutrient cycling in the north atlantic ocean. *Science* **254**:1356-1358.

18. Hoffmann L, Castenholz RW. 2015. Bergey's Manual of Systematics of Archaea and Bacteria, Subsection V. Chichester, UK: John Wiley & Sons, Ltd.

19. Chakravarty R, Manna S, Ghosh A, *et al*. 2007. Morphological Changes in an *Acidocella* Strain in Response to Heavy Metal Stress. *Res J Microbiol* **2**:742-748.

20. Claessen D, Rozen DE, Kuipers OP, *et al*. 2014. Bacterial solutions to multicellularity: a tale of biofilms, filaments and fruiting bodies. *Nat Rev Microbiol* **12**:115-124.

21. Dagan T, Roettger M, Stucken K, *et al*. 2013. Genomes of Stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. *Genome Biol Evol* **5**:31-44.

22. Delcher AL, Harmon D, Kasif S, *et al*. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**:4636-4641.

23. Dugar G, Herbig A, Förstner KU, *et al*. 2013. High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter jejuni* isolates. *PLoS Genet* **9**:e1003495-e1003495.

24. Dühring U, Axmann IM, Hess WR, *et al*. 2006. An internal antisense RNA regulates expression of the photosynthesis gene *isiA*. *Proc Natl Acad Sci USA* **103**:7054-7058.

25. Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**:1575-1584.

26. Evans HE, Foulds I, Carr NG. 1976. Environmental Conditions and Morphological Variation in the Blue-Green Alga *Chlorogloea fritschii*. *J Gen Microbiol* **92**:147-155.

27. Ewing B, Hillier L, Wendl MC, *et al*. 1998. Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. *Genome Res* **8**:175-185.

28. Flores E, Herrero A. 2010. Compartmentalized function through cell differentiation in filamentous cyanobacteria. *Nat Rev Microbiol* **8**:39-50.

29. Garcia-Pichel F, Belnap J, Neuer S, *et al*. 2003. Estimates of global cyanobacterial biomass and its distribution. *Algo Stud* **109**:213-227.

30. Garcia-Pichel F, Pringault O. 2001. Microbiology. Cyanobacteria track water in desert soils. *Nature* **413**:380-381.

31. Giddings TH Jr., Staehelin LA. 1981. Observation of microplasmodesmata in both heterocyst-forming and non-heterocyst forming filamentous cyanobacteria by freeze-fracture electron microscopy. *Arch Microbiol* **129**:295-298.

32. Güell M, van Noort V, Yus E, *et al*. 2009. Transcriptome complexity in a genome-reduced bacterium. *Science* **326**:1268-1271.

33. Güell M, Yus E, Lluch-Senar M, *et al*. 2011. Bacterial transcriptomics: what is beyond the RNA horiz-ome? *Nat Rev Microbiol* **9**:658-669.

34. Harvey SC, Reynolds RP. 1987. A common structural feature in promoter sequences of *E. coli*. *Nucleic Acids Res* **15**:4973-4985.

35. Haugen SP, Berkmen MB, Ross W, *et al*. 2006. rRNA Promoter Regulation by Nonoptimal Binding of σ Region 1.2: An Additional Recognition Element for RNA Polymerase. *Cell* **125**:1069-1082.

36. Haugen SP, Ross W, Gourse RL. 2008. Advances in bacterial promoter recognition and its control by factors that do not bind DNA. *Nat Rev Microbiol* **6**:507-519.

37. Hoen PAC', Friedländer MR, Almlöf J, *et al*. 2013. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat Biotechnol* **31**:1015-1022.

38. Imamura S, Asayama M. 2009. Sigma factors for cyanobacterial transcription. *Gene Regul Syst Bio* **3**:65-87.

39. Jacob F, Monod J. 1961. Genetic Regulatory Mechanisms in the Synthesis of Proteins. *J Mol Biol* **3**:318-356.

40. Jäger D, Sharma CM, Thomsen J, *et al*. 2009. Deep sequencing analysis of the *Methanosarcina mazei* Go1 transcriptome in response to nitrogen availability. *Proc Natl Acad Sci USA* **106**:21878-21882.

41. Jorjani H, Zavolan M. 2014. TSSer: an automated method to identify transcription start sites in prokaryotic genomes from differential RNA sequencing data. *Bioinformatics* **30**:971-974.

42. Kisand V, Lettieri T. 2013. Genome sequencing of bacteria: sequencing, *de novo* assembly and rapid analysis using open source tools. *BMC Genomics* **14**:211.

43. Kopf M, Hess WR. 2015. Regulatory RNAs in photosynthetic cyanobacteria. *FEMS Microbiol Rev* **39**:301-315.

44. Kopf M, Klähn S, Scholz I, *et al*. 2014. Comparative analysis of the primary transcriptome of *Synechocystis* sp. PCC 6803. *DNA Res* **21**:527-539.

45. Kopf M, Möke F, Bauwe H, *et al*. 2015. Expression profiling of the bloom-forming cyanobacterium *Nodularia* CCY9414 under light and oxidative stress conditions. *ISME J* **9**:2139-2152.

46. Krogh A, Larsson B, Heijne von G, *et al*. 2001. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J Mol Biol* **305**:567-580.

47. Kruse T, Bork-Jensen J, Gerdes K. 2004. The morphogenetic *Mre*BCD proteins of *Escherichia coli* form an essential membrane-bound complex. *Mol Microbiol* **55**:78-89.

48. Kumar K, Mella-Herrera RA, Golden JW. 2010. Cyanobacterial heterocysts. *CSH Perspect Biol* **2**:a000315-a000315.

49. Lehner J, Zhang Y, Berendt S, *et al*. 2011. The morphogene *ami*C2 is pivotal for multicellular development in the cyanobacterium *Nostoc punctiforme*. *Mol Microbiol* **79**:1655-1669.

50. Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**:1851-1858.

51. Martin WF, Stoebe B, Goremykin V, *et al*. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* **393**:162-165.

52. Merino-Puerto V, Mariscal V, Mullineaux CW, *et al*. 2010. Fra proteins influencing filament integrity, diazotrophy and localization of septal protein SepJ in the heterocyst-forming cyanobacterium *Anabaena* sp. *Mol Microbiol* **75**:1159-1170.

53. Meyer F, Goesmann A, McHardy AC, *et al*. 2003. GenDB--an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res* **31**:2187-2195.

54. Mitchell JE, Zheng D, Busby JW, et al. 2003. Identification and analysis of “extended -10” promoters in *Escherichia coli*. *Nucleic Acids Res* **31**:4689-4695.

55. Mitschke J, Georg J, Scholz I, *et al*. 2011. An experimentally anchored map of transcriptional start sites in the model cyanobacterium *Synechocystis* sp. PCC 6803. *Proc Natl Acad Sci USA* **108**:2124-2129.

56. Mitschke J, Vioque A, Haas F, *et al*. 2011. Dynamics of transcriptional start site selection during nitrogen stress-induced cell differentiation in *Anabaena* sp. PCC 7120. *Proc Natl Acad Sci USA* **108**:20130-20135.

57. Miyagishima S, Wolk CP, Osteryoung KW. 2005. Identification of cyanobacterial cell division genes by comparative and mutational analyses. *Mol Microbiol* **56**:126-143.

58. Mori T, Johnson CH. 2001. Independence of circadian timing from cell division in cyanobacteria. *J Bacteriol* **183**:2439–2444.

59. Nicol JW, Helt GA, Blanchard SG, *et al*. 2009. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics* **25**:2730–2731.

60. Nürnberg DJ, Mariscal V, Bornikoel J, *et al*. 2015. Intercellular diffusion of a fluorescent sucrose analog via the septal junctions in a filamentous cyanobacterium. *MBio* **6**: e02109- e02114

61. Nürnberg DJ, Mariscal V, Parker J, *et al*. 2014. Branching and intercellular communication in the Section V cyanobacterium *Mastigocladus laminosus*, a complex multicellular prokaryote. *Mol Microbiol* **91**:935–949.

62. Omairi-Nasser A, Mariscal V, Austin JR, *et al*. 2015. Requirement of Fra proteins for communication channels between cells in the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. PCC 7120. *Proc Natl Acad Sci USA* **112**:E4458-E4464.

63. Petersen TN, Brunak S, Heijne von G, *et al*. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* **8**:785-786.

64. Pfreundt U, Kopf M, Belkin N, *et al*. 2014. The primary transcriptome of the marine diazotroph *Trichodesmium erythraeum* IMS101. *Sci Rep* **4**:6187.

65. Planavsky NJ, Asael D, Hofmann A, *et al*. 2014. Evidence for oxygenic photosynthesis half a billion years before the Great Oxidation Event. *Nat Geosci* **7**:283-286.

66. Pointing SB, Chan Y, Lacap DC, *et al*. 2009. Highly specialized microbial diversity in hyper-arid polar desert. *Proc Natl Acad Sci USA* **106**:19964-19969.

67. Santangelo TJ, Artsimovitch I. 2011. Termination and antitermination: RNA polymerase runs a stop sign. *Nat Rev Microbiol* **9**:319-329.

68. Sharp M, Li WH. 1987. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**:1281-1295.

69. Schirrmeister BE, de Vos JM, Antonelli A, *et al*. 2013. Evolution of multicellularity coincided with increased diversification of cyanobacteria and the Great Oxidation Event. *Proc Natl Acad Sci USA* **110**:1791-1796.

70. Sesto N, Wurtzel O, Archambaud C, *et al*. 2013. The excludon: a new concept in bacterial antisense RNA-mediated gene regulation. *Nat Rev Microbiol* **11**:75-82.

71. Shao W, Price MN, Deutschbauer AM, *et al*. 2014. Conservation of Transcription Start Sites within Genes across a Bacterial Genus. *MBio* **5**:e01398-e01412.

72. Sharma CM, Hoffmann S, Darfeuille F, *et al*. 2010. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* **464**:250-255.

73. Sharma CM, Vogel J. 2014. Differential RNA-seq: the approach behind and the biological insight gained. *Curr Opin Microbiol* **19**:97-105.

74. Shi T, Falkowski PG. 2008. Genome evolution in cyanobacteria: the stable core and the variable shell. *Proc Natl Acad Sci USA* **105**:2510–2515.

75. Shih PM, Wu D, Latifi A, Axen SD, *et al*. 2013. Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc Natl Acad Sci USA* **110**:1053–1058.

76. Singh SP, Montgomery BL. 2011. Determining cell shape: adaptive regulation of cyanobacterial cellular differentiation and morphology. *Trends Microbiol* **19**:278-285.

77. Singh SP, Montgomery BL. 2014. Morphogenes *bolA* and *mreB* mediate the photoregulation of cellular morphology during complementary chromatic acclimation in *Fremyella diplosiphon*. *Mol Microbiol* **93**:167-182.

78. Sorek R, Cossart P. 2010. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat Rev Genet* **11**:9-16.

79. Rippka R, Deruelles J, Waterbury JB, *et al*. 1979. Generic Assignments, Strain Histories and Properties of Pure Cultures of Cyanobacteria. *J Gen Microbiol* **111**:1-61.

80. Stazic D, Lindell D, Steglich C. 2011. Antisense RNA protects mRNA from RNase E degradation by RNA-RNA duplex formation during phage infection. *Nucleic Acids Res* **39**:4890-4899.

81. Stork M, Di Lorenzo M, Welch TJ, Crosa JH. 2007. Transcription termination within the iron transport-biosynthesis operon of *Vibrio anguillarum* requires an antisense RNA. *J Bacteriol* **189**:3479-3488.

82. Stucken K, Ilhan J, Roettger M, *et al*. 2012. Transformation and conjugal transfer of foreign genes into the filamentous multicellular cyanobacteria (subsection V) *Fischerella* and *Chlorogloeopsis*. *Curr Microbiol* **65**:552-560.

83. Stucken K, John U, Cembella A, *et al*. 2010. The smallest known genomes of multicellular and toxic cyanobacteria: comparison, minimal gene sets for linked traits and the evolutionary implications. *PLoS One* **5**:e9235–e9235.

84. Tatusov RL, Koonin EV, Lipman DJ. 1997. A Genomic Perspective on Protein Families. *Science* **278**:631-637.

85. Thomason MK, Bischler T, Eisenbart SK, *et al*. 2014. Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in *Escherichia coli*. *J Bacteriol* **197**:18-28.

86. Thompson AW, Foster RA, Krupke A, *et al*. 2012. Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. *Science* **337**:1546-1550.

87. Typas A, Sourjik V. 2015. Bacterial protein networks: properties and functions. *Nat Rev Microbiol* **13**:559-572.

88. Vogel J. 2003. Experimental and computational analysis of transcriptional start sites in the cyanobacterium *Prochlorococcus* MED4. *Nucleic Acids Res* **31**:2890-2899.

89. Voigt K, Sharma CM, Mitschke J, *et al*. 2014. Comparative transcriptomics of two environmentally relevant cyanobacteria reveals unexpected transcriptome diversity. *ISME J* **8**:2056-2068.

90. Wade JT, Grainger DC. 2014. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nat Rev Microbiol* **12**:647-653.

91. Wang D, Lloyd AH, Timmis JN. 2012. Environmental stress increases the entry of cytoplasmic organellar DNA into the nucleus in plants. *Proc Natl Acad Sci USA* **109**:2444-2448.

92. Wilk L, Strauss M, Rudolf M, *et al*. 2011. Outer membrane continuity and septosome formation between vegetative cells in the filaments of *Anabaena* sp. PCC 7120. *Cell Microbiol* **13**:1744–1754.

93. Wurtzel O, Sapra R, Chen F, *et al*. 2009. A single-base resolution map of an archaeal transcriptome. *Genome Res* **20**:133-141.

94. Wurtzel O, Sesto N, Mellin JR, *et al*. 2012. Comparative transcriptomics of pathogenic and non-pathogenic *Listeria* species. *Mol Syst Biol* **8**:583.

95. Zhelyazkova P, Sharma CM, Förstner KU, *et al*. 2012. The primary transcriptome of barley chloroplasts: numerous noncoding RNAs and the dominating role of the plastid-encoded RNA polymerase. *Plant Cell* **24**:123-136.

10 Anhang

10.1 Wachstumsbedingungen

Alle im folgenden beschriebenen Laborschritte wurden von Frau Dr. Karina Stucken, Frau MSc. Judith Ilhan und Frau BSc. Katharina Brandstädter durchgeführt. Die Stämme der Spezies *Fischerella muscicola* PCC 7414, *Fischerella thermalis* PCC 7521 und *Chlorogloeopsis fritschii* PCC 6912 wurden vom Pasteur Culture Collection (PCC) Institut für Cyanobakterien aus Frankreich erhalten. Die Ausgangskulturen sind unter photoautotrophen Bedingungen bei einer Lichtintensität von 30 µmol m^{-2} s^{-1} in einem 12 Stunden Licht- zu 12 Stunden Dunkelheitsrhythmus in BG11 oder BG11o Medien (Rippka et al. 1979) bei 37°C kultiviert worden.

Für alle experimentellen Kulturen wurden mindestens drei biologische Replikate der Ausgangskulturen von *F. muscicola*, *F. thermalis* und *C. fritschii* erstellt. Die biologischen Replikate wurden 30 Tage lang in BG11 und BG11o Medien kultiviert. Um einen synchronisierten Morphotypen innerhalb der Spezieskulturen zu erzeugen, wurden die Kulturen steigenden Saccharose- bzw. NaCl-Konzentrationen ausgesetzt. Vor der Konzentrationserhöhung sind Kontrollproben jeder Spezieskultur entnommen worden, um RNA aus den Ausgangsbedingungen (t = 0) zu extrahieren. In den *F. muscicola* und den *F. thermalis* Kulturen wurde anschließend eine ansteigende Saccharosekonzentration (0, 10, 100 mM) und in den *C. fritschii* Kulturen eine ansteigende NaCl-Konzentration (0, 50, 100, 250, 450, 600 mM) induziert. Während dieser 30 Tage sind die induzierten Kulturen mikroskopisch auf deren synchronisierten Morphotypen hin analysiert worden. Proben, welche diesen innerhalb einer Spezieskultur aufwiesen, wurden für die Extraktion der RNA ausgewählt.

Insgesamt wiesen in *F. muscicola* sechs biologische Replikate bei einer Saccharoseinduktion von 10 mM einen unverzweigten, filamentösen Morphotypen auf. Je eine Probe wurde für die Sequenzierung entnommen. In *F. thermalis* wiesen insgesamt drei biologische Replikate einen unverzweigten, filamentösen Morphotypen auf. Auch hier wurde je eine Probe für die Sequenzierung entnommen. Kulturen der Spezies C. *fritschii* zeigten in nicht induzierter Ausgangslösung unizelluläres bzw. multiseriates Wachstum in Form von Kolonien. Bei einer Induktion von 100 mM Natriumchlorid wurde jedoch ein synchronisierter Morphotyp von Zellaggregaten aufgrund osmotisch induzierten Stresses durch die NaCl-Konzentration hervorgerufen. Insgesamt zeigten drei biologische Replikate diese Zellaggregate und jeweils eine Probe wurde für die RNA-Extraktion entnommen.

10.2 RNA-Extraktion und Sequenzierung

Das Zellmaterial der entnommenen Proben ist mithilfe eines 8 µm Zellulosenitratfilters geerntet worden. Anschließend wurden die Proben in einem RNA Lysis Reagenz (Invitrogen) [15] resuspendiert, direkt in flüssigen Stickstoff eingefroren und bei -80°C gelagert. Die RNA wurde anschließend mit dem Concert Plant™ RNA Reagenz (Invitrogen)[13], wie in (Stucken et al. 2012) beschrieben, isoliert. Einige wenige Änderungen im Isolationsablauf wurden vorgenommen. Die gefrorenen Proben wurden auf Eis aufgetaut und in sechs Wiederholungen mit Intervallen von 30 Sekunden zusammen mit Glasperlen eines Durchmessers von 212-300 µm in einem Homogenisator aufgeschlossen (SpeedMill)[16]. Der Überstand wurde von den Glasperlen und den Zellresten mittels Zentrifugation separiert (10 Minuten, 12,000 x g, 4 °C). Der DNA Verdau wurde nach der RNA-Isolation durch Zugabe RNase-freier DNase (Thermo Scientific)[13] durchgeführt. Die Abwesenheit genomischer DNA ist mittels PCR getestet und die Quantität der RNA mittels Nanodrop (Thermo Scientific) ermittelt worden. Die RNA-Integrität wurde mit einem RNA Nanochip des Bioanalyzer 2100 (Agilent) überprüft. Bei geeigneter Qualität und Quantität wurden die Proben zur Sequenzierung übergeben.

Die cDNA-Bibliothekserstellung und Sequenzierung sind für die drei Spezies mit den jeweils zwei Wachstumsbedingungen (t = 0 und synchronisierter Morphotyp) von der Firma Vertis Biotechnologie AG [17] durchgeführt worden. Für jede Bedingung ist jeweils eine TEX(-)-und eine TEX(+)-Probe erstellt worden. Dabei wurden die TEX(-)-Proben abweichend im Vergleich zu den TEX(+)-Proben behandelt.

Bei den TEX(-) Proben wurde ribosomale RNA mit dem RiboZero rRNA Removal Kit (Epicentre) abgebaut, während dies für die TEX(+)-Proben nicht durchgeführt wurde. Im nächsten Schritt sind mit einer Poly(A)-Polymerase an den TEX(-)-Transkripten 3'-Poly(A)-Enden ligiert worden. Anschließend sind die Transkripte mit einer Tobacco acid pyrophosphatase (TAP) behandelt worden, um die 5'-Triphosphatreste der Transkripte in 5'-Monophosphatreste zu überführen. Danach wurde am 5'-Monophosphatrest ein RNA-Adapter Primer ligiert. Im cDNA-Syntheseschritt wurde ein oligo(dT)-Adapter Primer am 3'-Ende der Transkripte ligiert und eine M-MLV reverse Transkriptase eingesetzt, um die RNA in cDNA umzuschreiben. Die resultierende cDNA wurde anschließend mit dem Agencourt AMPure XP Kit[18] aufgereinigt.

[15] https://www.thermofisher.com/de/de/home.html
[16] https://www.analytik-jena.de/de.html
[17] http://www.vertis-biotech.com/
[18] http://www.beckmangenomics.com

Die TEX(+)-Transkripte wurden vor der Polyadenylierung mit einer Terminator Exonuklease (TEX) behandelt, um bereits Transkripte mit 5'-Monophosphatrest abzubauen. Direkt im Anschluss erfolgte die TAP-Behandlung und erst danach wurde das 3'-Ende polyadenyliert. Danach wurde am 5'-Ende ein RNA-Adapter Primer ligiert. Die Erstellung der cDNA mittels der M-MLV reversen Transkriptase und dem 3'-oligo(dT)-Adapter Primer sind analog zu den TEX(-)-Proben verlaufen. In den TEX(-)-Proben wurden folgende Adapter für die PCR Amplifikation, welche die Transkripte am 5'-Ende und am 3'-Ende flankieren, verwendet:

TrueSeq Sense Primer:

5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'

TrueSeq Antisense Primer:

5'-CAAGCAGAAGACGGCATACGAGAT-NNNNNN-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC(dT25)-3'

Bei den N's handelt es sich um Barcodesequenz, welche in jeder cDNA-Bibliothek variiert. Insgesamt wird das Transkript von 146 Nukleotiden flankiert. Die flankierenden Adapter der TEX(+)-Proben unterschieden sich im Vergleich zu den TEX(-) Proben in zwei zusätzlichen variablen Nukleotiden am 3'-Ende des TrueSeq Sense Primers.

TrueSeq Sense Primer:

5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNN-3'

TrueSeq Antisense Primer:

5'-CAAGCAGAAGACGGCATACGAGAT-NNNNNN-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC(dT25)-3'

Die Gesamtanzahl der flankierenden Nukleotide ist daher 148. Mit einer HiSeq2000 Sequenziermaschine (Illumina)[19] und einer Readlänge von 50 Nukleotiden wurden die cDNA-Bibliotheken sequenziert.

[19] http://www.illumina.com

10.3 Cyanobakterielle Spezies

Tab. 11: Cyanobakterielle Spezies in dieser Analyse.

Sektion	Spezies	JGI Taxon ID	Genom (bp)	Proteom	In Proteinfamilien	Ohne Proteinfamilie
I	*Acaryochloris marina* MBIC11017	641228474	8.361.599	8.409	6.551	1.858
I	*Acaryochloris sp.* CCMEE 5410	2513237397	7.875.477	7.512	6.571	941
I	Candidatus *Atelocyanobacterium thalassa*	646311970	1.443.806	1.199	1.138	61
I	*Chamaesiphon minutus* PCC 6605	2510436000	6.761.765	6.345	5.120	1.225
I	*Crocosphaera watsonii* WH 8501	638341074	6.238.156	5.958	5.324	634
I	*Cyanobacterium aponinum* PCC 10605	2503707009	4.176.973	3.559	3.253	306
I	*Cyanobacterium stanieri* PCC 7202	2503283023	3.163.381	2.886	2.682	204
I	*Cyanobium gracile* PCC 6307	2508501011	3.342.364	3.383	2.844	539
I	*Cyanobium sp.* PCC 7001	647533126	2.834.252	2.771	2.293	478
I	*Cyanothece sp.* BH63E, ATCC 51472	2507262054	5.460.476	5.109	4.938	171
I	*Cyanothece sp.* BH68, ATCC 51142	641522622	5.460.377	5.304	4.929	375
I	*Cyanothece sp.* CCY 0110	640612201	5.880.532	6.475	4.966	1.509
I	*Cyanothece sp.* PCC 7424	643348533	6.554.169	5.880	5.214	666
I	*Cyanothece sp.* PCC 7425	643348534	5.786.110	5.428	4.596	832
I	*Cyanothece sp.* PCC 7822	648028021	7.841.948	6.981	5.950	1.031
I	*Cyanothece sp.* PCC 8801	643348535	4.787.694	4.566	4.378	188
I	*Cyanothece sp.* PCC 8802	644736348	4.803.347	4.648	4.413	235
I	*Dactylococcopsis salina* PCC 8305	2509276056	3.781.008	3.594	3.140	454
I	*Geminocystis herdmanii* PCC 6308	2509601046	4.263.418	4.140	3.648	492
I	*Gloeobacter violaceus* PCC 7421	637000121	4.659.019	4.430	3.352	1.078
I	*Gloeocapsa sp.* PCC 73106	2508501033	4.025.114	4.087	3.612	475
I	*Gloeocapsa sp.* PCC 7428	2503754017	5.882.710	5.254	4.769	485
I	*Halothece sp.* PCC 7418	2503538028	4.179.170	3.862	3.498	364
I	*Microcystis aeruginosa* NIES-843	641522640	5.842.795	6.312	4.988	1.324
I	*Prochlorococcus marinus* AS9601	640069321	1.669.886	1.939	1.778	161
I	*Prochlorococcus marinus marinus* CCMP 1375	637000213	1.751.080	1.883	1.648	235
I	*Prochlorococcus marinus* MIT 9202	647533199	1.691.453	1.890	1.702	188
I	*Prochlorococcus marinus* MIT 9211	641228501	1.688.963	1.856	1.599	257
I	*Prochlorococcus marinus* MIT 9215	640753041	1.738.790	2.014	1.824	190
I	*Prochlorococcus marinus* MIT 9301	640069322	1.641.879	1.921	1.763	158
I	*Prochlorococcus marinus* MIT 9303	640069323	2.682.675	3.075	2.245	830
I	*Prochlorococcus marinus* MIT 9312	637000210	1.709.204	1.811	1.707	104
I	*Prochlorococcus marinus* MIT 9313	637000211	2.410.873	2.275	2.123	152
I	*Prochlorococcus marinus* MIT 9515	640069324	1.704.176	1.922	1.719	203
I	*Prochlorococcus marinus* NATL1A	640069325	1.864.731	2.204	1.846	358
I	*Prochlorococcus marinus* NATL2A	637000212	1.842.899	1.896	1.833	63
I	*Prochlorococcus marinus pastoris* CCMP 1986	637000214	1.657.990	1.719	1.656	63
I	*Synechococcus elongatus* PCC 6301	637000307	2.696.255	2.527	2.492	35
I	*Synechococcus elongatus* PCC 7942	637000308	2.742.269	2.662	2.612	50
I	*Synechococcus elongatus* PCC 7942 re	-	2.742.269	2.720	2.611	109
I	*Synechococcus sp.* BL107	639857006	2.283.377	2.507	2.278	229
I	*Synechococcus sp.* CB0101	649990022	2.686.395	3.010	2.483	527
I	*Synechococcus sp.* CB0205	649990023	2.427.308	2.719	2.269	450
I	*Synechococcus sp.* CC9311	637000309	2.606.748	2.893	2.408	485
I	*Synechococcus sp.* CC9605	637000310	2.510.659	2.701	2.435	266
I	*Synechococcus sp.* CC9616	2517093019	2.645.910	2.892	2.436	456
I	*Synechococcus sp.* CC9902	637000311	2.234.828	2.307	2.206	101

Tab. 12: Cyanobakterielle Spezies in dieser Analyse (Fortsetzung).

I	*Synechococcus sp.* JA-2-3B'a(2-13)	637000312	3.046.682	2.886	2.605	281
I	*Synechococcus sp.* JA-3-3Ab	637000313	2.932.766	2.836	2.645	191
I	*Synechococcus sp.* PCC 6312	2509276030	3.720.499	3.746	3.141	605
I	*Synechococcus sp.* PCC 7002	641522654	3.409.935	3.186	2.872	314
I	*Synechococcus sp.* PCC 7335	647533236	5.973.558	5.586	4.302	1.284
I	*Synechococcus sp.* PCC 7336	2506520048	5.140.668	4.715	3.785	930
I	*Synechococcus sp.* PE A1 60AY6Li	-	2.983.476	2.751	2.661	90
I	*Synechococcus sp.* PE A1-1 60AY4M2	-	3.162.791	2.705	2.648	57
I	*Synechococcus sp.* PE A1-1 63AY4M1	-	3.175.936	2.699	2.649	50
I	*Synechococcus sp.* PE A1-1 65AY640	-	3.155.434	2.656	2.587	69
I	*Synechococcus sp.* PE A4 65AY6A5	-	3.129.263	2.645	2.595	50
I	*Synechococcus sp.* PE A6 63AY4M2	-	3.093.920	2.628	2.578	50
I	*Synechococcus sp.* RCC 307	640427148	2.224.914	2.535	2.107	428
I	*Synechococcus sp.* RS9916	639857007	2.664.465	2.961	2.398	563
I	*Synechococcus sp.* RS9917	638341213	2.579.542	2.770	2.434	336
I	*Synechococcus sp.* WH 7803	640427149	2.366.980	2.533	2.372	161
I	*Synechococcus sp.* WH 8016	2507262052	2.706.690	2.990	2.534	456
I	*Synechococcus sp.* WH 8109	2563366603	2.118.903	2.577	2.126	451
I	*Synechococcus sp.* WH5701	638341214	3.043.834	3.346	2.741	605
I	*Synechococcus sp.* WH7805	638341215	2.620.367	2.883	2.404	479
I	*Synechococcus sp.* WH8102	637000314	2.434.428	2.528	2.275	253
I	*Synechocystis sp.* PCC 6803	2561511183	3.947.019	3.569	3.465	104
I	*Synechocystis sp.* PCC 6803, GT-I	2513237196	3.570.103	3.168	3.148	20
I	*Synechocystis sp.* PCC 6803, GT-S	651053076	3.571.103	3.171	3.152	19
I	*Synechocystis sp.* PCC 6803, PCC-N	2513237195	3.570.114	3.168	3.149	19
I	*Synechocystis sp.* PCC 7509	2517572074	4.908.825	4.859	4.325	534
I	*Thermosynechococcus elongatus* BP-1	637000320	2.593.857	2.476	2.239	237
II	*Chroococcidiopsis sp.* PCC 6712	2505679029	5.720.887	5.116	4.463	653
II	*Chroococcidiopsis thermalis* PCC 7203	2503538021	6.689.401	5.975	5.314	661
II	*Pleurocapsa sp.* PCC 7319	2509601013	7.386.997	6.690	5.594	1.096
II	Pleurocapsa sp. PCC 7327	2509276061	4.986.817	4.609	4.078	531
II	*Stanieria cyanosphaera* PCC 7437	2503754019	5.544.990	4.989	4.477	512
II	*Xenococcus sp.* PCC 7305	2508501034	5.929.641	5.373	4.514	859
III	*Arthrospira maxima* CS-328	642979357	6.003.314	5.690	5.329	361
III	*Arthrospira platensis* C1	2507262036	6.089.210	6.108	5.263	845
III	*Arthrospira platensis* NIES-39	650377906	6.788.435	6.630	6.165	465
III	*Arthrospira platensis* Paraca	645951858	4.997.563	5.370	4.748	622
III	*Arthrospira sp.* PCC 8005	648276619	6.145.553	5.675	5.292	383
III	*Crinalium epipsammum* PCC 9333	2504643013	5.620.407	5.002	4.347	655
III	*Filamentous Cyanobacterium* ESFC-1	2517572024	5.632.035	4.914	4.336	578
III	*Geitlerinema sp.* PCC 7105	2510065011	6.152.351	5.338	4.472	866
III	*Geitlerinema sp.* PCC 7407	2503538020	4.681.111	3.854	3.514	340
III	*Leptolyngbya boryana* PCC 6306	2509601031	7.262.454	6.827	5.610	1.217
III	*Leptolyngbya sp.* JSC-1	-	6.335.039	5.606	4.467	1.139
III	*Leptolyngbya sp.* PCC 6406	2517572073	5.776.957	5.261	4.434	827
III	*Leptolyngbya sp.* PCC 7375	2509601039	9.422.968	8.366	6.433	1.933
III	*Leptolyngbya sp.* PCC 7376	2503754048	5.125.950	4.601	3.927	674
III	*Lyngbya sp.* CCY 8106	639857035	7.037.511	6.142	5.162	980

Tab. 13: Cyanobakterielle Spezies in dieser Analyse (Fortsetzung).

III	*Microcoleus chthonoplastes* PCC 7420	647533184	8.679.041	8.294	5.939	2.355
III	*Microcoleus sp.* PCC 7113	2509276031	7.966.510	6.734	5.774	960
III	*Nodosilinea nodulosa* PCC 7104	2509601026	6.891.283	6.414	5.423	991
III	*Oscillatoria acuminata* PCC 6304	2509276028	7.804.270	6.004	5.165	839
III	*Oscillatoria formosa* PCC 6407	2508501075	6.894.060	5.693	5.413	280
III	*Oscillatoria nigro-viridis* PCC 7112	2503982035	8.272.254	6.925	5.944	981
III	*Oscillatoria sp.* PCC 10802	2509276047	8.594.406	7.012	5.523	1.489
III	*Oscillatoria sp.* PCC 6506	648276706	6.676.705	5.822	5.197	625
III	*Oscillatoriales sp.* JSC-12	2510065010	5.530.391	5.024	4.207	817
III	*Prochlorothrix hollandica* PCC 9006	2509276045	5.646.343	4.770	3.788	982
III	*Pseudanabaena sp.* PCC 6802	2506783054	5.621.883	5.363	4.473	890
III	*Pseudanabaena sp.* PCC 7367	2504643012	4.885.680	3.960	3.273	687
III	*Pseudanabaena sp.* PCC 7429	2504557005	5.476.421	4.774	3.990	784
III	*Spirulina major* PCC 6313	2506520014	5.050.153	4.408	3.881	527
III	*Spirulina subsalsa* PCC 9445	2506520011	5.323.600	4.580	4.077	503
III	*Trichodesmium erythraeum* IMS101	637000329	7.750.108	5.076	4.207	869
IV	*Anabaena cylindrica* PCC 7122	2503982047	7.063.285	6.182	5.579	603
IV	*Anabaena sp.* PCC 7108	2506485002	5.886.741	5.169	4.708	461
IV	*Anabaena variabilis* ATCC 29413	646564504	7.105.752	5.710	5.501	209
IV	*Calothrix sp.* PCC 6303	2503982036	6.960.392	5.785	5.210	575
IV	*Calothrix sp.* PCC 7103	2507262048	11.584.393	10.231	8.424	1.807
IV	*Calothrix sp.* PCC 7507	2505679032	7.023.215	6.166	5.579	587
IV	*Cylindrospermopsis raciborskii* CS-505	647000233	3.879.030	3.449	3.169	280
IV	*Cylindrospermum stagnale* PCC 7417	2509601025	7.610.589	6.642	5.749	893
IV	*Microchaete sp.* PCC 7126	2509601027	5.742.264	5.192	4.746	446
IV	*Nodularia spumigena* CCY9414	639857037	5.316.258	4.860	4.269	591
IV	*Nostoc azollae* 0708	648028001	5.486.145	5.321	4.166	1.155
IV	*Nostoc punctiforme* PCC 73102	642555144	9.059.191	6.690	6.164	526
IV	*Nostoc sp.* PCC 7107	2503707008	6.329.823	5.446	5.021	425
IV	*Nostoc sp.* PCC 7120	637000199	7.211.789	6.130	5.725	405
IV	*Nostoc sp.* PCC 7524	2509601032	6.718.869	5.603	5.138	465
IV	*Raphidiopsis brookii* D9	647000303	3.186.511	3.007	2.738	269
IV	*Rivularia sp.* PCC 7116	2510065008	8.728.773	6.880	6.015	865
IV	***Scytonema hofmanni* PCC 7110**	**-**	**11.924.780**	**12.356**	**8.823**	**3.533**
IV	*Scytonema hofmanni* UTEX 2349	2507262016	8.182.091	7.397	6.431	966
V	***Chlorogloeopsis fritschii* PCC 6912**	**-**	**7.514.644**	**7.439**	**6.631**	**808**
V	***Chlorogloeopsis fritschii* PCC 9212**	**-**	**7.643.933**	**7.571**	**6.765**	**806**
V	*Chlorogloeopsis sp.* PCC 7702	2512564012	4.885.827	4.283	3.957	326
V	***Fischerella muscicola* PCC 73103**	**-**	**7.573.719**	**7.231**	**5.652**	**1.579**
V	**Fischerella muscicola PCC 7414**	**-**	**7.008.764**	**7.167**	**5.450**	**1.717**
V	*Fischerella sp.* JSC-11	2505679024	5.380.000	4.627	4.401	226
V	*Fischerella sp.* PCC 9339	2516653082	8.008.257	6.720	6.063	657
V	*Fischerella sp.* PCC 9431	2512875027	7.167.426	6.104	5.618	486
V	*Fischerella sp.* PCC 9605	2516143000	8.079.181	7.060	6.193	867
V	***Fischerella thermalis* PCC 7521**	**-**	**5.508.398**	**5.340**	**4.418**	**922**
V	*Mastigocladopsis repens* PCC 10914	2517093042	6.465.655	5.846	5.121	725

10.4 UTR-Schwellenwert Analyse

Tab. 14: TSS-Verteilung auf die Klassen bei variierenden UTR-Schwellenwerten.

UTR-Schwellenwert (nt)	Klasse	*F. muscicola*	*F. thermalis*	*C. fritschii*	%
2000	gTSS	3.998	3.257	4.615	**25,51%**
	aTSS	2.612	2.114	3.588	**17,87%**
	iTSS	1.168	981	1.957	**8,83%**
	gaTSS	3.032	2.266	3.637	**19,20%**
	giTSS	3.340	2.559	4.599	**22,56%**
	aiTSS	19	6	8	**0,07%**
	gaiTSS	27	20	31	**0,17%**
	nTSS	941	765	985	**5,78%**
1000	gTSS	3.519	2.819	4.084	**22,40%**
	aTSS	4.004	3.147	5.217	**26,58%**
	iTSS	2.258	1.839	3.431	**16,18%**
	gaTSS	1.640	1.233	2.008	**10,49%**
	giTSS	2.250	1.701	3.125	**15,21%**
	aiTSS	28	12	14	**0,12%**
	gaiTSS	18	14	24	**0,12%**
	nTSS	1.420	1.203	1.516	**8,90%**
750	gTSS	3.326	2.678	3.891	**21,27%**
	aTSS	4.419	3.489	5.766	**29,39%**
	iTSS	2.708	2.202	4.086	**19,34%**
	gaTSS	1.225	891	1.459	**7,68%**
	giTSS	1.800	1.338	2.470	**12,05%**
	aiTSS	32	12	17	**0,13%**
	gaiTSS	14	14	22	**0,11%**
	nTSS	1.613	1.344	1.709	**10,03%**
500	gTSS	3.031	2.449	3.548	**19,41%**
	aTSS	4.913	3.869	6.319	**32,46%**
	iTSS	3.257	2.631	4.896	**23,18%**
	gaTSS	731	511	906	**4,62%**
	giTSS	1.251	909	1.660	**8,21%**
	aiTSS	34	14	17	**0,14%**
	gaiTSS	12	12	22	**0,10%**
	nTSS	1.908	1.571	2.052	**11,89%**
250	gTSS	2.327	1.869	2.798	**15,03%**
	aTSS	5.346	4.197	6.882	**35,30%**
	iTSS	3.926	3.114	5.859	**27,72%**
	gaTSS	298	183	343	**1,77%**
	giTSS	582	426	697	**3,66%**
	aiTSS	36	18	25	**0,17%**
	gaiTSS	10	8	14	**0,07%**
	nTSS	2.612	2.153	2.802	**16,26%**

10.5 CDSe mit hoher TSS-Anzahl

Tab. 15: Top-15 CDSe mit höchster gTSS-Anzahl.

*Nur das GenDB-Annotationssystem hat diesen CDS gefunden.

Spezies	NCBI Locus Tag	gTSSe	NCBI Produkt Annotation
F. muscicola	UYI_RS0119350	8	succinyl-CoA synthetase subunit beta
	UYI_RS0129730	7	chemotaxis protein CheY
	*FisPCC7414_1042	6	-
	*FisPCC7414_2761	6	-
	UYI_RS0101635	6	amine oxidase
	UYI_RS0127405	6	hypothetical protein
	UYI_RS0114085	6	hypothetical protein
	UYI_RS0120335	6	N-acetylglucosamine transferase
	*FisPCC7414_5540	6	-
	*FisPCC7414_5916	6	-
	UYI_RS0107680	6	aspartate-semialdehyde dehydrogenase
	UYI_RS0107200	6	chromosome segregation protein SMC
	*FisPCC7414_7600	6	-
	UYI_RS0104170	5	hypothetical protein
	UYI_RS0104475	5	rhomboid family protein
F. thermalis	UYK_RS0104710	7	hypothetical protein
	UYK_RS0109175	6	polysaccharide deacetylase
	UYK_RS0103790	6	NagC family transcriptional regulator
	UYK_RS0122520	6	helicase
	UYK_RS0115550	6	ferredoxin
	UYK_RS0114290	5	3-dehydroquinate synthase
	UYK_RS0113380	5	hypothetical protein
	UYK_RS0112700	5	ribosome maturation protein RimP
	UYK_RS0112140	5	putative transcriptional regulator, XRE family
	UYK_RS0111965	5	esterase
	UYK_RS0105720	5	8-amino-7-oxononanoate synthase
	UYK_RS0105770	5	hypothetical protein
	*Fischerella_sp._PCC7521_2146	5	-
	*Fischerella_sp._PCC7521_2513	5	-
	UYK_RS0108720	5	urease accessory protein UreG
C. fritschii	UYC_RS0107585	7	histidine kinase
	UYC_RS0109530	7	CTP synthetase
	UYC_RS0127780	7	hypothetical protein
	UYC_RS0118010	6	hypothetical protein
	*UYC_01350	6	-
	UYC_RS0132390	6	hypothetical protein
	UYC_RS0103070	6	Cl- channel, voltage gated
	UYC_RS0105425	6	hypothetical protein
	UYC_RS0105695	6	mannose-1-phosphate guanyltransferase
	UYC_RS0106290	6	hypothetical protein
	UYC_RS0106790	6	hypothetical protein
	*UYC_02928	6	-
	UYC_RS0110830	6	hypothetical protein
	UYC_RS0116800	6	hypothetical protein
	*UYC_05989	6	-

Tab. 16: Top-15 CDSe mit höchster aTSS-Anzahl.

*Nur das GenDB-Annotationssystem hat diesen CDS gefunden.

Spezies	NCBI Locus Tag	aTSSe	NCBI Produkt Annotation
F. muscicola	UYI_RS0119155	8	hypothetical protein
	UYI_RS0109875	8	hypothetical protein
	UYI_RS0121530	8	hypothetical protein
	*FisPCC7414_5865	8	-
	UYI_RS0102840	8	hypothetical protein
	*FisPCC7414_1024	7	-
	UYI_RS0105605	7	hypothetical protein
	*FisPCC7414_1895	7	-
	UYI_RS0109670	7	hypothetical protein
	UYI_RS0110625	7	6-deoxyerythronolide-B synthase
	*FisPCC7414_4948	7	-
	UYI_RS0123680	7	hypothetical protein
	UYI_RS0127310	7	hypothetical protein
	UYI_RS0108215	7	penicillin-binding protein
	*FisPCC7414_6599	7	-
F. thermalis	*Fischerella_sp._PCC7521_4240	9	-
	UYK_RS0121385	9	hypothetical protein
	UYK_RS0100700	8	chemotaxis protein CheY
	UYK_RS0108255	8	ABC transporter substrate-binding protein
	UYK_RS0123495	8	AAA ATPase
	UYK_RS0100515	7	hypothetical protein
	UYK_RS0119240	7	dynamin family protein
	UYK_RS0116495	7	transcriptional regulator
	UYK_RS0111605	7	porin
	*Fischerella_sp._PCC7521_4849	7	-
	UYK_RS0101025	7	hypothetical protein
	UYK_RS0118985	6	protein of unknown function DUF490
	UYK_RS0104050	6	hypothetical protein
	UYK_RS0114180	6	phosphoketolase
	UYK_RS0109185	6	hypothetical protein
C. fritschii	*UYC_03142	18	-
	UYC_RS0111205	13	hypothetical protein
	*UYC_07109	12	-
	*UYC_05229	11	-
	UYC_RS0103465	10	hypothetical protein
	UYC_RS0132505	9	hypothetical protein
	UYC_RS0106975	9	hypothetical protein
	*UYC_03937	9	-
	UYC_RS0116425	9	carbamoyl phosphate synthase large subunit
	UYC_RS0128295	9	histidine kinase
	*UYC_07806	9	-
	UYC_RS0130825	8	hypothetical protein
	UYC_RS0133385	8	hypothetical protein
	UYC_RS0133065	8	hypothetical protein
	UYC_RS0109535	8	N-acetylmuramoyl-L-alanine amidase

Tab. 17: Top-15 CDSe mit höchster iTSS-Anzahl.

*Nur das GenDB-Annotationssystem hat diesen CDS gefunden.

Spezies	NCBI Locus Tag	iTSSe	NCBI Produkt Annotation
F. muscicola	UYI_RS0111685	8	hypothetical protein
	UYI_RS0102620	8	hypothetical protein
	*FisPCC7414_6599	8	-
	*FisPCC7414_6797	8	-
	UYI_RS0103560	7	hypothetical protein
	UYI_RS0107315	7	hypothetical protein
	UYI_RS0120660	6	DNA topoisomerase I
	UYI_RS0119975	6	peptidase C39
	UYI_RS0113780	6	sodium:proton antiporter
	UYI_RS0104180	6	serine/threonine protein kinase
	*FisPCC7414_1024	6	-
	UYI_RS0111750	5	hypothetical protein
	UYI_RS0128300	5	hypothetical protein
	UYI_RS0119240	5	cell division protein FtsW
	UYI_RS0127985	5	alpha-glucan phosphorylase
F. thermalis	UYK_RS0122685	10	hypothetical protein
	UYK_RS0116840	7	serine/threonine protein kinase
	UYK_RS0108790	6	long-chain fatty acid--CoA ligase
	UYK_RS0100040	6	hypothetical protein
	UYK_RS0104880	6	competence protein
	*Fischerella_sp._PCC7521_4136	6	-
	UYK_RS0104370	6	mercuric reductase
	UYK_RS0111740	5	segregation protein A
	UYK_RS0113140	5	Peptidoglycan-binding domain 1 protein
	UYK_RS0119750	5	ligand-gated channel
	UYK_RS0119270	5	multi-sensor signal transduction histidine kinase
	*Fischerella_sp._PCC7521_5309	5	-
	UYK_RS0105790	5	glycosyltransferase
	*Fischerella_sp._PCC7521_4923	5	-
	UYK_RS0108230	5	hypothetical protein
C. fritschii	UYC_RS0125555	10	hypothetical protein
	*UYC_01239	9	-
	*UYC_03099	9	-
	UYC_RS0112335	9	hypothetical protein
	UYC_RS0118430	8	hypothetical protein
	UYC_RS0134535	8	hypothetical protein
	*UYC_01919	8	-
	UYC_RS0113005	8	adenylate cyclase
	UYC_RS0113630	8	hypothetical protein
	UYC_RS0117765	7	hypothetical protein
	UYC_RS0106270	7	metallophosphoesterase
	UYC_RS0106415	7	phosphoglucosamine mutase
	UYC_RS0113505	7	hypothetical protein
	UYC_RS0115265	7	ligand-gated channel
	UYC_RS0129210	7	ATPase

10.6 Promotoreigenschaften intermediärer TSS-Klassen

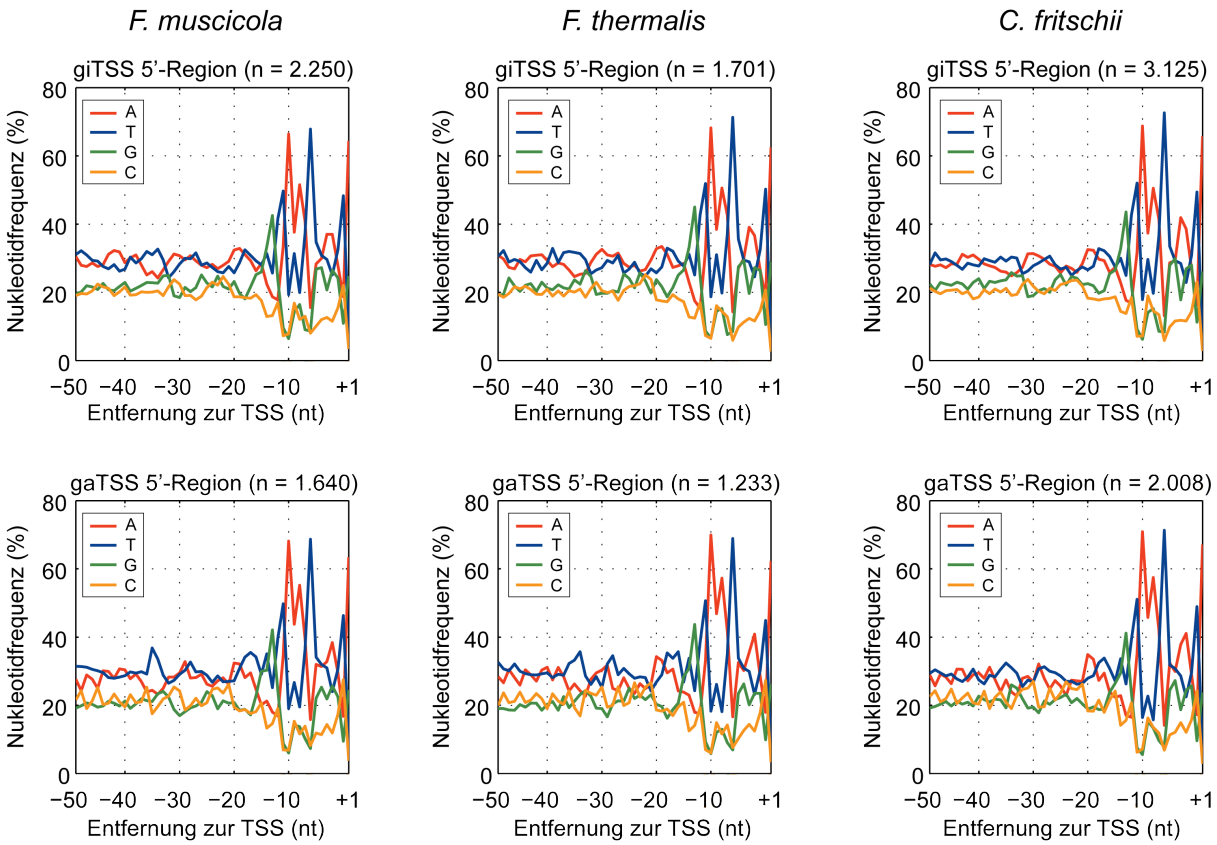


Abb. 32: Promotorregionen - intermediäre TSS-Klassen.

10.7 Exakter Test nach Fisher auf orthologe TSSe

Tab. 18: Einseitiger exakter Test nach Fisher - Intragenerischer Vergleich.

In dieser Tabelle wird der Vergleich der Anzahl orthologer TSSe in den puren TSS-Klassen gTSS, aTSS und iTSS für den *Fischerella* Vergleich dargestellt. Ein einseitiger exakter Test nach Fischer wurde durchgeführt. Es wurde getestet, ob eine der TSS-Klassen signifikant mehr oder weniger orthologe TSSe als die jeweils anderen aufweist. Durch das multiple Testverfahren wurden die p-Werte am Signifikanzniveau alpha = 5% korrigiert und sind in **schwarz** hervorgehoben.

A) Pure orthologe gTSSe vs. pure orthologe aTSSe

				Einseitiger exakter Test nach Fisher		
Kategorie	**pure gTSSe**	**pure aTSSe**	**Gesamt**		**H1 "grösser"**	**H2 "kleiner"**
Ortholog	796	994	1.790	p-Wert	9,91E-01	1,08E-02
Spezifisch	1.639	1.786	3.425	p-Wert (FDR-Korrektur)	9,91E-01	**3,23E-02**
Gesamt	2.435	2.780	5.215			

B) Pure orthologe gTSSe vs. pure orthologe iTSSe

				Einseitiger exakter Test nach Fisher		
Kategorie	**pure gTSSe**	**pure iTSSe**	**Gesamt**		**H1 "grösser"**	**H2 "kleiner"**
Ortholog	796	513	1.309	p-Wert	3,32E-01	6,93E-01
Spezifisch	1.639	1.091	2.730	p-Wert (FDR-Korrektur)	4,98E-01	9,95E-01
Gesamt	2.435	1.604	4.039			

C) Pure orthologe aTSSe vs. pure orthologe iTSSe

				Einseitiger exakter Test nach Fisher		
Kategorie	**pure aTSSe**	**pure iTSSe**	**Gesamt**		**H1 "grösser"**	**H2 "kleiner"**
Ortholog	994	513	1.507	p-Wert	6,11E-03	9,95E-01
Spezifisch	1.786	1.091	2.877	p-Wert (FDR-Korrektur)	**1,83E-02**	9,95E-01
Gesamt	2.780	1.604	4.384			

Interpretation: orthologe aTSSe > orthologe iTSSe, orthologe gTSSe

Tab. 19: Einseitiger exakter Test nach Fisher - Intergenerischer Vergleich.

In dieser Tabelle wird der Vergleich der Anzahl orthologer TSSe in den puren TSS-Klassen gTSS, aTSS und iTSS für den intergenerischen Vergleich dargestellt. Ein einseitiger exakter Test nach Fischer wurde durchgeführt. Es wurde getestet, ob eine der TSS-Klassen signifikant mehr oder weniger orthologe TSSe als die jeweils anderen aufweist. Durch das multiple Testverfahren wurden die p-Werte am Signifikanzniveau alpha = 5% korrigiert und sind in **schwarz** hervorgehoben.

A) Pure orthologe gTSSe vs. pure orthologe aTSSe

				Einseitiger exakter Test nach Fisher		
Kategorie	**pure gTSSe**	**pure aTSSe**	**Gesamt**		**H1 "grösser"**	**H2 "kleiner"**
Ortholog	449	314	763	p-Wert	2,20E-16	1,00E+00
Spezifisch	2.397	3.422	5.819	p-Wert (FDR-Korrektur)	**3,30E-16**	1,00E+00
Gesamt	2.846	3.736	6.582			

B) Pure orthologe gTSSe vs. pure orthologe iTSSe

				Einseitiger exakter Test nach Fisher		
Kategorie	**pure gTSSe**	**pure iTSSe**	**Gesamt**		**H1 "grösser"**	**H2 "kleiner"**
Ortholog	449	147	596	p-Wert	2,20E-16	1,00E+00
Spezifisch	2.397	2.195	4.592	p-Wert (FDR-Korrektur)	**3,30E-16**	1,00E+00
Gesamt	2.846	2.342	5.188			

C) Pure orthologe aTSSe vs. pure orthologe iTSSe

				Einseitiger exakter Test nach Fisher		
Kategorie	**pure aTSSe**	**pure iTSSe**	**Gesamt**		**H1 "grösser"**	**H2 "kleiner"**
Ortholog	314	147	461	p-Wert	1,22E-03	9,99E-01
Spezifisch	3.422	2.195	5.617	p-Wert (FDR-Korrektur)	**1,22E-03**	1,00E+00
Gesamt	3.736	2.342	6.078			

Interpretation: orthologe gTSSe > orthologe aTSSe > orthologe iTSSe

10.8 Informationen zur beigelegten CD

Microsoft Excel Tabelle.

„Tab20 - Proteinfamilien orthologer TSSe mit korrelierender Änderung der Transkriptabundanz.xlsx“

Diese Tabelle listet alle orthologen Transkriptionsstartpunkte mit korrelierender Änderung der Transkriptabundanz (Abb. 28, rote Punkte). Die Tabelle enthält zwei Arbeitsblätter (A-Intragenerischer Vergleich, B-Intergenerischer Vergleich). Ein orthologer TSS ist einer Proteinfamilie zugeordnet. Alle weiteren TSSe innerhalb dieser Proteinfamilie werden ebenfalls aufgelistet. Die Spaltendefinition lautet wie folgt:

Pfam = Laufnummer einer Einzelkopie-Proteinfamilie.

Spezies = Spezies in der ein TSS gefunden wurde.

Exp. Änderung = Änderung der Transkriptabundanz. Dieser Wert ist negativ, wenn der TSS in der Kontrollbedingung stärker transkribiert wurde. Dieser Wert ist positiv, wenn der TSS im synchronisierten Morphotypen stärker transkribiert wurde. Ein Wert von „-9“ bedeutet, dass eine Spezies innerhalb der Bedingung des synchronisierten Morphotypen keine Readinformation für diesen TSS aufwies. Ein Wert von „9“ bedeutet, dass eine Spezies innerhalb der Kontrollbedingung keine Readinformation für diesen TSS aufwies.

TSS-Typ = Die TSS-Klasse (Abb. 18), in der ein TSS eingeordnet wurde.

Entfernung zum 3’ CDS = Die Entfernung zum nächsten Adenin des Startkodon eines CDS, welcher Stromabwärts vom TSS liegt.

TSS ID = Innerhalb einer Proteinfamilie konnten orthologe (c) und speziesspezifische (u) TSSe vorliegen. Zusätzlich können für eine Proteinfamilie orthologe gTSSe (g), orthologe aTSSe (a) und orthologe iTSSe (i) vorhanden sein. Um ein orthologes TSS-Paar zu kennzeichnen, wird eine Laufnummer innerhalb einer Proteinfamilie verwendet. So steht beispielsweise „c1g“ für ein orthologes gTSS-Paar, „c1a“ für ein orthologes aTSS-Paar, „c1i“ für eine orthologes iTSS-Paar, „u1g“ für einen speziesspezifischen gTSS, „u1a“ für einen speziesspezifischen aTSS und „u1i“ für einen speziesspezifischen iTSS.

CDS ID = GenDB-Laufnummer einer kodierenden Sequenz auf dem jeweiligen Genom.

Locus Tag = NCBI Locus Tag. Für einige CDSe konnte ein NCBI Locus Tag ermittelt werden. Diese Locus Tags können verwendet werden, um den entsprechenden CDS beim NCBI wiederzufinden.

CDS Annotation = NCBI Annotation des CDS-Produkts.

Danke

Zuallererst möchte ich mich bei Frau Prof. Dr. Tal Dagan für die Möglichkeit bedanken, dass ich an diesem Projekt arbeiten durfte. Ihre Zeit, ihre Ideen, ihre Ratschläge und ihre Geduld während der gesamten Projektzeit waren letztendlich wichtige Bestandteile, welche diese Arbeit erst ermöglichten. Ich bedanke mich dafür, dass mir die Möglichkeit gegeben wurde meine Arbeit auf internationalen Konferenzen vorzustellen.

Ich danke Frau Prof. Dr. Ruth Schmitz-Streit für das meiner Arbeit entgegengebrachte Interesse und ihre Bereitschaft das Korreferat übernommen zu haben.

Ich möchte mich bei Frau Dr. Karina Stucken bedanken. Ohne ihre Hilfe wäre dieses Projekt nicht zustande gekommen. Danke für die Unterstützung, Hilfe und Diskussionsbereitschaft von Beginn an.

Frau Dr. Anne Kupczok danke ich für die Betreuung und Hilfestellung besonders in der Schlussphase dieses Projekts. Danke für das Korrekturlesen und die nützlichen Vorschläge, welche sehr geholfen haben.

Danke Frau Dr. Alexandra-Sophie Roy für einen wichtigen Rat in einer wichtigen Phase der Arbeit.

Ich danke Herrn Dr. Giddy Landan für die kritische Analyse meiner auf Konferenzen präsentierten wissenschaftlichen Poster und Hilfe in MatLab Fragen.

Ich möchte mich bei Herr Dr. Christian Wöhle für erhellende Gespräche im Dunkel des Themas Next Generation Sequencing bedanken.

Ich danke Dr. David Bogumil, Dr. Ovidiu Popa, Dr. Julia Weißenbach und Judith Ilhan. Ich kenne euch jetzt seit Beginn meines F-Praktikums in Düsseldorf und möchte mich bei euch sehr herzlich für eine schöne Zeit bedanken. Danke Judith und Julia für das kritische Korrekturlesen.

Ich danke allen Mitarbeitern des Instituts für Genomische Mikrobiologie für die nette Arbeitsatmosphäre.

Zuletzt möchte ich mich bei meiner Familie bedanken. Ein besonderer Dank geht an Cedrik Koch und Hendrike Mentler für die Unterstützung während der letzten drei Jahre.