Tiefencharakterisierung des intestinalen Transkriptoms der Maus mittels RNA-Seq

Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Ulrich C. Klostermeier

Kiel

Juli 2012

Referent:	Prof. Dr. Dr. h.c. Thomas C. G. Bosch
Korreferent:	Prof. Dr. Philip Rosenstiel
Tag der mündlichen Prüfung:	19. September 2012
Zum Druck genehmigt:	19. September 2012

gez. Prof. Dr. Wolfgang J. Duschl, Dekan

Inhaltsverzeichnis

Abkürzungsverzeichnis

3‘ UTR	3‘ *untranslated region*
5‘ UTR	5‘ *untranslated region*
Abb.	Abbildung
ApoB	Apolipoprotein B
AMP	Adenosinmonophosphat
APS	Adenosin-5‘-Phosphosulfat
ATP	Adenosintriphosphat
BMP	*bone morphogenetic protein*
CAGE	*cap analysis gene expression*
cDNA	Komplementär-DNA (*complementary DNA*)
ChIPseq	*chromatin immunoprecipitation sequencing*
cRNA	Komplementär-RNA (*complementary* RNA)
dATP	Desoxy-Adenosintriphosphat
dATP□S	Desoxy-Adenosin-5'-O-(1-Thio-Triphosphat)
ddATP	Didesoxy-Adenosintriphosphat
ddNTP	Didesoxy-Nukleotidtriphosphat
DNA	Desoxyribonukleinsäure (*desoxyribonucleic acid*)
DNase	Desoxyribonuklease
dNMP	Desoxy-Nukleotidmonophosphat
dNTP	Desoxy-Nukleotidtriphosphat
ds-cDNA	doppelsträngige cDNA (*double stranded* cDNA)
DSS	Natriumdextransulfat (*dextran sodium sulfate*)
DTT	Dithiothreitol
ePCR	Emulsions-Polymerasekettenreaktion (emulsion polymerase chain reaction)
EST	*expressed sequence tag*
FPKM	*fragments per kilobase of transcript per million fragments mapped*
GALT	*gut associated lymphoid tissue*
lincRNA	*long intergenic noncoding* RNA
MAMP	*microbe associated molecular pattern*
mRNA	Boten-RNA (*messenger* RNA)
miRNA	*micro* RNA
ncRNA	nicht-kodierende RNA (*non coding* RNA)
n. exp.	nicht exprimiert

NOD2	*nucleotide-binding oligomerization domain containing 2*
nt	Nukleotide
nTAR	nicht-annotierte, transkriptionell-aktive Region
P_i	anorganisches Phosphat (*phosphate, inorganic*)
piRNA	*piwi interacting* RNA
PP_i	Pyrophosphat (*pyrophosphate, inorganic*)
PSAP	Prosaposin
RIN	*RNA integrity number*
RMA	*robust multi-array average*
RNA	Ribonukleinsäure (*ribonucleic acid*)
RNase	Ribonuklease
RNA-Seq	RNA-Sequenzierung (hier im engeren Sinne mittels *next generation sequencing*)
RTq-PCR	*real time* quantitative Polymerasekettenreaktion
SAGE	*serielle analysis of gene expression*
SHH	*sonic hedgehog*
siRNA	*small interfering* RNA
snRNA	*small nuclear* RNA
snoRNA	small nucleolar RNA
SO_4^{2-}	Sulfat
SPRI	*solid phase reversible immobilisation*
tRNA	Transfer-RNA
WTAK	*whole transcriptome analysis kit*

Abbildungsverzeichnis

Tabellenverzeichnis

Formelverzeichnis

1. Einleitung

1.1. Der Aufbau des Darmtrakts von Säugetieren

Das Intestinum höherer Tiere und des Menschen ist ein komplexes Organ unter Beteiligung einer ganzen Reihe unterschiedlicher Zelltypen. Es erstreckt sich vom Pförtner des Magens bis zum After und wird grob in Dünndarm (*Intestinum tenue*) und Dickdarm (*Intestinum crassum*) gegliedert. In einer feineren Abstufung wird innerhalb des Dünndarms zwischen dem Zwölffingerdarm (*Duodenum*) als ersten Abschnitt direkt im Anschluss an den Magen, dem Leerdarm (*Jejunum*) als Mittelstück und dem Krummdarm (*Ileum*) unterschieden, der mit der Iliozäkalklappe in den Dickdarm übergeht. Im Dickdarm findet sich zunächst der Blinddarm (*Caecum*) mit dem Wurmfortsatz (*Appendix*). Das weite Mittelstück wird als Grimmdarm (*Colon*) bezeichnet, insbesondere beim Menschen wird hier aus der anatomischen Lokalisation noch das aufsteigende Colon (*Colon ascendens*), das Quercolon (*Colon transversum*), das absteigende Colon (*Colon descendens*) sowie die Sigma-Schlinge (*Colon sigmoideum*) unterschieden. Der Dickdarm endet mit dem Mastdarm (*Rectum*) als Übergang in den After. Die primäre Funktion des Dünndarms ist die Aufnahme von Nährstoffen aus der Umwelt. Dies bedingt insbesondere die Ausbildung einer großen Oberfläche, um einen Austausch effektiv zu gestalten. Das Colon fungiert insbesondere zur (Rück-) Resorption von Wasser und Elektrolyten.

Auch wenn der anatomische Feinaufbau der einzelnen Darmabschnitte im Detail voneinander abweicht, ist bei typischen Säugetieren wie Maus oder Mensch der prinzipielle Wandaufbau in allen Bereichen des Darms identisch. Zum Lumen hin wird die Darmwand durch die Schleimhaut oder Mukosa (*Tunika mucosa*) abgegrenzt, die aus einem einschichtigen Zylinderepithel (Enterozyten) und einer schmalen Schicht Bindegewebe (*Lamina propia*) besteht. Die nächste Schicht, die Submukosa (*Tela submucosa*), besteht primär aus lockerem Bindegewebe, beinhaltet aber auch viele Blutgefäße und Nervengewebe (*Plexus submucosa*) sowie in vielen Bereichen auch Drüsengewebe (z.B. die Brunner-Drüsen im *Duodenum*). Darauf folgt eine Muskelschicht (*Tunica muscularis*), die im Darmtrakt immer aus glatten Muskelzellen aufgebaut ist, dabei aber auch nervöse Strukturen (*Plexus myentericus*) sowie Lymph- und Blutgefäße enthält. Intraperitoneal gelegene Bereiche des Darms grenzen sich gegen die Bauchhöhle durch die *Tunica serosa* ab. Hier findet sich neben bindegewebigen Anteilen ein einschichtiges Plattenepithel. Die äußere Schicht des Darms, die nicht direkt frei in der Bauchhöhle liegt, wird als *Tunica adventitia* bezeichnet (Lüllmann-Rauch 2003). In **Abb. 1** ist der Aufbau der Darmwand mit den wichtigsten Strukturen graphisch skizziert.

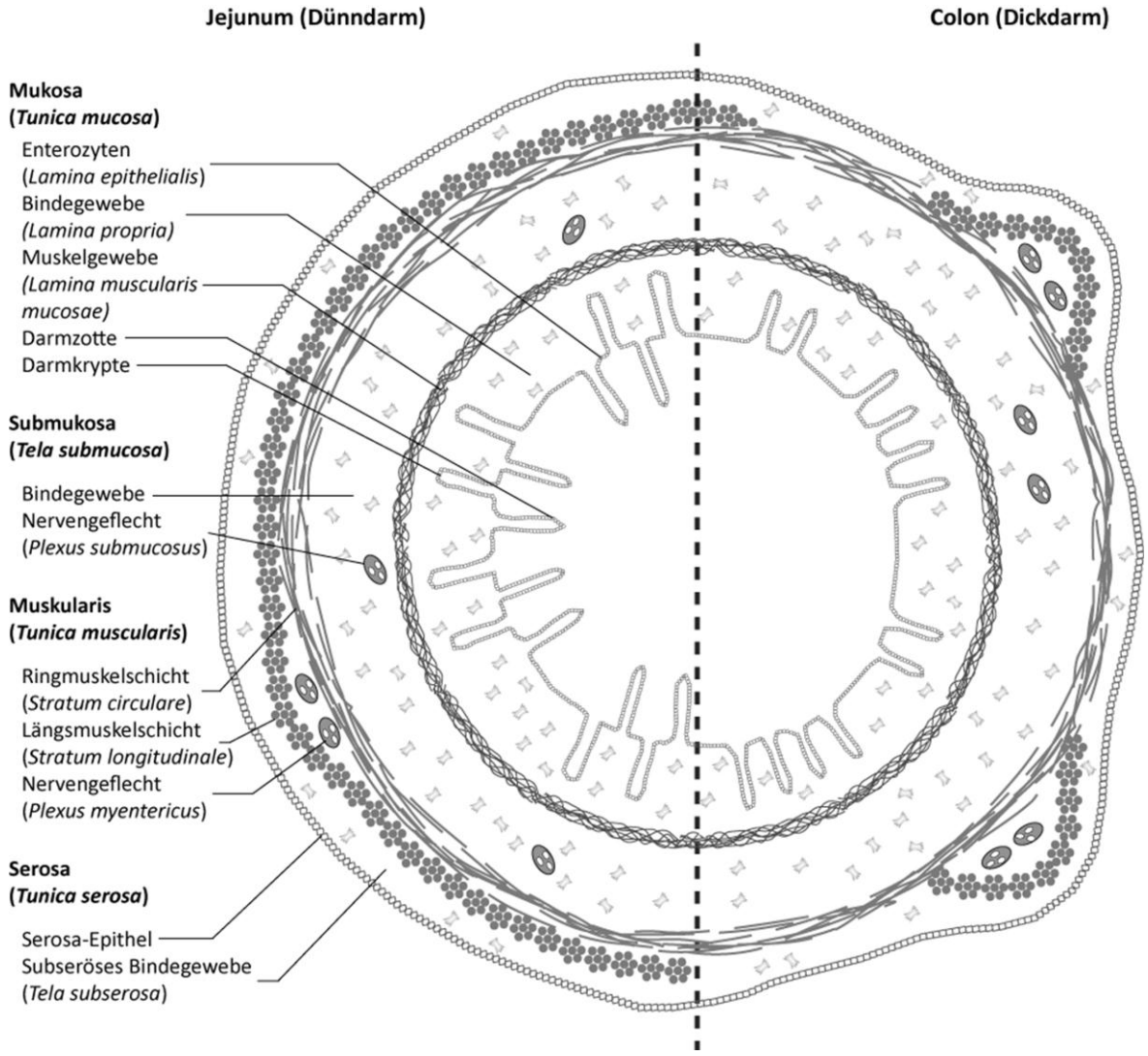


Abb. 1: Übersicht Aufbau des Darms.

In der Abbildung sind Skizzen des Querschnitts der in dieser Arbeit untersuchten Gewebetypen dargestellt: dem Jejunum (linke Hälfte) und dem Colon (rechte Hälfte). Wesentliche Unterscheidungsmerkmale des Colons gegenüber dem gesamten Dünndarm im histologischen Befund sind das Fehlen von Darmzotten sowie die nur partiell ausgebildete Längsmuskulatur (Tänien). Nicht dargestellt sind die Blutgefäße (sie finden sich im Bindegewebe aller Schichten) sowie die Aufhängebänder (Mesenterien) des Darms (in Anlehnung an Lüllmann Rauch 2003).

1.1.1. Molekulare Mechanismen bestimmen die Identität

In der Entwicklung des Darms kommt es zu einer engen Kooperation von Zellen. Schon in der Embryonalentwicklung ist das Wechselspiel zwischen Endoderm und Mesoderm entscheidend für die korrekte Bildung des Organismus. Während das Endoderm für die Induktion von verschiedenen mesodermalen Organen wie Herz oder Blutgefäßen entscheidend ist, kann es selbst auf unterschiedliche, regionale Strukturen mesodermalen Mesenchyms differenziert reagieren. Dadurch

werden schon früh in der Embryonalentwicklung Strukturen wie Ösophagus, Magen, Dünn- und Dickdarm festgelegt. Eine entscheidende Rolle nehmen hierbei Vertreter von Transkriptionsfaktoren wie *sonic hedgehog* (SHH) im Endoderm oder *bone morphogenetic protein* (BMP) im Mesoderm ein, welche auch schon in ursprünglicheren Lebewesen wie beispielsweise dem Zebrafisch diese Funktion zeigen (Gilbert 2003). Auch im adulten Intestinum finden sich in den einzelnen anatomischen Regionen Muster in der Genexpression, die für die Physis des jeweiligen Darmabschnitts entscheidend sind, darunter Gene für Transportvorgänge und Zell-Zellkommunikation (Bates u. a. 2002).

Neben der kranial-kaudalen Achse ist auch die Krypten-Villus-Achse innerhalb des Epithels entscheidend für die Funktion des Darms. In Studien konnten hoch spezifische, räumlich in unterschiedlichen Mustern exprimierte Regulatoren für Struktur und Funktion des intestinalen Epithels beobachtet werden (Schröder u. a. 2006; Mariadason u. a. 2005).

1.1.2. Das bedeutende Immunorgan Darm

Während die Mehrheit der Zellverbände von körpereigenen Strukturen umgeben ist, steht der Darm über seine große Oberfläche in direktem Kontakt mit der Umwelt. Dabei kann er bedingt durch seine Funktion nicht auf starke physikalische (Diffusions-) Barrieren zurückgreifen, wie es z.B. der Haut mit ihrem verhornenden Plattenepithel möglich ist. Zugleich finden Mikroorganismen im Darmtrakt nahezu optimale Wachstumsbedingungen vor, so dass auch unter physiologischen Bedingungen im Darmlumen anders als z.B. in der Lunge (Gerok 2006) eine kommensale Mikroflora existiert, deren Zellzahl die Zahl der körpereigenen Zellen übersteigt. Die hohe Anzahl an Mikroorganismen auf seiner Oberfläche und die durch die Funktion bedingte Schwäche der physikalischen Barriere erfordert, dass im Intestinum das körpereigene Abwehrsystem eine umfangreiche Manifestation erfährt. Nur eine enge Regulation der Mikroorganismen kann eine Invasion über die Darmwand oder andere nachteilige Prozesse verhindern. So ist der Darm das größte Immunorgan des menschlichen Körpers. Dabei besitzt das Immunsystem des Intestinums nicht nur Mechanismen, um Mikroorganismen abzuwehren. Es stellt sich immer mehr heraus, dass die Darmwand aktiv die Besiedelung ihrer Oberfläche durch eine günstige Mikroflora mitgestaltet (Rakoff-Nahoum u. a. 2004), zugleich aber auch die Zusammensetzung der Mikrobiota weitreichenden Einfluss auf Ernährung und Gesundheit des Wirtes haben (Sekirov u. a. 2010). Ist dieses Konzept der aktiven Gestaltung des mikrobiellen Oberflächenbewuchses durch den Wirt in der Wissenschaft noch neu, konnten erste Studien eine in der Evolution frühe Anwendung der aktiven Gestaltung nachweisen. So findet sich bereits für basale Metazoen wie den Süßwasserpolypen *Hydra* eine spezies-spezifische Zusammensetzung von Bakterien auf seiner Oberfläche (Fraune und Bosch 2007).

Mit der Erkennung und Überwachung der mukosa-assoziierten Mikroflora ist nicht allein das adaptive Immunsystem betraut. Zwar finden sich mit den Peyer-Plaques spezielle Bereiche der Darmwand, welche an die Begebenheiten des Darms angepasste Lymphfollikel darstellen. Sie werden auch als *gut associated lymphoid tissue* (GALT) bezeichnet. Eine wesentliche Rolle scheinen aber keimbahnkodierte Mustererkennungs-Rezeptoren (*pattern recognition receptors*, PRR) zu besitzen, welche an für Mikroben spezifische molekulare Muster (*microbe associated molecular pattern*, MAMP) binden können. Es konnte gezeigt werden, dass diese Rezeptoren im Darm eine vielfältige Rolle einnehmen. Das Produkt des Gens NOD2 (*Nucleotide-binding oligomerization domain containing 2*) dient beispielsweise als intrazellulärer Sensor für die Bakterienzellwandkomponente Muramyldipeptid (MDP). Der genaue Mechanismus der Interaktion ist zurzeit noch nicht verstanden. Solche Bestandteile des angeborenen Immunsystems sind dabei nicht auf einzelne (Immun-) Zellen beschränkt, sondern werden im gesamten Epithel des Darms gebildet.

1.1.3. Eine Störung der Selbstregulation führt zu schweren Erkrankungen

Störungen der Barrierefunktion des Darms können zu ernsthaften Gesundheitsbeeinträchtigungen führen, beispielsweise durch Ingestion von Pathogenen wie bei der Amöbenruhr (Stanley 2003). Zu schweren Störungen kann es aber auch kommen, weil die Regulation der Immunantwort auf die kommensale Mikrobiota des Darms nicht regelrecht funktioniert. Bei den beschriebenen chronisch-entzündlichen Darmerkrankungen Morbus Crohn oder Colitis ulcerosa handelt es sich um Erkrankungen mit massiven Entzündungsprozessen, die bereits in jungen Jahren auftreten können und mit schwerwiegenden Beeinträchtigungen einhergehen (Schreiber u. a. 2005). Für diese Erkrankungen konnte mittels genetischer Studien gezeigt werden, dass sie zu einem erheblichen Anteil eine erbliche Komponente besitzen (Franke u. a. 2010; Hampe u. a. 2007). Insbesondere konnten Abweichungen in Genen gefunden werden, die wie der zuvor erwähnte PRR NOD2 ebenfalls zentrale Prozesse in der Immunabwehr ausüben. So fanden sich krankheitsassoziierte Einzelnukleotid-Polymorphismen (*single nucleotide polymorphism*, SNP) in Genen mit Funktion in der Autophagozytose oder der Zytokin-Signalübertragung. Die über Zwillingsstudien ermittelte Gesamt-Erblichkeit dieser Krankheit konnte bisher aber nur zu einem geringen Grad (ca. 25%) mit dezidierten Variationen des Erbgutes in Einklang gebracht werden. Dies könnte durch sehr seltene SNP, die mit bisherigen Untersuchungen nicht analysiert wurden, oder auch durch andere Effekte erklärt werden, z.B. epigenetische Einflüsse (Debatte zur *missing heritability*) (Maher u. a. 2009). Neben fehlenden Informationen über krankheitsbegünstigende Erbgutveränderungen ist aber auch für die Mehrheit der in einen Zusammenhang gebrachten Gene nicht verstanden, wie und warum sie in den Pathomechanismus eingreifen, der zur Ausprägung einer chronisch-entzündlichen Darmerkrankung führt.

Bisherige Therapie-Konzepte basieren auf dem Einsatz von herkömmlichen Immunsuppressiva wie 5-Aminosalicylsäure (Mesalazin), Glukokortikoiden (Prednisolon) und Purinantagonisten (Azathioprin) (Ruß 2009). Bei schweren, therapieresistenten Formen des Morbus Crohn werden zunehmend rekombinant gewonnene Antikörper gegen den stark proinflammatorisch wirkenden, extrazellulären Liganden *Tumor Necrosis Factor α* (TNFα) eingesetzt (Infliximab, Adalimumab). Diese auch als Biologika bezeichnete neue Gruppe von Medikamenten ist bei prognostisch ungünstigem Verlauf schon früh indiziert und kann eine deutliche Milderung der Beschwerden mit weniger chirurgischen Interventionen und Hospitalisationen sowie den Erhalt der Arbeitsfähigkeit bedeuten (Herold 2011). Weitere Antikörper-Therapien gegen andere Zytokine sind in Vorbereitung (Baumgart und Sandborn 2007).

Diesen Therapieansätzen ist gemein, dass sie die Symptome einer entgleisten Immunantwort lindern, dabei aber mit globalen Effektoren des Immunsystems interagieren und zum Teil deutliche Nebenwirkungen mit sich bringen. Für die Entwicklung innovativer Therapiekonzepte, die die Immunantwort bereits selektiv bei der Erkennung und Einschätzung von potentiellen Gefahren durch die Mikroflora auf ein akzeptables Maß modulieren, ist ein besseres Verständnis der beteiligten Pathomechanismen auf molekularer Ebene erforderlich.

1.1.4. Chronische Entzündung im Tiermodell

Zum Verständnis der Pathomechanismen der chronischen Entzündung ist es hilfreich, wenn diese im Tiermodell bestmöglich simulierbar sind. Für chronisch-entzündliche Prozesse sind mehrere Tiermodelle etabliert worden, so kann z.B. durch die Gabe von Natriumdextransulfat (*dextrane sodium sulfate*, DSS) ein Entzündungsprozess des Darmes in der Maus simuliert werden und durch die Konzentration der Substanz (2% w/v vs. 4% w/v) sogar akuter oder chronischer Verlauf der Entzündung festgelegt werden (Sina u. a. 2010). Auch gibt es Standards zur Haltung von Mäusen, die zumindest die Abwesenheit von spezifisch-pathogenen Mikroorganismen in Mausmodellen sicherstellen (*standard of care for specific pathogen free* (SPF) mice). Darüber hinaus existieren Mauslinien, deren intestinale Mikroflora komplett eliminiert wurde (Wostmann u. a. 1970). Diese Verfahren erlauben weitgehend die experimentelle Beeinflussung der Interaktion von Wirt und Mikrobiota im Mausmodell.

Dabei erlaubt die Maus als in der biomedizinischen Forschung etabliertes und ausgiebig genutztes Tiermodell gentechnische Eingriffe in das Erbgut. Sowohl Gen-*knock out* als auch transgene Mäuse sind etabliert. Zusätzlich existieren Inzuchtstämme, deren Erbgut umfangreich erforscht und in hoher Auflösung bekannt ist.

Neben diesen technischen Vorteilen bietet das Modellsystem Maus auch eine ausreichende Nähe zum Menschen. Die Morphologie und Physiologie des Darmtrakts ist gegenüber dem Menschen

weitgehend identisch, so dass in der Maus gewonnene Erkenntnisse zu einem hohen Grad auf den Menschen übertragbar sind.

1.2. Das Transkriptom, Mittler genetischer Information

Für das Verständnis der Entwicklung, Differenzierung, Struktur und Funktion des Intestinums, aber auch für die Fähigkeit zur Diagnose und Therapie von intestinalen Krankheiten ist die detaillierte Kenntnis über die Expression und Regulation der genetischen Information in den Geweben ein wertvolles Hilfsmittel. Die Gesamtheit aller genetischen Informationen, die zu einem gegebenen Zeitpunkt von einer Zelle oder einer Population aus Zellen verwendet wird, wird dabei als Transkriptom bezeichnet.

Der Ausdruck Transkriptom wird in Anlehnung an den Begriff Genom genutzt. Dieser wurde von Hans Winkler aus den Worten Gen und Chromosom geprägt (Winkler 1920). Im heutigen Sprachgebrauch wird unter dem Begriff Genom die Gesamtheit der genetischen Information oder das Erbgut eines Organismus verstanden. Abgeleitet vom Begriff Transkript, welcher Ausdruck der molekularen Expression eines einzelnen Genes ist, wird analog unter dem Transkriptom die Gesamtheit aller Transkripte z.B. einer Zelle verstanden. Während das Genom eines Individuums weitgehend einer statischen Entität entspricht, gilt für die Zusammensetzung des Transkriptoms einer Zelle eine hohe Variabilität geprägt durch interne Prozesse (z.B. Zelldifferenzierung) oder ausgelöst durch externe Reize (z.B. in der Abwehr von Mikroorganismen). Auch ist im Transkriptom ein quantitativer Aspekt enthalten. Die Häufigkeit einzelner Transkriptformen beeinflusst direkt die zelluläre Ausprägung der genetischen Information. So wird angenommen, dass in der Gesamtheit weniger strukturelle Unterschiede von Genen als vielmehr die Varianz der Expressionsstärke die unterschiedliche Biologie von zumindest nah verwandten Lebensformen bedingt (King und Wilson 1975).

Das Genom einer Zelle bestimmt das Potential, hingegen stellt das Transkriptom bereits eine dynamische Verbindung zwischen genetischer Information und physischer Charakteristik einer Zelle dar (Velculescu u. a. 1997). Durch diese Fokussierung der genetischen Information kommt der Analyse des Transkriptoms eine besondere Bedeutung zu. Die sogenannte Transkriptomik (*transcriptomics*) erlaubt gegenüber der Betrachtung kompletter Genome eine Reduzierung der Komplexität des Erbguts auf funktionell entscheidende Informationen, zugleich werden aber durch die Expressionsstärke und posttranskriptionelle Modifizierungen von Transkripten Zusatzinformationen gewonnen, die sich derzeit nicht oder nur unvollständig aus genomischen Daten ableiten lassen (Farajollahi und Maas 2010; Chen und Manley 2009). In diesem Zusammenhang ist nicht außer Acht zu lassen, dass auf Ebene des Transkriptoms die Kontrolle und Regulation der Genexpression nicht abgeschlossen ist. So können beispielsweise Proteine in ihrer Aktivität,

Lokalisation, Interaktion oder Halbwertszeit beeinflusst werden, ohne dass auf Ebene des Transkriptoms eine Änderung zu beobachten ist (Toledo-Arana und Solano 2010).

In der Folge kann durch die Untersuchung des Transkriptoms nicht die volle Bandbreite der Interaktion zwischen Umwelt und Erbinformation abgebildet werden. Dennoch können Transkripte auf dieser Ebene direkt quantifiziert und in ihrer Zusammensetzung analysiert werden. Sie erlauben zumeist in guter Näherung eine Antizipation nachrangiger Ebenen der Genregulation. Geprägt auch durch die methodische Zugänglichkeit ist die Transkriptomik gegenwärtig eine Kernkomponente zur Untersuchung der ablaufenden, sehr komplexen Vorgänge in der Genregulation und verspricht sowohl ein besseres Verständnis fundamentaler Fragen der Biologie als auch tiefere Einsicht in Mechanismen der Krankheitsentstehung mit Aussicht auf neue Therapiekonzepte.

Molekulare Grundlage des Transkriptoms ist die Ribonukleinsäure (*ribonucleic acid*, RNA), ein Copolymer aus verschiedenen Ribonukleotiden, welche wiederum im Allgemeinen aus einer Phosphat-Gruppe, einer Ribose und je einer Purinbase (Adenin oder Guanin) oder einer Pyrimidinbase (Cytosin oder Uracil) aufgebaut sind (siehe **Abb. 2A**) und über Phosphodiesterbindungen der C5- und C3-Kohlenstoffatome der Ribose miteinander verknüpft werden. Zugleich erhält die RNA durch diesen Vorgang eine Polarität, das Nukleotid ohne Phosphodiesterbindung am C5-Kohlenstoffatom wird als 5‘-Nukleotid bezeichnet, entsprechend

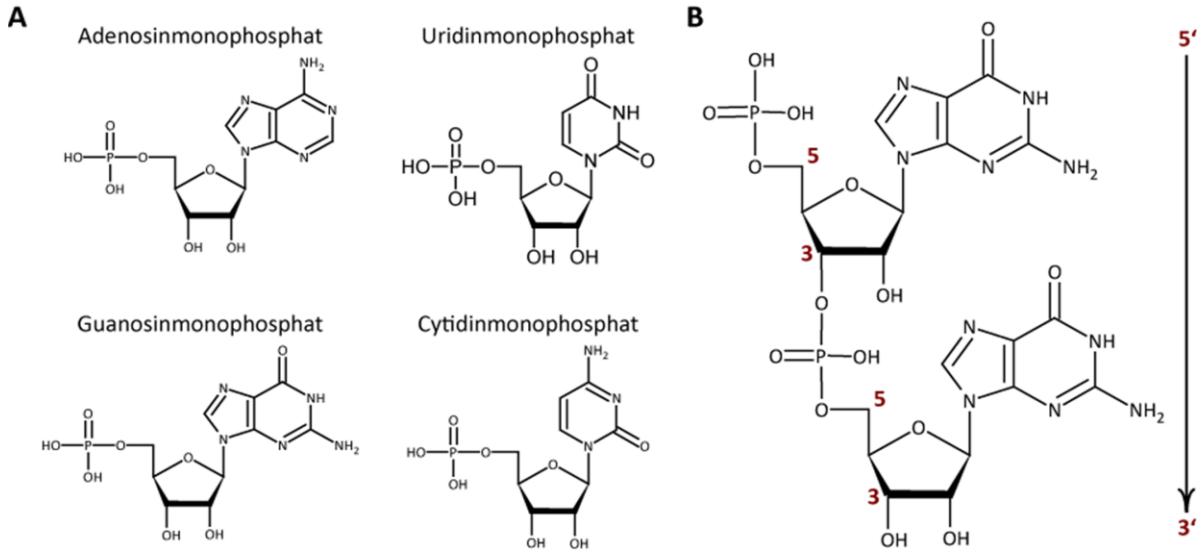


Abb. 2: Chemische Struktur der RNA-Nukleotide.

(A) Die Grundeinheiten der RNA sind Adenosin-, Uridin-, Guanosin- und Cytidinmonophosphat, bestehend je aus einer Phosphat- bzw. Ribosegruppe und der entsprechenden Nukleinbase. Entscheidend für das Bindungsverhalten sind die Substituenten der Nukleinbasen. Nicht dargestellt ist die Vielzahl von Modifizierungen der Nukleinbasen, die innerhalb der RNA auftreten können. **(B)** Die einzelnen Nukleotide werden unter Bildung eines Phosphodiesters miteinander verbunden, hier gezeigt anhand zweier Guanosinnukleotide. Dabei reagiert die jeweils am C5-Atom der Ribose verankerte Phosphat-Gruppe mit der Hydroxylgruppe des C3-Atoms der vorherigen Gruppe. Durch diese lineare Anordnung der Ribonukleotide besitzt die synthetisierte RNA eine festgelegte Orientierung, es lassen sich 5‘- und 3‘-Ende unterscheiden.

wird das Nukleotid mit fehlender Phosphodiesterbindung am C3-Kohlenstoffatom als 3‘-Nukleotid bezeichnet (siehe **Abb. 2B**).

1.2.1. RNA-Moleküle lassen sich anhand der Funktion klassifizieren

Um einen besseren Überblick über die verschiedenen RNA-Moleküle zu ermöglichen, können diese funktionellen Klassen zugeordnet werden.

Die innerhalb der Zelle am häufigsten vorkommende RNA-Gruppe ist die ribosomale RNA (rRNA). In Eukaryoten sind vier verschiedene rRNA am Aufbau des Ribosoms beteiligt. Drei finden sich in der großen Untereinheit des Ribosoms, die größte davon besteht aus ca. 4.700 Nukleotiden (nt), die beiden anderen sind mit Größen zwischen 120 und 160 nt deutlich kleiner. Die kleine Untereinheit beinhaltet nur eine rRNA von ca. 1900 nt. Diese nimmt eine entscheidende Rolle bei der Erkennung der Boten-RNA (s.u.) in der Proteinbiosynthese wahr.

Eine weitere Klasse mit einer Funktion in der Proteinbiosynthese stellt die Gruppe der Transfer-RNA (tRNA) dar. Diese aus 73-94 Nukleotiden bestehenden Moleküle unterliegen erheblichen posttranskriptionellen Modifikationen und binden je nach Art der tRNA spezifisch Aminosäuren und stellen sie in der Proteinbiosynthese für den Einbau in das Polypeptid zur Verfügung. Transfer-RNA ähneln in der dreidimensionalen Struktur einem Bumerang (Shi und Moore 2000). Am Ende des einen Schenkels bindet die Aminosäure. Das Ende des anderen Schenkels dient in der Proteinbiosynthese als Identifizierungssignal für das Ribosom, welches als Anti-Codon bezeichnet wird. Auch wenn die überragende Bedeutung der tRNA in der Proteinbiosynthese liegt, nehmen sie weitere Funktionen ein, so z.B. in der Signaltransduktion oder Proteindegradierung (Phizicky und Hopper 2010).

Während die beiden erstgenannten Gruppen funktionell-strukturell an der Umsetzung des genetischen Programms beteiligt und somit in jeder Proteinbiosynthese betreibenden Zelle in ihrer Gesamtheit unverzichtbar sind und kontinuierlich gebildet werden, ist die Gruppe der Boten-RNA (*messenger* RNA, mRNA) gegenüber den anderen Gruppen hoch divers und in der Expressionsstärke der einzelnen Mitglieder stark variabel. Die mRNA liefert im Rahmen der Translation die Blaupausen für die Synthese der Proteine. Je nach Anforderungen des Zelltypus oder der Stoffwechsellage kann sich daher die Zusammensetzung dieser RNA-Gruppe - gesteuert durch die Genregulation - erheblich unterscheiden. Dabei unterliegt die mRNA einem Reifungsprozess, der zu einer umfangreichen Modifizierung der prä-mRNA führen kann (Soller 2006).

Neben diesen umfangreich untersuchten Mitgliedern der RNA konnten weitere Klassen identifiziert und näher beschrieben werden, die häufig als nicht-kodierende RNA betrachtet werden (*non coding* RNA, ncRNA), so z.B. *small nuclear* RNA (snRNA) und *small nucleolar* RNA (snoRNA), welche

Tab. 1: Überblick verschiedener RNA-Klassen und deren Funktion

Bezeichnung	Abkürzung	Länge (nt)	Funktion
ribosomale RNA (*ribosomal RNA*)	rRNA	120 - > 4.800	Katalytisches Zentrum der Proteinbiosynthese
Transfer-RNA (*transfer RNA*)	tRNA	73-94	Bindung und Bereitstellung von Aminosäuren
Boten-RNA (*messenger RNA*)	mRNA	variabel	Informationsträger und Modulator
*long interfering non-coding RNA**	lincRNA	variabel	Chromatinmodifikationen, epigenetische Vererbung, Transkriptionskontrolle
*micro RNA**	miRNA	ca. 22	Genregulation über Modulation der mRNA
*small interfering RNA**	siRNA	21-22	Genregulation über Modulation der mRNA
*piwi interacting RNA**	piRNA	26-31	Unterdrückung von Transposonen, Keimzellentwicklung
*small nucleolar RNA**	snoRNA	variabel	rRNA-Editierung, RNA-Spleißen
*small nuclear RNA**	snRNA	variabel	RNA-Spleißen, Genregulation

* Die englischen Fachbegriffe werden auch im deutschen Sprachgebrauch überwiegend genutzt. Auf eine Übersetzung wurde verzichtet.

Funktionen in der RNA-Editierung und beim Spleißen einnehmen (Jawdekar und Henry 2008). In jüngerer Zeit zunehmend in den Fokus der Wissenschaft geraten, ist der Mechanismus der sogenannten RNA-Interferenz, der die Aktivität und Stabilität von RNA-Molekülen beeinflusst (Fire u. a. 1998) und in Zellen experimentell die gerichtete Ausschaltung bestimmter Transkripte erlaubt (Elbashir u. a. 2001). Zunächst wurde dieser Mechanismus im Zusammenhang mit z.B. der Abwehr viraler Infektionen diskutiert. Mittlerweile gibt es auch Anhaltspunkte, dass endogen kodierte RNA-Varianten genutzt werden, um direkt in die Genregulation der Zelle einzugreifen. In dieser Gruppe finden sich sogenannte *small interfering* RNA (siRNA) (Fagegaltier u. a. 2009), *micro* RNA (miRNA) (Liston, Linterman, und Lu 2010) und *piwi interacting* RNA (piRNA). Auch nicht-kodierende, langkettige RNA-Formen konnten identifiziert werden. Für die sogenannten *long intergenic noncoding* RNA (lincRNA) konnte gezeigt werden, dass sie eine bedeutende Rolle auch in epigenetischen Kontrollmechanismen besitzen (Gupta u. a. 2010). Eine Übersicht der verschiedenen RNA-Klassen ist in **Tab. 1** gezeigt.

Zusammenfassend lässt sich sagen, dass die RNA Aufgaben wahrnimmt, die sowohl für die Struktur als auch für die Funktion von Zellen unverzichtbar sind. Die RNA ist nicht nur Zwischenspeicher genetischer Informationen, sondern wirkt auch als Katalysator in elementaren Funktionen des Zellstoffwechsels.

1.2.2. Posttranskriptionelle Prozessierung der RNA

In der Transkription wird ein RNA-Molekül mit der identischen Basenfolge der als Vorlage dienenden DNA (ausgenommen des Thymin/Uracil-Basenaustausches) synthetisiert. Bevor die RNA den Kern verlassen kann, oftmals sogar direkt mit der Transkription assoziiert, wird die RNA in der Regel umfangreich modifiziert und somit die genetische Information moduliert. Insbesondere für die

mRNA, welche durch die hohe Variabilität der Genexpression besonderes Interesse erregt, ist dieser Vorgang gut untersucht. Hier sind vier verschiedene Prozesse für die Reifung der mRNA beschrieben: *capping*, RNA-Editierung, Polyadenylierung und Spleißen.

Beim *capping* wird das 5‘-Ende eines Transkripts modifiziert. Am besten untersucht ist die Bildung eines 5‘-5‘-Phosphodiesters unter Anlagerung eines methylierten Guanosinmonophosphats im Rahmen der Transkription durch die RNA-Polymerase II (vorwiegend mRNA). Diese bildet zugleich eine Schutzgruppe gegen enzymatischen Abbau der RNA und ein Erkennungsmotiv für den Kernexport (Knippers 2001). Im Falle der snRNA konnten auch abweichende *capping*-Strukturen gezeigt werden (Mattaj, Tollervey und Séraphin 1993). Für die snRNA U6 konnte eine Modifizierung des 5‘-Endes gezeigt werden, obwohl sie durch die RNA-Polymerase III synthetisiert wird (Kwan, Gerlach und Brow 2000).

Unter RNA-Editierung wird die posttranskriptionelle Modifizierung einzelner Basen eines Transkripts verstanden. Ein bekanntes Beispiel für diesen Fall ist die Konvertierung eines Cytidin-Restes zu einem Uracil in der ApoB (Apolipoprotein B)-mRNA. Durch diese Modifikation wird ein neues Stopcodon in die mRNA eingebaut, die mit einer deutlichen Verkürzung des resultierenden Proteins einhergeht (ApoB48 gegenüber ApoB100). Die verkürzte Variante tritt gewebespezifisch im Dünndarm von Säugetieren auf und wird durch das Enzym *Apolipoprotein B mRNA editing enzyme* (Apobec) gebildet (Davidson und Shelness 2000). Im Einzelfall kann die RNA-Editierung umfangreiche Ausmaße annehmen. So konnte für mitochondriale RNA von Trypanosomen gezeigt werden, dass nahezu jede zweite Uracil-Base durch RNA-Editierung eingeführt wird (Stuart u. a. 2005).

Nahezu alle mRNA-Transkripte werden polyadenyliert. Dies bedeutet, dass im Kern durch einen Protein-Komplex die mRNA an der Position eines sogenannten Polyadenylierungssignals gespalten wird. Ausgehend von dieser Bruchstelle wird in der Folge durch eine im Komplex enthaltene Polyadenylat-Polymerase eine Folge von Adenosinresten an das 3‘-Ende des Transkripts angefügt, der sogenannte Poly(A)-Schwanz (*poly(A) tail*). Durch diesen Prozess wird die Stabilität der mRNA im Zytoplasma erhöht. Zugleich ist dies ein wichtiges Erkennungsmotiv für den erfolgreichen Kernexport (Guhaniyogi und Brewer 2001). Dieser Prozess ist auch nicht auf die mRNA beschränkt. Faktisch alle Syntheseprodukte der RNA-Polymerase II wie z.B. miRNA können zumindest zwischenzeitlich polyadenyliert werden (Saini, Griffiths-Jones, und Enright 2007). Auch sind viele lincRNA polyadenyliert (Amaral und Mattick 2008). Es konnte gezeigt werden, dass ein unreifes Transkript nicht zwingend nur ein einzelnes Polyadenylierungssignal beinhalten muss, sondern je nach physiologischen Umständen verschiedene Polyadenylierungssignale gewählt werden können (Shell u. a. 2005).

Als letzte posttranskriptionelle Modifikation soll das Spleißen kurz vorgestellt werden. Hierbei werden in eukaryotischen Zellen Teile der unreifen RNA, die sogenannten Introne, aus dem

bestehenden Transkript herausgeschnitten und anschließend die verbliebenen Enden des Transkripts (die Exone) wieder verbunden. Dieser im Kern lokalisierte Prozess wird maßgeblich durch das Spleißosom vorangetrieben. Weiterhin sind aber auch Fälle von selbstspleißenden RNA-Spezies bekannt. Das Spleißosom besteht aus *small nuclear ribonucleoproteins* (snRNP), in denen neben Proteinen die bereits genannten snRNA eine Kernrolle bei der Katalyse des Spleiß-Prozesses einnehmen (Alberts u. a. 2003). Das Spleißosom erkennt die korrekten Positionen des Spleißens an konservierten Sequenzmotiven, den Spleißdonoren und –akzeptoren. Die Mehrheit der Gene in höheren Eukaryoten besitzt eine Multi-Exon-Struktur, so dass erst durch das Spleißen reife mRNA-Formen entstehen, die als Matrize für funktionstüchtige Proteine dienen können. Interessant im Rahmen der Genexpression ist dieser Prozess, weil auch hier nicht immer festgelegte Wege beschritten werden. So können durch alternatives Spleißen aus einer prä-mRNA eine Vielzahl verschiedener mRNA und somit unterschiedliche Proteine aus einem einzigen Gen entstehen, beispielsweise ausgelöst durch die gewebsspezifische Expression von Spleißmodulatoren (Black 2003).

Diese posttranskriptionellen Modifikationen sorgen für Unterschiede zwischen der Sequenz des Erbgutes und der Sequenz der RNA, die sich im Falle der Boten-RNA auch auf die Ebene der Proteine erstrecken. Die detaillierte Aufklärung der Zusammensetzung des Transkriptoms und der Basensequenz der einzelnen Transkripte verspricht daher einen zentralen Einblick in das genetische Programm einer Zelle oder einer Zellpopulation, wie es bei alleiniger Betrachtung des Erbgutes zurzeit nicht möglich ist.

1.3. Etablierte Methoden zur Charakterisierung des Transkriptoms

In der Vergangenheit wurden im Wesentlichen drei methodische Ansätze genutzt, um das Transkriptom zu analysieren: *microarray*-Technologien, quantitative Polymerasekettenreaktion (PCR) und Sanger-Sequenzierung.

Unter der Nutzung von *microarrays* wird eine Technologie verstanden, die auf der Hybridisierung komplementärer Nukleinsäuren beruht. Dazu kann beispielsweise ein Trägermedium mit einer Vielzahl von bekannten Oligonukleotiden an definierten Positionen beschichtet werden, die der Sequenz bekannter Transkripte entspricht. Anschließend wird aus der zu untersuchenden RNA generierte, fluoreszenzmarkierte cDNA zum Versuchsansatz hinzugegeben, die mit den Oligonukleotiden hybridisieren und dadurch lokal immobilisiert werden (Ehrenreich 2006). Die Menge an cDNA eines bestimmten Transkripts kann nun anhand des Fluoreszenz-Signals nach Anregung für jede Oligonukleotidposition ausgelesen werden. Für Modellorganismen und den Menschen sind standardisierte *microarrays* erhältlich, die die Gesamtheit der bekannten Transkripte in einem Experiment abfragen können. Zwar ist diese Methode auch durch die hohe Anzahl an

Abfragen anfällig gegenüber Verzerrungen, z.B. durch die Bedingungen bei der RNA-Isolation. Durch die Wahl günstiger Oligonukleotide und die Abfrage mehrerer Positionen pro Transkript sowie der Anwendung von Normalisierungsverfahren, die z.B. die Expression von stabil exprimierten Genen berücksichtigen, sogenannten *housekeeping*-Genen, kann die Validität der Experimente erhöht werden. Die Möglichkeit der genomweiten Messung von Expressionsstärken kombiniert mit einem moderaten Preis insbesondere für Modellorganismen erhoben in der Vergangenheit das *microarray* zur Standardmethode zum Auffinden differentiell regulierter Gene zwischen unterschiedlichen Geweben, in der Entwicklung von Organismen oder im Rahmen pathologischer Veränderungen.

Die quantitative PCR ist die aktuell meist genutzte Methode, um eine genaue Mengenbestimmung von einzelnen Transkripten durchzuführen. Da die Effizienz einer PCR in nur wenigen Zyklen hoch ist, eignet sich die Endbestimmung der amplifizierten cDNA-Konzentration nur unzureichend, um die genaue Häufigkeit von Transkripten zu bestimmen. Daher hat es sich bei der quantitativen PCR durchgesetzt, die DNA-Konzentration während der laufenden PCR ständig zu überwachen und so die Zyklen zu überblicken, bei denen die Amplifizierung exponentiell und damit sehr effizient verläuft (sogenannte quantitative Echtzeit-PCR, *real time quantitative PCR* (RTq-PCR)). Auch in der RTq-PCR werden üblicherweise Fluoreszenzfarbstoffe genutzt, um die Amplifizierung des Zieltranskripts zu quantifizieren. Das methodische Spektrum umfasst einfache Fluoreszenzfarbstoffe (z.B. *SYBR Green I*), die die Gesamtheit doppelsträngiger DNA misst und dazu führen kann, dass auch unerwünschte Produkte der PCR gemessen werden. Andere Anwendungen vermeiden solche Verzerrungen, z.B. durch Verwendung von Oligonukleotiden, die sowohl einen Fluoreszenzfarbstoff als auch einen Farblöscher (*quencher*) beinhalten. In dieser auch als *Taqman Assay* bezeichneten Modifikation der qPCR kann der Fluoreszenzfarbstoff erst angeregt werden, wenn eine lokale Trennung vom *quencher* erfolgt. Dies geschieht durch den Abbau des Oligonukleotids, das an das Zieltranskript bindet und durch die 5‘-3‘-Nukleaseaktivität der Taq-Polymerase abgebaut wird. Bei dieser Methode verfälschen etwaige Nebenprodukte der PCR nicht das Ergebnis der Quantifizierung (Kubista u. a. 2006)

Diese Methode der quantitativen Polymerasekettenreaktion ist in der Lage, bekannte Transkripte sehr genau zu bestimmen, im Regelfall kann aber nur ein Transkript pro Ansatz quantifiziert werden. Zwar gibt es hier auch Anwendungen, die die Messung der Expressionsstärke mehrerer Gene parallel erlauben (z.B. *micro fluidic cards* (Keys, Au-Young und Fekete 2010)). Generell sind die Kosten aber relativ hoch und schon die Quantifizierung weniger Gene ist gegenüber einer genomweiten Untersuchung durch *microarrays* nicht ökonomisch.

Mit der Einführung von Methoden zur Sequenzierung von DNA durch Allan Maxam, Walter Gilbert (Maxam und Gilbert 1977) und Frederick Sanger (Sanger u. a. 1977) bestand erstmals die Möglichkeit, die genaue Basenabfolge von Ribonukleinsäuren zu bestimmen. Auch in der Erforschung des Transkriptoms ist die Sequenzierung eine wichtige Komponente, da die im vorherigen Abschnitt

genannten Methoden das Wissen um die Nukleinsäuresequenz der zu untersuchenden Transkripte voraussetzen. Um die Sequenz zu erhalten, wird eine *expressed sequence tag* (EST)-Sequenzierung durchgeführt. Hierbei werden Transkripte mittels einer reversen Transkriptase in Komplementär-DNA (*complementary* DNA, cDNA) umgeschrieben, in Plasmide ligiert und über bakterielle Replikation vermehrt. Die isolierten Plasmide werden anschließend mittels der Didesoxy-Methode nach Sanger sequenziert (Adams u. a. 1991). Für die Maus konnte so durch das FANTOM-Konsortiums (*Functional Annotation of the Mammalian Genome*) eine umfangreiche Sammlung von Transkripten beschrieben werden, welche die Basis für die Lokalisation und Annotation der Gene im Maus-Genom darstellten (Kawai u. a. 2001; Okazaki u. a. 2002). Um Aussagen über die Expressionsstärke bestimmter Gene zu treffen, ist diese Methode aufgrund von Verzerrungseffekten bei der bakteriellen Klonierung und durch die enormen Kosten, die durch die hohe Anzahl der zu sequenzierenden ESTs anfallen würden, nicht praktikabel. Diese Nachteile der Sanger-Sequenzierung können zwar durch spezielle Modifikationen der Methode reduziert werden. So werden eine Vielzahl von kleinen Sequenzsegmente des 3‘-Endes (*serielle analysis of gene expression*, SAGE) (Velculescu u. a. 1995) oder vom 5‘-*Cap* (*cap analysis of gene expression*, CAGE) (Shiraki u. a. 2003) zufällig hintereinander ligiert. Dies erlaubt mit einem einzelnen *read* der Sanger-Methode die Identifikation einer Vielzahl exprimierter Transkripte. Zusätzlich kann eine potentielle Verzerrung der Expressionslevel hier erkannt werden, da in diesem Fall identische Kombinationen von Transkriptfragmenten wiederholt auftreten. Diese Methoden sind auch geeignet, zuvor nicht beschriebene Transkripte zu identifizieren. Bedingt durch die nur kurze Fragmentlänge (initial z.B. nur 10 Nukleotide im SAGE) lassen sich nur Aussagen über sehr kurze Bereiche der Transkripte treffen. Zusätzlich ist der mit diesen Methoden behaftete Aufwand hoch, so dass sie sich nicht als Standardprozedur durchsetzen konnten.

1.4. *next generation sequencing*

Um zu geringeren Kosten erheblich mehr Sequenzinformationen zu erhalten, wurden in der jüngeren Zeit neue Methoden der DNA-Sequenzierung entwickelt. Aktuell werden drei kommerziell erhältliche Systeme weltweit in größerem Umfang genutzt: der *Genome Analyzer* der Firma Illumina (entwickelt von Solexa), der *Genome Sequencer* der Firma Roche (entwickelt von 454) und der SOLiD (*Sequencing by Oligonucleotide Ligation and Detection*) von Life Technologies (ehemals Applied Biosystems). Mit der Einführung dieser *next generation sequencing* (NGS)-Technologien ist es Molekulargenetikern nun möglich, Informationen aus dem Erbgut in immer größerer Zahl auszulesen. Bei überschaubarem Aufwand ist es nun ökonomisch denkbar, sowohl die DNA von einer Vielzahl von Organismen zu entziffern, zugleich aber auch das vollständige Genom einer großen Zahl an Individuen einer Art zu untersuchen (1000 Genomes Project Consortium. 2010). So könnten z.B. Bereiche des Genoms

verstanden werden, die das zurzeit beobachtete Missverhältnis zwischen der Vererbung bestimmter Krankheiten und der durch herkömmliche Methoden tatsächlich aufgedeckten Variationen des Genoms erklären. Aufgrund der zentralen Natur der Fragestellungen, die nun Biologen und medizinischen Forschern durch *next generation sequencing* zugänglich sind, wurde diese Methode 2007 durch die Editoren der Zeitschrift *Nature Methods* zur erstmals benannten Methode des Jahres (*Method of the year*) erkoren (Chi 2008).

In der Folge sollen die NGS-Verfahren, welche in dieser Arbeit eine zentrale Rolle einnehmen, kurz vorgestellt werden.

1.4.1. Parallele Pyrophosphatsequenzierung

Der *Genome Sequencer* FLX kann über massive Parallelisierung gegenüber dem Sanger-Verfahren größere Mengen an Sequenzinformationen erzeugen und unterscheidet sich in wesentlichen Punkten von der Dideoxymethode Sangers. Zunächst werden eine Vielzahl von unterschiedlichen Nukleotidfragmenten für die Sequenzierung aufbereitet, indem die fraktionierten und mit Adaptoren versehenden DNA-Fragmente zusammen über Oligonukleotide an kleine Kugeln (*beads*) gebunden werden. Anschließend wird unter Zugabe von Nukleotiden, Polymerase und geeigneten Primern eine sogenannte Emulsions-Polymerasekettenreaktion (ePCR) durchgeführt. Dies bedeutet, dass sich kleine Wassertröpfchen mit den erforderlichen PCR-Reagenzien innerhalb eines Ölgemisches bilden (die Emulsion). Durch die Wahl der Konzentrationen finden sich bei einem hohen Anteil der Reaktionsräume nur jeweils ein *bead* und ein DNA-Fragment. Über die Adaptorensequenzen wird nun eine PCR durchgeführt. Durch die räumliche Trennung aufgrund der Emulsion entstehen auf jedem *bead* identische DNA-Fragmente. Danach werden die *beads* aus der Emulsion zurückgewonnen (die Emulsion wird „gebrochen“) und nur mit DNA beladene *beads* über Hybridisierung der bekannten Adaptorsequenz ausgewählt. Mit diesem Verfahren lassen sich in der Regel *beads* gewinnen, deren DNA-Fragmente zu ca. 85% monoklonalen Ursprungs sind und somit erfolgreich sequenziert werden können. Diese *beads* werden nun auf einer kleinste Vertiefungen tragenden, durchsichtigen Trägerplatte immobilisiert. Nicht nur die Erstellung der Sequenzier-*libraries* unterscheidet sich gegenüber der Dideoxy-Methode nach Sanger, auch die für die eigentliche Sequenzierung verwandte Methode weicht deutlich vom Prinzip des Kettenabbruchs ab. Zwar handelt es sich auch hier um ein Verfahren, das auf der Nutzung eines Startprimers und einer Polymerase bei der Zweitstrangsynthese beruht (*sequencing by synthesis*). Zur Identifikation der eingebauten Basen werden aber keine an Dideoxynukleotide geknüpften Fluoreszenzfarbstoffe genutzt, sondern ein Verfahren, welches als Pyrophophat-Sequenzierung (*pyrosequencing*) bezeichnet wird. Hierbei wird das bei der Kettenverlängerung eines DNA-Strangs freiwerdende Pyrophosphat mit Adenosin-5‘-Phosphosulfat (APS) katalysiert von einer Sulfurylase zu

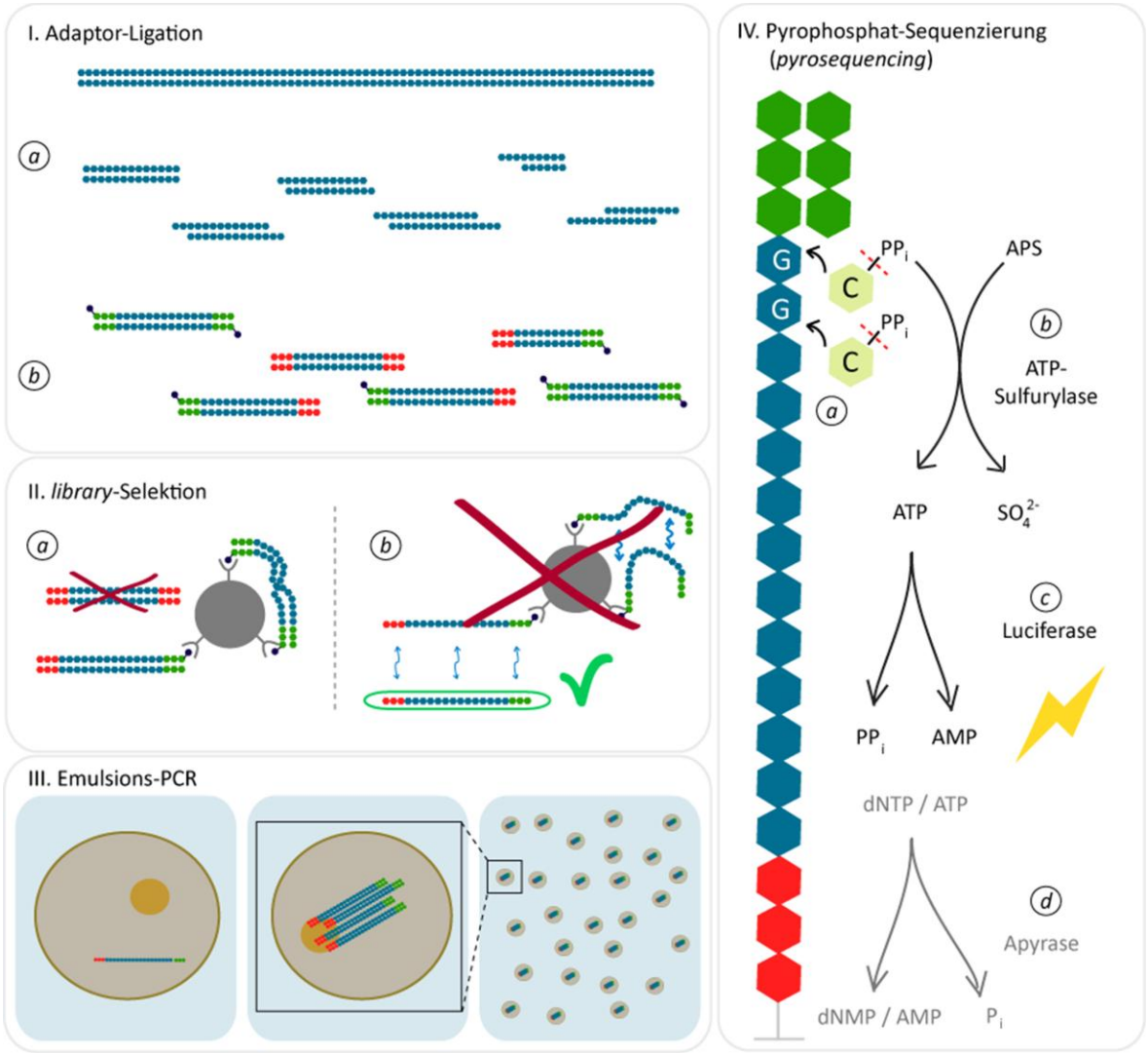


Abb. 3: Ablauf der Roche FLX Pyrophosphat-Sequenzierung

(I) Zunächst wird die Ziel-DNA geschert (a). Anschließend werden die Enden der fragmentierten DNA repariert und Adaptoren angefügt, so dass Fragmente mit definierten Enden entstehen (b). (II) Unter den generierten DNA-Fragmenten finden sich neben der gewünschten Kombination aus beiden Adaptoren zusätzlich Fragmente, die nur einen Adaptor doppelt tragen. Um die Fragmente mit der gewünschten Kombination zu selektionieren, werden zunächst die doppelsträngigen Fragmente über eine Biotingruppe des einen Adaptors an Streptavidin-tragende *beads* gebunden. Wird der Überstand nun verworfen, verbleiben nur Fragmente mit dem Biotin-markierten Adaptor (a). In einem zweiten Schritt werden nun die Fragmente denaturiert, so dass ein Einzelstrang frei in der Lösung vorliegt. Besitzen sie an beiden Enden den Biotin-markierten Adaptor, binden dabei beide Einzelstränge weiterhin an die Streptavidin-tragenden *beads*. Diese werden aus der Reaktion entfernt, so dass nur noch freie Einzelstrang-Fragmente mit beiden Adaptoren verbleiben (b). (III) Die selektionierten Einzelstränge werden nun in einer sogenannten Emulsions-PCR amplifiziert. Dabei werden die Fragmente in eine Öl/Wasser-Emulsion gebracht, die dazu dient, eine Vielzahl von wässrigen Reaktionsräumen zu schaffen. Liegen in einem Wassertröpfchen innerhalb dieser Emulsion neben den üblichen Reagenzien einer PCR sowohl ein Oligonukleotid tragendes *bead* als auch ein DNA-Fragment vor, welches über die Adaptorsequenz an das Oligonukleotid des *bead* hybridisiert, kann dieses Fragment amplifiziert werden, bis das ganze *bead* mit einer Vielzahl von DNA-Fragmenten überzogen ist. Auch wenn zur Vermeidung von *beads* mit polyklonalen DNA-Fragmenten die Konzentration der Ausgangs-DNA sehr gering gewählt wird und dadurch nur knapp 15% der *beads* überhaupt beladen werden, können so sehr viele *beads* in einem Ablauf mit unterschiedlichen DNA-Fragmenten beladen werden. Nicht beladene *beads* können einfach anhand der fehlenden Adaptoren erkannt und vom weiteren Sequenziervorgang

ausgeschlossen werden (nicht gezeigt). (IV) Die mit den klonalen DNA-Fragmenten beladenen *beads* werden für die eigentliche Sequenzierung auf eine Mikrotiterplatte überführt, in der jedes *bead* in einer kleinen Vertiefung liegt. Bei der Pyrosequenzierung handelt es sich um einen *sequencing by synthesis*-Ansatz, in diesem Fall wird mit unveränderten Desoxy-Nukleotiden gearbeitet. Allein dATP wird modifiziert verwendet (2'-Deoxy-Adenosin-5'-O-(1-Thio-Triphosphat) [dATPαS]), damit dieses nicht als Substrat für die Luciferase dienen kann (s.u.). Der Einbau eines Nukleotids wird über den Nachweis des freiwerdenden Pyrophosphats detektiert (a). Um zwischen den einzelnen Nukleotiden zu unterscheiden, werden diese in einer festen Reihenfolge nacheinander in den Reaktionsraum gegeben. Zum Nachweis, ob ein Nukleotid inkorporiert wurde, wird freigesetztes Pyrophosphat zunächst mit Adenosinphosphosulfat katalysiert durch eine Sulfurylase in ATP und anorganisches Phosphat (P_i) umgesetzt (b). Gebildetes ATP wird nun mit Luciferin durch eine Luciferase in eine aktivierte Form umgewandelt (nicht gezeigt), die unter Lichtemission und Bildung von Adenosinmonophosphat (AMP) und Pyrophosphat zerfällt. Die Stärke dieses Lichtsignals ist proportional zur Menge der inkorporierten Nukleotide und kann über einen Photosensor ermittelt werden (c). Die Anwesenheit einer Apyrase stellt sicher, dass nicht inkorporierte dNTP und überschüssiges ATP zersetzt wird, bevor durch Zugabe des nächsten dNTP ein neuer Zyklus begonnen wird (d). [basierend auf Informationen von www.roche.com]

Adenosintriphosphat (ATP) und Sulfat umgesetzt. In einer zweiten Reaktion wird das gebildete ATP dann mittels eines Luziferins und katalysiert durch eine Luziferase unter Bildung eines Lichtblitzes wieder zersetzt. Die Menge der eingebauten Nukleotide ist dabei proportional zum freigesetzten Pyrophosphat, aus diesem wird in gleichem Verhältnis ATP gebildet. Das durch die Luziferase emittierte Licht hängt wiederum von der Menge des verfügbaren ATP ab, so dass mittelbar über die Stärke des Lichtes auf die Menge der inkorporierten Nukleotide rückgeschlossen werden kann. Da anhand des Lichtblitzes nicht auf die Natur des eingebauten dNTP geschlossen werden kann, müssen diese zyklisch in bekannter Reihenfolge zur DNA zugefügt werden und vor Beginn des nächsten Zyklus wieder abgebaut werden (Margulies u. a. 2005). Zugleich kann aber auch der Einbau mehrerer Basen identischer Natur in einem Zyklus beobachtet werden. Bei zunehmend längeren Abschnitten von Homopolymeren (vielfacher Wiederholung einer einzelnen Base) kann diese Technologie die Stärke des Lichtblitzes aber nicht mehr sicher auflösen, so dass ab sieben Basen gleichen Typs in Folge die Genauigkeit abnimmt, mehr als zehn Basen lassen sich nicht mehr trennen. Neben diesem Homopolymer-Problem ist auch die generierte Menge an entschlüsselten Basen verglichen zu anderen *next generation sequencing*-Systemen relativ gering, so werden mit der aktuellen Version (*Genome Sequencer* FLX *Titanium*) etwa eine Million Sequenzen generiert. Allerdings ist die Leselänge der erhaltenen *reads* mit im Durchschnitt 500 Nukleotiden (basierend auf der Technik ist die Leselänge in diesem Fall nicht konstant) deutlich höher als bei anderen aktuellen Systemen. Wenn so auch pro Lauf mit 500 Megabasen der Sequenzgewinn für *next generation sequencing*-Systeme gering anmutet, so muss dennoch bedacht werden, dass die höhere Leselänge die anschließende Analyse enorm erleichtern kann. Dies gilt insbesondere dann, wenn Organismen untersucht werden, für die noch kein Genom annotiert ist, die komplette Struktur also aufklärt werden muss (*de novo*-Assemblierung) und nicht die grobe Struktur eines bereits annotierten Genoms genutzt werden kann

(Resequenzierung) (Mardis 2008). Eine graphische Darstellung zur Methode der parallelen Pyrosequenzierung findet sich in **Abb. 3**.

1.4.2. sequencing by ligation

Das zweite System der in dieser Arbeit genutzten *next generation sequencing*-Systeme wird von Life Technologies unter dem Namen SOLiD (*Sequencing by Oligonucleotide Ligation and Detection*) vertrieben. Die Erstellung der *libraries* erfolgt ähnlich des zuvor genannten Verfahrens. Die DNA wird per Ultraschall fraktioniert und mit Adaptoren versehen. Auch werden hier in einer ePCR *beads* mit DNA-Fragmenten klonalen Ursprungs erzeugt und diese im Anschluss auf einer Glasplatte immobilisiert.

Die eigentliche Sequenzierung erfolgt hier aber nicht nach einem *sequencing by synthesis*-Ansatz, sondern unter Einsatz einer Ligase (*sequencing by ligation*). Auch hier bindet zunächst ein Primer, der den Sequenzierstart festlegt. Dieser wird aber nicht mittels einer Polymerase verlängert, sondern hier befinden sich kurze Oligonukleotide, die in Abhängigkeit von den ersten beiden Basen mit einer von vier Fluoreszenzfarben markiert sind. Die Verknüpfung zweier Basen erlaubt prinzipiell 16 Kombinationen, so dass mit vier Fluoreszenzfarben nicht alle Möglichkeiten eindeutig abgebildet werden können. Im Gegensatz zu anderen System verzichtet die SOLiD-Technologie darauf, eine Base eindeutig zu identifizieren. Die erhaltene Information bezieht sich immer auf den Basenübergang. So ist die farbliche Markierung der Kombination identischer Basen für alle vier Nukleotide (TT/AA/GG/CC) identisch, erst aus der Kenntnis der ersten Base lässt sich die Identität der zweiten Base ableiten. Sobald eines dieser Oligonukleotide direkt anschließend an den Sequenzierprimer bindet, kann durch eine Ligase die Verknüpfung erstellt werden. Ähnlich einer Polymerase ist eine erfolgreiche Ligation des Oligonukleotids nur bei korrekter Basenpaarung wahrscheinlich. Fehlgepaarte Oligonukleotide werden als solche erkannt und dissoziieren wieder vom Strang ab. Anschließend werden nicht gebundene Oligonukleotide entfernt und das inkorporierte Oligonukleotid anhand des Fluoreszenzfarbstoffes mittels einer hochauflösenden Kamera identifiziert. Abgeschlossen wird der Zyklus durch die Entfernung des Farbstoffes. Bei dieser Reaktion wird das inkorporierte Oligonukleotid auf fünf Basen gekürzt und kann im nächsten Zyklus als verlängerter Primer für die Ligation eines weiteren Oligonukleotids verwendet werden.

Die zyklische Ligation/Detektion wird mehrfach mit einen um eine Base versetzten Anfangsprimer wiederholt. Erst wenn dieses Verfahren für fünf um jeweils eine Base verschobene Primer durchgeführt wurde, kann die Sequenz vollständig ermittelt werden. Die Farbkodierung des Basenüberganges gegenüber der direkten Farbkodierung der einzelnen Basen hat dabei den Vorteil, dass im Abgleich zu einem Referenzgenom Sequenzierfehler gegenüber SNP besser abgegrenzt werden können. Während diese bei den anderen Systemen nicht von einem SNP zu unterscheiden

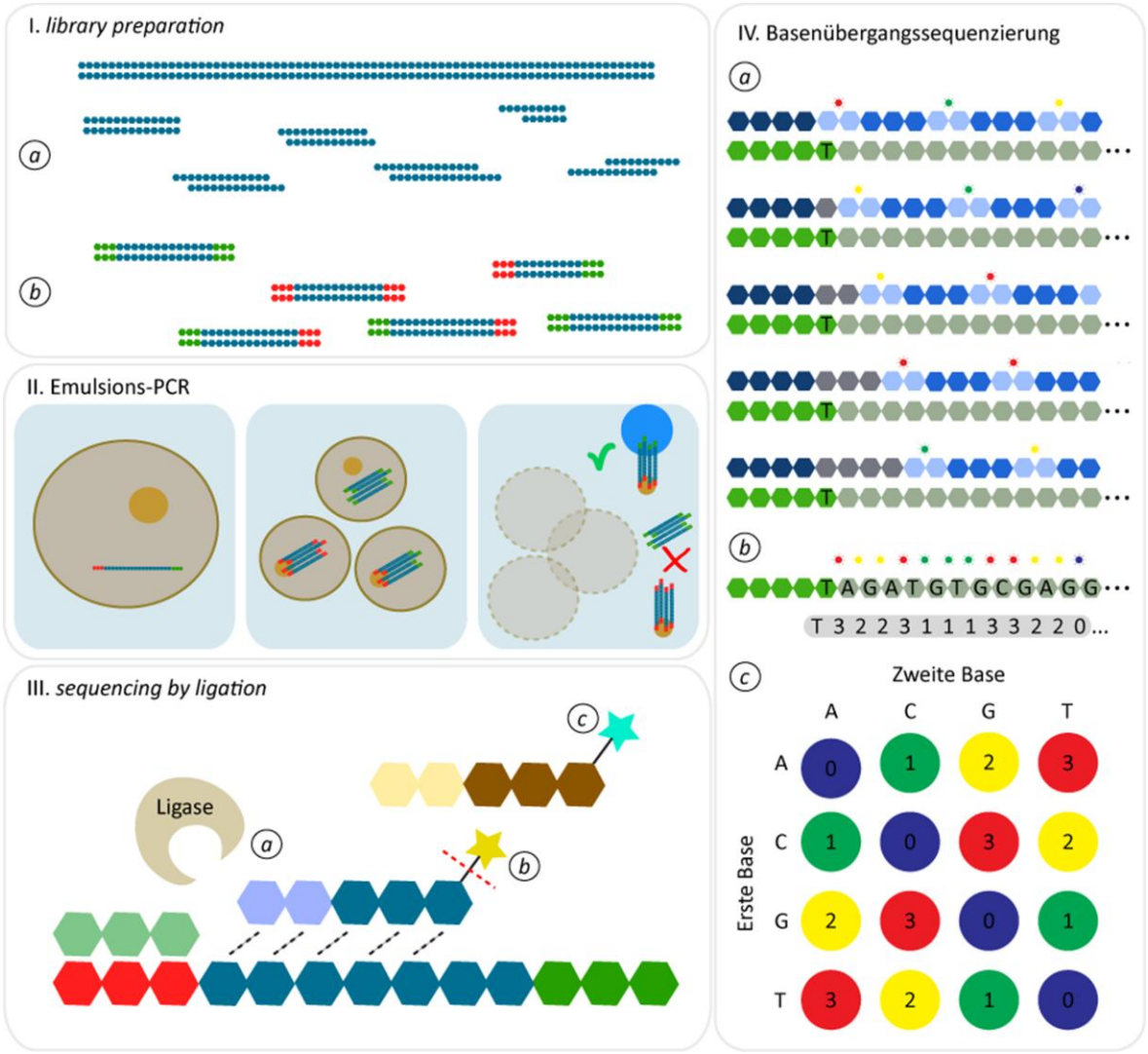


Abb. 4: Life Technologies SOLiD-System.

(I) Die ersten Schritte der *library preparation* sind im Wesentlichen mit der Pyrosequenzierungstechnologien vergleichbar: die DNA wird mittels Ultraschall in kleine Fragmente geschert (a) und an die Enden werden zwei unterschiedliche Adaptoren ligiert (b). Das SOLiD-System verzichtet zunächst auf eine Selektion von Fragmenten, die beide Adaptoren tragen. (II) Die generierten Fragmente werden nun in einer Öl-Wasser-Emulsion über einen der Adaptoren amplifiziert und an *beads* immobilisiert, auch hier ist das richtige Verhältnis zwischen *beads* und DNA-Fragmenten entscheidend, um eine möglichst hohe Ausbeute an monoklonal beladenen *beads* zu erhalten. Im Anschluss an die Emulsions-PCR wird die Emulsion aufgebrochen und beladene *beads* werden über die Bindung an aufschwimmende Partikel (blau) selektioniert. So können *beads* angereichert werden, die nur mit beiden Adaptoren versehende DNA-Fragmente tragen. (III) Die eigentliche Sequenzierung erfolgt hier über die Ligation von kurzen Oligonukleotiden und nicht einzelner Nukleotide, Voraussetzung ist auch hier die erfolgreiche Basenpaarung des Oligonukleotids mit der Matrize. Nur dann kann eine Ligation erfolgen (a). Jedes Oligonukleotid trägt dabei eine von vier Fluoreszenzfarbstoffen, mittels einer hochauflösenden Kamera können diese detektiert und unterschieden werden (b). Nachdem die Art des Fluoreszenzfarbstoffes ausgelesen wurde, wird dieser abgespalten, der verbliebene Anteil des Oligonukleotids dient nun als Ankerpunkt für die Ligation des nächsten Oligonukleotids (c). (III) Dieser Vorgang der Ligation kann wiederholt werden (aktuell zehnfach), über die Verwendung von vier Fluoreszenzfarbstoffen kann aber nicht die Sequenz des gesamten Oligonukleotids kodiert werden. So gibt der Fluoreszenzfarbstoff nur eine Auskunft über die Zusammensetzung der ersten zwei Basen. Daher werden nach erfolgreicher Ligation und Identifikation alle Oligonukleotide entfernt und der Vorgang erneut gestartet, diesmal allerdings um eine Base

verschoben (a). Daraus folgt, dass jede Position der Matrize doppelt ausgelesen wird. Dies ist auch erforderlich, da nur vier Fluoreszenzfarbstoffe pro Oligonukleotid keine eindeutige Kodierung von zwei Nukleotiden erlauben. Ausgehend vom (bekannten) Startprimer kann die Sequenz rekonstruiert werden. Nicht die einzelne Base wird sequenziert, sondern der Basenübergang (b), d. h. bei bekannter Base kann nach einem bekannten Übergangsschema die nächste Base identifiziert werden (c). Dies erleichtert insbesondere die Detektion von einzelnen Nukleotiden gegenüber einer bekannten Referenzsequenz (SNP), da so Fehler im Sequenziervorgang (eine Abweichung) von echten SNP (zwei Abweichungen) unterschieden werden können. [basierend auf Informationen von www.appliedbiosystems.com]

sind, wird beim SOLiD-System ein Sequenzierfehler durch Inkorporation eines falschen Oligonukleotids durch eine einzelne Abweichung gegenüber der Referenz kenntlich. Ein SNP zeigt hingegen zwei Abweichungen gegenüber der Referenz (der Basenübergang hin zum SNP wird durch einen anderen Farbstoff markiert, aber auch der folgende Übergang vom SNP zur nächsten Referenzbase ist ein anderer). **Abb. 4** gibt einen graphischen Einblick über wesentliche Abläufe der SOLiD-Sequenzierung.

Momentan ist der SOLiD V4 in der Lage, *read*-Längen von 50 Basenpaaren zu erreichen, auch besteht hier die Möglichkeit, zwei Fragmente in bekannter räumlicher Anordnung zu sequenzieren (*mate-pair* und *paired-end sequencing*). In der neusten Version (5500xl SOLiD System) verspricht der Hersteller *read*-Längen von bis zu 75 bp (*mate-pair*: 2x60). Insgesamt sollen pro Lauf bis zu 300 Gigabasen Sequenzdaten erhalten werden können.

1.5. Zielsetzung der Arbeit

Next generation sequencing ist jedoch nicht beschränkt auf die Entzifferung des Erbgutes, sondern kann auch zur funktionellen Analyse des Genoms genutzt werden. So kann z.B. durch Bisulfitbehandlung der DNA die methylierte und unmethylierte Form der Base Cytidin unterschieden werden (Frommer u. a. 1992) oder durch Präzipitation von DNA-bindenden Proteinen und anschließender Sequenzierung der copräzipitierten DNA-Abschnitte der Bindungsort von Transkriptionsfaktoren oder modifizierten Histonen im Genom lokalisiert werden (*chromatin immunoprecipitation sequencing*, ChIPseq) (Park 2009).

Wie beschrieben, konnte auch das Transkriptom in der Vergangenheit durch herkömmliche Sequenziermethoden untersucht werden. Zumindest für den Roche *Genome Sequencer* FLX wurde bereits vor Aufnahme dieser Arbeit beschrieben, dass *next generation sequencing* geeignet ist, um einen tiefen Einblick in das Transkriptom zu erzielen (Cheung u. a. 2006; Emrich u. a. 2007).

In dieser Arbeit sollen Methoden zur *next generation sequencing* basierten RNA-Sequenzierung (RNA-Seq) etabliert und erprobt werden. Im nächsten Schritt soll das Transkriptom intestinalen Gewebes der Maus als Modellorganismus für die Erforschung des Darms untersucht werden, um die molekularen Prozesse auf Ebene des Transkriptoms näher zu betrachten, ein besseres Verständnis

der Entwicklung und der Funktion als bedeutendes Barriere- und Immunorgan zu entwickeln, die in der Zukunft für die Diagnose und Therapie von Erkrankungen des Intestinums hilfreich sein können. Folgende Fragestellungen bzw. Hypothesen sollen bearbeitet werden:

- RNA-Seq ist geeignet, um das Transkriptom zu untersuchen und dabei anderen Methoden überlegen.
- Die Komplexität der intestinalen Genexpression soll ermittelt und mit geeigneten Methoden weiter untersucht werden.
- Posttranskriptionelle Modifikationen wie alternatives Spleißen oder zusätzliche Polyadenylierungssignale von Genen in intestinalen Geweben sollen untersucht werden.
- Die bisherige Annotation von Genen ist unvollständig. RNA-Seq-Daten erlauben die Identifizierung von nicht-annotierten, transkriptionell aktiven Regionen (nTAR) im Genom.
- Die nähere Charakterisierung potentiell identifizierbarer nTAR bezüglich ihrer Lage und Orientierung im Genom, ihrer Expression sowie ihrer möglichen Funktion sollen untersucht werden.

2. Material und Methoden

2.1. Molekularbiologische Methoden für RNA-Seq

Im Anhang findet sich eine Liste mit detaillierten Angaben zur Herkunft der verwendeten Chemikalien, Enzyme und zu weiteren für diese Arbeit genutzten Materialien.

2.1.1. Verwendete Mauslinien

Für die in dieser Doktorarbeit durchgeführten Experimente wurden Gewebe des Maus-Inzuchtstammes C57/BL6 verwendet. Die Tiere wurden im Tierhaus der CAU Kiel (Viktor-Hensen-Haus) unter Berücksichtigung des *„standard of care for specific pathogen free mice“* (SPF-Konditionen) aufgezogen. Experimente am lebenden Organismus wurden nicht vorgenommen, so dass es sich gemäß §7 des deutschen Tierschutzgesetzes nicht um Tierversuche im juristischen Sinn handelte. Die Tiere wurden im Alter von 9 bis 10 Wochen von einer geschulten Tierärztin durch Genickbruch getötet. Anschließend wurde umgehend die Bauchhöhle eröffnet und die Organe entnommen. Um eine schnelle Durchkühlung der Proben zu sichern, wurden jeweils das Jejunum und das Colon in kleinere Bestandteile separiert, in vorbereitete Cryoröhrchen transferiert und in flüssigem Stickstoff schockgefroren. Anschließend wurden die Proben bei -80°C bis zur weiteren Bearbeitung gelagert.

2.1.2. Isolation von Gesamt-RNA

Zur Isolation der Gesamt-RNA aus den intestinalen Mausgeweben wurden zwei Systeme gemäß den Angaben des Herstellers eingesetzt: (a) das *„RNeasy mini“*-Kit von Qiagen für das SMART-RNA-Seq-Protokoll, sowie (b) das *„mirVana miRNA Isolation“*–Kit für das WTAK-RNA-Seq-Protokoll. Für beide Ansätze wurde zunächst das stickstoffgefrorene Gewebe aufgeschlossen und homogenisiert, indem das gefrorene Gewebe in einen Tieftemperaturmörser mit integrierter Stickstoffkühlung überführt und mit einem vorgekühlten Pistill zügig zerrieben wurde. Durch die konsequente Kühlung wurde gewährleistet, dass die RNA nicht durch im Gewebe vorhandene Ribonukleasen (RNasen) degradiert werden konnte.

Das *„RNeasy mini“*-Kit basiert auf der Nutzung von hochkonzentrierten Salzlösungen und Ethanol, um die Löslichkeit von Nukleinsäuren in Wasser soweit zu reduzieren, dass diese gefällt und an einer Säulenmatrix gebunden werden können. Zur Isolation mit dem *„RNeasy mini“*-Kit wurden bis zu 30 mg des homogenisierten Gewebes in 550 µl RLT-Puffer überführt, welcher zuvor mit 10 µl/ml 2-Sulfanylethanol (ß-Mercaptoethanol) versetzt wurde. Eine gründliche Durchmischung unter Nutzung eines Vortex stellte sicher, dass das homogenisierte Gewebe nicht verklumpte und

gleichmäßig im RLT-Puffer verteilt war. Durch die Stabilisierung des Gewebes im Puffer konnten sämtliche weitere Schritte bei Raumtemperatur erfolgen. Zunächst wurde das Gemisch auf „Qiashredder"-Säulen übertragen und bei einer relativen Zentrifugalbeschleunigung (*relative centrifugal force*, RCF) von 16.100 × *g* für 2 min zentrifugiert. Das Homogenat wurde in ein Mikrolitergefäß überführt und erneut bei einer RCF von 16.100 × *g* für 3 min zentrifugiert. Nichtlösliche Zellbestandteile im Sediment wurden verworfen, der Überstand in ein neues Mikrolitergefäß übertragen und mit 550 µl 70 % Ethanol versetzt. Nach gründlicher Durchmischung wurde die Lösung auf eine RNeasy-Säule übertragen und bei einer RCF von 16.100× *g* für 1 min zentrifugiert (unter Beachtung der maximalen Volumenbeladung von 700 µl, daher wurde diese Prozedur in zwei Schritten durchgeführt). Im nächsten Schritt wurde die Säule in ein neues Sammelgefäß übertragen, es erfolgte ein Verdau der DNA durch Zugabe von Desoxyribonuklease I (DNase I) (pro Säule 10 µl DNase I in 70 µl RDD-Puffer, Qiagen). Anschließend wurde die Säule mit RW1-Puffer gewaschen, indem 700 µl des Puffers auf die Säule gegeben, mit einer RCF von 16.100× *g* für 30 s zentrifugiert und der Durchfluss verworfen wurde. Zur weiteren Aufreinigung wurde dieser Schritt zweifach mit 500 µl des Waschpuffers RPE wiederholt, im letzten Waschschritt wurde dabei allerdings die Zentifugalzeit auf 2 min erhöht. Um sicherzustellen, dass sich nach dem letzten Waschschritt kein Puffer mehr auf der Säule befand, wurden die Säulen erneut in ein neues Sammelgefäß übertragen und zusätzlich ohne Zugabe eines Puffers für 2 min (RCF 16.100× *g*) zentrifugiert. Anschließend konnte die RNA mit 50 µl RNase-freiem Wasser eluiert werden, dazu wurde die Säule in ein vorbereitetes Mikrolitergefäß überführt und nach Zugabe des Wassers für 1 min (RCF 16.100× *g*) zentrifugiert. Bis zur weiteren Verwendung wurde die RNA bei -80 °C gelagert.
Das „mirVana miRNA Isolation" -Kit beruht auf einer Kombination von Phenol-Chloroform-Extraktion und anschließender Ethanolfällung und Aufreinigung über eine Säule. Hierzu wurden bis zu 100 mg des im Tiefkühlmörser zerriebenen Gewebes in 1 ml *Lysis buffer* aufgenommen und durch Nutzung eines Vortex kräftig geschüttelt, bis sich das Pulver gleichmäßig verteilt hat. Anschließend wurden 100 µl *miRNA Homogenate Additive* der Gewebssuspension zugesetzt, erneut mit einem Vortex kräftig durchmischt und der Ansatz für 10 min auf Eis inkubiert. Danach wurde 1 ml einer gebrauchsfertigen Phenol-Chloroform-Isoamylalkohol-Lösung (im Verhältnis 25:24:1; pH 6,7) in den Ansatz gegeben und dieser für 60 s auf einem Vortex durchmischt. Nun wurde der Reaktionsansatz mit einer RCF von 16.100× g für 5 min bei Raumtemperatur zentrifugiert. Für die weitere RNA-Isolation wurde die obere, wässrige Phase vorsichtig abpipettiert und auf zwei Mikroliterreaktionsgefäße verteilt, mit jeweils 625 µl Ethanol versetzt und kurz durchmischt. In Schritten von bis zu 700 µl wurde mit dieser Lösung eine mirVana-Säule beladen (für jeden Schritt für 30 s bei 10.000× *g* RCF zentrifugiert, der Durchfluss wurde verworfen). Anschließend wurde die RNA-bindende Säulen gewaschen, zunächst mit 700 µl *miRNA Wash Solution 1*, darauf zweifach mit 500 µl

miRNA Wash Solution 2/3, wobei jeweils nach Zugabe des Waschpuffers für 30 s bei 10.000× *g* RCF zentrifugiert und der Durchfluss verworfen wurde. Im letzten Schritt wurde die Säule in ein neues Mikrolitergefäß übertragen, mit 100 µl auf 95 °C vorgeheizte *Elution Solution* versetzt und bei 16.100× *g* RCF für 30 s zentrifugiert. Die eluierte RNA wurde bis zur weiteren Verwendung bei -80 °C gelagert.

2.1.3. Quantitäts- und Qualitätskontrolle der isolierten RNA

Um die Quantität und Qualität der RNA zu bestimmen, wurden jeweils drei Untersuchungen vorgenommen:

(I) Als zügiges Verfahren, um Verunreinigungen durch z.B. Proteine auszuschließen und die Menge an RNA zu bestimmen, wurden 2 µl der RNA-Lösung mittels eines Nanodrop-ND-1000 Spektrophotometers vermessen (Ziel $E_{260}/_{E280} > 1{,}8$; $E_{260} = 1$ entspricht einer RNA-Konzentration von 40 ng/µl).

(II) Als zweites Verfahren wurde die RNA-Lösung kapillar-elektrophoretisch aufgetrennt und vermessen. Für diese Aufgabe wurde das Agilent *RNA 6000 Nano Kit* gemäß den Angaben des Herstellers genutzt. Dazu wurde 1 µl der Probe (mind. 25 ng RNA) für 2 min auf 72 °C erwärmt, um Sekundärstrukturen aufzulösen, und anschließend auf den vorbereiteten Chip geladen. Danach wurde der Chip in einen Agilent 2100 *Bioanalyzer* geladen und die Probe kapillar-elektrophoretisch aufgetrennt und mittels eines Sensors ein Elektropherogramm angefertigt. **Abb. 5** zeigt exemplarisch ein solches Elektropherogramm, welches zur Beurteilung der Qualität der RNA genutzt wurde. Insbesondere kann abgeschätzt werden, ob die RNA bereits vollständig oder partiell degradiert ist.

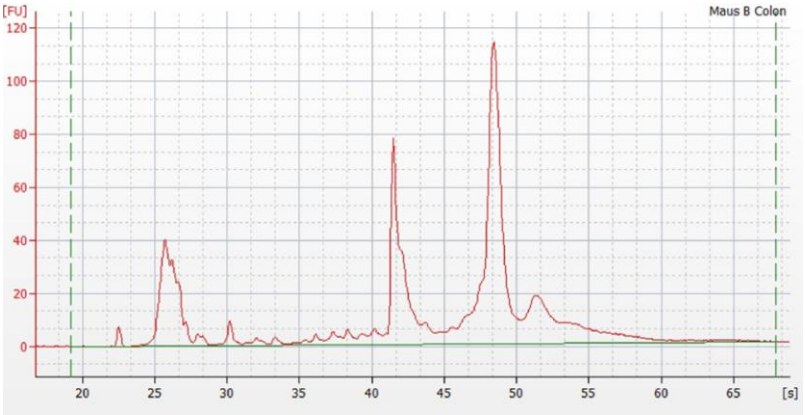


Abb. 5: Agilent Bioanalyzer RNA 6000 Nano-Elektropherogramm

Die Abbildung zeigt exemplarisch das Elektropherogramm für eine in dieser Arbeit mittels der mirVana-Prozedur isolierten RNA-Probe (gewonnen aus Colon-Gewebe). Das Elektropherogramm zeigt die Emission eines an RNA bindenden Fluoreszenzfarbstoffes (proportional zur RNA-Stoffmenge, FU) in Abhängigkeit der Zeit, die die RNA bis zum Erreichen des Sensors benötigt (kleine RNA-Nukleotide durchlaufen die Kapillare schneller als große Moleküle). Die beiden großen Fluoreszenzgipfel entsprechen der ribosomalen RNA (42 s 18 S RNA, 48 s 28 S RNA). Für die hier untersuchte Probe wurde eine Konzentration von 854 ng/µl und eine RNA Integrity Number (RIN) von 9,1 ermittelt.

Dies würde zu sehr kurzen Fragmenten und zur Auflösung der typischen Doppelgipfel-Struktur für intakte RNA-Isolate führen. Zusätzlich errechnet das Agilent-System auf Basis der RNA-Größenverteilung einen numerischen Qualitätswert, die *RNA Integrity Number* (RIN), welche Werte von 0 (komplett degradierte RNA) bis 10 (intakte RNA ohne Anzeichen von Degradation) annehmen kann. In dieser Arbeit wurde für sämtliche RNA-Seq-Proben ein RIN von mind. 7 vorausgesetzt.

(III) Als drittes Verfahren wurde eine (RT-)PCR gegen die Glycerinaldehyd-3-Phosphat-Dehydrogenase (GAPDH) durchgeführt, um die Expression dieses ubiquitär gebildeten Transkripts in der Probe nachzuweisen und eine Kontamination durch genomische DNA auszuschließen. Dazu wurde zunächst mittels reverser Transkription RNA in cDNA umgeschrieben (siehe **Tab. 2**).

Tab. 2: Reverse Transkription

100 ng	Total RNA
1 µl	Oligo-dT-Primer
Ad 6,75 µl	H_2O (bidest.)
70°C für 2min	
2 µl	5x MMLV-RT-Puffer
0,5 µl	dNTP (je 10mM)
0,25 µl	RNase-Inhibitor
0,5 µl	MMLV Reverse Transkriptase
42 °C für 60 min	
95 °C für 5 min	
10 °C für 10 min	
40 µl	H_2O (bidest.)

Lagerung bei -20 °C bis zur weiteren Verwendung.

Anschließend wurde sowohl unbehandelte Total-RNA als auch umgeschriebene cDNA für eine PCR eingesetzt (siehe **Tab. 3**). Die PCR-Ansätze wurden gelelektrophoretisch aufgetrennt (1% Agarosegel, 80 Volt für 30min, BioRad-System). Eine positive Qualitätskontrolle erfordert hier den

Tab. 3: GAPDH-PCR

2 µl		Total RNA oder cDNA
4 µl		GoTaq 5x Puffer
0,5 µl		dNTP (je 10mM)
0,5 µl		GAPDH-Primer (Clontech)
0,2 µl		GoTaq-Polymerase
Ad 20 µl		H_2O (bidest.)
PCR-Block:		
		96 °C für 4 min
	┌	96 °C für 30 s
35x	│	54 °C für 30 s
	└	72 °C für 1 min
		72 °C für 4 min
		4 °C bis Ende

Lagerung bei -20 °C bis zur weiteren Verwendung.

Nachweis eines Amplikons der ubiquitär exprimierten GAPDH in der cDNA-Probe (GAPDH-mRNA vorhanden), nicht aber in der Total-RNA. Dies würde bedeuten, dass in der Total-RNA bereits DNA vorliegen würde, was auf eine Kontamination mit genomischer DNA hindeuten würde.

2.1.4. Natriumacetat-Fällung von RNA

Lag die RNA in zu großen Volumina gelöst vor, um weiter verarbeitet zu werden, wurde die RNA durch eine Natriumacetat-Fällung aufkonzentriert. Dazu wurde die Probe mit einem Zehntel Volumen 3 M Natriumacetat-Lösung (pH 5,2) versetzt und gut gemischt. Anschließend wurde das Vierfache des Volumens an Ethanol (100 %) zu der Probe gegeben. Nach erneuter Durchmischung wurde die RNA über Nacht (mind. für vier Stunden) bei -20 °C gefällt. Am folgenden Tag wurde die Lösung für 30 min bei 4 °C und einer RCF von 16.100× *g* zentrifugiert. Anschließend wurde der Überstand vorsichtig abpipettiert und verworfen. Das Sediment mit der RNA wurde zweifach mit 500 µl kaltem Ethanol (80 %, 4 °C) gewaschen. Dabei erfolgte die Durchmischung durch mehrmaliges Auf- und Abpipettieren gefolgt von einer Zentrifugation (10 min, 4 °C, 16.100× *g*). Um nach dem zweiten Waschschritt verbliebenes Ethanol zu entfernen, wurden die Proben vorsichtig in den offenen Mikrolitergefäßen bei Raumtemperatur getrocknet (abhängig von der Menge des verbliebenen Ethanols, in der Regel für ca. 5-10 min), bis keine Flüssigkeit mehr am Sediment zu beobachten war. Anschließend konnte die RNA in einem geeigneten Volumen RNase-freien Wassers aufgenommen werden.

2.1.5. Aufreinigung von mRNA

Zur Aufreinigung von mRNA aus Total-RNA wurde das *Oligotex Direct mRNA Mini Kit* von Qiagen genutzt. Hierbei wird polyadenylierte mRNA über Oligo-dT-Nukleotide an feine Resin-Partikel (Oligotex) gebunden und damit von der restlichen RNA separiert, nach mehreren Waschschritten wird die RNA von den Partikeln eluiert und kann für folgende Schritte eingesetzt werden.

Bis zu 250 µg Total-RNA wurden auf ein Volumen von 250 µl eingestellt, anschließend wurde zu dem Ansatz 250 µl OBB-Puffer und 15 µl Oligotex-Suspension gegeben. Dieser Ansatz wurde nun für 3 min in einem vorgewärmten Heizblock auf 70 °C erwärmt. Anschließend wurden die Proben langsam abgekühlt, indem sie für 10 min bei Raumtemperatur belassen wurden. Mittels Zentrifugation bei einer RCF von 16.100× *g* für 2 min wurden die Oligotex-Partikel sedimentiert und der Überstand verworfen. 400 µl OW2-Puffer wurden eingesetzt, um das Sediment unter kräftiger Vortex-Durchmischung zu resuspendieren. Mit dieser Suspension wurden die im Kit enthaltenen Säulen beladen. Die Reaktionsansätze wurden für 1 min mit einer RCF von 16.100× *g* zentrifugiert und die Säulen auf ein neues Mikrolitergefäß übertragen. Zum Waschen wurden 400 µl OW2-Puffer auf die Säulen gegeben und abzentrifugiert (RCF 16.100× *g*, 1 min). Die Säulen wurden erneut auf ein neues

Mikrolitergefäß übertragen und mit 20 µl im Heizblock auf 70 °C vorgewärmten OEB-Puffer beladen. Zügig wurden die auf der Säule liegenden Oligotex-Partikel durch Auf- und Abpipettieren im OEB-Puffer resuspendiert. Die Elution erfolgte durch Zentrifugation (RCF 16.100× *g*, 1 min).

2.1.6. Modifizierte SMART cDNA-Synthese

Für die cDNA-Synthese wurde ein Verfahren angepasst, dass bereits erfolgreich für den Roche *Genome Analyzer* eingesetzt werden konnte und im Wesentlichen auf der *template-switch*-Technologie des Clontech SMART cDNA *synthesis*-Kit beruhte. Dazu wurde für die Erststrangsynthese ein Primer gewählt, der erneut an die Polyadenylierung bindet (zweifache Selektion für polyadenylierte Transkripte zusätzlich zur mRNA-Aufreinigung, siehe 2.1.5). Dieses Verfahren sollte sicherstellen, dass insbesondere die intrazellulär in großen Mengen vorliegende, aber kaum differentiell regulierte ribosomale RNA effektiv entfernt wurde. Zusätzlich wurde der Primer am 5'-Ende durch eine definierte Sequenz flankiert.

Als *template-switch* wird dabei ein Verfahren bezeichnet, dass auf der Verwendung einer modifizierten reversen Transkriptase und einem zusätzlichen RNA-Oligonukleotid beruht. Die verwendete reverse Transkriptase (genetisch modifiziert ausgehend von einem murinen Leukämievirus) ist in der Lage, am Ende des Matrizenstranges zusätzlich ohne Vorlage zusätzliche Cytidin-Nukleotide an den synthetisierten Strang anzuhängen. Diese ergänzten Basen erlauben in der Folge die Hybridisierung und Bindung des zusätzlichen Oligonukleotids. Neben einer Guanosin-reichen Nukleotidfolge zur Bindung an den Erststrang besitzt dieses Oligonukleotid zusätzlich eine flankierende Sequenz, die nach Bindung als Matrize dient und das Erststrang-cDNA-Produkt am 3'-Ende um eine definierte Sequenz ergänzt. In der Folge wurde eine Erststrang-cDNA generiert, die an beiden Enden durch definierte, zueinander revers komplementäre Sequenzen ergänzt wurde. In der Regel wurden 500 ng aufgereinigte mRNA eingesetzt (siehe **Tab. 4**).

Tab. 4: SMART-Erststrang-Synthese

500 ng (1 µg)	mRNA (Total-RNA)
1 µl	3'-SMART CDS Primer II A (12 µM)
1 µl	SMART II A Oligonucleotide (12 µM)
Ad 5 µl	$H_2O_{(bidest.)}$
72 °C für 2 min	
4 °C für 2 min	
2 µl	*First strand*-Puffer
1 µl	Dithiothreitol (DTT, 20 mM)
1 µl	dNTP (je 10 mM)
1 µl	Superscript II Reverse Transkriptase
42 °C für 60 min	

Lagerung bei -20 °C bis zur weiteren Verwendung.

Diese Prozedur erlaubte durch die revers-komplementären Enden der cDNA zum einen eine Amplifikation der cDNA über einen einzigen Primer. Daneben entscheidend war, dass nur die cDNA amplifiziert wurde, in denen die reverse Transkription über die gesamte Länge des Transkripts erfolgte, unvollständig revers-transkribierte cDNA besaß nur an einem Ende die für die Amplifikation benötigte Sequenz. Die generierte Erststrang-cDNA wurde in einer PCR mittels des *Advantage 2 PCR* Kits amplifiziert, dabei wurde abweichend vom SMART-Protokoll die Primer durch eine biotinylierte Variante ersetzt (siehe **Tab. 5**).

Tab. 5: SMART-cDNA-Amplifikation

Modifizierter 5' PCR Primer II A: 5'-Biotin-AAG CAG TGG TAT CAA CGC AGA GT-3'

	2 µl	Erststrangprodukt
	2 µl	50x dNTP-Mix (je 10 mM)
	4 µl	5' PCR-Primer II A (12 µM, modifiziert)
	10 µl	10x Advantage 2 PCR-Puffer4 µl
	2 µl	Advantage 2 PCR Polymerase
	80 µl	$H_2O_{(bidest.)}$
		95 °C für 1 min
	┌	95 °C für 15 s
13x (15x)	|	65 °C für 30 s
	└	68 °C für 7 min
		4 °C bis Ende

Lagerung bei -20 °C bis zur weiteren Verwendung. Für Total-RNA wurden 15 Amplifikationszyklen angewandt.

Die Konzentration der amplifizierten cDNA-Bibliothek wurde mittels Nanodrop bestimmt, zusätzlich wurden 2 µl des Ansatzes mittels Gelelektrophorese aufgetrennt und die Größenverteilung der cDNA-Bibliothek dokumentiert. Für jede Probe wurden zwei cDNA-Amplifikationsansätze erstellt, üblicherweise betrugen die Ausbeuten 3-4 µg cDNA. Das verbliebene Erststrangprodukt wurde bei -20 °C verwahrt.

2.1.7. cDNA-Sequenzierung auf dem SOLiD V2

3 µg cDNA wurden als Startmaterial für die Sequenzierung auf der institutseigenen Sequenzierplattform mittels des SOLiD V2-Systems eingesetzt. Zur Erstellung der zu sequenzierenden *library* wurde das V2 35 bp *fragment library*-Protokoll genutzt. In Abweichung von diesem Protokoll wurden die biotinylierten cDNA-Enden mittels immobilisiertem Streptavidin (*Dynabeads*) entfernt. Dazu wurde direkt nach dem *end polishing* (im Anschluss an die Fragmentierung) das gleiche Volumen Streptavidin-Magnetkügelchen (5 µg/µl in 2x B&W-Puffer) zu den Proben gegeben und für 15 min bei Raumtemperatur leicht geschüttelt. Die vorsichtig geschätzte Menge biotinylierter cDNA-Enden (Annahme: 3 µg Einsatz, durchschnittliche Nukleotidlänge von mind. 500 bp) liegt bei unter 20 pmol. Die vom Hersteller angegebene Bindungskapazität der Streptavidin-Kügelchen liegt in der in diesem Ansatz verwandten Menge bei ca. 200 pmol. Der große Überschuss sollte eine möglichst

vollständige Entfernung der biotinylierten cDNA-Enden sicherstellen. Anschließend wurden die mit cDNA-Enden beladenen Streptavidin-Partikel über einen Magneten sedimentiert. Der Überstand wurde vorsichtig abpipettiert und für den nächsten Schritt der *fragment library*-Erstellung genutzt (Aufreinigung der fragmentierten cDNA). Die so erstellten *fragment libraries* wurden auf je einer halben SOLiD V2 *flow cell* sequenziert. Die wesentlichen Prinzipien des *sequencing by ligation* sind in Kapitel 1.4.2 erläutert. Detaillierte Angaben zur Probenhandhabung und dem schrittweisen Ablauf können beim Hersteller (http://www.appliedbiosystems.com) und auf der Sequenzierplattform des Instituts für Klinische Molekularbiologie erfragt werden (http://www.ikmb.uni-kiel.de/cms/technologien/sequenzierung/).

2.1.8. *microarray*-Expressionsanalysen

Die auf *microarrays* basierenden Expressionsanalysen wurden standardisiert in Zusammenarbeit mit der institutsinternen Technologieplattform Systematische Expressionsanalyse durchgeführt. Hierzu wurden Affymetrix *Mouse 430 2.0-Microarrays* sowie die vom Hersteller für die Hybridisierungs-, Wasch- und Färbeprozedur empfohlene Komponenten unter Einhaltung des MIAME-Standards (*Minimum Information About a Microarray Experiment*) genutzt. Kurz zusammengefasst wurde ausgehend von 15 µg Total-RNA zunächst Erst- und Zweitstrangsynthese mit einem T7-Promotor-Sequenz flankierten Oligo-dT-Primer durchgeführt. Die so erzeugte cDNA wurde genutzt, um mittels *in vitro*-Transkription Biotin-markierte Komplementär-RNA (*complementary* RNA, cRNA) zu erzeugen. Die cRNA wurde fragmentiert und zur Hybridisierung mit dem *microarray* genutzt. Nachdem nicht gebundene Fragmente vom *microarray* über Waschschritte entfernt wurden und über die Biotin-Markierung eine antikörpervermittelte Fluoreszenzfärbung vorgenommen wurde, wurde die Menge an hybridisierter cRNA für die einzelnen Oligonukleotidpositionen über den *microarray*-Scanner ausgelesen. Die für die einzelnen Genpositionen ermittelte Expressionsstärke wurde mittels der Methode *robust multiarray average* (RMA) normalisiert (Irizarry u. a. 2003). Die Durchführung erfolgte nach Herstellerangaben (http://www.affymetrix.com).

2.1.9. *Whole transcriptome*-Sequenzierung mit dem SOLiD V4

Das von Applied Biosystems für RNA-Sequenzierung unterstützte SOLiD *Total RNA-Seq Kit* wurde in enger Zusammenarbeit mit der institutseigenen Sequenzierplattform angewandt, um RNA-Proben auf dem weiter entwickelten SOLiD V4-System zu analysieren. Nach Empfehlungen des Herstellers wurden aus 20 µg Total-RNA mittels des *MicroPoly(A)Purist*-Kits von Ambion polyadenylierte RNA-Formen selektioniert und für das *Whole Transcriptome Library Preparation*-Protokoll eingesetzt. Dieses Protokoll beginnt direkt mit der Fragmentierung der RNA. Von den beiden im Protokoll unterstützten Methoden wurde hier die chemisch-thermische Fragmentierung gewählt. Die

fragmentierte RNA wurde über das *Purelink RNA Micro Kit* aufgereinigt. Im nächsten Schritt wurden die SOLiD-Adaptoren über Hybridisierung und Ligation an die RNA-Fragmente angefügt, um anschließend mittels reverser Transkription die RNA in cDNA umzuschreiben. Nun erfolgte eine Größenselektion der cDNA-Fragmente mittels *Agencourt AMPure XP Reagent*. In zwei Schritten wurden zunächst Fragmente kleiner 100 bp und anschließend Fragmente größer 160 bp mittels dieser *solid phase reversible immobilisation* (SPRI)-Technik aus der cDNA-Bibliothek entfernt. Die selektionierte cDNA wurde mittels PCR amplifiziert und die Ausbeute und Größenverteilung kontrolliert.

Die erzeugte cDNA-Bibliothek wurde zusammen mit P1 DNA-*beads*, einer wässrigen Lösung, die alle erforderlichen Komponenten einer PCR beinhaltet, sowie einem Öl in ein 50 ml-Röhrchen gegeben und mittels des *Ultra-Turrax Tube Drive* in eine feine Emulsion verwandelt. Für die Emulsions-PCR wurde die Emulsion in eine 96-Loch-Platte übertragen. Nach erfolgter ePCR wurde die Emulsion aus der 96-Loch-Platte mittels Zentrifugation im SOLiD *emulsion collection tape* zusammengeführt und mit 2-Butanol durch Auf- und Abpipettieren aufgebrochen. Nach Überführung in ein neues 50 ml-Röhrchen und erneuter Zentrifugation konnte das 2-Butanol-Öl-Gemisch vorsichtig abgekippt werden. Um die im Sediment verbliebenen *beads* zu trocknen, wurde das 50 ml-Röhrchen vorsichtig nach unten gekippt und auf ein Papiertuch gestellt. Nachdem die Emulsion gebrochen wurde, wurden die *beads* gewaschen und quantifiziert.

Um *beads* aus der Lösung zu entfernen, die keine oder nur wenig amplifizierte DNA trugen, wurden diese mit P2-Polystyrol-*beads* in einer 60 % Glycerinlösung zum Ansatz gegeben. Korrekt beladene *beads* bilden mit den P2-Polystyrol-*beads* ein an der Oberfläche schwimmendes Netz und können so leicht von nichtbeladenen *beads* getrennt werden. Zuletzt wurde die Anzahl der DNA-beladenen *beads* quantifiziert und im Anschluss auf je einer Viertel *flow cell* des SOLiD V4 zur Sequenzierung deponiert. Die Durchführung erfolgte nach Herstellerangaben (http://www.appliedbiosystems.com).

2.1.10. Polyadenylierungsnachweis mittels *pyrosequencing*

Um direkt polyadenylierte Transkripte zu bestimmen, wurde eine modifizierte Variante des Roche FLX *pyrosequencings* eingesetzt (Torres u. a. 2008). Ausgehend von 3 µg Total-RNA wurde unter Nutzung des *RevertAid H Minus First Strand cDNA Synthesis*-Kits von Fermentas nach den Angaben des Herstellers die Synthese der cDNA durchgeführt. Anstelle des Oligo-dT-Primers wurde eine modifizierte Variante genutzt, die bereits den Sequenzier-Adapter B des Roche FLX-Systems enthielt. Die cDNA-Erststrangsynthese ist in **Tab. 6** beschrieben.

Tab. 6: Modifizierte Erststrang-cDNA-Synthese für *pyrosequencing*

Modifizierter Oligo-dT-Primer (mit Roche FLX B-Adaptor):

5'-Biotin-GCC TGG CCA GCC CGC TCA $G(T)_{17}V$-3'

3 µg	mRNA aus intestinalen Geweben
1 µl	modifizierter Oligo-dT-Primer (s.o.; 100 µM)
Ad 12 µl	$H_2O_{(bidest.)}$
	70 °C für 5 min
	4 °C für 2 min
4 µl	5x Reaktionspuffer
1 µl	RiboLock Ribonukleaseinhibitor
2 µl	dNTP-Mix (je 10 mM)
	37 °C für 5 min
1µl	*RevertAid H Minus* M-MuLV Reverse Transkriptase
	42 °C für 60 min
	70 °C für 10 min
	4 °C bis Ende

Lagerung bei -20 °C bis zur weiteren Verwendung.

Ausgehend von der Erststrangsynthese wurde im nächsten Schritt eine Zweitstrangsynthese nach den Empfehlungen des *RevertAid H Minus cDNA Synthesis*-Kits durchgeführt (die dafür benötigten Komponenten sind nicht im Umfang des Kits enthalten, siehe **Tab. 7**).

Mittels dieser Methode konnten ca. 3 µg ds (*double stranded,* doppelsträngige)-cDNA gewonnen werden, die in Zusammenarbeit mit der Sequenzierplattform des Instituts für das *pyrosequencing*–Verfahren genutzt wurden. Dabei wurde entsprechend der Angaben des Herstellers die cDNA mittels Nebulisierung fragmentiert (Margulies u. a. 2005), die biotinylierten 3'-Enden über M-270 *Streptavidin beads* selektioniert, mittels T4 DNA-Polymerase ein *end polishing* durchgeführt und der doppelsträngige Adaptor A an die 3'-Fragmente ligiert. Die Sequenzierung erfolgte auf einem Roche *Genome Sequencer* GS FLX. Die Durchführung erfolgte nach Herstellerangaben (https://www.roche-applied-science.com).

Tab. 7: RevertAid-Zweitstrangsynthese

20 µl	Erststrang-Syntheseprodukt (kompletter Ansatz)
8 µl	10x Reaktionspuffer für DNA-Polymerase I
0,2 µl	*E. coli* Ribonucelase H (5 U/µl)
4 µl	*E. coli* DNA-Polymerase I
Ad 100 µl	$H_2O_{(bidest.)}$
	15 °C für 120 min
2,5 µl	T4 DNA-Polymerase (5 U/µl)
	15 °C für 5 min
4 µl	Ethylendiamintetraessigsäure (EDTA, 500 mM, pH 8)

Lagerung bei -20 °C bis zur weiteren Verwendung.

2.2. Bioinformatische Methoden des RNA-Seq

Bioinformatische Fragestellungen wurden in enger Zusammenarbeit mit Mitarbeitern der institutsinternen Arbeitsgruppe Bioinformatik bearbeitet. Insbesondere die Steuerung des *clusters* im Rechenzentrum und die erforderliche Übersetzung der Algorithmen in funktionstüchtige Computerprogramme wären ohne diese Unterstützung nicht möglich gewesen.

2.2.1. Genomische Zuordnung der *reads* (*mapping*)

Um die mittels RNA-Seq ermittelten Nukleotidsequenzen mit der genomischen Referenz abzugleichen (*alignment* oder *mapping*), wurde das von Applied Biosystems für den SOLiD entwickelte Programm *BioScope* in der Version 1.2.1 genutzt. Für das *mapping* wurden die vom SOLiD-System erstellten Rohdaten (*colour space reads*, .csfasta) zunächst mit einem Filter abgeglichen, in denen unter anderem *reads* beruhend auf Adaptorsequenzen, ribosomaler oder Transfer-RNA aus der Analyse entfernt wurden. Im nächsten Schritt erfolgte der Abgleich mit dem murinen Genom, indem für mittels SMART-RNA-Seq erzeugte *reads* zunächst Übereinstimmungen von 30 der 35 Basenpaare (*seed*) mit bis zu 3 Fehlern ermittelt wurden und diese nach Möglichkeit um die verbliebenen Basen verlängert wurden (Extension). Für die 50 bp-*reads* des WTAK-Protokolls wurden zunächst nach Übereinstimmungen von 38 bp mit maximal 3 Fehlern gesucht und ebenfalls, falls möglich, um die verbliebenen Basen ergänzt. Konnte im Falle des WTAK-Protokolls für einen *read* mit diesen Parametern kein *mapping* im Genom erfolgen, wurde dieser Schritt mit einem verkürzten *seed* von 25 bp und maximal zwei erlaubten Fehlern wiederholt. Bei der Extension der *reads* wurde ein Punktesystem genutzt. Für jede zusätzliche Übereinstimmung der Basenabfolge wurde ein Punkt vergeben, für Insertionen, Deletionen oder Substitutionen zwei Punkte abgezogen. Wurden für den gleichen genomischen Bereich mehrere ähnliche *mappings* gefunden, so wurde das *mapping* mit der höchsten Punktzahl für die Analysen weiter verwendet. Erreichten mehrere *mappings* eine identische Punktzahl, wurde die kürzeste Variante für die weiteren Analysen verwendet. Zu Beginn der Arbeit wurde vor der Veröffentlichung von *BioScope* mit *Corona light* ein weiteres Programm für die Zuordnung der *reads* zu den jeweiligen Positionen im Genom genutzt. Dieses Programm gibt direkt die Anzahl der *mismatches* für einen *read* gegenüber seiner zugeordneten Position aus. Hier wurde ein Maximum von 3 *mismatches* erlaubt. Dieser Algorithmus wurde im Verlauf der Arbeit verlassen und durch das komplexere *BioScope* ersetzt. Detaillierte Informationen zur Nutzung von *BioScope* und *Corona light* können vom Herausgeber zur Verfügung gestellt werden (www.appliedbiosystems.com).

2.2.2. Bestimmung der Abdeckung in annotierten/nicht-annotierten Bereichen

Für die Unterscheidung, ob ein *read* ganz oder teilweise in zuvor nicht-annotierte Bereiche fiel, d.h. in Bereichen liegt, für die zuvor keine transkriptionelle Aktivität beschrieben wurde, wurden zwei aktuelle Genannotationslisten (Stand August 2010) als Referenzdatenbank genutzt: (I) *Reference Sequence* (*RefSeq*) *Gene* des *National Center for Biotechnology Information* (NCBI) und (II) *Ensembl Gene*, gemeinsam betrieben vom *European Bioinformatics Institute* (EBI, Teil des *European Molecular Biology Laboratory*, EMBL) und dem *Wellcome Trust Sanger Institute*. Basierend auf diesen Genannotationen wurde betrachtet, welche Basen eine Abdeckung (*coverage*) durch RNA-Seq-*reads* in annotierten Bereichen besaßen und welche Basen nicht in annotierten Bereichen lagen, aber dennoch eine *coverage* durch RNA-Seq-*reads* zeigten. Dabei wurde es als ausreichend für die Zuordnung zu annotierten Bereich angesehen, wenn mindestens eine der beiden Gendatenbänke diese entsprechend auswies. Der prozentuale Anteil der nicht-annotierten bzw. annotierten RNA-Seq-Sequenzdaten wurde berechnet, indem die jeweilige *coverage* als Anteil der Gesamt-*coverage* berechnet wurde. Dazu wurde für jede Base des Genoms die Menge der *reads* bestimmt, die dieser Position eindeutig im *mapping* zugeordnet werden konnte. Diese wurden genomweit aufaddiert, jeweils für nicht-annotierte Bereiche, für annotierte Bereiche und für das gesamte Genom.

2.2.3. Erzeugung zufällig generierter *reads*

Um zu testen, inwiefern zufällig *reads* dem Genom zugeordnet wurden, wurden *reads* künstlich erzeugt. Dazu wurde mittels eines Computeralgorithmus jeder Position der 35 bp-*reads* mit einer Wahrscheinlichkeit von P = 0,25 eine der vier Basen Desoxyadenosin, Desoxythymidin, Desoxycytidin oder Desoxyguanosin zugeordnet.

2.2.4. Ermittlung der Genexpression mit *Cufflinks*

Für die Berechnung der Expressionsstärke der einzelnen Gene wurde das frei zugängliche Software-Paket *Cufflinks* genutzt (Trapnell u. a. 2010). Dieses Programm berechnete den relativen Anteil einzelner Transkripte an der Gesamtheit des untersuchten Transkriptoms anhand der Verteilung der *reads* auf die einzelnen Transkripte. Für genomische Regionen, in denen die *reads* z.B. durch überlappende Transkripte nicht zweifelsfrei einem Transkript zugeordnet werden konnten, erfolgte eine Abschätzung der Häufigkeit der beteiligten Transkripte anhand eindeutiger Bereiche und fragmentierter *reads*, d.h. durch Spleißen entstandene, über mehrere Exone verlaufene *reads*. Die Ausgabe der Expressionsstärke erfolgte dabei basierend auf den zugeordneten *reads* eines Transkripts in *fragments per kilobase of transcript per million fragments mapped* (FPKM). Detaillierte

Informationen zu diesem Programm und dessen Handhabung finden sich unter http://cufflinks.cbcb.umd.edu/.

2.2.5. Berechnung der Sättigungskinetik der detektierten Transkripte

Ausgehend von der *RefSeq*-Datenbank wurden sämtliche Einträge, die identische *gene symbols* besaßen, künstlich fusioniert, so dass *reads* an einer beliebigen Position der beteiligten *RefSeq*-Einträge zum Nachweis dieses „Supertranskripts“ führten. Nun wurden aus dem Datensatz des Jejunums zufällig je 10.000 der eindeutig zugeordnete *reads* entnommen und bestimmt, wie hoch die Anzahl der damit nachweisbaren *gene symbols* war. Dabei wurde gefordert, dass der Nachweis der Expression eines *gene symbols* durch fünf *reads* zu erfolgen hat (häufig publizierte Referenzgröße, siehe z.B. Tang u. a. 2009). Nachdem die Anzahl der *gene symbols* bestimmt und neben der Anzahl der *reads* dokumentiert wurde, wurden weitere 10.000 *reads* zufällig aus dem Datensatz entnommen und die Prozedur mit der erweiterten Datenmenge erneut durchgeführt. Dieser Schritt wurde solange wiederholt, bis sämtliche *reads* des Datensatzes in die Untersuchung eingeflossen waren. Anschließend wurden die so ermittelten Daten graphisch dargestellt, indem die Anzahl der nachgewiesenen *gene symbol*-Einträge in Abhängigkeit der Anzahl der *reads* aufgetragen wurden.
Die nicht-lineare Regressionsanalyse wurde basierend auf einer Hyperbel-Funktion vergleichbar der Enzymsättigungskinetik der Michaelis-Menten-Theorie (siehe **F. 1**) durchgeführt, indem unter Nutzung des Statistik-Programms R nach der Methode der kleinsten Quadrate die bestmögliche Annäherung an die experimentell bestimmten Daten ermittelt wurde.

F. 1: Basisfunktion der nicht-linearen Regression

$$f(n_{reads}) = \frac{a \cdot n_{reads}}{b + n_{reads}}$$

n_{reads} = Anzahl der *reads*

Ausgehend von dieser ersten Regressionsanalyse wurden weitere Regressionsanalysen basierend auf dem gleichen Modell durchgeführt, die sich auf die experimentellen Daten für hohe *read*-Mengen beschränkten. Für die folgende Regressionsanalyse wurden so nur die Datenpunkte verwendet, die von der vorherigen Regressionsanalyse dauerhaft unterlaufen wurden. Dies wurde sooft wiederholt, bis das Bestimmtheitsmaß der Regressionsanalyse über 0,99 lag. Diese Regressionsanalyse wurde genutzt, um die Gesamtmenge an auffindbaren Transkripten im untersuchten Gewebe über Extrapolation zu schätzen.

2.2.6. Ermittlung differentiell regulierter Transkripte

Als zwischen den beiden untersuchten Geweben differentiell exprimierte Transkripte wurden solche angesehen, die in einem Gewebe eine Abweichung des Expressionsverhältnisses um das Dreifache des FPKM-Wertes gegenüber dem anderen Gewebe erreichten. Da in beiden Geweben sehr schwach

exprimierte Transkripte nur durch wenige *reads* repräsentiert und starke Abweichungen allein durch die zufällige Verteilung dieser möglich waren, wurde der FPKM-Wert des jeweiligen Transkripts im Gewebe mit der schwächeren Expression um 1 erhöht, so dass für schwach exprimierte Transkripte zufällige Schwankungen minimiert wurden, stärker exprimierte Transkripte aber kaum betroffen sind (siehe **F. 2**).

F. 2: Berechnung der differentiellen Expression

$$EV_{Transkript_n} = \frac{FPKM_H}{(FPKM_L + 1)}$$

EV = Expressionsverhältnis, Transkript_n = Transkript n ϵ *RefSeq*, FPKM = Expressionsstärke in *fragments per kilobase of transcript per million fragments mapped*, H = Gewebe mit höherer Expression für Transkript n, L = Gewebe mit niedrigerer Expression für Transkript n,

2.2.7. Expressionscharakterisierung mittels *gene ontology*-Analyse

Beim *gene ontology*-Projekt handelt es sich um die 1998 ins Leben gerufene Bemühung eines Konsortiums von Wissenschaftlern, Gene und deren Genprodukte einheitlich und strukturiert auch über Speziesgrenzen hinweg zu annotieren, indem jedes Gen in den drei großen Domänen (a) molekulare Funktion, (b) biologischer Prozess und (c) zelluläre Komponente durch ein festes Vokabular definiert wurde (z.B. Immunabwehr für ein Defensin als biologischer Prozess) (Gene Ontology Consortium 2008).

Im Detail wurde in dieser Arbeit zunächst für jede *gene ontology*-Kategorie der Subkategorie „Biologische Prozesse“ (http://www.geneontology.org) der Erwartungswert in der Menge der gewebsspezifisch-exprimierten Transkripte berechnet. Dieser besagt, wie hoch die zu erwartende Anzahl der Elemente dieser *gene ontology*-Kategorie im Falle einer zufälligen Verteilung ist (Agresti 1992). Lag in diesem zweiseitigen Test der Erwartungswert unter der Anzahl der tatsächlich im Experiment beobachteten Elemente einer *gene ontology*-Kategorie, so wurde die Wahrscheinlichkeit für eine zufällige Überrepräsentierung dieser Kategorie geprüft, andernfalls die Wahrscheinlichkeit für eine zufällige Unterrepräsentierung.

Für diese Untersuchung wurde ein statistisches Modell basierend auf der hypergeometrischen Verteilung genutzt, welche die Wahrscheinlichkeit ermittelt, mit der eine Menge von Elementen zufällig aus einer dichotomen Grundgesamtheit entnommen wurde (Tavazoie u. a. 1999). Die ermittelte Wahrscheinlichkeit wurde für multiples Testen nach Benjamini-Hochberg korrigiert (Benjamini und Hochberg 1995). Kategorien, die nach Korrektur für zweiseitiges und multiples Testen mit einer Wahrscheinlichkeit von unter 5% (Signifikanzniveau α von 0,05) zufällig eine Über- bzw. Unterrepräsentierung von gewebespezifisch exprimierten Genen zeigten, wurden als statistisch signifikant betrachtet.

2.2.8. Berechnung der Reliabilität

Für den Vergleich der technischen und biologischen Replikate sowie des Vergleichs von mehrfacher Anreicherung für polyadenylierte Transkripte und des Plattformvergleichs zwischen RNA-Seq und *microarrays* wurde für jedes einzelne *RefSeq*-Transkript die Expressionsstärke in FPKM errechnet (bzw. im Fall des *microarrays* die Stärke des RMA-normalisierten Fluoreszenzniveaus). Die Transkripte, deren Expression in beiden der gegenübergestellten Datensätze nachgewiesen wurde, wurden graphisch doppelt logarithmisch aufgetragen. Zur besseren Trennung von Bereichen mit wenigen und vielen Datenpunkten wurden die Datenpunkte in Abhängigkeit der Dichte der Datenpunkte in der jeweiligen Region farblich markiert.

Zusätzlich wurde der Rangkorrelationskoeffizient nach Spearman für die jeweiligen Paare berechnet, dazu wurde die folgende Formel **F. 3** verwandt (Rudolf und Kuhlisch 2008):

F. 3: Berechnung des Spearman Rangkorrelationskoeffizienten

$$r_{xy} = \left(\sum_{i=1}^{n}\left(r_{x_i} - \bar{r}_x\right)\cdot\left(r_{y_i} - \bar{r}_y\right)\right)\cdot\left((n-1)\cdot s_{r_x}\cdot s_{r_y}\right)^{-1}$$

r_{xy} = Rangkorrelationskoeffizient nach Spearman, n = Anzahl der Messwertpaare, r_{x_i} bzw. r_{y_i} = Rangplätze der Messwerte, $\bar{r}_x$ bzw. $\bar{r}_y$ = arithmetischer Mittelwert der Rangplätze, s_{r_x} bzw. s_{r_y} Standardabweichung der Rangplätze.

2.2.9. Darstellung der Transkriptabdeckung

Für die exemplarische Darstellung der Abdeckung einzelner Transkripte wurde das Statistikprogramm R genutzt. Für jeweils die erste Base eines zugeordneten *reads* wurde die entsprechende Base des Transkripts markiert. Dies wurde für alle dem Transkript zugeordneten *reads* wiederholt. Anschließend wurde die Exonstruktur des Transkripts mit den markierten Basen graphisch dargestellt, so dass sich ein Bild der Verteilung der *reads* über das Transkript ergab. Existierten mehrere *reads* mit der gleichen Startbase, wurde die Anzahl der *reads* zur besseren Darstellbarkeit über der Farbe der Markierung kodiert und anhand einer Legende kenntlich gemacht.

Für die genomweite Darstellung der Abdeckung in verschiedenen Klassen von Transkripten basierend auf ihrer Länge wurden zunächst alle Transkripte einer Klasse auf eine fiktive Länge normalisiert. Dazu wurde für jedes Element eines in 1000 Teile gegliederten Transkripts die mittlere Abdeckung berechnet. Um sicher zu stellen, dass einzelne Transkripte mit starker Expression das Gesamtergebnis verzerren, wurde zusätzlich auch die Expressionsstärke der einzelnen Transkripte normalisiert, so dass alle Transkripte mit gleichem Gewicht in die Analyse eingingen. Dazu wurde die Abdeckung des Fragments eines Transkripts mit der höchsten Abdeckung gleich 100 gesetzt und alle weiteren Elemente dieses Transkripts mit dem aus dieser Normalisierung errechneten Anpassungskoeffizienten korrigiert. Die so bezüglich ihrer Länge und Expressionsstärke normalisierten Transkripte wurden anschließend in Abhängigkeit der ursprünglichen Länge in verschiedene Klassen sortiert und für die einzelnen Klassen die mittlere Abdeckung für jedes Element

berechnet, indem die normalisierten Abdeckungen aufsummiert und durch die Anzahl der Elemente geteilt wurden.

Abschließend wurde zur besseren Interpretation vergleichbar der Anpassung der Expressionsstärke einzelner Transkripte das Element mit der höchsten Abdeckung für eine Transkriptklasse gleich 100 gesetzt und die anderen Elemente entsprechend korrigiert. Die Darstellung der Ergebnisse erfolgte in Form eines Diagramms.

2.2.10. Erstellung putativer Spleißbindungen

Putative Spleißbindungen wurden erzeugt, indem sämtliche Exone eines annotierten *RefSeq*-Transkripts entsprechend ihrer Position im Transkript geordnet wurden (E_1, E_2,…, E_n) und jeweils die letzten 25 Basen eines Exons E_i mit den 25 ersten Basen der folgenden Exone E_f (mit $i < f$) kombiniert wurden. Aus dieser Prozedur hervorgegangene Duplikate mit vollständiger Sequenzübereinstimmung wurden entfernt. Die verbliebenen derart generierten Spleißbindungen wurden getrennt durch 50 bp lange Abschnitte von nicht-definierten Nukleotidabfolgen (N_{50}) hintereinander gefügt und so als artifizielles Chromosom für das *mapping* von *reads* genutzt, die zuvor keiner Position im Genom zugeordnet werden konnte. Spleiß-Ereignisse, die zum Analysezeitpunkt nicht als Exone annotierter Bereiche des Genoms bekannt waren, wurden in dieser Untersuchung nicht berücksichtigt.

2.2.11. Algorithmus zur Detektion von nTAR

Zur Detektion zuvor nicht-annotierter, transkriptionell aktiver Regionen (nTAR) wurden zunächst bekannte, annotierte Regionen transkriptioneller Aktivität für die weitere Analyse ausgeschlossen. Im nächsten Schritt wurde dieser um bekannte Gene reduzierte Sequenzdatensatz mit den erhobenen cDNA-Sequenzierdaten abgeglichen und abgedeckte Regionen als potentiell transkriptionell-aktive Regionen angenommen. Für diese Analyse wurden sämtliche cDNA-Daten aus den beiden untersuchten Geweben vereint. Zur Erhöhung der Spezifität wurden zwei Qualitätskriterien festgelegt: (a) Die nicht annotierten, transkriptionell-aktiven Regionen mussten über eine Länge von mindestens 50 Basenpaaren eine cDNA-Abdeckung aufweisen, dadurch mussten mindestens zwei *reads* nicht-klonalen Ursprungs und einem nicht-überlappenden Anteil von mindesten 15 Basen die Beobachtung stützen. (b) Die Abdeckung über den Bereich von 50 Basenpaaren musste im Durchschnitt mindestens einer dreifachen Basenabdeckung entsprechen. Dieser Schwellenwert wurde in Anlehnung an die aktuelle Literatur gewählt, wo fünf *reads* als sicheres Zeichen der Genexpression angenommen werden (vgl. Tang u. a. 2009). Dies entsprach in dieser auf der *coverage* basierender Analyse einem Schwellenwert von bis zu drei.

Zur unabhängigen Verifizierung der Daten wurden die gefundenen Regionen des Genoms mittels WTAK-Sequenzierung bestätigt. Nur genomische Bereiche, die durch mindestens drei WTAK-*reads* bestätigt werden konnten, wurden für die folgenden Analysen berücksichtigt.

2.2.12. Klassifizierung der nTAR

Die beobachteten und den Qualitätskriterien entsprechenden nTAR wurden im nächsten Schritt entsprechend ihrer Position zu annotierten Genen bestimmten Klassen zugeordnet. Fand sich kein annotiertes Gen im Umkreis von 10.000 Basenpaaren um die Position des nTAR, wurde es als nicht-genassoziiertes nTAR (NGA) bezeichnet. In einem Abstand von weniger als 10.000 Basen, die sich aber nicht mit dem annotierten Transkriptionsstart oder dem Transkriptionsende überschnitten, wurden nTAR als in Gennachbarschaft klassifiziert. Zusätzlich wurde unterschieden, ob sie stromaufwärts oder stromabwärts des Gens lagen: vorgelagerte Gennachbarschaft (*upstream gene neighborhood*, UGN) bzw. nachgelagerte Gennachbarschaft (downstream gene neighborhood, DGN). Lag die Region direkt angrenzend an ein annotiertes Gen, so dass dieser Bereich in eine bereits als transkriptionell-aktiv annotierte Region überging, so wurden diese nTAR als Genüberschneidung betrachtet. Für diese Gruppen wurde zwischen vor- und nachgelagerten Elementen unterschieden (*upstream gene intersection*, UGI bzw. *downstream gene intersection*, DGI).

Auch im Bereich annotierter Gene können zuvor nicht-annotierte nTAR in den Intronen vorkommen. Hier wurden folgende Klassen für die nTAR gebildet: Intron-überspannende Elemente (*intron spanning element*, ISE) sind Introne, die lückenlos mit den RNA-Seq-Daten abgedeckt wurden. Exon-nachgelagerte Überschneidungen (*exon-linked (downstream)*, ELD) bzw. Exon-vorgelagerte Überschneidungen (*exon-linked (upstream)*, ELU) zeigen nahtlose Übergänge von bzw. hin zu annotierten Exonen. Alle anderen nTAR, die innerhalb eines Genes liegen, aber keine direkten Überschneidungen zu bekannten Transkripten zeigten, wurden als intragenische Elemente (, IGE) bezeichnet.

Für nTAR, die z.B. durch Nähe zu mehreren Genen oder Isoformen eines Transkripts die Anforderungen mehrerer Klassifizierungen erfüllten, wurde immer nur zu einem Gen eine Zuordnung vorgenommen. Dabei wurde die Zuordnung zu den einzelnen Klassifizierungen in folgender Weise priorisiert: Intron überspannende Elemente hatten die höchste Priorität, gefolgt von den Exon-Überschneidungen. Die nächsthöhere Priorität besaßen die Genüberschneidungen, so dass nTAR mit direkter Überschneidung auch diesem Gen zugeordnet wurden. Lag allerdings eine Überschneidung mit mehreren Genen oder Isoformen vor, erfolgte die Zuordnung zufällig.

Innerhalb der Gruppe der nTAR in Nachbarschaft, aber ohne direkte Anlagerung an annotierte Gene, wurden intragenetische Elemente mit der höchsten Priorität versehen, gefolgt von den Elementen, die auf dem Chromosom vor (UGN) bzw. hinter (DGN) dem Gen angeordnet sind. Bei diesen

Elementen wurde als nachrangiges Kriterium noch die Entfernung zum nächsten annotierten Bereich in die Zuordnung einbezogen, so wurde bei mehrdeutiger Zuordnung immer das näher gelegene Gen als assoziiertes Gen betrachtet. Für nTAR im Bereich annotierter Gene, die trotz der benannten Kriterien nicht eindeutig einem Gen bzw. einer Klasse zugeteilt werden konnten, erfolgte eine zufällige Zuordnung. Nicht-genassoziierte nTAR waren per Definition immer ohne assoziiertes Gen.

2.2.13. Feststellung der Orientierung der nTAR

Zur Feststellung der Orientierung der nTAR wurde die Orientierung der WTAK-*reads* ausgewertet. Es wurden nur die Fälle von nTAR berücksichtig, wo alle WTAK-*reads* die gleiche Orientierung aufwiesen. Zeigten alle WTAK-*reads* die gleiche Orientierung wie das assoziierte Gen, so wurde die Orientierung des nTAR als in *sense* bezeichnet, bei antiparalleler Orientierung entsprechend als in *antisense*. Für NGA-nTAR konnte mangels assoziierten Gens keine Orientierung festgelegt werden.

2.2.14. Berechnung der Expressionsstärke von nTAR

Für die Berechnung der Genexpression der nTAR wurde aufgrund der teilweise sehr kurzen Fragmente ein Algorithmus genutzt, der auf der gemittelten Basenabdeckung über das nTAR beruht. Die mittlere Basenabdeckung (*average base coverage*, abc) wurde berechnet, indem die Abdeckung jeder Base des nTAR bestimmt, aufsummiert und anschließend durch die Anzahl der nTAR-Basen geteilt wurde. Dieser Wert wurde zur Vergleichbarkeit von verschiedenen Sequenzdatensätzen auf die Menge der erhobenen Sequenzdaten eines Experiments angeglichen, indem die einzelnen Datensätze auf 10^9 sequenzierte Basen normalisiert wurden. In dieser Arbeit wurden für die einzelnen Experimente Sequenzdaten im Umfang von ca. 10^9 Basen erhoben, so dass sich die Größenordnung des Ergebnisses nach Normalisierung nicht änderte (abc_{Gb}, *average base coverage per gigabase sequence space*). Für die assoziierten Gene wurde die Expression entsprechend durchgeführt, um hier einen direkten Vergleich von nTAR zu assoziiertem Gen zu erlauben.

Für die Berechnung des Expressionsverhältnis von nTAR und assoziiertem Gen wurde die mittlere Basenabdeckung in beiden Geweben addiert und durch die Summe der mittleren Basenabdeckung des assoziierten Gens geteilt (siehe **F. 4**).

F. 4: nTAR-Expression gegenüber assoziiertem Gen

$$Expressionsverhältnis = \frac{[abc_{GB}\ (nTAR_{Gewebe\ a})] + [abc_{GB}\ (nTAR_{Gewebe\ b})]}{[abc_{GB}\ (Transkript_{Gewebe\ a})] + [abc_{GB}\ (Transkript_{Gewebe\ b})]}$$

2.2.15. Feststellung polyadenylierter nTAR

Die mittels Pyrosequenzierung von polyadenylierten Fragmenten (siehe 2.1.10) erhobenen Daten wurden mit der Liste der beobachteten nTAR abglichen. Fand sich eine Übereinstimmung zwischen

nTAR-Lokalisation und der Zuordnung des Pyrosequenzierungs-*reads*, so wurde dies als eine Polyadenylierung des nTAR interpretiert.

2.2.16. Festlegung gewebsspezifisch differentiell regulierter nTAR

Um die gewebsspezifische Regulierung von NGA-nTAR zu analysieren, wurde das Verhältnis von normalisierter nTAR-Expression zwischen den Geweben berechnet, wobei der Divisor um 1 erhöht wurde, um zufällige Schwankungen bei sehr schwach exprimierten Genen abzufangen. Diese Operation wurde aufgrund des Korrekturfaktors unabhängig für beide Gewebe durchgeführt (jeweils als Dividend und Divisor). Ein Wert von > 3 wurde in Anlehnung an (konservative) *microarray*-Studien als differentielle Regulierung interpretiert.

Für genassoziierte nTAR wurde diese Berechnung modifiziert. Unter der Annahme, dass nTAR und assoziiertes Gen zusammen reguliert werden und/oder in vielen Fällen sogar nTAR strukturell in Form von z.B. Exonen oder alternativen Polyadenylierungssignalen dem Transkript zugehören, wurde hier eine Analyse gewählt, die die Expression des assoziierten Gens beinhaltet.

Dazu wurde der Quotient aus dem Verhältnis der Genexpression von nTAR zu assoziiertem Gen beider Gewebe berechnet. Wie zuvor wurden zufällige Schwankungen schwach exprimierter nTAR bzw. ihrer assoziierten Gene durch die Erhöhung der Divisoren um 1 reduziert. Die Analyse wurde für beide Gewebe als Dividend bzw. Divisor durchgeführt. Ein Quotient von > 3 wurde als differentielle Regulierung interpretiert. Die für die Auswertung verwendete Formel findet sich in **F. 5**.

F. 5: Differentiell regulierte nTAR

$$Expressionsverhältnis = \frac{[abc_{GB}\ (nTAR_{Gewebe\ a})][abc_{GB}\ (Transkript_{Gewebe\ b})]}{[abc_{GB}\ (nTAR_{Gewebe\ b}) + 1][abc_{GB}\ (Transkript_{Gewebe\ a}) + 1]}$$

3. Ergebnisse

3.1. Darstellung von cDNA für RNA-Seq

Für die ABI SOLiD-Sequenziertechnologie waren zu Beginn dieser Arbeit keine Protokolle zur Sequenzierung von RNA verfügbar. Um dennoch das Transkriptom intestinalen Gewebes analysieren zu können, wurde zunächst eine Methode etabliert, um RNA möglichst vollständig und die quantitative Zusammensetzung der einzelnen Transkriptvarianten repräsentierend in DNA umzuschreiben. Dabei wurde ein Verfahren angepasst, dass bereits erfolgreich für den Roche *Genome Analyzer* eingesetzt werden konnte (Hemmrich u. a. 2012) und im Wesentlichen auf der *template-switch*-Technologie des Clontech SMART cDNA *synthesis*-Kit beruhte. Abweichend von den Vorgaben des Herstellers sind biotinylierte Primer während der Amplifikation genutzt wurden, um nachträglich etwaige Primerkontaminationen sowie die durch die Art der cDNA-Synthese erzeugten, artifiziellen Transkriptenden nach Fraktionierung der cDNA aus der SOLiD cDNA *library* zu entfernen. Dies wurde bereits dadurch erreicht, dass biotinylierte Primer nicht mit SOLiD-Sequenz-Adaptoren

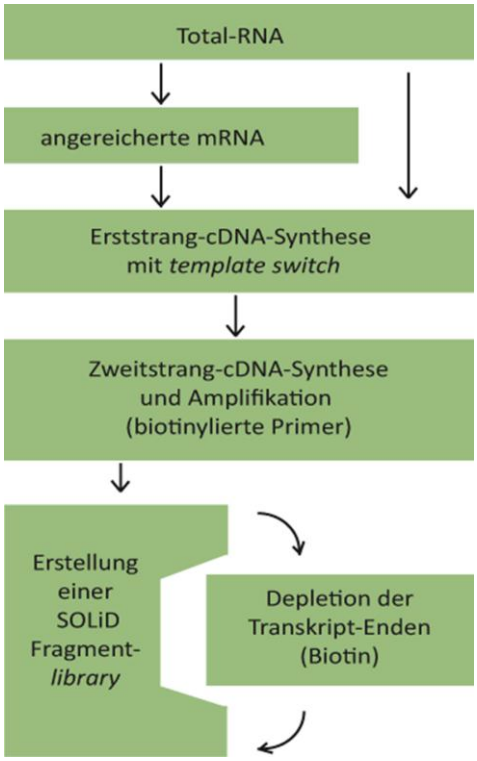


Abb. 6: Flussdiagramm cDNA-Synthese für SOLiD RNA-Seq

Für die Erststrangsynthese kann entweder bereits für polyadenylierte Transkripte selektionierte mRNA oder direkt isolierte Total-RNA eingesetzt werden. Nach der Erststrangsynthese erfolgte eine Amplifikation der cDNA-Fragmente über 13 (für mRNA) bzw. 15 (für Total-RNA) Zyklen mit biotinylierten Primern. In der *library preparation* des SOLiD *fragment protocols* wurden über die Biotin-Markierung die Transkriptenden mit den definierten Sequenzen wieder entfernt.

ligiert werden können. Um unvorhergesehene Einflüsse auf nachfolgende Reaktionsschritte zu vermeiden, wurden hier die biotinylierten DNA-Fragmente dennoch mittels immobilisierten Streptavidin-*beads* während der *library*-Erstellung aus der Lösung abgereichert (eine Übersicht der cDNA-Synthese als Flussdiagramm ist in **Abb. 6** dargestellt). Dies führte bei 500ng mRNA zu cDNA-Ausbeuten von 3-4 µg. Eine gelelektrophoretische Auftrennung der gewonnenen cDNA zeigte typischerweise ein diverses Gemisch von Fragmenten unterschiedlicher Größe. Der Großteil der erzeugten Fragmente fand sich im Falle der aus intestinalem Mausgewebe gewonnenen cDNA im Bereich von 200 bis 6.000 bp (siehe **Abb. 7**). Diese gewonnene Menge an cDNA war ausreichend, um im SOLiD *library preparation*-Protokoll eingesetzt zu werden. Während initial mit 500ng mRNA noch große Mengen von RNA erforderlich waren, konnte die erforderliche Menge an RNA im Laufe der Arbeit zunehmend reduziert werden. So waren in späteren Kooperationsprojekten selbst RNA-Isolate basierend auf wenigen Zellen für die Analyse ausreichend (Autran u. a. 2011).

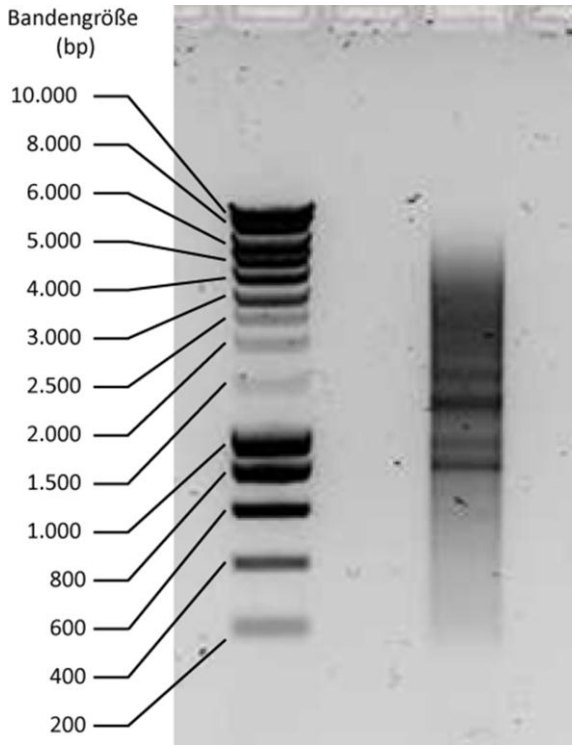


Abb. 7: Größenverteilung erstellter cDNA-Bibliotheken

Exemplarische dargestellt ist eine SMART cDNA-Bibliothek hergestellt aus 500 ng aufgereinigter mRNA des murinen Jejunums. Diese wurde auf einem 1,5 % Agarosegel aufgetrennt. Der Großteil der amplifizierten cDNA besitzt eine Größe von ca. 200 bis 6.000 bp.

3.1.1. cDNA-basiertes RNA-Seq auf dem SOLiD V2

Je eine Probe aus dem Jejunum und dem Colon wurden auf dem SOLiD V2 (35 bp *read*-Länge) auf jeweils einer halben *flow cell* (Reaktionsraum des Gerätes) sequenziert. Im Falle des Jejunums

konnten 77.504.263 Sequenzen vom System detektiert werden. Für die Colon-Probe wurde dieser Wert mit 83.202.412 Sequenzen übertroffen. Die gewonnenen Rohdaten sind auf der NCBI-Seite *Gene Expression Omnibus* (GEO, http://www.ncbi.nlm.nih.gov/geo/) deponiert und können unter der Zugangsnummer GSE21746 heruntergeladen werden.

Als erster Schritt der Datenanalyse wurde für die gewonnenen Sequenzdaten die bestmögliche Übereinstimmung im Maus-Genom basierend auf der Version NCBI 37/mm9 des Maus-Erbgutes (Rhead et al. 2010)) ermittelt. Dieser auch als *mapping* bezeichnete Schritt wurde mit der Software *BioScope* des SOLiD-Herstellers durchgeführt. Unter stringent gewählten Qualitätskriterien (siehe 2.2.1) konnten mit 40.774.275 (Jejunum) bzw. 40.810.609 (Colon) *reads* für ca. die Hälfte der Sequenzdaten eine Position im murinen Genom zugeordnet werden (52,61% bzw. 48,99%), insgesamt wurden 1,335 (Jejunum) bzw. 1,343 (Colon) Milliarden Basen sequenziert und ihrem genomischen Ursprung zugeordnet. Für einen Teil der *reads* wurden mehrere Positionen im Genom gefunden, d.h. es fanden sich mehrere Positionen im Genom, die für den *read* bezüglich der Qualität mit gleicher Wahrscheinlichkeit als Ursprung dienten. In diesem Fall wurde der *read* komplett aus der Analyse entfernt. Nach Ausschluss dieser nicht-eindeutig zuordbaren Sequenzen verblieben 28.178.017 (Jejunum) bzw. 28.439.386 (Colon) *reads* für weitere Analysen (siehe **Tab. 8**).

Tab. 8: Überblick initiale Sequenzdaten.

Die Tabelle zeigt die wichtigsten Kennziffern für die SMART cDNA SOLiD-Sequenzierung.

cDNA-Fragmentierung		**Jejunum**		**Colon**	
Gesamtzahl reads		77.504.263		83.302.412	
Gesamtzahl zugeordnete reads	Anzahl (reads)	40.774.275	52,61%	40.810.609	48,99%
	Einzelbasen(10^9)	1,335		1,343	
Eindeutig zuzuordnen	Anzahl (reads)	28.178.017	36,36%	28.439.386	34,14%
	Einzelbasen(10^9)	0,922		0,935	

Um abzuschätzen, inwieweit *reads* mit dieser Methode fälschlicherweise einem Transkript zugeordnet wurden, wurden in einem Vorversuch zufällig generierte *reads* in vergleichbarer Anzahl ($n > 77 \cdot 10^6$) erzeugt und mittels einer früheren Version des *mapping*-Algorithmus (Corona light) mit den annotierten Transkripten der Maus abgeglichen. Beruhend auf dieser Vorgehensweise konnte nur für insgesamt 33 der artifiziellen *reads* eine Zuordnung zu einem Transkript beobachtet werden. In allen Fällen wurden die Qualitätskriterien für eine erfolgreiche Zuordnung nur knapp erreicht. (siehe **Tab. 9**).

Tab. 9: Genomweite Zuordnung künstlich erzeugter *reads*

Die Tabelle zeigt die Anzahl zufällig erzeugter *reads*, die einem annotierten Transkript erfolgreich zugeordnet werden konnten.

Gesamtzahl artifiziell erzeugter *reads*	77.504.263	
keine *mismatches*	-	-
Anzahl *mismatches* = 1	-	-
Anzahl *mismatches* = 2	-	-
Anzahl *mismatches* = 3	33	$4 \cdot 10^{-5}$ %

In einem weiteren Vorversuch, indem ribosomal depletierte, fragmentierte RNA für die cDNA-Synthese eingesetzt wurde, wurde untersucht, inwiefern die stringente Aufreinigung von polyadenylierter mRNA im untersuchten Versuchsaufbau entscheidend ist. Hier fand sich sowohl für das Jejunum als auch für das Colon ein deutlich reduzierter Anteil der zuordbaren *reads.* Während für mehrfach aufgereinigte polyadenylierte mRNA deutlich über 40% der *reads* einer oder mehrerer Positionen im Genom zugeordnet werden konnten, lag der Wert hier bei unter 10 % (siehe **Tab. 10**).

Tab. 10: Überblick cDNA-Synthese mit ribosomaler Depletierung

Die Tabelle zeigt die *mapping*-Effizienz für die Sequenzdaten ribosomal depletierter RNA.

	Jejunum		**Colon**	
Gesamtzahl *reads*	36.419.942		37.466.674	
Anzahl *mismatches* = 0	783.567	2,15 %	1.504.810	4,02 %
Anzahl *mismatches* = 1	555.795	1,53 %	813.348	2,34 %
Anzahl *mismatches* = 2	393.917	1,08 %	626.624	1,67 %
Anzahl *mismatches* = 3	377.499	1,04 %	564.556	1,51 %
Gesamt	**2.110.778**	**5,8 %**	**3.572.185**	**9,53** %

Nach der Zuordnung der *reads* wurde betrachtet, inwiefern bereits annotierte Gene durch die gewonnenen RNA-Seq-Daten abgebildet werden. Zunächst wurde ermittelt, wie groß der Anteil der Sequenzdaten ist, der mit existierenden Genmodellen übereinstimmt. Für diesen Abgleich wurden aktuelle Genannotationen der Annotationsdatenbanken *RefSeq* (NCBI) und *Ensembl* (EMBL-EBI) genutzt (Stand 25. April 2010). Über das gesamte Genom betrachtet fand sich der Großteil der sequenzierten Basen erwartungsgemäß in Regionen, die bereits als Exon eines Genes annotiert waren. Dennoch fand sich mit 6,79 % (Jejunum) bzw. 8,2 % (Colon) ein signifikanter Anteil an Sequenzinformationen, der nicht mit bekannten Genmodellen übereinstimmte. Dies beruhte z.B. auf der Erweiterung der Exonstruktur beschriebener Gene, aber auch auf Regionen, für die in der bisherigen Genannotation auch im weiteren Umfeld keine annotierten Gene bekannt waren (siehe **Tab. 11**).

Tab. 11: Zuordnung der *reads* im Genom

Die Tabelle zeigt den Anteil der sequenzierten Basen, die auf bekannten Exonen bzw. im Bereich von Intronen oder zwischen annotierten Genen liegen.

	Jejunum		Colon	
cDNA-Fragmentierung	Sequenzierte Basen	**Prozent**	Sequenzierte Basen	**Prozent**
Exon (*RefSeq* + *Ensembl*)	$0{,}859 \cdot 10^9$	**93,21%**	$0{,}859 \cdot 10^9$	**91,80%**
Intron oder intergenisch	$0{,}063 \cdot 10^9$	**6,79%**	$0{,}077 \cdot 10^9$	**8,20%**

Zusammenfassend zeigten die Ergebnisse, dass basierend auf der SOLiD-Sequenziertechnologie der RNA-Sequenzierraum erfasst werden konnte und die überwiegende Anzahl der *reads* sich dabei mit der gegenwärtigen Genannotation deckte. Weiter zeigten die Ergebnisse, dass der Einfluss per Zufall zugeordneter *reads* vernachlässigbar gering ist und der relative Anteil der zuordbaren *reads* stark von der Qualität und Aufreinigung der eingesetzten RNA abhängig ist.

3.2. Genexpression annotierter Gene im Darm

Um die Expression von einzelnen Genen zu berechnen, wurde die veröffentlichte Software *Cufflinks* (Trapnell u. a. 2010) genutzt. Hierbei wird die Genexpression als *fragments per kilobase of transcript per million fragments mapped* (FPKM) ausgegeben. Für diese Betrachtung der Expression annotierter Gene wurde zur Vermeidung doppelter Einträge die Analyse auf die Gen-Annotation der *RefSeq*-Datenbank begrenzt. Prinzipiell wurde dabei berechnet, wie viele *reads* auf ein Transkript entfallen und eine um die Länge des Transkripts korrigierte, relative Häufigkeit ermittelt. Dabei wurden folgende Sonderfälle berücksichtigt: (I) In den seltenen Fällen, wo zwei verschiedene Gene überlappend im gleichen Bereich des Genoms kodiert sind, wird die Aufteilung der *reads* in diesem Bereich anhand der Verteilung der eindeutigen *reads* korrigiert. (II) Verschiedene Isoformen von Genen, die über alternatives Spleißen oder unterschiedliche Polyadenylierungssignale entstehen, wurden *in silico* zu einem Gesamt-Transkript zusammengefasst („projiziert“). In der Folge wurde die Genexpression über dieses Gesamt-Transkript berechnet. In **Abb. 8** wird dies anhand eines hypothetischen Beispiels verdeutlicht. Basierend auf den gewählten Schwellenwerten der *Cufflinks*-Veröffentlichung wurde ein Transkript als exprimiert angesehen, wenn der FPKM-Wert mindestens 0,01 betrug.

Von den 27.722 Genen in der Datenbank war für 20.541 Gene zumindest in einem der untersuchten Gewebe der Schwellenwert von 0,01 FPKM überschritten, 7.181 annotierte Gene lagen unter der gewählten Detektionsschwelle. Davon konnte für den Großteil der Gene (6.509; 90,4%) nicht ein einzelner *read* gefunden werden. Innerhalb der Gruppe der Gene, die den Schwellenwert überschritten, wurden 817 Gene nur im Jejunum detektiert, 1.735 lagen exklusiv im Colon vor. Als differentiell exprimierte Gene wurden solche angesehen, die entweder nur in einem Gewebe exprimiert wurden oder eine 3-fach höhere, normalisierte Basenabdeckung in einem Gewebe

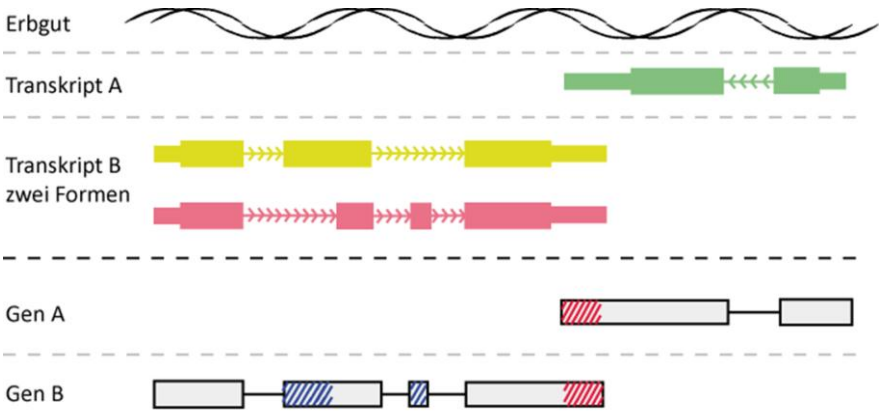


Abb. 8: Sonderfälle der Expressionsbestimmung von Transkripten

Die Abbildung zeigt den speziellen Fall der Lokalisierung zweier Gene bzw. der daraus resultierenden Transkripte im Erbgut, die sich antiparallel am 3'-Ende überkreuzen. Zusätzlich wird das zweite Gen durch alternatives Spleißen in unterschiedliche Isoformen transkribiert. Da die mittels SMART RNA-Seq gewonnenen Daten in bestimmten Bereichen nicht zweifelsfrei einem Gen bzw. einer Transkriptisoform zugeordnet werden konnten, wurde folgendermaßen vorgegangen: Für den Fall, dass sich zwei Gene überschneiden (rote Schraffur), wurden die *reads* innerhalb der roten Schraffur zwischen den beiden Genen so aufgeteilt, dass das Verhältnis dem der eindeutig zuordbaren *reads* entspricht. Im Falle des Auftretens mehrere Transkriptisoformen eines Gens wurden diese aufeinander projiziert (blaue Schraffur) und die Expression anhand aller *reads* dieser Genprojektion ermittelt.

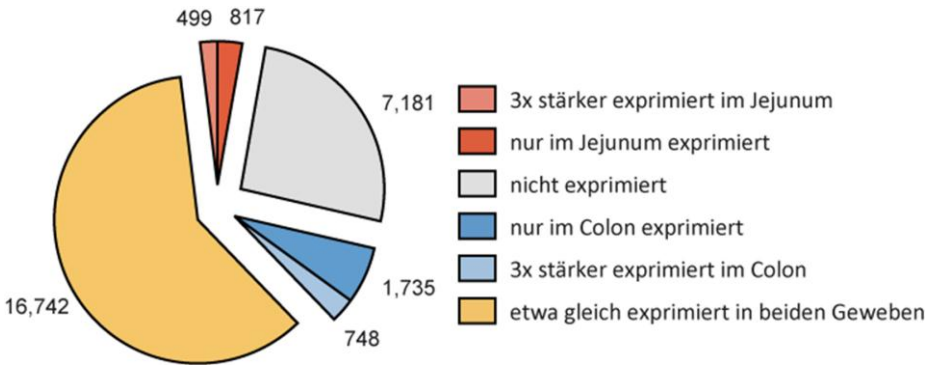


Abb. 9: Expression annotierter Gene

Übersicht der Anzahl der detektierten Gene in beiden Geweben bei einer Transkript-Detektionsschwelle von 0,01 FPKM pro Transkript. Annotationsdaten basieren auf 27.722 Einträgen der *RefSeq*-Gendatenbank (NCBI, Stand 25. April 2010). Für die Zuordnung zu den untersuchten Geweben siehe Legende.

gegenüber dem anderen Gewebe aufwiesen. Davon fanden sich im Jejunum 499 Gene und im Colon 748. Die Mehrheit mit 16.742 der annotierten Gene zeigten Expression in beiden Geweben, ohne im Verhältnis um mehr als das 3-fache abzuweichen (siehe **Abb. 9**).

Die Expressionsstärke der einzelnen *RefSeq*-Transkripte konnte dabei über mehrere Zehnerpotenzen variieren, neben sehr raren Transkripten, die nur knapp die Schwelle von 0,01 FPKM überschritten, bis hin zu einzelnen Transkripten, die einen FPKM-Wert über 1000 erreichten. Unter den sehr stark exprimierten Genen fanden sich so z.B. im Jejunum solche mit Bezug zur Immunabwehr (Defensin-6α, Lysozyme 1) oder Funktion in der Nährstoffaufnahme (*Fatty acid binding protein* 2, *Cysteine rich protein* 1). Im Colon sind hoch abundante Gene am Ionenaustausch (Carboanhydrase 1) oder der

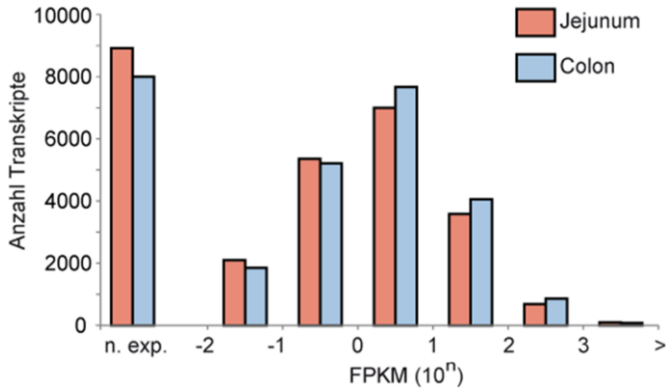


Abb. 10: Dynamische Breite der Genexpression

Die Abbildung zeigt die Verteilung der Expressionsstärken der einzelnen Transkripte in beiden Geweben. Dazu wurden jeweils eine Zehnerpotenz umfassende Klassen geschaffen und die einzelnen Transkripte anhand ihrer Expressionsstärke in FPKM diesen zugeordnet. Die Legende zeigt die Zuordnung der einzelnen Gewebe (n. exp. = nicht exprimiert (FPKM < 0,01).

Mukusbildung (*Anterior gradient* 2, *Serine protease inhibitor Kazal-type* 4) beteiligt. Die Mehrheit der Gene zeigte hingegen moderate Expressionsstärken, so sind ca. 43 % aller Transkripte mit einer durchschnittlichen Basenabdeckung von 1 bis 10 (Jejunum: 42,62 %, Colon: 43,8 % aller Transkripte) exprimiert. Für 91,5 % aller Transkripte der *RefSeq*-Datenbank wurde eine Expression zwischen 0,1- bis 100-facher Basenabdeckung (Jejunum: 91,51 %, Colon 91,61 %) je untersuchter 10^6 *reads* beobachtet (siehe **Abb. 10**). Eine komplette Liste der ermittelten Expressionsstärken für die einzelnen untersuchten Gene findet sich unter http://www.ikmb.uni-kiel.de/murine-transcriptomes.

3.2.1. Abdeckung der Transkripte entlang der 5‘-3‘-Achse

RNA-Seq ist nicht nur geeignet, um genomweit Expressionslevel von Transkripten in unterschiedlichen Geweben zu vergleichen, sondern offenbart auch Informationen über die Feinstruktur der einzelnen Transkripte (Spleißen, posttranskriptionelle Modifikationen, Polyadenylierung etc.). Hierfür ist entscheidend, dass die *coverage* über das gesamte Transkript bekannt und nach Möglichkeit gleichmäßig ist. Daher wurde die Verteilung der *reads* bei ausgewählten Transkripten betrachtet. Hier zeigte sich in der Regel eine ausgewogene Verteilung. Die Transkripte wurden weitgehend mit einer konstanten Abdeckung erfasst. In manchen Fällen zeigte sich aber auch, dass Exone kaum oder überhaupt nicht abgedeckt wurden. Insbesondere zeigte die Abdeckung der Transkripte überlappende *reads*, d. h. die Abdeckung der Transkripte basierte nicht auf wenigen, identischen *reads*, sondern auf einer Vielzahl von *reads* mit unterschiedlichen Startpunkten. Zur Veranschaulichung wurde in **Abb. 11** exemplarisch die Abdeckung durch die einzelnen *reads* für die intestinale Alkalische Phosphatase gezeigt.

Während die *coverage* eines spezifischen Transkripts an ganz unterschiedlichen Positionen variieren konnte, zeigte sich anhand ausgewählter Transkripte bereits eine Tendenz, dass das 3‘-Ende eines

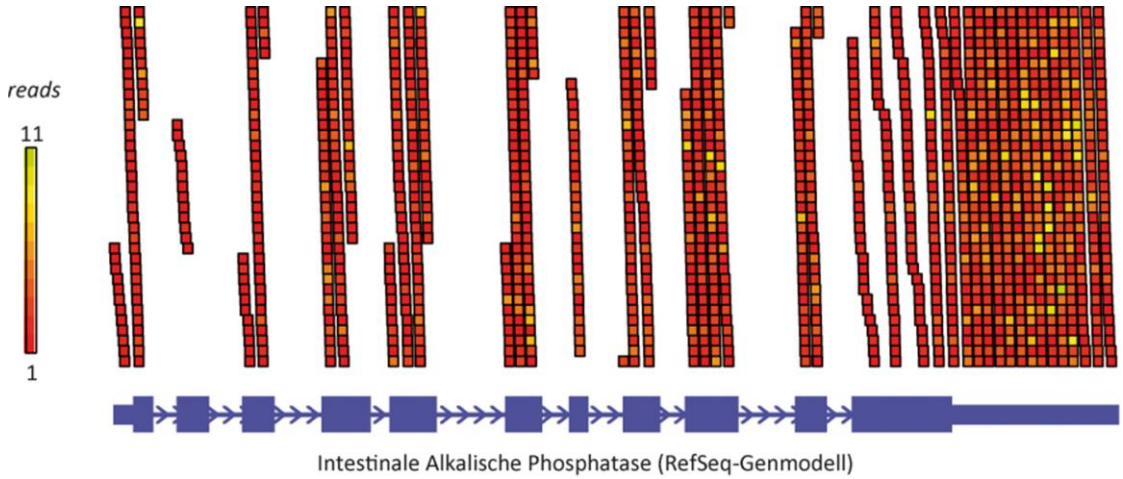


Abb. 11: Abdeckung in *read*-Auflösung der Alkalischen Phosphatase

Dargestellt sind die Startpunkte der *reads* über das *RefSeq*-Genmodell (blau). Dienen Positionen als Startpunkt für mehrere *reads*, kann die Anzahl über die farbliche Unterlegung abgelesen werden.

Transkripts gegenüber des 5‘-Endes eine höhere *coverage* aufwies. Um diese Beobachtung aus wenigen Kandidatengenen auf die Gesamtheit der Transkripte zu erweitern, wurden die *RefSeq*-Transkripte entsprechend ihrer Länge Klassen zugeordnet. Anschließend wurde für jede Klasse die durchschnittliche Transkriptabdeckung entlang der 5‘-3‘-Achse berechnet und graphisch dargestellt. Dieser Vergleich wurde sowohl für die cDNA-Synthese direkt aus Total-RNA als auch für mittels Oligotex vorangereicherter mRNA durchgeführt (siehe **Abb. 12**). Um dabei die Vergleichbarkeit von Transkripten zu gewährleisten, wurde sowohl auf die Transkriptlänge als auch auf die Expressionsstärke normalisiert, so dass alle Transkripte mit gleicher Gewichtung in die Untersuchung einflossen. Für sehr kurze Transkripte zeigte sich so, dass diese an den Transkriptenden eher unterrepräsentiert abgebildet sind, dieser Effekt wurde mit zunehmender Transkriptlänge weniger bedeutsam. Sehr lange Transkripte von mehr als 5000 Basen zeigten hingegen deutlichere Anzeichen der vermuteten Überrepräsentierung der 3‘-Enden. So zeigen hier die 5‘-Enden nur noch ca. 20% der Abdeckung im Vergleich zu den 3‘-Enden, wenn Total-RNA direkt in die cDNA-Synthese einsetzt wurde. Wurde die RNA zuvor für polyadenylierte Formen vorgereinigt, war die Unterrepräsentierung der 5‘-Enden mit 40 % geringer.

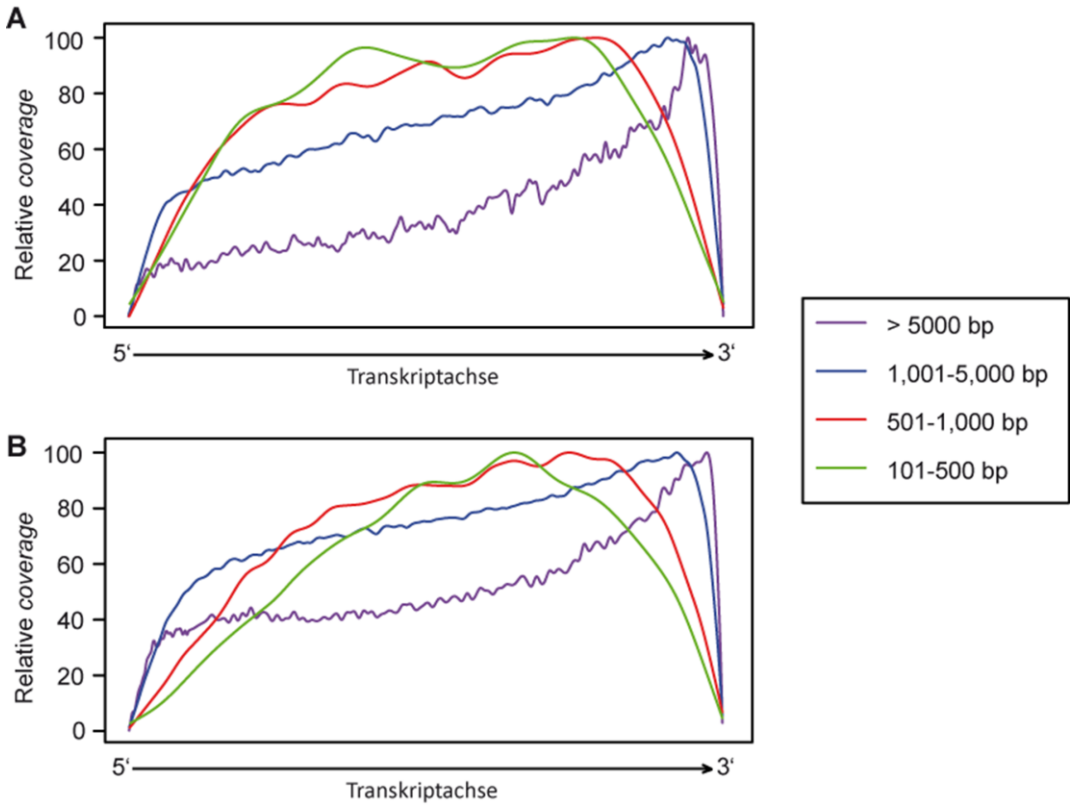


Abb. 12: RNA-Seq-Abdeckung in Abhängigkeit der Transkriptlänge

Globale Ansicht der Transkriptabdeckung (*coverage*) für alle *RefSeq*-Transkripte ausgehend von (A) einfach bzw. (B) doppelt für Polyadenylierung aufgereinigter mRNA. Dazu wurden die Transkripte zu Klassen basierend auf der Transkriptlänge zugeordnet.

3.2.2. Expressionsdaten – Reliabilität und *microarray*-Abgleich

Um die Reliabilität der ermittelten Expressionsdaten zu überprüfen, wurden für aus dem murinen Jejunum isolierte RNA zwei technisch unabhängige Experimente durchgeführt und die Expressionslevel der einzelnen Transkripte bestimmt. Anschließend wurde der Rangkorrelationskoeffizient nach Spearman (*Spearman Rho*) berechnet. Für dieses technische Replikat wurde dabei eine Korrelation von 0,92 beobachtet (siehe **Abb. 13A**). Um eine Aussage über die biologische Reliabilität zu ermöglichen, wurden zusätzlich die Expressionslevel der Transkripte in jejunaler RNA aus unterschiedlichen Mäusen bestimmt, die aus dem gleichen Wurf stammten und unter identischen Bedingungen gehältert wurden. Hier ergab sich ein Rangkorrelationseffizient von 0,82 (siehe **Abb. 13B**). Die Isolation von mRNA in ausreichenden Mengen für eine mRNA-Anreicherung ist nicht immer möglich. Auch wenn in dieser Arbeit die Menge an isolierter RNA kein limitierender Faktor war, sollte für zukünftige Projekte (z.B. basierend auf nur begrenzt verfügbaren Patientenbiopsien) ermittelt werden, ob die RNA zwingend für mRNA angereichert werden muss.

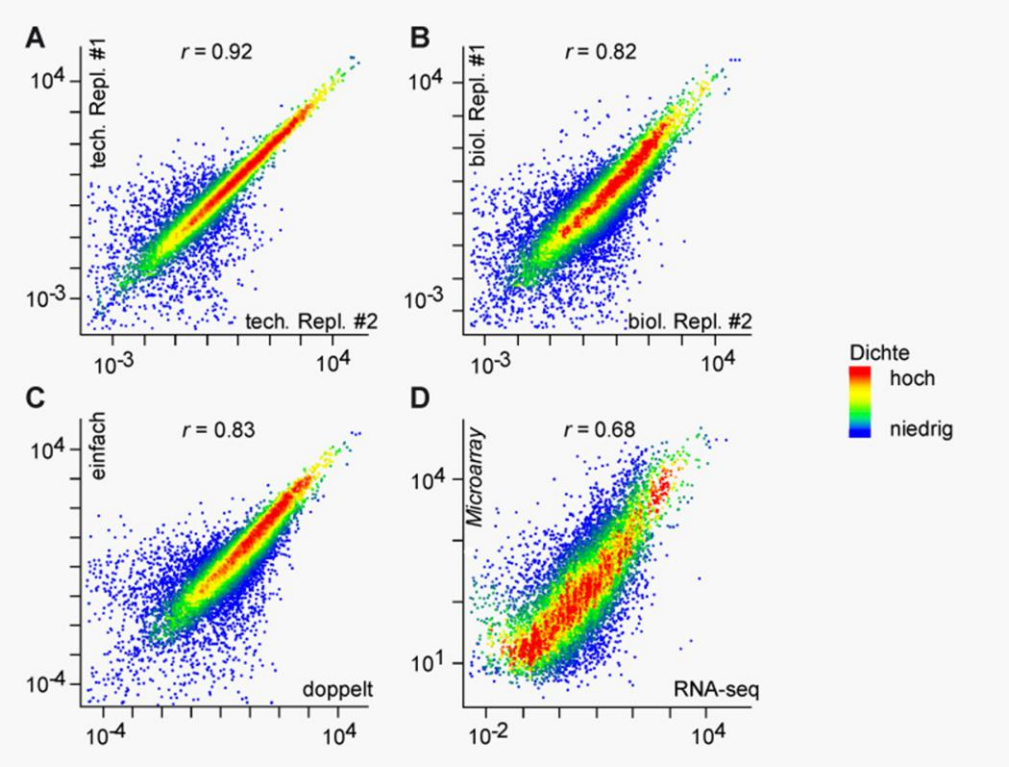


Abb. 13: Stärke der Genexpression – Reliabilität und *microarray*-Abgleich

Um eine Aussage über die Zuverlässigkeit der Ergebnisse zu erlauben, wurden verschiedene Versuche durchgeführt und die dabei ermittelten Expressionsstärken einzelner Gene gegeneinander aufgetragen. Zusätzlich wurde die Dichte der Punktwolke farblich markiert, um bei der enormen Anzahl an Datenpunkten in der Graphik Bereiche mit vielen Transkripten von solchen mit wenigen Transkripten kenntlich zu machen. Als numerisches Maß der Korrelation wurde der Rangkorrelationskoeffizient nach Spearman (*Spearman rho*, *r*) berechnet. Die einzelnen Graphen zeigen die relativen Expressionsstärken für **(A)** technische Replikate, ausgehend von identischer mRNA; **(B)** biologische Replikate, ausgehend von unterschiedlichen Mäusen des gleichen Wurfes bei gleichen Hälterungsbedingungen; **(C)** einen Vergleich von einfach gegenüber doppelt poly-A angereicherter RNA sowie **(D)** den Vergleich eines RNA-Seq- gegenüber eines *microarray*-Experiments. Sämtliche Skalen sind logarithmisch.

Dazu wurde der Einfluss der Anreicherung untersucht, indem 500ng angereicherte mRNA (Oligotex) sowie 1000ng nicht weiter vorbehandelte Total-RNA einer RNA-Probe als Ausgangspunkt für die SMART-cDNA-Synthese genutzt wurde, so dass die einfache Selektion für polyadenylierte Transkripte (nur Oligo-dT-Primer der cDNA-Synthese) der doppelten Selektion (Oligotex und cDNA-Synthese) gegenübersteht. Der Rangkorrelationskoeffizient für den Vergleich dieser beiden Verfahren war 0,83 (siehe **Abb. 13C**).

Um eine Aussage über die Validität der ermittelten RNA-Seq-Expressionsdaten zu erlauben, wurde zusätzlich zu einer RNA-Seq-Probe mit der identischen RNA die Genexpression basierend auf einem *microarray* untersucht. Um eine Vergleichbarkeit der beiden Versuche zu ermöglichen, wurden nur die Transkripte für den Vergleich herangezogen, die mit dem *microarray* als präsent (exprimiert) beobachtet wurden. Für diese Gruppe betrug der Rangkorrelationskoeffizient 0,73 (siehe **Abb. 13D**).

3.2.3. Beurteilung der Sequenziertiefe

Zur Abschätzung, inwieweit die gewählte Sequenziertiefe in der Lage ist, die Gesamtheit der exprimierten Transkripte abzubilden, wurde die Anzahl der detektierten *RefSeq*-Transkripte in Abhängigkeit der Anzahl der zugrundeliegenden *reads* (in Millionen eindeutig zuordbaren *reads*) ermittelt. Um Fehler durch die im Absatz 3.2 beschriebenen Sonderfälle zu vermeiden, wurde nicht eine reine *RefSeq*-Annotation genutzt, sondern eine modifizierte Version basierend auf den Gensymbolen, in der *RefSeq*-Transkripte mit überlappender genomischer Lokalisation künstlich vereint wurden (siehe 2.2.5). Durch diese Abwandlung der *RefSeq*-Transkriptdatenbank reduzierte sich die Anzahl der Eintragungen auf 21.923 (zuvor 27.722). Zugleich konnte ausgeschlossen werden, dass korrekt im Genom lokalisierte *reads* aufgrund von Überlappungseffekten einem falschen Transkript zugeordnet wurden.

Unter Verwendung einer zunehmenden Zahl von SMART-RNA-Seq-*reads* des Jejunums zeigte sich eine Sättigungskinetik mit zunächst steil steigender Anzahl an detektierten Transkripten, die dann zunehmend abflachte und unter Einsatz aller experimentell gewonnenen *reads* (> $77 \cdot 10$ reads) bereits nahe der Sättigung ist (siehe **Abb. 14**). Im Jejunum wurden so 14.801 der Transkripte der modifizierten *RefSeq*-Datenbank sicher (> 5 *reads*) detektiert (68% der Transkripte in der zugrunde liegenden Datenbank).

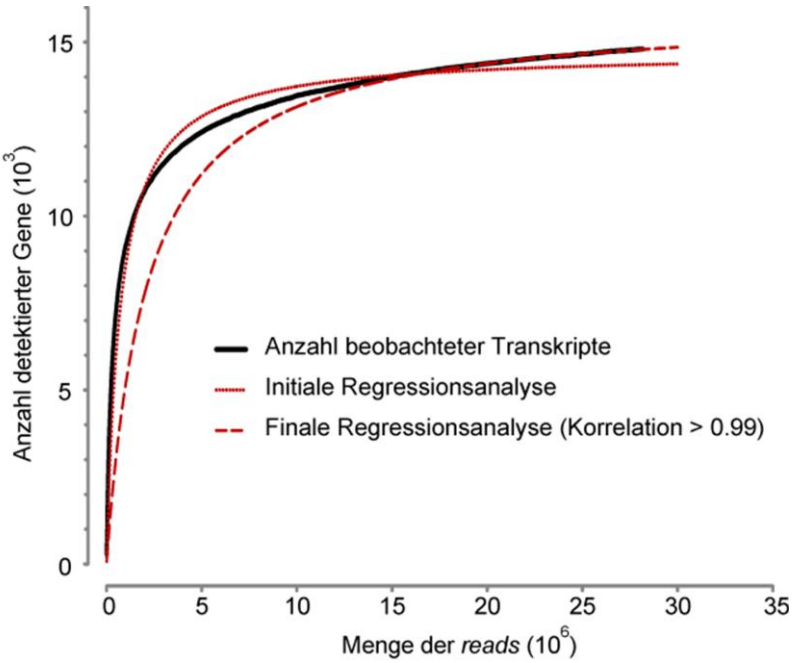


Abb. 14: Detektion der exprimierten Gene in Abhängigkeit der Anzahl genutzter *reads*

Die Abbildung gibt die Anzahl der mittels SMART-RNA-Seq als exprimiert detektierten Gene in Abhängigkeit der zugrunde liegenden Sequenziertiefe (in *reads*) wieder (schwarze Linie). Zu beachten ist, dass nicht die Gesamtheit der *RefSeq*-Transkripte genutzt wurde, sondern eine modifizierte *RefSeq*-Liste (siehe 2.2.5). Ergänzend wurde eine serielle Regressionsanalyse durchgeführt, um auf den bestehenden Daten die Anzahl der zu erwartender Gene für eine sich gegen unendlich nähernden Sequenziertiefe zu schätzen (rote Linie).

Zur Schätzung der Anzahl der tatsächlich exprimierten Transkripte bei einer sich unendlich nähernden Sequenziertiefe (Anzahl *reads* → ∞) wurden mehrere, nicht-lineare Regressionsanalysen durchgeführt, bis die Korrelation für den Bereich hoher *read*-Zahlen über 0,99 lag (für Details siehe 2.2.5). Die Gesamtanzahl an exprimierten Genen wurde so auf 15.884 geschätzt (72% der modifizierten *RefSeq*-Datenbank). Dies entspricht einer Zunahme der geschätzten Gesamtzahl exprimierter Transkripte um 1.083 (7%) gegenüber der Anzahl experimentell detektierter Transkripte.

3.2.4. Funktionelle Gliederung spezifisch exprimierter Gene

Im folgenden Abschnitt sollen die Unterschiede zwischen der Expression in den beiden untersuchten Geweben näher betrachtet werden. Aus den Daten ist ein direkter Vergleich der Expressionslevel eines festgelegten Transkripts in beiden Geweben möglich. Dieser Ansatz ist aber ungeeignet, um die hier vorgestellten, sehr komplexen Datensätze überschaubar zu vergleichen. Um dies zu erreichen, wurden die Transkripte näher betrachtet, die als spezifisch für jeweils eines der beiden untersuchten Gewebe beobachtet wurden (vgl. **Abb. 9**). Für alle spezifischen Transkripte wurden so mittels des bioinformatischen Hilfsmittels „*gene ontology*“ (siehe 2.2.7 und Gene Ontology Consortium 2008)) die beschriebenen molekularen Funktionen des im Transkript kodierten Proteins ermittelt. Anschließend wurde untersucht, für welche molekularen Funktionen auffällig viele oder aber gegenüber einer zufälligen Verteilung zu wenige Transkripte gefunden wurden, die in der Untersuchung eine gewebespezifische Expression zeigten. Die wesentlichen Funde sollen in der Folge kurz vorgestellt werden.

Für metabolische Prozesse der Makromoleküle fanden sich deutlich weniger gewebsspezifische Transkripte als durch Zufall zu erwarten war. Dies erscheint auch sinnvoll, da die Prozesse zum Auf- und Abbau der Makromoleküle in allen Zellen kaum unterschiedlich sind. Bei Betrachtung von Transkripten, die mit dem Ionentransport assoziiert sind, wurde im Colon ein hoher Anteil an gewebsspezifischen Transkripten beobachtet, die wesentliche Funktionen der Elektrolyt- und Wasserresorption des Dickdarms widerspiegeln. Analog finden sich im Jejunum viele gewebsspezifische Transkripte, die eine Rolle in der Immunabwehr entfalten. Hier findet sich ein Korrelat der immensen Bedeutung des Darms für die Immunabwehr.

Bei Betrachtung der gewebsspezifischen Expression von Genen mit Funktion in der Zell-Zell-Kommunikation fand sich sowohl im Jejunum als auch im Colon eine größere Anzahl von Vertretern dieser Klasse, als durch Zufall zu erwarten war. Dieses Ergebnis legt nahe, dass trotz vieler Ähnlichkeiten in Anatomie, Physiologie und Entwicklung Gewebe wie Jejunum und Colon unterschiedliche Wege der Zellkommunikation nutzen. Die Ergebnisse für die vorgestellten *gene*

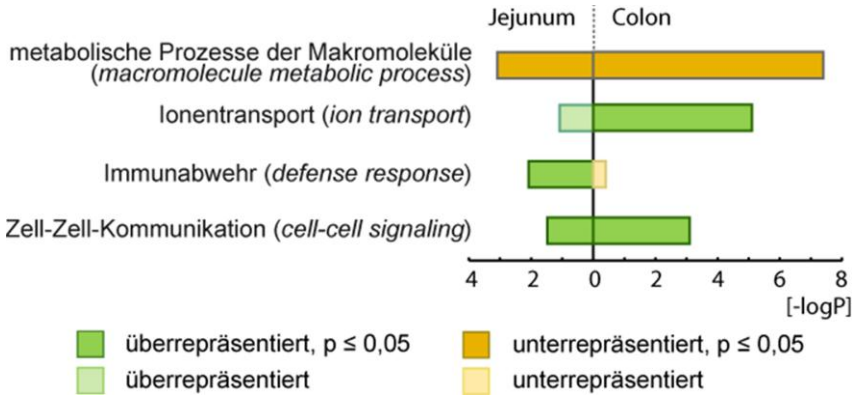


Abb. 15: *gene ontology*-Analyse ausgewählter biologischer Prozesse

Die Abbildung zeigt graphische die Ergebnisse der im Text vorgestellten „gene ontology"-Klassen. Anhand der Legende lässt sich ablesen, ob die entsprechende Klasse eine Über- bzw. Unterrepräsentierung innerhalb der im Gewebe spezifisch exprimierten Transkripte erfuhr und ob das Signifikanzniveau von 0,05 erreicht wurde. Die Skala gibt logarithmisch an, mit welcher Überschreitungswahrscheinlichkeit das Signifikanzniveau erreicht bzw. überschritten wurde.

ontology-Klassen finden sich in **Abb. 15**, eine vollständige Liste aller untersuchten *gene ontology*-Kategorien findet sich unter http://www.ikmb.uni-kiel.de/murine-transcriptomes.

3.2.5. RNA-Seq als Werkzeug zur Analyse des Spleißens

Um Spleiß-Mechanismen im Allgemeinen und insbesondere alternatives Spleißen zu beobachten, kann wie zuvor schon angedeutet beispielsweise die Abdeckung einzelner Exone verglichen werden. Dies hat den Nachteil, dass zwar die Frequenz des Auftretens des Exons in den Transkriptvarianten eines Genes geschätzt werden kann. Anhand der Abdeckung eines Exons kann aber keine Aussage getroffen werden, welches die benachbarten Exone im Transkript sind. Um eine bessere Auflösung dieser Feinstrukturen zu erreichen, wurde daher versucht, die Übergänge einzelner Exone genauer zu untersuchen. Dazu wurde ausgehend von der *RefSeq*-Datenbank eine Liste von potentiellen Spleißübergängen von einem Exon zum nächsten *in silico* berechnet. Dies erfolgte, indem je 25 Basenpaare des 3'-Endes eines Exons mit den 25 Basen des 5'-Endes des nächsten Exon verbunden wurden. Dies wurde für alle denkbaren Exonkombinationen eines Transkripts durchgeführt (siehe **Abb. 16**). In der Summe wurden 1.712.061 potentielle Spleißübergänge erzeugt. Zunächst wurde dieser Katalog auf das Auftreten von Duplikaten geprüft. Dies traf für 12.036 der potentiellen Spleißübergänge zu, welche für die weitere Analyse ausgeschlossen wurden. Der Großteil der generierten Spleißübergängen (1.700.025) war nicht redundant (siehe **Abb. 17A**). Sie dienten in der

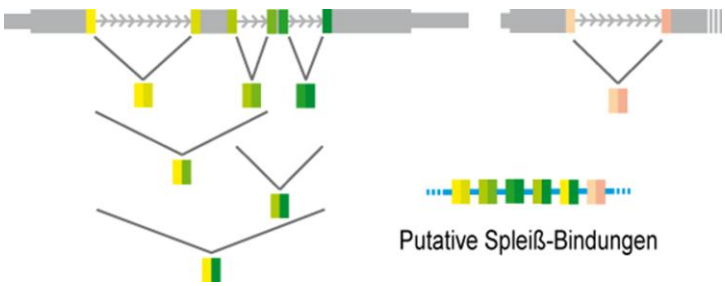


Abb. 16: Putative Spleißbindungen

Putative Spleißbindungen wurden erzeugt, indem 3‘-Enden und 5‘-Enden verschiedener Exone verbunden wurden. Dabei wurden folgende Regeln beachtet: (1) Das 3‘-Ende eines Exons wurde nur mit 5‘-Enden weiterer Exone verbunden, die hinter dem 3‘-Ende lagen. (2) Es wurden nur die Exone innerhalb eines annotierten Transkripts verbunden. Die Gesamtheit der so generierten Spleiß-Bindungen diente in der Folge als Rückgrat, um *reads* zu identifizieren, die die einzelnen Exone verbinden.

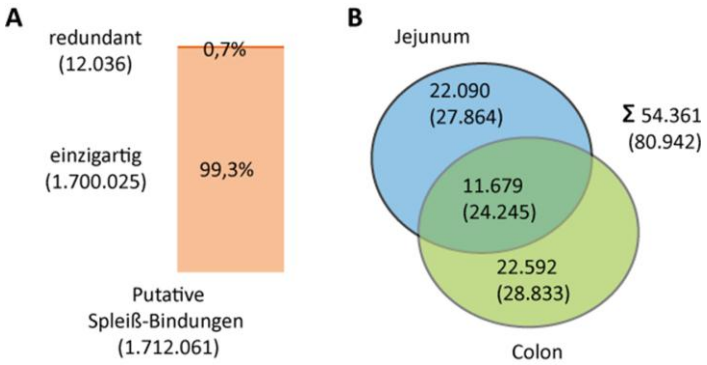


Abb. 17 Erzeugte und beobachtete Spleiß-Bindungen

(A) Vergleich redundanter zu einzigartigen Spleiß-Bindungen, basierend auf der *RefSeq*-Genannotation. **(B)** Tatsächlich innerhalb der gewonnenen RNA-Seq-Daten realisierte Spleißbindungen, die entweder im Jejunum (blau), Colon (grün) oder jeweils in beiden Geweben beobachtet wurden (Schnittmenge) sowie der Summe aller Teilmengen. Es wurden nur Spleißübergänge beachtet, die mindestens dreifach durch *reads* abgedeckt wurden, die in Klammern stehende Zahl gibt die Anzahl beobachteter Spleißbindungen bereits ab einem einzelnen *read* wieder.

Folge als Rückgrat, um *reads* abzugleichen, die bisher nicht einer genomischen Region zugeordnet werden konnten. Für insgesamt 80.942 der putativen Spleiß-Bindungen wurde tatsächlich zumindest ein *read* in den RNA-Seq-Daten gefunden. Für 24.245 Spleiß-Bindungen war dies unabhängig sowohl im Jejunum als auch im Colon der Fall. Viele dieser Bindungen sind nur durch einen *read* beobachtet worden. Wurden mindestens drei *reads* pro Gewebe für einen sicheren Nachweis gefordert, so reduzierte sich die Anzahl der beobachteten Bindungen auf 54.361 (siehe **Abb. 17B**).

Exemplarisch als interessantes Beispiel für potentielle Auswirkungen des alternativen Spleißens wurde Prosaposin (PSAP) beobachtet, welches das Vorläufer-Molekül für die Saposine (*sphingolipid activator pro(s)teins*) ist. Saposine sind wichtige Bestandteile des Sphingolipidstoffwechsels und sind Bestandteil der Lysosomen in den Zellen. Sie gehen aus der Vorläufer-Variante durch proteolytische Spaltung hervor (Kishimoto, Hiraiwa, und O’Brien 1992).

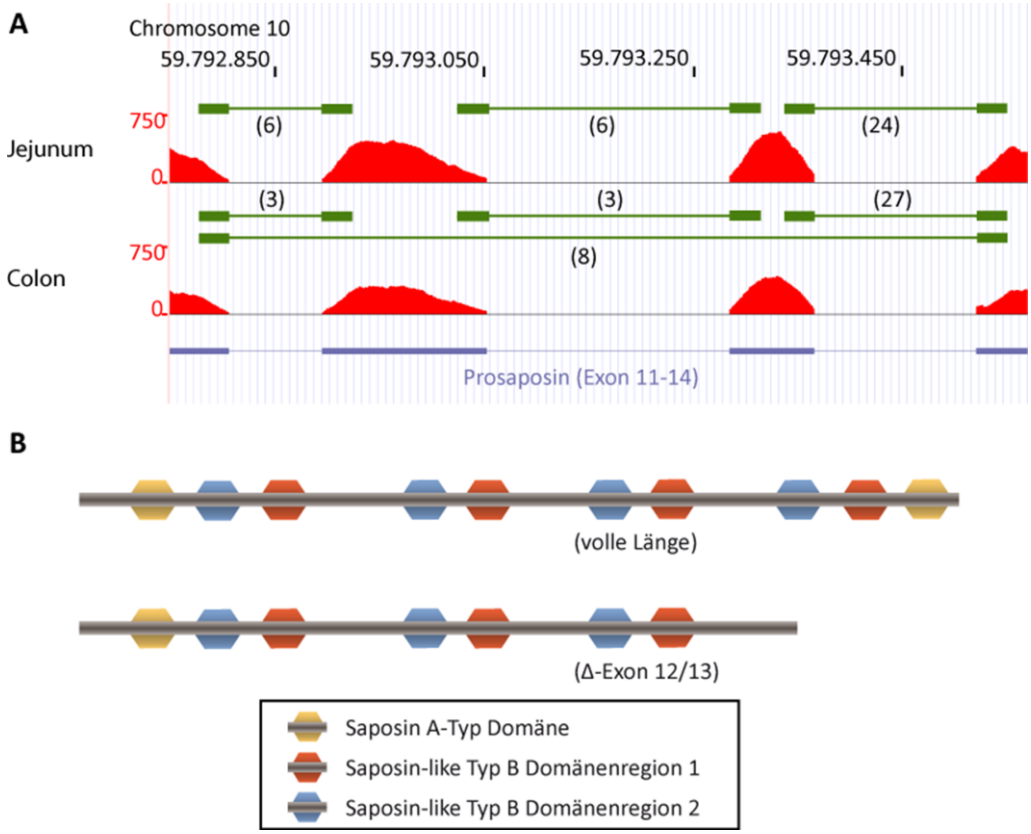


Abb. 18: Prosaposin als Beispiel für Alternatives Spleißen.

(A) Darstellung der beobachteten Spleiß-Bindungen (grün, in Klammern die Anzahl der beobachteten *reads*) für die Exone 11 bis 14 des Prosaposins, auffallend ist, dass im Jejunum keine direkten Spleißbindungen von Exon 11 zu 14 beobachtet wurden, während im Colon neben den im Jejunum realisierten Varianten noch eine direkte Verkettung von Exon 11 und 14 bestand. **(B)** Bei Betrachtung der Nukleotidsequenz dieser Δ-Exon 12/13-Variante fand sich der Einbau eines zusätzlichen, frühen Stopcodons. Das Transkript wird nicht mehr vollständig translatiert. Das resultierende Protein verliert in der Pfam-Domänenprädiktion seine letzten drei Domänen. Während Pro-Saposin in der Regel durch proteolytische Spaltung in seine aktiven Formen Saposin A, B, C und D überführt wird, legt diese Spleiß-Bindung die Existenz einer Variante nahe, die nur Saposin A, B und C enthält. Das aus den drei letzten Domänen bestehende Saposin D wird durch diese Transkriptvariante hingegen nicht kodiert.

Während die Normvariante des Prosaposins für die vier Saposine A, B, C und D kodiert, fand sich selektiv in den RNA-Seq-Daten, die aus den Colon-Proben erzeugt wurden, eine Spleißbindung, die auf eine verkürzte Transkriptvariante hindeutete, in der die Exone 12 und 13 nicht enthalten sind (siehe **Abb. 18A**). Bei Nutzung von Domänenvorhersage-Programme wie Pfam (Finn u. a. 2008) entspricht diese Spleißvariante einer Form des Prosaposin-Transkripts, in der durch die Einführung eines vorzeitigen Stopcodons das Saposin D nicht mehr kodiert ist (siehe **Abb. 18B**).

Zusammenfassend zeigen die Ergebnisse, dass die mittels RNA-Seq gemessene Genexpression im intestinalen Gewebe die Mehrheit der bekannten Transkripte umfasste, dabei die einzelnen Transkripte im Wesentlichen ausgeglichen und reproduzierbar abgebildet wurden. Eine noch tiefere Sequenzierung lässt die Detektion der Expression weiterer Transkripte vermuten. Des Weiteren zeigten die Ergebnisse, dass die Expression von Genen für molekulare Schlüsselfunktionen in den

jeweiligen untersuchten intestinalen Geweben nachweisbar ist und wesentliche Punkte in der Zusammensetzung des Transkriptoms wie z.B. alternatives Spleißen der Methode zugänglich sind.

3.3. Nicht-annotierte, transkriptionell aktive Regionen im Genom

Als wesentlicher Bestandteil dieser Arbeit sollte untersucht werden, ob mit der Methode des RNA-Seq Regionen im Genom entdeckt werden können, die zuvor nicht als transkriptionell aktiv galten. Um dieses Vorhaben umzusetzen, wurden ausgehend von der genomweiten *read*-Zuordnung des cDNA-Protokolls zunächst alle bereits annotierten Bereiche entfernt (basierend auf der *RefSeq*- und *Ensembl*-Annotation). Die verbliebenen Bereiche wurden darauf geprüft, ob sie noch durch die RNA-Seq-Daten abgedeckt werden. Dazu wurden die Datensätze aus Jejunum und Colon vereinigt, um eine bessere Abdeckung zu schaffen. Konnten nicht-annotierte Bereiche im Genom mit den RNA-Seq-Daten abgedeckt werden, so wurden die einzelnen Basen in größeren Gruppen von durchgängig abgedeckten Basen zusammengefasst und als Kandidaten für nicht-annotierte, transkriptionell aktive Regionen (nTAR) bezeichnet. Mit dieser Vorgehensweise wurde eine Vielzahl von Regionen im Genom identifiziert. Viele dieser Positionen waren aber nur wenige Basen lang oder nur durch wenige *reads* belegt.

Um die Zahl falsch-positiver Treffer einzuschränken und die Qualität der einzelnen Funde zu erhöhen, wurden zusätzlich zwei Qualitätskriterien erhoben und als Bemessungsgrundlage gewählt. Zum einen musste die durchschnittliche Basendeckung über den gesamten Bereich > 3 sein, zum anderen mussten mindestens 50 Nukleotide des Erbgut zusammenhängend abgedeckt sein. Damit wurde zugleich auch sichergestellt, dass die Abdeckung nicht auf der klonalen Amplifikation eines einzelnen *reads* beruhte. Alle Kandidaten für nicht-annotierte, transkriptionell aktive Regionen (nTAR) im Genom, die diesen Standard einhielten, wurden in einem zweiten Experiment einer Verifikation unterzogen. Dazu wurde ein zweites RNA-Seq-Experiment durchgeführt, da zwischenzeitlich durch den Hersteller des SOLiD-Systems, Life Technology, ein Protokoll für RNA-Seq auf den Markt gebracht wurde. Dieses Protokoll unterschied sich in erheblichen Maß von der in dieser Arbeit zuvor vorgestellten Vorgehensweise. Die Fragmentierung erfolgte hier bereits auf RNA-Ebene vor der reversen Transkription. Zuletzt erlaubte es dieses Protokoll auch, die Orientierung der ursprünglichen

Tab. 12: Überblick der wichtigsten Kennziffern für die WTAK-cDNA-Sequenzierung
Angegeben sind die Anzahl der *reads* sowie der prozentuale Anteil der Validierungs-Sequenzdaten.

cDNA-Fragmentierung		**Jejunum**		**Colon**	
Gesamtzahl reads		227.802.832		234.732.979	
Gesamtzahl zugeordnete reads	Anzahl (*reads*)	151.970.014	66,92%	128.017-728	54,54%
Eindeutig zuzuordnen	Anzahl (*reads*)	116.790.095	51,27%	94.395.393	40,21%

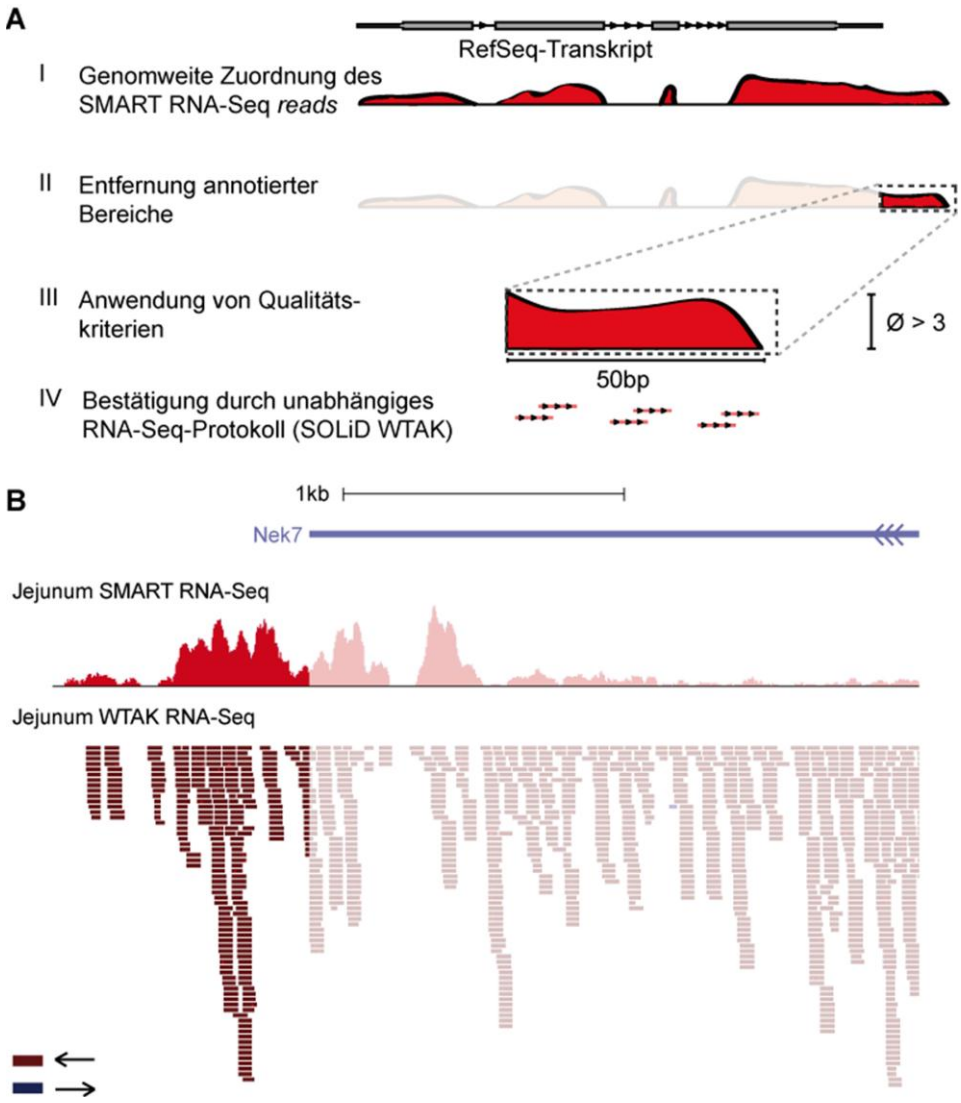


Abb. 19: Identifizierung nicht-annotierter, transkriptionell aktiver Regionen

(A) Zur Identifikation neuer transkriptionell aktiver Regionen (nTAR) wurde die bereits zuvor verwandte genomweite Zuordnung der *reads* des SMART RNA-Seq-Protokolls genutzt (I). Bereiche, in denen bereits annotierte Gene vorliegen (*RefSeq*- und *Ensembl*-Datenbank), wurden von der Betrachtung ausgeschlossen (II). Nicht-annotierte Bereiche des Genoms, die dennoch eine Abdeckung aufwiesen, mussten Qualitätskriterien erfüllen: eine durchschnittliche Abdeckung > 3 und eine Mindestlänge von 50 Basenpaaren (III). Elemente, die diese Kriterien erfüllten, wurden durch ein zweites RNA-Seq-Protokoll (WTAK RNA-Seq) verifiziert. Zusätzlich konnte in diesem Schritt die Orientierung des ursprünglichen Transkripts bestimmt werden. **(B)** Beispiel eines nTAR direkt am 3'-Ende des Gens *NIMA (never in mitosis gene a)-related kinase 7* (*Nek7*). Das Transkript zeigte sich deutlich über das annotierte Transkriptionsende hinaus verlängert. Das Niveau der Abdeckung blieb auf dem Niveau des annotierten Transkripts. Dieser Befund konnte mit der zweiten Methode verifiziert werden, zusätzlich zeigte sich, dass die *reads* die gleiche Orientierung besaßen, wie das benachbarte Gen, so dass eine elongierte Variante des Transkripts mit alternativem Polyadenylierungssignal naheliegend ist.

RNA zu bestimmen, so dass eine Aussage über die Orientierung der transkriptionellen Aktivität möglich ist. Die Eckdaten dieses Validierungsexperiments sind in **Tab. 12** zusammengefasst**.**

Um ein nTAR zu bestätigen, wurden mindestens drei unterstützende *reads* im Datensatz des mit dem zweiten Protokoll durchgeführten Experiments gefordert. **Abb. 19** zeigt einen graphischen Überblick

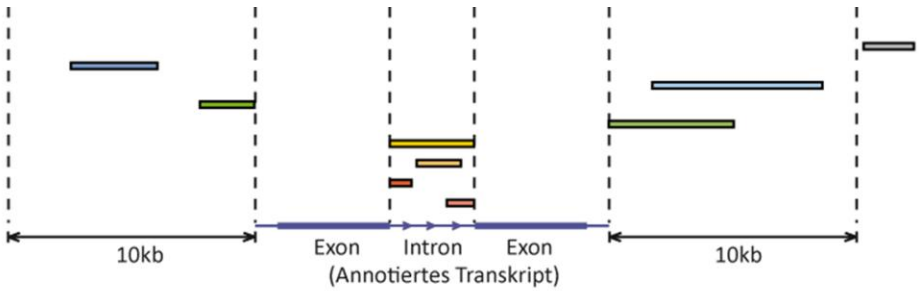
(A) und ein Beispiel **(B)** der Strategie zur Entdeckung nicht-annotierter, transkriptionell aktiver Regionen.

Mit dieser Vorgehensweise fanden sich in den Daten des SMART RNA-Seq-Protokolls 27.543 Bereiche im Genom, die die Qualitätskriterien erfüllten. 20.966 dieser Funde wurden durch das unabhängige, zweite RNA-Seq-Experiment bestätigt (76,12%).

Um diese Elemente näher zu betrachten, wurden sie abhängig von ihrer Position in der Nachbarschaft von bekannten, annotierten Genen diesen zugeordnet und aufgrund ihrer relativen Position klassifiziert. Dazu wurden folgende Klassen definiert: NGA (*non-gene associated*), UGN (*upstream gene neighborhood*), DGN (downstream gene neighborhood), UGI (*upstream gene intersection*), DGI (*downstream gene intersection*), ISE (*intron spanning element*), ELD (*exon-linked (downstream))*, ELU (*exon-linked (upstream))* sowie IGE (*intragenic element*). Die für die Zuordnung zu einer der Klassen erforderliche Position des nTAR im Bezug zu annotierten Genen ist im Methodenteil (siehe 2.2.12) beschrieben. **Abb. 20** zeigt eine graphische Zusammenfassung der definierten Klassen.

Im Fall eines nTAR, welches die Anforderungen für mehrere Klassen erfüllte, wurde jedem nTAR immer nur eine Klasse zugeordnet. Dabei wurden direkte Überschneidungen, Position und Entfernung zum nächsten Gen als Grundlage für die Priorisierung einer Klasse genutzt.

Basierend auf dieser Klassifizierung erfolgte eine Zuordnung für alle 20.966 nTAR, die durch das unabhängige RNA-Seq-Experiment bestätigt werden konnten. Dabei zeigte sich, dass die



Bezeichnung	Kürzel	Definition
Non-gene associated	NGA	Fund ohne annotiertes Transkripts im Abstand von 10 Kilobasen (kb)
Upstream gene neighborhood	UGN	Fund in Nachbarschaft zum 5'-Ende eines annotierten Transkripts
Downstream gene neighborhood	DGN	Fund in Nachbarschaft zum 3'-Ende eines annotierten Transkripts
Upstream gene intersection	UGI	Fund in direkter Überschneidung zum 5'-Ende eines annotierten Transkripts
Downstream gene intersection	DGI	Fund in direkter Überschneidung zum 3'-Ende eines annotierten Transkripts
Intron spanning element	ISE	Fund deckt das komplette Intron eines annotierten Transkripts ab
Intragenic element	IGE	Fund deckt einen Teil des Introns eines annotierten Transkripts ab
Exon-linked (downstream)	ELD	Fund deckt den Spleißakzeptor-Bereich eines Introns ab (3'-Ende des Introns)
Exon-linked (upstream)	ELU	Fund deckt den Spleißdonator-Bereich eines Introns ab (5'-Ende des Introns)

Abb. 20: Definition und graphische Darstellung der nTAR-Klassen

Erfüllte ein nTAR die Bedingungen gleich für mehrere Zuordnungen, wurden diese entsprechend einer Prioritätenregel (siehe Text) immer nur jeweils einer Klasse zugeordnet.

überragende Mehrheit der nTAR in Nachbarschaft, innerhalb oder sogar direkt angrenzend an Exone bekannter Gene lagen. Nur 2.780 der nTAR (13,3%) lagen mit mehr als 10.000 Basen Abstand nicht im direkten Umfeld bekannter Gene und wurden folglich als nicht-genassoziiert klassifiziert. Mit 55,8% knapp über die Hälfte der nTAR fanden sich in Intronen bekannter Gene, davon mit 8.232 die Mehrheit ohne direkt an annotierte Exone anzugrenzen (IGE). 2.914 nTAR zeigten eine Überschneidung mit beschriebenen Exonen. Dabei wurden Exon vorgelagerte nTAR (ELU) häufiger als nachgelagerte (ELD) beobachtet (1747 gegenüber 1167). Intron überspannende Elemente (ISE) waren mit 559 Beobachtungen die mit der geringsten Frequenz beobachtete Klasse im intronischen Bereich von Genen.

In der näheren Umgebung annotierter Gene fanden sich 6.481 nTAR (30,9%). Mehrheitlich mit 4.547 nTAR waren dabei die beobachteten Elemente den assoziierten Genen nachgelagert. In 1.934 Fällen befanden sie sich vor dem benachbarten Gen. In den meisten Fällen zeigte sich dabei keine direkte Überschneidung mit den assoziierten Genen. 430 nTAR grenzten direkt an die 5‘-UTR (UGI) bzw. 837 an die 3‘-UTR (DGI) eines Gens gegenüber 1.504 ohne direkten Kontakt zu einer 5‘-UTR (UGN) bzw. 3.700 zu einer 3‘-UTR (DGN). **Abb. 21** zeigt eine graphische Darstellung der Verteilung der nTAR.

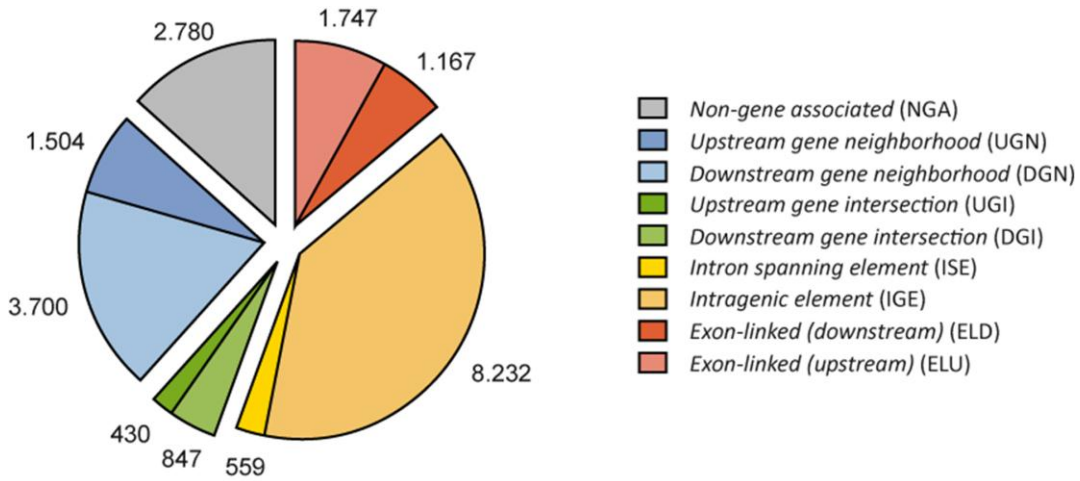


Abb. 21: Klassenverteilung der Gesamtzahl beobachteter nTAR

3.3.1. Orientierung und Expressionsstärke

Um die beobachteten nTAR näher zu charakterisieren, wurde zunächst betrachtet, wie die gefundenen nTAR im Vergleich zu ihren assoziierten Genen orientiert sind. Für diese Untersuchung wurden nur nTAR betrachtet, deren zugrunde liegenden *reads* eine einheitliche Orientierung besaßen. Bedingt durch diese Selektion und durch den Ausschluss der NGA-nTAR reduzierte sich die Anzahl der verwendeten nTAR auf 9.972. Dabei zeigte der überwiegende Anteil die gleiche Orientierung („*sense*“) wie das assoziierte Gen (8.334, 83,5%). Nur eine Minderheit (1.638, 16,5%) ist

entgegen der Orientierung des assoziierten Gens ausgerichtet („*antisense*“). So errechnet sich ein *sense/antisense*-Verhältnis über die Gesamtheit der genassoziierten nTAR von 5,09:1. Innerhalb der definierten Klassen waren aber zum Teil gravierende Unterschiede zu beobachten. So zeigten einzig nTAR in vorgelagerter Gennachbarschaft (UGN) ein nahezu ausgeglichenes Verhältnis von *sense*- zu *antisense*-Orientierung (1,09:1). Als intragenische Elemente (IGE) klassifizierte nTAR zeigten ein Verhältnis von 5,03:1. Aufgrund der großen Zahl dieser nTAR konnte absolut hier mit 769 der größte Anteil an *antisense*-nTAR beobachtet werden. Für nTAR in nachgelagerter Gennachbarschaft (DGN) fand sich ein den IGE vergleichbares Verhältnis mit 5,85:1. Als nachgelagerte Gennachbarschaft (DGN) klassifizierte nTAR zeigten so im Gegensatz zu UGN auch eine Begünstigung der *sense*-Orientierung. Alle anderen Klassen, die direkt an annotierte Gene grenzten, zeigten eine deutlich ausgeprägtere Neigung zur *sense*-Orientierung: 10,25:1 für vorgelagerte Genüberschneidungen (UGI), 14,26:1 für Intron überspannende Elemente (ISE), 16,75:1 für Exon vorgelagerte nTAR (ELD), 17,38:1 für nachgelagerte Genüberschneidungen (DGI) sowie 18,57:1 für Exon nachgelagerte nTAR (ELU). Zusammenfassend lässt sich sagen, dass eine *sense*-Orientierung bei den beobachteten nTAR deutlich häufiger auftritt. Dies gilt insbesondere für Regionen, die direkt an annotierte Gene angelagert sind. Das absolute Auftreten von *sense*- und *antisense*-nTAR für die einzelnen Klassen ist in **Abb. 22** dargestellt. Als Kontrolle des für diese Untersuchung genutzten Algorithmus wurde die Orientierung der NGA-nTAR auf ihrem Chromosom untersucht. Hier fand sich mit 963 gegenüber 941 ein nahezu ausgeglichenes Verhältnis von 1,02:1.

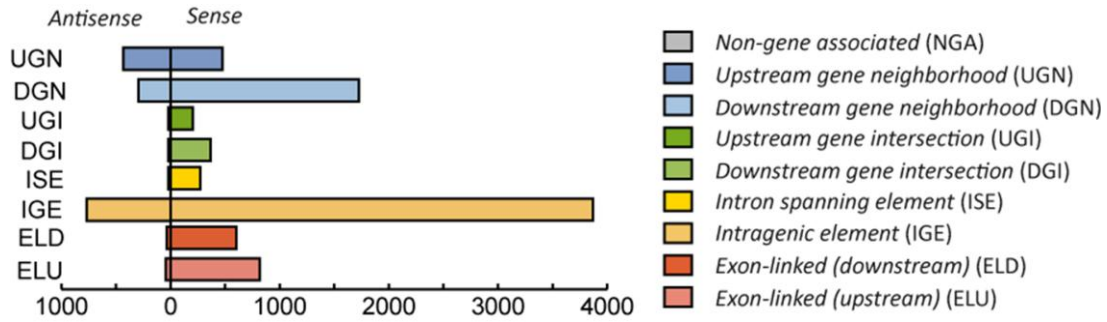


Abb. 22: Vorkommen und Orientierung genassoziierter nTAR

Es wurden nur nTAR herangezogen, die eindeutig in *sense*- bzw. *antisense*-Orientierung zu ihrem assoziierten Gen stehen.

Neben der Orientierung wurde auch die Expression der nTAR im Zusammenhang mit der Expression des assoziierten Genes betrachtet. Von 18.186 genassoziierten nTAR zeigte in 408 Fällen das assoziierte Gen keine Expression (2,24 % der Fälle). Für den Anteil der nTAR, für die eine Expression des verknüpften Genes zu beobachten war, wurde der Quotient aus Expression des nTAR und der Expression des assoziierten Genes gebildet. Für die Mehrheit (53,33 %) dieser beobachteten nTAR fand sich nur eine geringe relative Expression von nicht mehr als einem Viertel der Expression des assoziierten Gens. Für 24,6 % der nTAR wurde eine vergleichbare oder sogar höhere Expression

ermittelt. Für ISE-, ELD- und ELU-nTAR fanden sich nur selten bedeutende Expressionslevel erreichten. Bis zu 75 % der beobachteten nTAR zeigten eine Expressionsstärke von unter einem Viertel gegenüber dem assoziierten Gen. Für nTAR im Bereich des Transkriptionsstartes und Transkriptionsendes fanden sich sehr viel häufiger relevante Expressionslevel gleich oder höher des verknüpften Gens, so bei DGI-nTAR in 32,47%, bei UGN-nTAR in 38,03% und bei DGN-nTAR in 41,65% der Fälle. IGE-nTAR bildeten mit den UGI-nTAR das Mittelfeld mit 21,77% bzw. 16,51% der Fälle (siehe **Abb. 23**).

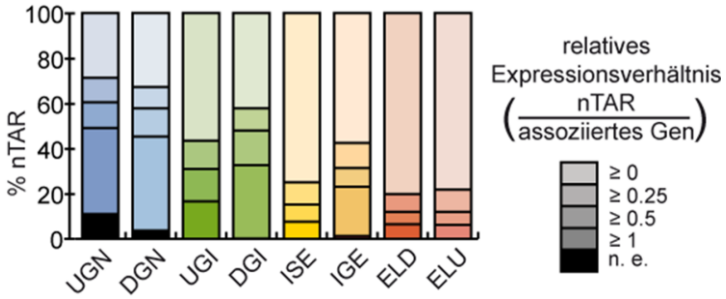


Abb. 23: Darstellung der relativen Expressionsstärke und der Anteile genassoziierter nTAR

Diese Abbildung zeigt den relativen Anteil der nTAR, die mind. 25%, 50% oder 100% des Expressionslevels des assoziierten Genes erreichen. Dazu wurde die mittlere Basenabdeckung des nTAR mit dem assoziierten Gen ins Verhältnis gesetzt. Einige nTAR zeigten auch eine Expression, obwohl ihr assoziiertes Gen nicht detektiert werden konnte (schwarz). n. e.: nicht exprimiert.

3.3.2. Polyadenylierungs-Nachweis über *pyrosequencing*

Zwar basierten die hier vorgestellten Daten ausschließlich auf polyadenylierter RNA, welches durch die Verfahrensweise der experimentellen Aufbereitung sichergestellt wurde. Da die nTAR aber überwiegend im Bereich bereits annotierter Transkripte liegen, war es von hohem Interesse, etwas über den Polyadenylierungs-Status direkt der einzelnen nTAR zu erfahren. Um dieses experimentell umzusetzen, wurde ein weiteres RNA-Seq-Verfahren basierend auf der Roche FLX *pyrosequencing*-Technologie im Labor etabliert. Zwar werden mittels *pyrosequencing* deutlich geringere Sequenzdatenmengen erhoben. In dieser Arbeit lag so die Menge an nutzbaren Sequenzdaten für das *pyrosequencing* bei $4 \cdot 10^6$ bp gegenüber $1{,}8 \cdot 10^9$ bp für *sequencing by ligation*. Durch die bessere Leselänge lassen sich aber Polyadenylierungen zweifelsfrei erkennen. Um die geringere Datenmenge dennoch gut nutzen zu können, wurde ein Protokoll gewählt, welches ausschließlich die Sequenzierung des 3‘-Endes eines Transkripts erlaubt (Torres u. a. 2008). Mit dieser Methode wurde für insgesamt 488 der nTAR eine Polyadenylierung gefunden. In **Abb. 24A** ist die Verteilung in die einzelnen Klassen gezeigt. Aufgrund der unterschiedlichen Sequenziertiefe ist ein direkter Vergleich mit den SOLiD-Daten schwierig. Bei Betrachtung des relativen Anteils, den die verschiedenen Klassen an der Gesamtheit der polyadenylierten nTAR besaßen, fanden sich deutliche

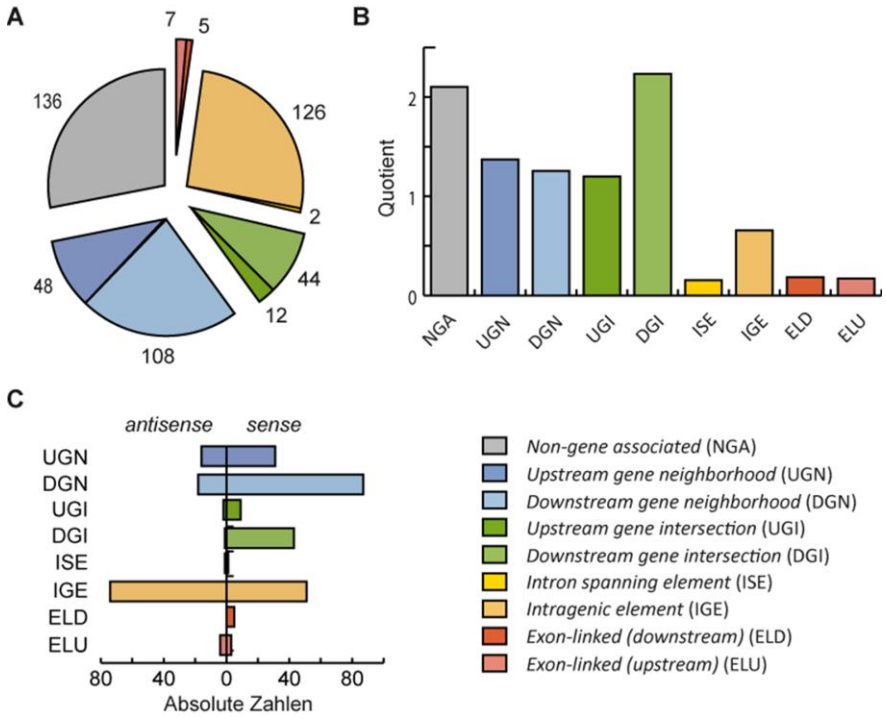


Abb. 24: Häufigkeit und Verteilung polyadenylierter nTAR

Polyadenylierung von nTAR wurde mittels Roche FLX *pyrosequencing* nachgewiesen. **(A)** Anzahl und Verteilung von polyadenylierten nTAR. **(B)** Quotienten aus der relativen Häufigkeit der untersuchten Klassen zwischen polyadenylierten nTAR und der Gesamtheit der nTAR (vergleiche **Abb. 21**). **(C)** Verteilung der polyadenylierten nTAR in Abhängigkeit ihrer Orientierung zum assoziierten Gen.

Unterschiede (siehe **Abb. 24B)**. So sind NGA- und DGI-nTAR mehr als doppelt so häufig in der polyadenylierten Gruppe anzutreffen. UGN-, DGN-, und UGI-nTAR sind relativ ausgeglichen. Unter den polyadenylierten nTAR finden sich hingegen aber kaum solche innerhalb von Genen: ISE-, ELD- und ELU-nTAR sind sehr selten, bei IGE-nTAR ist dieser Effekt weniger stark. Bei Betrachtung der Orientierung der polyadenylierten nTAR fällt für innerhalb von Genen gelegene nTAR auf, das diese häufig eine Polyadenylierung in gegensätzlicher Orientierung zum assoziierten Gen zeigten. Für die anderen Klassen sind fast ausschließlich nTAR in *sense*-Orientierung zum verknüpften Gen beobachtet worden. Insbesondere gilt dies für die DGI-nTAR, die bereits durch einen hohen Anteil polyadenylierter nTAR auffielen (siehe **Abb. 24C)**.

3.3.3. nTAR-Unterschiede zwischen den untersuchten Geweben

Während die bisherigen Beobachtungen der nTAR aus der Verbindung der beiden RNA-Seq-Datensätze aus Jejunum- und Colon-RNA getroffen worden sind, sollte abschließend überprüft werden, ob sich die Expression von nTAR zwischen den beiden untersuchten Geweben Colon und Jejunum unterschieden. Für NGA-nTAR wurde dazu das Verhältnis der Expressionsstärke in beiden Geweben ermittelt. Für die genassoziierten nTAR wurde dieses Verhältnis zusätzlich um

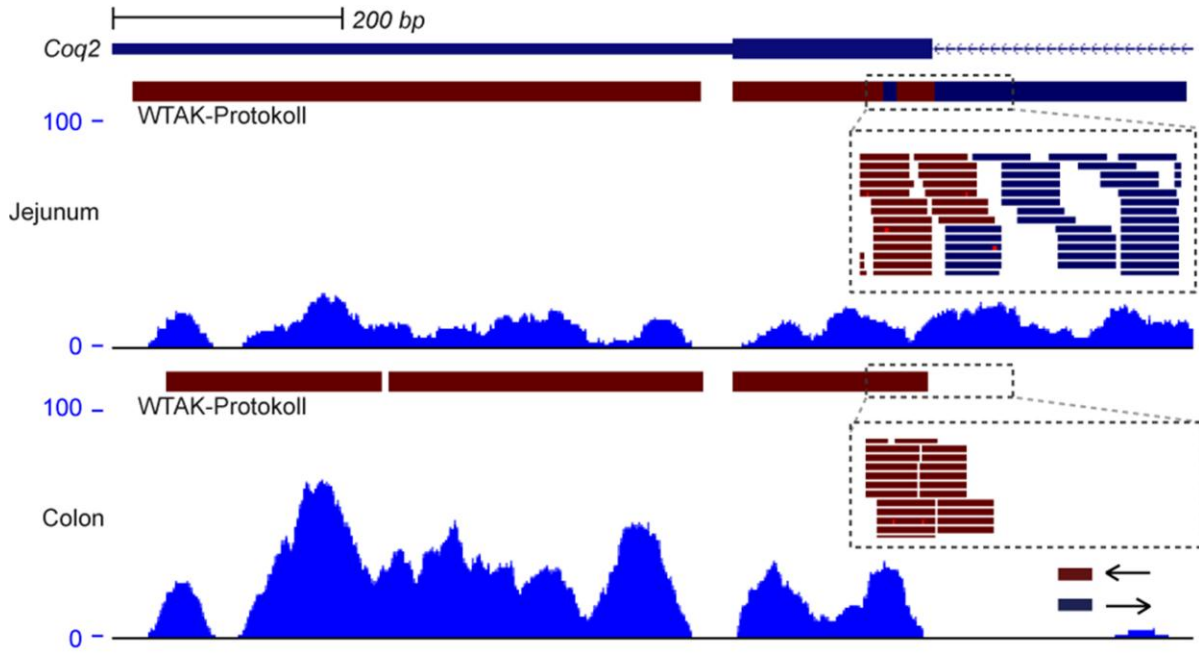


Abb. 25: *Coq2* als Beispiel für differentiell regulierte nTAR

Die Abbildung zeigt die SMART RNA-Seq-Abdeckung im Bereich des Gens *coenzyme Q2 homolog, prenyltransferase* (Coq2) für Jejunum und Colon. Zusätzlich wurde in einer Spur (WTAK-Protokoll) angedeutet, mit welcher Orientierung das WTAK-Protokoll die Expression ermittelt hat. Für einen Bereich, der im Jejunum eine Umkehrung der *read*-Orientierung zeigte, wurden zusätzlich die einzelnen *reads* angezeigt (gestrichelte Box). Abszisse: Position im Transkript, Ordinate: Abdeckung der Base.

Unterschiede in der Expression der assoziierten Gene korrigiert. Als differentielle Regulation wurde dabei eine mehr als dreifache Abweichung der Expressionsstärke angenommen. So konnte in einem Extrem ein genassoziiertes nTAR durch starke Schwankungen seiner eigenen Expression bei konstanter Expression des assoziierten Gens als differentiell reguliert betrachtet werden. Genauso denkbar ist aber auch, dass ein nTAR als differentiell reguliert betrachtet wurde, dass stabil exprimiert wurde und nur das verknüpfte Gen in seiner Expression stark schwankte. **Abb. 25** zeigt als ein Beispiel die transkriptionelle Aktivität rund um das letzte Exon des *Coq2*-Genes an. Hier ist das assoziierte Transkript in Jejunum deutlich schwächer exprimiert. Dafür findet sich vor dem letzten Exon eine transkriptionelle Aktivität, die im Colon nicht zu beobachten ist. Die Information über die Orientierung der hier gelegenen *reads* zeigte, dass es sich hier um eine Transkription in *antisense*-Orientierung zum annotierten Gen handelte.

Insgesamt wurden 1.483 nTAR gefunden, die unter den zuvor genannten Parametern als differentiell reguliert betrachtet wurden. NGA-nTAR besaßen mit 759 Ereignissen ein deutliches Übergewicht. Unter den restlichen Gruppen von 724 als differentiell-reguliert angenommenen nTAR dominierten die IGE-nTAR (siehe **Abb. 26A)**. Bei Betrachtung der Orientierung der differentiell-regulierten nTAR fand sich, dass die Anzahl der in *sense*-Orientierung vorliegenden nTAR deutlicher rückläufig waren als *antisense*-nTAR (siehe **Abb. 26B)**. Dieser Effekt ist bei UGN- und IGE-nTAR relativ schwach

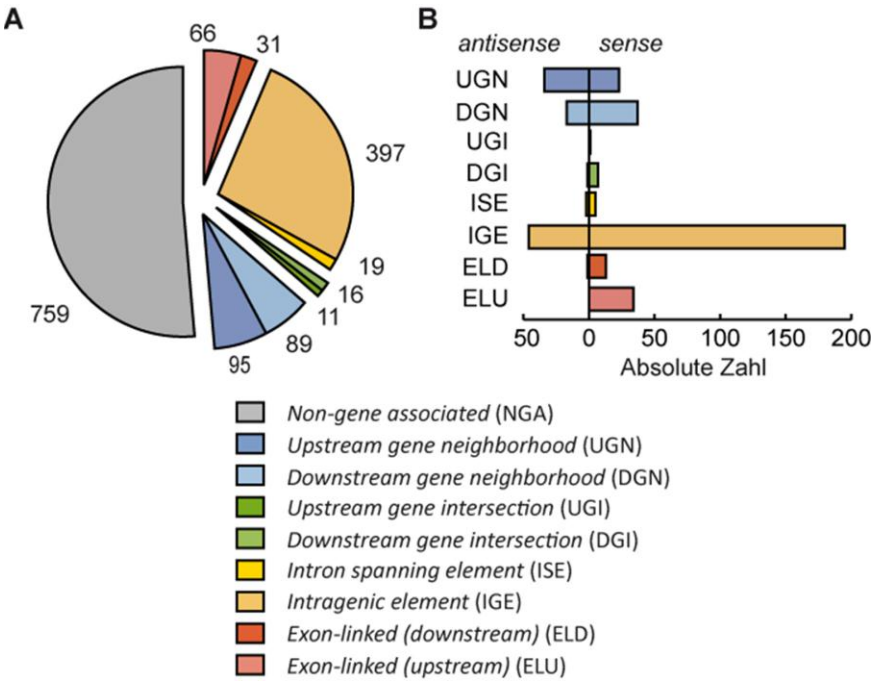


Abb. 26: Anzahl und Verteilung differentiell regulierter nTAR

(A) Anzahl der nTAR, die im Jejunum und Colon um mindestens das dreifache abweichende Expressionslevel aufzeigten. Für nTAR, die dabei mit einem annotierten Gen assoziiert waren, wurde eine Korrektur um den Wert vorgenommen, um den die Expressionsstärke des assoziierten Genes in beiden Geweben abweicht. **(B)** Zusätzlich wurde die Orientierung der differentiell exprimierten Gene für genassoziierte nTAR dargestellt, vergleiche dazu auch **Abb. 22**.

ausgeprägt. DGN-, ELD- und ELU-nTAR zeigen schon deutlich geringere Anteile der differentiell regulierten *sense*-nTAR. Im Fall der UGI-, ISE- und DGI-nTAR tendiert der Anteil der differentiell regulierten nTAR in *sense*-Orientierung gegen null.

Zusammenfassend wurde gezeigt, dass mittels umfangreicher Sequenzierung von cDNA aus intestinalen Geweben zusätzliche transkriptionelle aktive Bereiche des Genoms identifiziert werden können. Im Wesentlichen handelt es sich dabei um Bereich, die in direkter Nachbarschaft zu anderen Genen standen, dabei aber mehrheitlich in der Expressionsstärke gegenüber ihren assoziierten Genen abfielen. Weiter zeigten die Ergebnisse, dass in vielen Fällen für diese Befunde eine Vereinbarkeit als Transkriptvarianten bekannter Gene bestand, z.B. anhand der Orientierung und dem Auftreten von Polyadenylierungssignalen in Abhängigkeit zur Position der benachbarten Gene. Ein weiteres Ergebnis dieser Arbeit ist aber auch, dass es eine nennenswerte Anzahl von nicht annotierten Bereichen gibt, die entgegen der Orientierung bekannter Gene transkribiert wurden. Zusätzlich wurden Bereiche identifiziert, die im Versuch in den beiden untersuchten Geweben eine stark abweichende Expression zeigten.

4. Diskussion

4.1. SOLiD RNA-Sequenzierung

In dieser Arbeit wurde mit der Etablierung eines Protokolls zur RNA-Sequenzierung unter Nutzung der SOLiD-Sequenziertechnologie begonnen. Zu diesem Zeitpunkt waren dazu keine Methoden veröffentlicht. Daher wurde aufbauend auf Erfahrungen mit dem Roche FLX Sequenziersystem (Cheung u. a. 2006) eine Methode basierend auf der initialen Umschreibung von RNA in cDNA mit anschließender Fragmentierung etabliert (siehe Abschnitte 2.1.6, 2.1.7, 3.1, 3.1.1). Ausgehend von 500ng mRNA erfolgte die Erstellung der Sequenzier-cDNA *library,* welche bereits mit der SOLiD V2-Version zur erfolgreichen Erhebung von großen Mengen an Sequenzdaten führte (> $1{,}8 \cdot 10^9$ zuordbare, sequenzierte Basen, ca. $5 \cdot 10^7$ zuordbare *reads*). Zeitweilig wurden mit dieser Methode deutlich mehr Sequenzinformationen als in anderen Arbeiten basierend auf *next generation sequencing*-Systemen erhoben (z.B. < 10 Millionen *reads* von nur 27 bp Länge in Sultan u. a. 2008).
Die Selektion für polyadenylierte RNA-Formen basierte dabei auf einer Oligo-dT-Aufreinigung gefolgt von der direkten cDNA-Synthese über ein Oligo-dT-Primer. In beiden untersuchten Geweben wurden mit dieser Methode mehr als 90% der zuordbaren *reads* bekannten Exonen zugeordnet. Im Verlauf der Arbeit konnten durch andere Arbeitsgruppen weitere Methoden zur RNA-Sequenzierung mittels der SOLiD-Technologie veröffentlicht werden. So wurde in der Veröffentlichung einer australischen Forschungsgruppe (Cloonan u. a. 2008) für den SOLiD neben der mRNA-Aufreinigung über die Polyadenylierung die direkte Fragmentierung der RNA genutzt. Die cDNA-Synthese erfolgte hier erst im Anschluss an die Fragmentierung der RNA. Dabei wurden Nukleotidhexamere genutzt, die bereits den Sequenzieradaptor trugen. Die cDNA-Synthese erfolgte ohne direkte Selektion der Polyadenylierung. Hier lag der Anteil an zu bekannten Exonen zuordbaren *reads* an der Gesamtheit der im Genom lokalisierbaren *reads* bei unter 60%. Mit mehr als 40% fanden sich deutlich mehr *reads* in anderen, nicht als transkriptionell-aktiv beschriebenen Regionen des Genoms.
Auch wenn nicht auszuschließen ist, dass der Anteil der Exon-zuordbaren *reads* durch Unterschiede in den zu Grunde liegenden Datenbanken variiert, ist die Diskrepanz erheblich. Zwar wurde in der erwähnten Publikation mit der *UCSC genome browser*-Genannotation eine andere Datenbank verwendet. Diese Datenbank nutzt dabei neben *RefSeq, Genbank* und *Unitprot* als Quellen für die Annotation, so dass der Umfang z.B. der proteinkodierenden Einträge um 10 % gegenüber *RefSeq* erhöht ist (http://genome.ucsc.edu). In der hier vorgelegten Arbeit wurde neben *RefSeq*- die *Ensembl*-Genannotation gewählt, eine Kombination, die aufgrund ihrer Nutzung in anderen Veröffentlichungen verwandt wurde (z.B. Sultan u. a. 2008). Dabei beruht *Ensembl* auch ausschließlich auf experimentellen Daten von mRNA- und Proteinsequenzen

(http://www.ensembl.org). Weiter erfolgt eine regelmäßige Synchronisation mit anderen Datenbanken für Transkript- und Proteinsequenzen, darunter neben *RefSeq* auch *Genbank* und *Uniprot*. Im Wesentlichen beruhen die benutzten Datenbanken der hier verglichenen Experimente also auf dem Austausch der gleichen Ausgangsdaten. Eine erhebliche Abweichung in Umfang und Qualität dieser Datenbanken ist daher nicht naheliegend.

Die genutzten Methoden unterschieden sich hingegen in entscheidenden Punkten (s.o.). Während bei der getrennten Aufreinigung und cDNA-Synthese so eine nicht vollständige Aufreinigung von mRNA z.B. durch falsche Salzkonzentrationen oder zu tiefe Temperaturen denkbar ist, wurde in dem hier vorgestellten Verfahren die mRNA doppelt für eine Polyadenylierung aufgereinigt. Dabei ist der zweite Schritt direkt an ein enzymatisches Verfahren gekoppelt, welches für seine Funktion auf die genaue Übereinstimmung der molekularen Wechselwirkungen beruht. Die Anfälligkeit für akzidentell in der Probe vorhandene RNA-Formen ohne Polyadenylierung erscheint hier deutlich reduziert. Eine denkbare Erklärung für die Diskrepanz zwischen den verglichenen Methoden wird hier also in der unterschiedlichen Stringenz bezüglich der Poly-A-Aufreinigung gesehen, dabei muss es sich aber nicht um den einzigen bedeutsamen Faktor handeln. Eine weitere Abklärung würde aber die Aufarbeitung und Kenntnis des exakten experimentellen Ablaufes für die verglichenen Arbeiten erfordern. Im Vergleich mit anderen Experimenten beruhend auf der Illumina-Sequenziertechnologie fand sich ein Verhältnis vergleichbar der in der hier vorgelegten Arbeit beobachteten Verteilung von *reads* bzgl. annotierter/nicht-annotierter Bereiche (Mortazavi u. a. 2008).

Für die Quantifizierung der bekannten Gene, wie sie in den folgenden Absätzen diskutiert wird, ist ein höherer Anteil Sequenzdaten, der mittels der hier etablierten Methode für annotierte Exone erzeugt wurde, unzweifelhaft von Vorteil. Für den Nachweis von nicht-annotierten, transkriptionell aktiven Regionen im Genom kann hingegen die von Cloonan u.a. vorgestellte Methode das vielleicht vollständigere Bild zeigen. Dennoch wurde in dieser Arbeit auch für dieses Vorhaben die selbst eingeführte Methode vorgezogen, da basierend auf der vorherigen Einschätzung einer konservativen gegenüber einer vielleicht vollständigeren, potentiell aber weniger stringenten Identifizierung von polyadenylierten Transkripten der Vorzug gegeben wurde.

Darüber hinaus bietet die Methode von Cloonan u. a. aber auch Vorteile, so kann die Orientierung der sequenzierten Transkripte ausgelesen werden. Ein vergleichbares Verfahren basierend auf RNA-Fragmentierung wurde in dieser Arbeit daher zu einem späteren Zeitpunkt genutzt, um die Verifikation der hier beobachteten Funde vorzunehmen.

Neben dem hohen Anteil von *reads*, die bei der Fragmentierung von cDNA bekannten Exonen zugeordnet werden konnten, zeigte sich in einem Vorversuch, dass die Sequenzierung von ribosomal depletierter RNA zu nur wenigen *reads* führte (<10%), die überhaupt einer Position im Genom zugeordnet werden konnten. Dies wurde darauf zurückgeführt, dass der Darm als Grenzfläche zur

Umwelt auch viele Zellen der Mikrobiota enthält und eine Trennung der isolierten RNA bei ribosomal depletierter RNA nicht stattfindet, eventuell sogar der Anteil der mikrobiellen Gesamt-RNA einer Probe gesteigert wird, da die Depletionsverfahren auf eukaryotische RNA ausgerichtet waren. Durch die strikte Selektion von polyadenylierten RNA-Formen wurde diese mögliche Ursache für Fehler in der weiteren Arbeit umgangen, da prokaryotische mRNA-Formen überwiegend nicht polyadenyliert sind (Madigan und Martinko 2008). Während aber mikrobielle RNA-Formen in dieser Arbeit durch die Wahl der Methoden ausgeschlossen wurden, liegt hier zugleich die Möglichkeit verborgen, in Zukunft das Wechselspiel vom Transkriptom des Wirts mit seinen mikrobiellen Kommensalen direkt zu betrachten, auch wenn dafür umfangreiche Anpassungen der Methode erforderlich sind.
Im Verlauf der Arbeit wurden das Verfahren des RNA-Seq weiter entwickelt, insbesondere konnte die erforderliche Menge an Ausgangs-RNA gesenkt werden, so dass auf dieser Methode aufbauende Folgeprojekte (Schmid u. a. 2012; Autran u. a. 2011) deutlich geringere Mengen von RNA aus nur wenigen Zellen für die Erstellung von Transkriptom-Daten benötigten.

4.2. Quantifizierung der Genexpression mittels RNA-Seq

4.2.1. Genexpression im murinen Darmtrakt

Mit den in dieser Arbeit durchgeführten Experimenten wurde das erste Mal die Genexpression in intestinalen Geweben mittels *next generation sequencing* untersucht. Für 20.541 von 27.772 *RefSeq*-Genen wurde dabei eine Expression bei Nutzung des *Cufflinks*-Algorithmus (Trapnell u. a. 2010) in zumindest einem der analysierten Darmabschnitte beobachtet. Dies entspricht einem prozentualen Anteil von 74,1 % aller annotierten Gene, 64,9 % davon unabhängig in beiden Geweben. Auch wenn die Häufigkeit von Transkripten nicht zwingend mit der Menge des synthetisierten Proteins oder dessen Aktivität korreliert, finden sich unter den am stärksten exprimierten Transkripten z.B. Defensine im Dünndarm oder Carboanhydrasen als Wassertransporter im Dickdarm, also Transkripte, deren Genprodukt mit der Funktion des Organs gekoppelt sind.
Diese Form der Analyse der intestinalen Genexpression bestätigt und erweitert das in früheren Arbeiten erworbene Wissen basierend auf der Anwendung der *microarray*-Technologie (Bates u. a. 2002, siehe auch 4.2.5). Weiter zeigte das RNA-Seq zuvor beschriebenen Vorteile (Mortazavi u. a. 2008; Cloonan u. a. 2008; Sultan u. a. 2008) für die Analyse der Genexpression gegenüber anderen Methoden. Die Daten zeigten eine hohe dynamische Breite, durch den Umfang der erhobenen Sequenzmengen lässt sich die Auflösung nahezu beliebig steuern. Und es wurden fast überhaupt keine unspezifischen Signale gemessen, für mehr als 90% der nicht-detektierten Gene wurde nicht ein einzelner zugeordneter *read* beobachtet. In einem Vorversuch konnten für ca. 80 Millionen künstlicher, per Zufall generierter *reads* so nur für 33 eine Zuordnung im Genom erfolgen (siehe Abschnitt 3.1.1). Zurzeit werden zwar meist noch 5 *reads* für einen sicheren Nachweis gefordert (z.B.

Tang u. a. 2009). Da aber nicht ein indirektes Signal, sondern direkt die Sequenz des *reads* ermittelt wird, ist eine fehlerhafte Zuordnung unwahrscheinlich, so dass ein einzelner *read* die Expression eines Gens zumindest andeutet. Für die Minderheit der Transkripte, für die zwar einzelne *reads* beobachtet wurden, die aber nicht den gewählten Nachweisschwellenwert erreichten, erscheint so die Möglichkeit des Nachweises durch die weitere Erhöhung der generierten Sequenzdaten (siehe auch 4.2.4).Mögliche Fehlerquellen wie Kontaminationen müssen hingegen weiterhin sorgfältig erwogen und kontrolliert werden. Zumindest über Artgrenzen hinweg besteht potentiell über einen Sequenzabgleich aber die Möglichkeit, diese zu identifizieren und aus der Analyse zu entfernen.

Die Verteilung der Expressionsstärken einzelner Gene ähnelte dabei einer logarithmischen Normalverteilung. Sowohl wenige Gene mit sehr hohen Expressionswerten als auch wenige Gene mit sehr geringen Expressionswerten wurden beobachtet. Die überwiegende Zahl der Transkripte zeigten Expressionslevel, die um nicht mehr als den Faktor 1000 abwichen. Während einige Transkripte in großer Menge erforderlich sind, um die Funktion des Gewebes zu ermöglichen, gibt es ebenso eine Gruppe von Transkripten, die nur sehr geringe Expressionsstärken zeigten. Inwiefern diese entscheidend für den Aufbau und Unterhalt des Gewebes sind, kann anhand dieser Daten nicht diskutiert werden.

Insgesamt zeigte der Darm als Funktionsträger mit unterschiedlichsten Aufgaben wie Nahrungsaufnahme, Immunabwehr, aber auch als Organ bestehend aus Nerven- und Muskelgewebe eine hohe Komplexität der Genregulation. Zurzeit ist ein umfangreicher Vergleich mit RNA-Seq-Daten für andere Gewebe aufgrund der beschränkten Anzahl veröffentlichter Forschungsartikel nur unzureichend möglich. Zumindest für Gewebe aus dem Gehirn von neugeborenen Mäusen wurde die Expression von 58,7% der annotierten Gene berichtet (Han u. a. 2009). Hier muss aber bedacht werden, dass alternatives Spleißen von Genen nicht berücksichtigt wird, welches insbesondere im Gehirn die Vielfältigkeit des Transkriptoms moduliert (Mortazavi u. a. 2008). Dennoch zeigte sich der Darm in diesen Untersuchungen als ein Organ, dessen Funktion in seiner Gesamtheit auf der Aktivität eines bedeutenden Anteils der bekannten Gene beruht.

4.2.2. Abdeckung entlang der 5‘-3‘-Transkriptachse

Bei Betrachtung der Abdeckung entlang eines Transkripts fanden sich häufig Bereiche, die durch die erhobenen *reads* nicht abgedeckt wurden, z. B. einzelne Exone ohne jegliche Expression. In manchen Fällen wurden diese vermutlich aufgrund alternativen Spleißens im untersuchten Gewebe nicht exprimiert. Andere denkbare Möglichkeiten sind konservierte, in der Nukleotidsequenz mehrfach im Genom realisierte Bereiche, z.B. Genduplikationen, so dass *reads* nicht eindeutig zugeordnet und in den Analysen nicht berücksichtigt wurden. Initiale Versuche zur RNA-Seq durch andere

Forschungsgruppen zeigten aber auch eine methodische Verzerrung der Abdeckung entlang der Transkriptachse.

Primär fragmentierte RNA mit anschließender Ligation und Sequenzierung zeigte so eine Depletierung der Transkriptenden. Wurde die RNA hingegen zunächst in cDNA umgeschrieben, fragmentiert und mit den Sequenzligatoren versehen, ergab sich eine starke Anreicherung der 3‘-Enden zu Ungunsten der Abdeckung des restlichen Transkripts (Wang, Gerstein, und Snyder 2009). Auch in der hier vorgestellten Methode zeigten sich Verzerrungen entlang der 5‘-3‘-Transkriptachse. Für kurze RNA-Stränge wurden zwar nur geringe Abreicherungen der *coverage* an beiden Enden der Transkripte beobachtet, vermutlich basierend auf der Depletion der artifiziellen cDNA-Enden. Transkripte zwischen 1000 und 5000 bp Länge zeigten aber bereits eine leichte Überrepräsentierung von 3‘-Endfragmenten. Transkripte von einer Länge über 5000 bp zeigten eine deutliche Anreicherung der 3‘-Enden, auch wenn diese nicht das Ausmaß der von Wang u.a. geschilderten Dimension erreichte.

Als Ursache für die beschriebene hohe Abdeckung der 3‘-Enden in der cDNA langer Transkripte wurden in dieser Arbeit zwei Möglichkeiten bedacht. Beruht die cDNA-Synthese auf teilweise denaturierter RNA, so kann mittels eines Oligo-dT-Primers nur das endständige Fragment umgeschrieben werden. Degradierung von RNA erfolgt in unserer Umwelt zum Großteil über RNasen, die überall in der Natur vorkommen und z.B. auch auf der Haut des Menschen gebildet werden (Harder und Schröder 2002). Entscheidend ist hier die sorgfältige Durchführung der RNA-Isolation und weiterführender Schritte, um die Degradation der RNA weitestgehend zu verhindern. Zusätzlich wurde in dieser Arbeit die Qualität der RNA engmaschig überprüft, um Probleme frühzeitig zu identifizieren. Als zweite Ursache für eine starke Überrepräsentation des 3‘-Endes in der Transkriptabdeckung kann ein vorzeitiger Abbruch der reversen Transkription angenommen werden. Da diese bei Nutzung eines Oligo-dT-Primers immer vom 3‘-Ende des Transkripts ausgeht, würde hier zwangsläufig eine Anreicherung erfolgen. Dies konnte in dieser Arbeit durch die Nutzung der *template switch*-Technologie (SMART *cDNA synthesis*) methodisch verhindert werden, so dass hier eine mögliche Ursache der gleichmäßigeren Abdeckung gegenüber dem von Wang u. a. beschriebenen Bild besteht.

Eine mögliche Erklärung für die Überrepräsentierung der 3‘-Enden sehr langer Transkripte trotz Verwendung eines *template switch* ist, dass durch die Länge eine vollständige reverse Transkription nur ineffizient gelingt, somit denaturierte und damit kürzere Varianten langer Transkripte während der Amplifikation durch eine höhere Effizienz in den Vordergrund treten, während dies bei kürzeren Transkripten vernachlässigbar ist. Für weiterführende Arbeiten z.B. über die Quantifizierung von differentiell gespleißten Varianten muss dieser Umstand bedacht und entsprechend korrigiert werden.

Ausgeschlossen werden konnte, dass die Abdeckung einzelner Transkripte auf der klonalen Amplifikation weniger *reads* basiert. In der Regel fand sich für Transkripte ein vielfältiges Muster von *reads*. Für Transkripte mit hoher Abdeckung diente der überwiegende Anteil der Basen als Startpunkt der einzelnen *reads* (vergleiche **Abb. 11**).

4.2.3. Verlässlichkeit der Methode

In veröffentlichten Arbeiten konnte deutlich eine hohe Korrelation von RNA-Seq-Ergebnissen zu tatsächlich in der Probe vorhandenen RNA-Formen gezeigt werden (Mortazavi u. a. 2008). Um die Verlässlichkeit der in dieser Arbeit vorgestellten Methode einzuschätzen, wurden sowohl technische als auch biologische Replikate erzeugt und die ermittelten Genexpressionsstärken miteinander verglichen. Zusätzlich wurde noch ein Vergleich mit einer modifizierten Form der mRNA-Aufreinigung mit direkter cDNA-Synthese aus Total-RNA angestellt, also deutlich reduzierten Anforderungen an das Ausgangsmaterial. Abschließend erfolgte der Vergleich mit dem bisherigen Goldstandard für genomweite Expressionsanalysen, den *microarrays*.

Im technischen Replikat ausgehend von derselben RNA-Probe fand sich eine hohe Spearman-Rangkorrelation der ermittelten Genexpressionsstärken von 0,92. Dabei zeigten nur schwach exprimierte Transkripte höhere Variabilitäten, während die Genexpressionsstärke für mittel- und hochgradig exprimierte Gene in den beiden unabhängigen technischen Experimenten sehr viel stabiler war. Naheliegend ist, dass bei sehr gering exprimierten Transkripten bereits kleine, zufällige Unterschiede in der Effizienz der reversen Transkription zu erheblichen Abweichungen bei der Messung der Transkriptionsstärke führen können. Eine denkbare Lösung, um die genaue Quantifizierung von schwach exprimierten Genen weiter zu steigern, ist die parallele cDNA-Synthese mehrere Ansätze der gleichen Probe, um zufälligen Schwankungen auszugleichen.

Zwischen den beiden biologischen Replikaten fand sich eine Spearman-Rangkorrelation von 0,82. Schwankungen wurden beobachtet, die über die von technischen Replikaten hinausgehen. Bei Betrachtung der mittels einer Dichtefunktion gefärbten Verteilungen zwischen technischen und biologischen Replikat fällt auf, dass hier insbesondere die Variabilität nicht mehr auf schwach exprimierte Genen beschränkt ist, sondern auch zunehmend mittel- bis hochgradig exprimierte Gene stärkeren Schwankungen ausgesetzt sind. Insgesamt zeigt dieser Versuch dennoch, dass in der Regel das Transkriptom in definierten Geweben von genetisch identischen Individuen eine hohe Korrelation bzgl. der Expression der Gesamtheit der Gene unter stabilen Umweltbedingungen zeigt.

Der Vergleich von zweifacher Aufreinigung für eine Polyadenylierung und hohen Ausgangsmengen an RNA gegenüber nur einfacher Selektion für die Polyadenylierung bei deutlich reduzierten Anforderungen an die RNA-Menge zeigte mit 0,83 eine vergleichbare Korrelation wie zwischen biologischen Replikaten. Hier zeigten bei Ansicht der Dichteverteilung in den Korrelationsplots die

mittel- bis hochgradig exprimierten Gene eine höhere Variabilität als in den technischen Replikaten, wenn auch nicht so hoch wie bei den biologischen Replikaten. Nicht beobachtet wurde, dass die verschiedenen Techniken einen Einfluss auf die Erfassung von bestimmten Expressionsstärken zeigten. So finden sich für beide Protokolle sehr schwach exprimierte Transkripte, die mit der jeweils anderen Technik als deutlich stärker exprimiert gemessen wurden. Eine Aussage, welche Methode der RNA-Aufbereitung die tatsächlichen Transkript-Häufigkeiten besser abbildet, ist anhand dieser Daten nicht möglich. Allerdings zeigte sich für die aufwendiger aufgereinigte RNA eine bessere Kontinuität bei der Abdeckung entlang der 5‘-3‘-Achse, so dass diese Methode bei Projekten vorzuziehen ist, die von einer gleichmäßigen Abdeckung besonders profitieren, soweit es die erforderliche Verfügbarkeit von großen RNA-Mengen zulässt. Die Ursache, warum diese Verzerrung bei direkt eingesetzter Total-RNA stärker ausfiel, konnte in dieser Arbeit nicht sicher identifiziert werden. Denkbar wäre, dass die Stabilität der RNA durch die Aufreinigung erhöht wurde, indem z.B. die Anzahl potentieller RNasen in der wässrigen RNA-Lösung reduziert wurde.

Im Abgleich mit *microarray*-Untersuchungen fand sich eine Spearman-Rangkorrelation von 0,68, wobei die Anzahl der als exprimiert nachgewiesenen Transkripte im *microarray* gegenüber RNA-Seq deutlich reduziert war. Diese Korrelation ist vergleichbar mit anderen Analysen zwischen *microarray*- und RNA-Seq-Experimenten (Rosenkranz u. a. 2008; Wilhelm u. a. 2008) sowie dem Vergleich verschiedener *microarray*-Plattformen untereinander (Kuo u. a. 2006). Bei Betrachtung der Dichteverteilungs-Darstellung fällt auf, dass das *microarray* tendenziell bereits für im RNA-Seq-Experiment mittelgradig exprimierte Gene eine maximale Expression angibt. Dies ist im Einklang mit der beschriebenen, deutlich geringeren dynamischen Breite des *microarrays* (Mortazavi u. a. 2008).

4.2.4. Einfluss der Sequenziertiefe

Die Möglichkeiten der DNA-Sequenzierung sind mit Einführung der *next generation sequencing*-Technologieplattformen sprunghaft gestiegen. Aber auch nach der Einführung steigt die Kapazität weiter stark an. Die ersten SOLiD-Versionen erzeugten pro Lauf nur ca. 1 Gb Sequenz-Rohdaten, mit der neueren Version SOLiD 5500 sind laut Angaben des Herstellers 300Gb möglich. Konnten so zu Beginn dieser Arbeit nur relativ wenige Proben in dem hier vorgestellten Umfang sequenziert werden, hat sich dies in den letzten Jahren dramatisch geändert.

Dennoch muss zu Beginn eines Forschungsprojektes die Frage beantwortet werden, wie viel Sequenzdaten zu generieren sind, um ein Transkriptom im Hinblick auf die Fragestellung analysieren zu können. Sind nur die stark exprimierten Gene im untersuchten Gewebe von Interesse, reichen bereits wenige *reads.* Im Bereich von Nicht-Modell-Organismen konnte beispielsweise die FLX-Technologie, die mangels Referenzgenom hier eine außerordentliche Rolle trotz deutlich geringerer

read-Anzahl pro Lauf spielt (bis zu 10^6), sinnvoll für die Analyse der Zusammensetzung des Transkriptoms eingesetzt werden (Lange u. a. 2011; Hemmrich u. a. 2012).

Bei Betrachtung der Anzahl der detektierten Gene findet sich bei der Untersuchung eines komplexen Organs wie des hier untersuchten Darms nach über 25 Millionen eindeutig zuordbarer *reads* eine mit zunehmender Anzahl der analysierten *reads* abflachende Kinetik, die aber unter Einbeziehung aller *reads* noch nicht vollständig im Sättigungsbereich angelangt ist. Nach der hier vorgenommenen Schätzung mittels nichtlinearer Regressionsanalyse kann die Anzahl der nachweisbaren, annotierten Transkripte noch um bis zu 7% erhöht werden, wenn die zugrunde liegenden Sequenzdaten beliebig gesteigert werden. Dies deckt sich auch mit dem Fund von Transkripten, die zwar den Schwellenwert von 0,01 FPKM nicht übertreten konnten und somit als nicht exprimiert aufgefasst wurden, dennoch aber wenige *reads* ihre Existenz aufzeigten und sich damit von der Mehrheit der nicht exprimierten Gene abhoben, für deren Expression überhaupt kein Anhalt gefunden werden konnte.

Spezielle Fragestellungen, die über die einfache Quantifizierung von Transkripten hinausgehen, so beispielsweise die genomweite, quantitative Auswertung von Spleiß-Isoformen, können noch weitaus größere Mengen an Sequenzdaten erforderlich machen, versprechen aber auch wichtige Einblicke in zentrale Fragen der Biologie.

4.2.5. Untersuchung biologischer Prozesse mittels *gene ontology*

Da die Aussagekraft einzelner differenziell regulierter Gene für die unterschiedliche Identität der untersuchten Gewebe gering ist , wurde nach einer Alternative gesucht, um Unterschiede der zellulären Zusammensetzung auf Ebene der Transkripte in den Geweben zu untersuchen. Dazu bot sich die *gene ontology*-Klassifizierung von Genen zu wichtigen biologischen Prozessen an.

In dieser Arbeit wurden die Gene näher analysiert und deren Zuordnung in die einzelnen *gene ontology*-Klassen geprüft, die nur in einem der untersuchten Gewebe detektiert wurden. Für diese Methode wurde in dieser Arbeit gezeigt, dass sie in der Lage ist, Schlüsselfunktionen eines Gewebes darzustellen. So fanden sich Prozesse wie der Ionentransport im Dickdarm oder Immunabwehr im Dünndarm durch eine hohe Vielfalt an gewebespezifisch exprimierten Transkripten repräsentiert. Durch die Beschränkung auf für das jeweilige Gewebe spezifisch detektierte Gene wurden zugleich Gene mit weitreichenden Funktionen im Zellstoffwechsel, die ubiquitär in allen Zellen des Körpers exprimiert werden, aus der Betrachtung entfernt. So fanden sich in den gezeigten Beispielen metabolische Prozesse des Protein- und Nukleotidstoffwechsels deutlich herabreguliert, da in diesen Gruppen der Anteil an Transkripten überwiegt, die in allen Zellen essentiell sind. Mit der Zell-Zellkommunikation fand sich ein Beispiel für die Beobachtung, dass beide Gewebe eine über einen Zufallsbefund zu erwartende Anzahl von Nachweisen für diese Klasse aufwiesen. Dieser Prozess spielt in beiden Geweben eine wichtige Rolle, die hier erhobenen Daten auf Transkriptionsebene zeigten

aber, dass unterschiedliche Mediatoren und intrazelluläre Signalkaskaden in den beiden untersuchten Gewebe beteiligt sind, wie zuvor berichtet werden konnte (Bates u. a. 2002).

Eine wichtige Limitierung der *gene ontology*-Klassifizierung liegt dabei in der Datenbank selbst. Ist ein biologischer Prozess nicht als solcher in den Datenbankdefinitionen hinterlegt oder sind ihm nur wenige Gene zugeordnet, so kann eine An- oder Abreicherung dieses Prozesses nur schwer beobachtet werden. Dieser Effekt verstärkt sich, wenn viele *gene ontology*-Abfragen parallel durchgeführt werden und durch die Korrektur für multiples Testen hohe Hürden für eine statistische Signifikanz zu nehmen sind. Manche biologische Prozesse sind noch wenig charakterisiert oder kaum mit beteiligten Genen hinterlegt. So existiert beispielsweise keine *gene ontology*-Klasse, die sich primär an Transportvorgänge der Schleimhäute richtet. Entsprechend konnte keine Klasse diese unbestrittene Funktion der untersuchten Gewebe zeigen.

In dieser Arbeit wurden exprimierte gegen nicht-exprimierte Gene verglichen. Dies erlaubte aufgrund des fast nicht messbaren Hintergrundsignals eine klare Trennung. Zusätzlich konnten Probleme für die Auswertung von RNA-Seq-Experimenten umgangen werden, wie sie von Young u.a. für besonders stark exprimierte oder lange Transkripte und der Nutzung von *gene ontology*-Untersuchungen beschrieben wurden (Young u. a. 2010). Für umfassendere Analysen basierend auf *gene ontology* ist die Weiterentwicklung der Methodik unter Einbeziehung der Expressionsstärke dennoch erstrebenswert.

Für zukünftige Projekte müssen also die Fragen beantwortet werden, welche biologischen Prozesse untersucht werden sollen, ob diese ausreichend in der *gene ontology*-Datenbank ausgearbeitet sind und welche Schwellenwerte der Expressionsstärke für eine sinnvolle Analyse festzulegen sind. Dann kann diese Form der Analyse die Komplexität des molekularen Netzwerks weiter reduzieren und anschaulich die funktionelle Identität von Geweben darstellen.

4.2.6. Analyse des alternativen Spleißens

Zwar erlaubt RNA-Seq, Spleißformen unabhängig der bestehenden Genannotation *de novo* aus Sequenzdaten und Referenzgenom zu bestimmen, indem die Enden eines *reads* jeweils einer Position im Referenzgenom zugeordnet werden und anschließend geprüft wird, ob eine Kontinuität besteht oder eine Form des Spleißens vorliegen muss. Um eine solche Analyse hochwertig durchzuführen, müssen aber beide Enden sicher zugeordnet werden können. Da die *read*-Länge mit 35 bp in der hier vorliegenden Arbeit nicht ausreichend war und kürzere Fragmente von weniger als 18 bp zu einer Vielzahl von Zuordnungen im Genom führen würden, wurde auf eine genomweite Zuordnung der *reads* verzichtet und die bestehende Exon-Annotation als Ausgangsposition genutzt. Durch die Verknüpfung aller bekannten Exone eines Gens untereinander wurden mehr als $1{,}7 \cdot 10^6$ putative Spleißbindungen erstellt und ermittelt, wie viele durch die erhobenen *reads* abgedeckt wurden. Für

mehr als 80.000 dieser artifiziellen Spleißverbindungen wurde eine Abdeckung gefunden, in ca. 54.000 dieser Fälle durch mehrere *reads*. Dieser Befund befindet sich in der gleichen Größenordnung wie vergleichbare Untersuchungen zu diesem Thema (Pan u. a. 2008; Sultan u. a. 2008), darunter viele Befunde, die in der bisherigen Annotation von Spleißvorgängen nicht berücksichtigt wurden. Pan u. a., die verschiedene Gewebe des Menschen untersuchten, fanden so in einem Gewebe im Mittel ca. 60.000 Spleißbindungen, nahmen sie ein zweites Gewebe hinzu, erhöhte sich die Anzahl der beobachteten Spleißbindungen im Mittel auf ca. 100.000. Dabei wurde nicht das Transkriptom zumindest ähnlicher Gewebe wie in der hier vorgelegten Arbeit verglichen, sondern das Spleißverhalten der Transkripte so unterschiedlicher Gewebe wie Gehirn, Herz, Skelett, Muskel, Lunge und Leber. Für Gehirn und Leber bestätigte sich dabei das Bild eher vielfältiger Spleißbindungen. Auch wenn sich die methodische Herangehensweise in den Details unterscheidet, zeigt dies, dass auch im Darm Spleißen eine gewichtige Rolle einnimmt, die einen Umfang vergleichbar mit den Spleißvorgängen anderer Gewebe annimmt.

Exemplarisch wurden einige der Befunde näher untersucht. So fand sich für das in den Ergebnissen (siehe 3.2.5) dieser Arbeit näher vorgestellte Spleißen des Prosaposins eine Variante. Diese zeigte im Colon zusätzlich zum konventionellen Spleißen *reads*, die für eine alternative gespleißte Form des Prosaposins spricht. Diese entspricht basierend auf Pfam-Analysen einer gekürzten Variante des Vorläuferproteins, der das Saposin D fehlt. Dies zeigt, wie alternatives Spleißen die Funktion von Proteinen beeinträchtigen kann, hier z.B. potentiell dazu beitragen kann, dass aus einer Vorläuferform eines Proteins eine unterschiedliche Anzahl an reifen Proteinen hervorgehen kann.

Die Möglichkeiten, alternatives Spleißen in dieser Arbeit zu untersuchen, waren aufgrund der kurzen *read*-Längen bei zugleich enormen Mengen an benötigten *reads* begrenzt. Mit Fortschreiten der Sequenzier-Technologien und der Möglichkeit, *mate-pair* oder *paired-end libraries* zu sequenzieren, bietet sich hier aber enormes Potential für die Zukunft. Auch ist man der eingangs erwähnten Idealvorstellung näher, eine *de novo*-Analyse des alternativen Spleißens durchzuführen. So existieren bereits erste Programme, die nicht auf einer bestehenden Genannotation aufbauen (Guttman u. a. 2010).

4.3. Nicht-annotierte, transkriptionell-aktive Regionen

Die Anwendungsmöglichkeiten des RNA-Seq sind nicht auf die Erfassung bereits bekannter Transkripte beschränkt, sondern erlauben auch die Detektion zuvor nicht-bekannter Transkripte oder Transkriptmodifikationen. Anders als z.B. zur Nutzung von *microarrays* ist eine Festlegung *a priori* auf bestimmte Nukleotidsequenzen nicht erforderlich. Die Untersuchung von bisher nicht-annotierten Bereichen des Genoms, die im Darm eine transkriptionelle Aktivitität zeigten, wurde in dieser Arbeit besonders verfolgt.

Wie im ersten Unterkapitel der Diskussion bereits aufgegriffen, wurde etwa ein Zehntel der zuordbaren Sequenzdaten in Regionen des Genoms beobachtet, welche in den hier genutzten, gängigen Transkript-Datenbanken keine transkriptionelle Aktivität zeigten. Die meisten dieser Datenbanken beruhen dabei auf Informationen aus umfangreichen *expressed sequence tag* (EST)-Analysen (Carninci u. a. 2005; Kawai u. a. 2001; Okazaki u. a. 2002), die mittels Sanger-Sequenzierung revers-transkribierter Transkripte erfolgten. Die direkte Sequenzierung der Transkripte hat sich mit Einführung von RNA-Seq in diesem Fall nicht grundlegend geändert. Allerdings erlaubt die neue, verbesserte Technologie die einfache und direkte Sequenzierung in deutlich höherer Tiefe, als es zuvor wirtschaftlich möglich war.

4.3.1. Algorithmus und Verifikation

Aber nicht alle Basen mit einer RNA-Seq-*coverage*, die in Bereichen des Genoms liegen, für die keine transkriptionelle Aktivität beschrieben wurde, konnten als sichere Kandidaten für tatsächlich transkriptionell aktive Bereiche angesehen werden. Häufig finden sich so z.B. *reads*, die mit nur wenigen Basen bestehende Exongrenzen erweitern und sich die Frage aufdrängte, ob nicht Sequenzierfehler und/oder eine zufällige Identität mit der Sequenz des nächsten Exons zu diesem Befund führten. Daher wurden Kriterien definiert, die den Fund eines nTAR absichern sollten. Die Länge eines nTAR sollte mindestens 50 bp betragen. Dies stellt sicher, dass die Sequenz deutlich länger als ein einzelner *read* ist. Dadurch wird die Möglichkeit einer zufälligen Verlängerung eines *reads* durch eine Kombination aus Sequenzierfehlern/Sequenzidentität verhindert und sichergestellt, dass mehrere *reads*, die nachweislich nicht klonalen Ursprungs sind, die Entdeckung des nTAR belegen. Durch dieses Kriterium sind sehr kurze nTAR, deren Existenz nicht auszuschließen ist, mit der Methode nicht nachweisbar. Zukünftige Entwicklungen mit längeren *read*-Längen und besseren Algorithmen zum Nachweis von Spleißvorgängen könnten hier Abhilfe schaffen. Mit den in dieser Arbeit nutzbaren Ressourcen musste das Augenmerk zu Gunsten einer hohen Spezifität der Ergebnisse verschoben werden.

Zusätzlich wurde eine durchschnittliche Abdeckung von mindestens drei für ein nTAR gefordert. Hier wurde die aktuelle Sichtweise aus der Literatur übernommen, dass für den sicheren Nachweis der Genexpression eines Transkripts mehrere *reads* erforderlich sind. Die in dieser Arbeit untersuchten Befunde basierten so für jedes nTAR auf mehr als fünf *reads*. Bei Überschneidungen mit benachbarten Genen konnte es vorkommen, dass nur ein Teil eines *reads* die Existenz eines nTAR unterstützte. Da der hier verwandte Algorithmus die Abdeckung für Einzelnukleotide auflöst, mussten in diesem Fall entsprechend mehr *reads* vorhanden sein, um dennoch die durchschnittliche Mindestabdeckung von drei *reads* pro Base eines nTAR zu erfüllen.

Als letzten Schritt, um die beobachteten nTAR experimentell zu verifizieren, wurde ein zwischenzeitlich durch den Hersteller zur Verfügung gestelltes Protokoll für RNA-Seq genutzt. Gegenüber anderen denkbaren Methoden wie z.B. Nachweis der gefundenen nTAR mittels quantitativer PCR konnte dadurch die tatsächliche Sequenz bestätigt werden, zudem sind die Kosten weitaus geringer. Dabei unterschied sich das vom Hersteller zur Verfügung gestellte Protokoll in wesentlichen Punkten von dem in dieser Arbeit etablierten Protokoll und ist eher vergleichbar mit dem Ansatz der zuvor erwähnten Veröffentlichung von Cloonan u. a. Damit sollte neben der reinen Replikation auch erreicht werden, dass falsche Befunde durch einen systematischen Fehler im hier vorgestellten Verfahren erkannt wurden. Nur die eigentliche Sequenzierung erfolgte weitgehend identisch, wobei für das Verifikationsprotokoll eine neuere Version V4 der SOLiD-Sequenziertechnologie mit längeren *reads* und höherer Sequenzausbeute genutzt wurde. Andere Methoden wie die angesprochene qPCR hätten in der Methodik ähnliche Überschneidungen aufgewiesen, so dass sie in dieser Arbeit nicht als klarer Methodenwechsel interpretiert wurden, zumal als Ergebnis nur ein über- bzw. unterschreiten eines definierten Detektionsschwellenwertes gemessen worden wäre. Neben den wirtschaftlichen Erwägungen sind also auch methodische Überlegungen bedacht worden, die im Resultat zur Festlegung auf RNA-Seq als beste Methode der Wahl für die Verifikation der Ergebnisse führten. Zusätzlich erlaubt das genutzte Verifikationsprotokoll eine Aussage über die Orientierung der beobachteten nTAR, was als wertvolle Ressource betrachtet wurde.

4.3.2. Klassifizierung neuer Bereiche transkriptioneller Aktivität

Die große Mehrheit der in dieser Arbeit gefundenen nTAR zeigte eine nahe Lokalisation zu bereits bekannten Genen, entweder indem sie sich direkt mit den Exonen überschnitten, innerhalb der Introne oder unmittelbar vor oder hinter bekannten Genen zu liegen kamen. Dieser Befund fand sich auch in parallel zu dieser Arbeit vorgenommenen Untersuchungen und Veröffentlichungen. In der Arbeit von van Bakel u. a. wurde argumentiert, dass der im menschlichen Gehirn beobachtete Fund von transkriptioneller Aktivität in Regionen ohne Annotation mit der Nähe zu bekannten Genen assoziiert ist und ein zuvor vermutetes Hintergrundrauschen im Sinne von unspezifischer transkriptioneller Aktivität nicht zu belegen sei (van Bakel u. a. 2010). Dieser Befund wird durch die in der hier vorgelegten Arbeit beobachteten Befunde unterstützt. Zwar finden sich im Genom weitab von anderen Genen Bereiche transkriptioneller Aktivität, ohne dass diese derzeit als Transkripte annotiert sind. Mehrheitlich finden sich solche Bereiche aber in der Nähe bekannter Gene, insbesondere konnte auch keine Form einer unspezifischen Transkription beobachtet werden. Über weite Strecken des Genoms fand sich überhaupt keine Expression unter der Einschränkung, dass sehr kurze oder Transkripte mit fehlender Polyadenylierung hier nicht betrachtet wurden.

Um diesen Umstand näher zu beleuchten, wurden die gefundenen nTAR entsprechend ihrer Position für weitergehende Analysen klassifiziert. Dabei wurde die von van Bakel vorgenommene Zuordnung der nTAR um feinere Abstufungen erweitert. Nicht-annotierte, transkriptionell aktive Regionen wurden in dieser Arbeit in eine Reihe von Klassen unterteilt, die auf ihrer relativen Position gegenüber bekannten Genen beruhten. Der größte Anteil der gefundenen nTAR fand sich innerhalb bekannter Gene, entweder als Intron-überspannende Elemente, Exon vor- oder nachgelagerten Überschneidungen oder in der überwiegenden Zahl der Fälle als intragenische Elemente, also ohne direkte Überschneidung mit bekannten Exonstrukturen. Zusätzlich mit den Überschneidungen an 5‘- bzw. 3‘-Ende eines Transkripts findet sich also ein deutlicher Anteil der nTAR in Überschneidung zu bekannten Genen oder zwischen deren Exonen (62%), wie von van Bakel u. a. beschrieben wurde. Von den verbliebenen nTAR findet sich wiederum der Großteil in Nachbarschaft zu bekannten Genen (24,8 %), also innerhalb von 10.000 bp um ein bekanntes Gen herum. Der Anteil der restlichen nTAR in größeren Entfernungen zu annotierten Genen belief sich auf nur 13,2 %.

Bei Betrachtung der Orientierung der genassoziierten nTAR fiel auf, dass der Großteil der gefundenen nTAR mit direkter Überschneidung weit überwiegend die gleiche Orientierung wie das assoziiertes Gen zeigte, also die meisten nTAR durchaus direkter Bestandteil der Sequenz der reifen mRNA sein könnten. Für nTAR ohne direkte Überschneidung fand sich hingegen weitaus häufiger auch eine *antisense*-Orientierung gegenüber dem assoziierten Gen. Hier kann das nTAR nicht als struktureller Bestandteil vorliegen. Über die Funktion können an dieser Stelle nur Vermutungen geäußert werden. Denkbar ist eine regulatorische Modulierung des assoziierten Gens. Ein kausaler Zusammenhang muss aber nicht zwingend nur aufgrund der örtlichen Nähe im Genom vorliegen. Die detaillierte Untersuchung der transkriptionellen Aktivität um bekannte Gene verspricht im Einzelfall neue Einblicke über neue Transkriptisoformen und ggf. auch regulatorische RNA-Elemente, die für die Funktion des Gens bedeutsam sein können.

In der Regel fand sich gegenüber dem assoziierten Gen eine verminderte transkriptionelle Aktivität. Dies wurde in dieser Arbeit u.a. darauf zurückgeführt, dass bei höheren Expressionsstärken diese Transkripte bereits in eingangs genannten, konventionellen EST-Sequenzierprojekten entdeckt worden wären. Auch wenn im Einzelfall eine hohe Expression für die biologischen Funktion keine oder eine geringe Bedeutung haben kann, ist das Expressionslevel vieler der hier gefundenen nTAR im Vergleich mit der Expression bereits annotierter Bereiche nicht unerheblich. Die niedrigsten Expressionslevel wurden dabei an Überschneidungen innerhalb von Genen erhoben. Unter der Annahme, dass es sich bei der Mehrheit um eine Variante des assoziierten Gens handelt, können diese Befunde als alternative Spleißdonoren oder –akzeptoren gedeutet werden, die nur selten durch die zelluläre Spleißmaschinerie genutzt werden. Die Expression an Überschneidungen mit 5‘- oder 3‘-Enden zeigte dem gegenüber eine erhöhte Expression. Sie könnten als alternative

Transkriptionsstartpositionen oder Polyadenylierungssignale strukturell in die mRNA des assoziierten Gens einfließen. Warum die Expression im Mittel deutlich stärker ist, kann anhand der Daten nicht abgeleitet werden. Die etwas höhere Expression der DGI gegenüber den UGI könnte ein Effekt des protokollbezogenen Fehlers der Überrepräsentation von 3‘-Enden darstellen. Für nTAR in Nähe zu bekannten Genen fanden sich ebenfalls starke Expressionsmuster. Auch hier kommen alternative Transkriptionsstartinitiatoren oder Polyadenylierungssignale in Betracht, die über Spleißvorgänge mit dem assoziierten Transkript verbunden sind. Zusätzlich fanden sich hier auch erstmals mehrere nTAR, die ohne Expression des assoziierten Gens auftraten. Hierbei könnte es sich um Elemente handeln, die direkt die Expression des assoziierten Gens verhindern bzw. beeinflussen.

4.3.3. Identifizierung der Polyadenylierungssignale

Die vorherigen Ergebnisse zeigten, dass die direkte Inkorporation vieler der gefundenen nTAR in bekannte Transkripte wahrscheinlich ist. Neben völlig neuen Transkripten existieren vielseitige Modifikationen für bekannte Transkripte. Ein weiteres Indiz für diese Annahme fand sich in der Sequenzierung der dem Polyadenylierungssignal direkt vorgelagerten Nukleotidbasen. Für diese Technologie bestand nur ein Protokoll auf Basis der FLX-Sequenziertechnologie und somit konnten im direkten Vergleich gegenüber dem SOLiD nur wenige *reads* erhoben werden. Dieser Ansatz führte zum Nachweis von 488 nTAR, die direkt ein Polyadenylierungssignal trugen.

Aufgrund der stark reduzierten Sequenziertiefe ist davon auszugehen, dass nur ein Teil der direkt polyadenylierten nTAR nachgewiesen werden konnten. Unter der Annahme, dass die gefundenen nTAR unabhängig von den bekannten Transkripten exprimiert wurden, wäre aber eine gleichmäßige Verteilung innerhalb der hier definierten Klassen zu erwarten gewesen. Wurde aber der relative Anteil der beobachteten nTAR einer Klasse mit Polyadenylierungssignal ins Verhältnis zum relativen Anteil der Gesamtzahl der nTAR dieser Klasse gesetzt, so fanden sich erhöhte Quotienten vorrangig bei den NGA- und DGI-nTAR, intermediäre Quotienten für UGN, DGN, UGI und in abgeschwächter Form für IGE. Die Klassen der nTAR innerhalb eines Gens mit einer Überschneidung zu bekannten Genen zeigten einen deutlich geringeren Quotienten.

Aufgrund der für die Aufreinigung der untersuchten RNA verwandten Methode tragen NGA-nTAR in dem hier vorgestellten Modell entweder selbst direkt eine Polyadenylierung oder es sind benachbarte nTAR gegeben, mit denen sie zusammen transkribiert werden können. Entsprechend würde unter den NGA-nTAR ein hoher Quotient an direkt polyadenylierten nTAR erwartet werden. Dies konnte in den hier vorgestellten Daten beobachtet werden. Auch mit dem 3‘-Ende bekannter Transkripte überschneidende nTAR würden im Falle einer direkten gemeinsamen Transkription mit dem assoziierten Transkript ein alternatives Polyadenylierungssignal tragen müssen, entsprechend würde auch hier ein deutlich erhöhter Quotienten erwartet werden. Auch dies wurde beobachtet.

Innerhalb bekannter Gene fanden sich für ELU, ELD und ISE nahezu keine nTAR mit direkter Polyadenylierung, was deren Inkorporation in bestehende Transkripte stützt, da in dem Fall die Polyadenylierung des assoziierten Transkripts genutzt werden kann. IGE fanden sich deutlich häufiger mit einer direkten Polyadenylierung. Hierbei könnte es sich zwar um Exone handeln, die ein frühes Polyadenylierungssignal setzen und zu einem verkürzten Transkript führen. Bei Betrachtung der Orientierung der direkt polyadenylierten IGE-nTAR lagen diese aber mehrheitlich in *antisense*-Orientierung gegenüber dem assoziierten Gen vor, so dass in vielen dieser Fälle eine gemeinsame Transkription ausgeschlossen war.

Für UGN- und DGN-nTAR fand sich jeweils ein intermediärer Quotient aus polyadenylierten zur Gesamtzahl der beobachteten nTAR, wobei die beiden Klassen auch gehäuft eine *antisense*-Orientierung zeigten. Im Fall der DGN-nTAR würde bei gemeinsamer Transkription mit dem assoziierten Gen ein höherer Wert, für die UGN-nTAR einen geringerer erwartet werden. Auch für UGI, obwohl fast ausschließlich in *sense*-Orientierung gelagert, ist der relative Anteil der polyadenylierten nTAR dieser Klasse höher als der relative Anteil der gesamten UGI-nTAR. Basierend auf diesen Daten scheint für diese Klasse eine gemeinsame Transkription mit dem assoziierten Gen eher die Ausnahme darzustellen. Allerdings basiert diese Annahme auf nur zwölf UGI-nTAR mit Polyadenylierungssignal von insgesamt 430 UGI-nTAR, so dass diese Klasse durch die geringe Anzahl empfindlich für zufällige Schwankungen ist. Sollte sich dennoch bestätigen, dass UGI-nTAR häufig polyadenyliert sind, kann dies vielschichtige Ursachen haben. So kann die Transkription unabhängig von dem assoziierten Gen erfolgen. Denkbar wäre aber auch, dass aus UGI-nTAR und assoziiertem Gen in diesen Fällen ein gemeinsames Transkript hervorgeht, die Polyadenylierung aber bereits an einer frühen Signalsequenz angefügt wird. Mittels der zurzeit erhobenen Daten kann diese Frage noch nicht abschließend beantwortet werden.

4.3.4. Gewebespezifische Unterschiede der Expression

Abschließend wurde die Expression der einzelnen nTAR in den beiden analysierten Geweben separat untersucht. Eindrückliche Beispiele fanden sich für die unterschiedliche Expression von nTAR in beiden Geweben. So wurde z.B. das vorgestellte Beispiel des Transkripts *Coq2* gefunden, dessen Co-Expression mit einem gefundenen nTAR in *antisense*-Orientierung mit einer reduzierten Expression einherging. Die detaillierte Untersuchung und der Nachweis einer biologischen Funktion für solche nTAR im Einzelnen kann in zukünftigen Projekten zum besseren Verständnis der Feinregulation der Genexpression beisteuern.

Bei der Betrachtung der differentiellen Regulation der verschiedenen nTAR-Klassen zeigten überwiegend NGA-nTAR eine hohe Variabilität der Expression. Aufgrund der Korrektur der anderen Klassen für die Expression des assoziierten Gens ist es aber nicht zwingend, dass die NGA-nTAR

absolut die stärksten Schwankungen zeigten. Genauso zulässig ist die Interpretation, dass genassoziierte nTAR zwar in ihrer Expression schwanken, dabei aber vornehmlich eine ähnliche Regulation erfahren wie ihre assoziierten Gene.

Zusammengenommen sprechen die Funde bezüglich der nicht-annotierten, transkriptionell aktiven Regionen in vielen Fällen für eine Realisation als Isoformen bekannter Transkripte, basierend auf Lokalisation, Orientierung, Position von Polyadenylierungssignalen und der Expression. Dennoch fand sich auch eine erhebliche Anzahl von Funden, die z.B. aufgrund ihrer Orientierung unabhängig ihrer assoziierten Gene exprimiert wurden und vielleicht als Regulatoren der Genexpression dienen, für die zurzeit aber keine gesicherte Funktion bekannt ist.

4.4. Perspektive

In dieser Arbeit wurde das Transkriptom zweier Schlüsselgewebe des Darms untersucht. Um die Erkenntnisse weiter auszubauen, wird es in Zukunft entscheidend sein, diese grobe Unterteilung durch feinere Untergliederungen zu ersetzen und um andere Bereiche des gastrointestinalen Trakts zu ergänzen (z.B. Speiseröhre, Magen). Neben der Beobachtung des Intestinums zu einem statischen Zeitpunkt ist die Analyse der Genregulation in der Entwicklung des Darms ebenfalls von hohem Interesse für die Biologie, hieraus kann auch ein besseres Verständnis der Fehlbildungen in der Entwicklung des menschlichen Intestinums hervorgehen. Ebenfalls von Interesse für die medizinische Forschung ist die detaillierte Untersuchung der Genexpression in erkranktem Gewebe, so können mit der hier entwickelten Methode die molekularen Vorgänge in Mausmodellen z.B. der chronisch-entzündlichen Darmerkrankungen besser untersucht werden. In dieser Arbeit wurde basierend auf der experimentellen Methode explizit vermieden, die Genaktivität der intestinalen Mikroflora mitzubestimmen. Bei entsprechender Anpassung der Methode kann in Zukunft direkt die Interaktion von intestinalen Zellen und ihren Kommensalen auf Ebene der Genexpression beobachtet werden und entscheidend zu diesem neuen Forschungsfeld beitragen.

Die eigentliche Stärke des RNA-Seq liegt dabei nicht in der genauen, genomweiten Ermittlung der reinen Genexpression, sondern im enormen Potential detaillierterer Untersuchungen zur Architektur des Transkriptoms. RNA-Seq erlaubt Wissenschaftlern, die Komplexität der Genexpression auch im Hinblick auf posttranskriptionelle Modifikationen wie Spleißen oder Polyadenylierungsstatus exakt abzubilden. Eine vorhandene Genannotation wird in Zukunft zunehmend weniger erforderlich und ggf. zu einem bestimmten Grad kontraproduktiv sein, weil - wie für intestinales Gewebe in der hier vorgelegten Arbeit gezeigt - die Genannotation von selbst sehr gut charakterisierten Modellorganismen bis zur heutigen Zeit noch unvollständig ist. Erste Ansätzen gehen wie zuvor erwähnt bereits so weit, die bestehende Genannotation für RNA-Seq-Daten vollkommen zu ignorieren und aus den RNA-Seq-Daten und den Genomdaten die Genannotation für den jeweiligen

Datensatz neu zu erzeugen (Guttman u. a. 2010). Zusätzlich werden die Anforderungen an das Ausgangsmaterial zunehmend reduziert. Anpassungen der Verfahren erlauben mit guter Auflösung die Analyse des Transkriptoms einzelner Zellen (Tang u. a. 2009). Dies ist verbunden mit stetig fallenden Preisen, die innerhalb von nur zwei Jahren RNA-Seq zu einer kompetitiven Alternative gegenüber der *microarray*-Technologie machten.

Als herausfordernd erwies sich in dieser Arbeit der enorme Aufwand, RNA-Seq-Daten zu analysieren. Zum einen fehlten einheitliche Standards und Programme für selbst relativ einfache Ziele wie die Bestimmung der Genexpression. Zum anderen bedurfte es erheblicher Ressourcen an Rechenleistung und –zeit sowie der fachlichen Unterstützung von (Bio-)Informatikern, um die vorgenommenen Analysen zu realisieren. Zwar werden Programme wie das in der Arbeit genutzte *BioScope* zunehmend benutzerfreundlich und mit graphischen Benutzereingabemasken ausgestattet. Zusammen mit der fehlenden Standardisierung ist aber die Analyse im Vergleich zu den herkömmlichen *microarrays* für Forschungsgruppen ohne bioinformatischen Fokus schwierig bis nicht darstellbar.

Trotz dieser Hürden wird der Nutzen dieser Technologie vielfältig auch für die medizinische Forschung diskutiert, z.B. in der Alzheimerforschung (Sutherland, Janitz und Kril 2011). Bisher keine Diskussion konnte aber beobachtet werden, wie mit den erzeugten Daten im Zusammenhang mit humanen Proben umgegangen wird. Während *microarray*-Daten weitgehend als Phänotyp angesehen werden, muss bei RNA-Seq die Frage gestellt werden, ob nicht so weitreichende Informationen auch über den Genotypen des Patienten oder naher Verwandter besonders schützenswert sind, vergleichbar dem Schutz von Genomdaten. Sollten im Rahmen der gesellschaftlichen Debatte über die personalisierte Genomik verbindliche Standards entwickelt werden, wie mit der genetischen Informationen von Menschen umgegangen wird, empfiehlt es sich, diese auch auf RNA-Seq-Daten auszuweiten.

Bei weiterhin starkem Innovations- und Kostenwettbewerb zwischen bestehenden *next generation sequencing*-Plattformen und neuen Technologien (z.B. *ion semiconductor sequencing* von Life Technologies) werden RNA-Seq-Experimente von führenden Wissenschaftlern als zukünftiger Standard der Genexpressionsanalyse gesehen (Wang, Gerstein, und Snyder 2009). Die Umsetzung verschiedener RNA-Seq-Verfahren kann dabei im Einzelnen stark variieren, z.B. versprechen viele Hersteller für die Zukunft *single molecule sequencing* direkt der RNA (Schadt, Turner und Kasarskis 2010). Trotz offener Fragen bezüglich der bioinformatischen Analyse und dem datenschutzrechtlich-ethischen Umgang im Zusammenhang mit humanen Proben: beruhend auf den Erfahrungen dieser Arbeit ist die Etablierung des RNA-Seq als zukünftiger Standard für die qualitative und quantitative Untersuchung des Transkriptoms ein wesentlicher Schritt hin zu einem besseren Verständnis der Genregulation.

Literatur

1000 Genomes Project Consortium. 2010. „A map of human genome variation from population-scale sequencing“. *Nature* 467 (7319) (Oktober 28): 1061–1073. doi:10.1038/nature09534.

Adams, M D, J M Kelley, J D Gocayne, M Dubnick, M H Polymeropoulos, H Xiao, C R Merril, A Wu, B Olde, und R F Moreno. 1991. „Complementary DNA sequencing: expressed sequence tags and human genome project“. *Science (New York, N.Y.)* 252 (5013) (Juni 21): 1651–1656.

Agresti, Alan. 1992. „A Survey of Exact Inference for Contingency Tables“. *Statistical Science* 7 (1) (Februar 1): 131–153.

Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, und Peter Walter. 2003. *Molekularbiologie der Zelle*. 4. Aufl. Wiley-VCH Verlag GmbH & Co. KGaA.

Amaral, Paulo P, und John S Mattick. 2008. „Noncoding RNA in development“. *Mammalian Genome: Official Journal of the International Mammalian Genome Society* 19 (7-8) (August): 454–492. doi:10.1007/s00335-008-9136-7.

Autran, Daphné, Célia Baroux, Michael T Raissig, Thomas Lenormand, Michael Wittig, Stefan Grob, Andrea Steimer, u. a. 2011. „Maternal epigenetic pathways control parental contributions to Arabidopsis early embryogenesis“. *Cell* 145 (5) (Mai 27): 707–719. doi:10.1016/j.cell.2011.04.014.

van Bakel, Harm, Corey Nislow, Benjamin J Blencowe, und Timothy R Hughes. 2010. „Most ‚dark matter‘ transcripts are associated with known genes“. *PLoS Biology* 8 (5): e1000371. doi:10.1371/journal.pbio.1000371.

Bates, Michael D, Christopher R Erwin, L Philip Sanford, Dan Wiginton, Jorge A Bezerra, Lynn C Schatzman, Anil G Jegga, u. a. 2002. „Novel genes and functional relationships in the adult mouse gastrointestinal tract identified by microarray analysis“. *Gastroenterology* 122 (5) (Mai): 1467–1482.

Baumgart, Daniel C, und William J Sandborn. 2007. „Inflammatory bowel disease: clinical aspects and established and evolving therapies“. *Lancet* 369 (9573) (Mai 12): 1641–1657. doi:10.1016/S0140-6736(07)60751-X.

Benjamini, Yoav, und Yosef Hochberg. 1995. „Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing“. *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1) (Januar 1): 289–300.

Black, Douglas L. 2003. „Mechanisms of alternative pre-messenger RNA splicing“. *Annual Review of Biochemistry* 72: 291–336. doi:10.1146/annurev.biochem.72.121801.161720.

Carninci, P, T Kasukawa, S Katayama, J Gough, M C Frith, N Maeda, R Oyama, u. a. 2005. „The transcriptional landscape of the mammalian genome“. *Science (New York, N.Y.)* 309 (5740) (September 2): 1559–1563. doi:10.1126/science.1112014.

Chen, Mo, und James L. Manley. 2009. „Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches“. *Nat Rev Mol Cell Biol* 10 (11) (November): 741–754. doi:10.1038/nrm2777.

Cheung, Foo, Brian J Haas, Susanne M D Goldberg, Gregory D May, Yongli Xiao, und Christopher D Town. 2006. „Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology“. *BMC Genomics* 7: 272. doi:PMC1635983.

Chi, Kelly Rae. 2008. „The year of sequencing“. *Nat Meth* 5 (1) (Januar): 11–14. doi:10.1038/nmeth1154.

Cloonan, Nicole, Alistair R R Forrest, Gabriel Kolle, Brooke B A Gardiner, Geoffrey J Faulkner, Mellissa K Brown, Darrin F Taylor, u. a. 2008. „Stem cell transcriptome profiling via massive-scale mRNA sequencing“. *Nature Methods* 5 (7) (Juli): 613–9. doi:nmeth.1223.

Davidson, N O, und G S Shelness. 2000. „APOLIPOPROTEIN B: mRNA editing, lipoprotein assembly, and presecretory degradation“. *Annual Review of Nutrition* 20: 169–193. doi:10.1146/annurev.nutr.20.1.169.

Ehrenreich, Armin. 2006. „DNA microarray technology for the microbiologist: an overview“. *Applied Microbiology and Biotechnology* 73 (2) (November): 255–273. doi:10.1007/s00253-006-0584-2.

Elbashir, S M, J Harborth, W Lendeckel, A Yalcin, K Weber, und T Tuschl. 2001. „Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells“. *Nature* 411 (6836) (Mai 24): 494–498. doi:10.1038/35078107.

Emrich, Scott J, W Brad Barbazuk, Li Li, und Patrick S Schnable. 2007. „Gene discovery and annotation using LCM-454 transcriptome sequencing“. *Genome Research* 17 (1) (Januar): 69–73. doi:10.1101/gr.5145806.

Fagegaltier, Delphine, Anne-Laure Bougé, Bassam Berry, Emilie Poisot, Odile Sismeiro, Jean-Yves Coppée, Laurent Théodore, Olivier Voinnet, und Christophe Antoniewski. 2009. „The endogenous siRNA pathway is involved in heterochromatin formation in Drosophila“. *Proceedings of the National Academy of Sciences of the United States of America* 106 (50) (Dezember 15): 21258–21263. doi:10.1073/pnas.0809208105.

Farajollahi, Sanaz, und Stefan Maas. 2010. „Molecular diversity through RNA editing: a balancing act“. *Trends in Genetics: TIG* 26 (5) (Mai): 221–230. doi:10.1016/j.tig.2010.02.001.

Finn, Robert D, John Tate, Jaina Mistry, Penny C Coggill, Stephen John Sammut, Hans-Rudolf Hotz, Goran Ceric, u. a. 2008. „The Pfam protein families database“. *Nucleic Acids Research* 36 (Database issue) (Januar): D281–288. doi:10.1093/nar/gkm960.

Fire, A, S Xu, M K Montgomery, S A Kostas, S E Driver, und C C Mello. 1998. „Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans“. *Nature* 391 (6669) (Februar 19): 806–811. doi:10.1038/35888.

Franke, Andre, Dermot P B McGovern, Jeffrey C Barrett, Kai Wang, Graham L Radford-Smith, Tariq Ahmad, Charlie W Lees, u. a. 2010. „Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci". *Nature Genetics* (November 21). doi:10.1038/ng.717. http://www.ncbi.nlm.nih.gov/pubmed/21102463.

Fraune, Sebastian, und Thomas C G Bosch. 2007. „Long-term maintenance of species-specific bacterial microbiota in the basal metazoan Hydra". *Proceedings of the National Academy of Sciences of the United States of America* 104 (32) (August 7): 13146–13151. doi:10.1073/pnas.0703375104.

Frommer, M, L E McDonald, D S Millar, C M Collis, F Watt, G W Grigg, P L Molloy, und C L Paul. 1992. „A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands". *Proceedings of the National Academy of Sciences of the United States of America* 89 (5) (März 1): 1827–1831.

Gan, Qiang, Iouri Chepelev, Gang Wei, Lama Tarayrah, Kairong Cui, Keji Zhao, und Xin Chen. 2010. „Dynamic regulation of alternative splicing and chromatin structure in Drosophila gonads revealed by RNA-seq". *Cell Research* (Mai 4). doi:10.1038/cr.2010.64. http://www.ncbi.nlm.nih.gov/pubmed/20440302.

Gene Ontology Consortium. 2008. „The Gene Ontology project in 2008". *Nucleic Acids Research* 36 (Database issue) (Januar): D440–444. doi:10.1093/nar/gkm883.

Gerok, Wolfgang. 2006. *Die innere Medizin: Referenzwerk für den Facharzt*. Schattauer Verlag.

Gilbert, Scott F. 2003. *Developmental Biology*. 7. A. Palgrave Macmillan.

Guhaniyogi, J, und G Brewer. 2001. „Regulation of mRNA stability in mammalian cells". *Gene* 265 (1-2) (März 7): 11–23.

Gupta, Rajnish A, Nilay Shah, Kevin C Wang, Jeewon Kim, Hugo M Horlings, David J Wong, Miao-Chih Tsai, u. a. 2010. „Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis". *Nature* 464 (7291) (April 15): 1071–1076. doi:10.1038/nature08975.

Guttman, Mitchell, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, u. a. 2010. „Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs". *Nature Biotechnology* 28 (5) (Mai): 503–510. doi:10.1038/nbt.1633.

Hampe, Jochen, Andre Franke, Philip Rosenstiel, Andreas Till, Markus Teuber, Klaus Huse, Mario Albrecht, u. a. 2007. „A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1". *Nature Genetics* 39 (2) (Februar): 207–211. doi:10.1038/ng1954.

Han, Xinwei, Xia Wu, Wen-Yu Chung, Tao Li, Anton Nekrutenko, Naomi S Altman, Gong Chen, und Hong Ma. 2009. „Transcriptome of embryonic and neonatal mouse cortex by high-throughput RNA

sequencing“. *Proceedings of the National Academy of Sciences of the United States of America* (Juli 17). doi:10.1073/pnas.0902417106. http://www.ncbi.nlm.nih.gov/pubmed/19617558.

Harder, Jürgen, und Jens-Michael Schröder. 2002. „RNase 7, a Novel Innate Immune Defense Antimicrobial Protein of Healthy Human Skin“. *Journal of Biological Chemistry* 277 (48) (November 29): 46779–46784. doi:10.1074/jbc.M207587200.

Hemmrich, Georg, Konstantin Khalturin, Anna-Marei Boehm, Malte Puchert, Friederike Anton-Erxleben, Jörg Wittlieb, Ulrich C Klostermeier, u. a. 2012. „Molecular signatures of the three stem cell lineages in Hydra and the emergence of stem cell function at the base of multicellularity“. *Molecular biology and evolution* (Mai 16). doi:10.1093/molbev/mss134. http://www.ncbi.nlm.nih.gov/pubmed/22595987.

Herold, Gerd. 2011. *Innere Medizin 2012*. Herold, Gerd.

Irizarry, Rafael A., Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs, und Terence P. Speed. 2003. „Summaries of Affymetrix GeneChip probe level data“. *Nucleic Acids Research* 31 (4) (Februar 15): e15. doi:10.1093/nar/gng015.

Jawdekar, Gauri W, und R William Henry. 2008. „Transcriptional regulation of human small nuclear RNA genes“. *Biochimica Et Biophysica Acta* 1779 (5) (Mai): 295–305. doi:10.1016/j.bbagrm.2008.04.001.

Kawai, J, A Shinagawa, K Shibata, M Yoshino, M Itoh, Y Ishii, T Arakawa, u. a. 2001. „Functional annotation of a full-length mouse cDNA collection“. *Nature* 409 (6821) (Februar 8): 685–690. doi:10.1038/35055500.

Keys, David N., Janice K. Au-Young, und Richard A. Fekete. 2010. „TaqMan® Array Cards in Pharmaceutical Research“. In *Microarray Methods for Drug Discovery*, 632:87–97. Totowa, NJ: Humana Press. http://www.springerlink.com/content/g763585753507qg6/#section=671289&page=2&locus=59.

King, M C, und A C Wilson. 1975. „Evolution at two levels in humans and chimpanzees“. *Science (New York, N.Y.)* 188 (4184) (April 11): 107–116.

Kishimoto, Y, M Hiraiwa, und J S O'Brien. 1992. „Saposins: structure, function, distribution, and molecular genetics“. *Journal of Lipid Research* 33 (9) (September): 1255–1267.

Kubista, Mikael, José Manuel Andrade, Martin Bengtsson, Amin Forootan, Jiri Jonák, Kristina Lind, Radek Sindelka, u. a. 2006. „The real-time polymerase chain reaction“. *Molecular Aspects of Medicine* 27 (2-3) (Juni): 95–125. doi:10.1016/j.mam.2005.12.007.

Kuo, Winston Patrick, Fang Liu, Jeff Trimarchi, Claudio Punzo, Michael Lombardi, Jasjit Sarang, Mark E Whipple, u. a. 2006. „A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies“. *Nat Biotech* 24 (7) (Juli): 832–840. doi:10.1038/nbt1217.

Kwan, S, V L Gerlach, und D A Brow. 2000. „Disruption of the 5' stem-loop of yeast U6 RNA induces trimethylguanosine capping of this RNA polymerase III transcript in vivo". *RNA (New York, N.Y.)* 6 (12) (Dezember): 1859–1869.

Lange, Christina, Georg Hemmrich, Ulrich C Klostermeier, Javier A López-Quintero, David J Miller, Tasja Rahn, Yvonne Weiss, Thomas C G Bosch, und Philip Rosenstiel. 2011. „Defining the origins of the NOD-like receptor system at the base of animal evolution". *Molecular Biology and Evolution* 28 (5) (Mai): 1687–1702. doi:10.1093/molbev/msq349.

Liston, Adrian, Michelle Linterman, und Li-Fan Lu. 2010. „MicroRNA in the adaptive immune system, in sickness and in health". *Journal of Clinical Immunology* 30 (3) (Mai): 339–346. doi:10.1007/s10875-010-9378-5.

Lüllmann-Rauch, Renate. 2003. *Histologie. Verstehen - Lernen - Nachschlagen*. 1. Aufl. Thieme, Stuttgart.

Madigan, Michael T., und John M. Martinko. 2008. *Brock Mikrobiologie*. 11., aktualisierte Aufl. Pearson Studium.

Maher, Christopher A, Chandan Kumar-Sinha, Xuhong Cao, Shanker Kalyana-Sundaram, Bo Han, Xiaojun Jing, Lee Sam, Terrence Barrette, Nallasivam Palanisamy, und Arul M Chinnaiyan. 2009. „Transcriptome sequencing to detect gene fusions in cancer". *Nature* 458 (7234) (März 5): 97–101. doi:10.1038/nature07638.

Mardis, Elaine R. 2008. „Next-generation DNA sequencing methods". *Annual Review of Genomics and Human Genetics* 9: 387–402. doi:10.1146/annurev.genom.9.081307.164359.

Margulies, Marcel, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, u. a. 2005. „Genome sequencing in microfabricated high-density picolitre reactors". *Nature* 437 (7057) (September 15): 376–80. doi:PMC1464427.

Mariadason, John M, Courtney Nicholas, Kaitlin E L'Italien, Min Zhuang, Helena J M Smartt, Barbara G Heerdt, Wancai Yang, u. a. 2005. „Gene expression profiling of intestinal epithelial cell maturation along the crypt-villus axis". *Gastroenterology* 128 (4) (April): 1081–1088.

Mattaj, I W, D Tollervey, und B Séraphin. 1993. „Small nuclear RNAs in messenger RNA and ribosomal RNA processing". *The FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology* 7 (1) (Januar): 47–53.

Maxam, und Gilbert. 1977. „A new method for sequencing DNA". *Proceedings of the National Academy of Sciences of the United States of America* 74 (2) (Februar): 560–4. doi:PMC392330.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, und Barbara Wold. 2008. „Mapping and quantifying mammalian transcriptomes by RNA-Seq". *Nature Methods* 5 (7) (Juli): 621–8. doi:nmeth.1226.

O A. „Amaral und Mattick - 2008 - Noncoding RNA in development.pdf“. http://www.springerlink.com/content/t25m385772v56u4w/fulltext.pdf.

Okazaki, Y, M Furuno, T Kasukawa, J Adachi, H Bono, S Kondo, I Nikaido, u. a. 2002. „Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs“. *Nature* 420 (6915) (Dezember 5): 563–573. doi:10.1038/nature01266.

Pan, Qun, Ofer Shai, Leo J Lee, Brendan J Frey, und Benjamin J Blencowe. 2008. „Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing“. *Nature Genetics* 40 (12) (Dezember): 1413–5. doi:10.1038/ng.259.

Park, Peter J. 2009. „ChIP-seq: advantages and challenges of a maturing technology“. *Nature Reviews. Genetics* 10 (10) (Oktober): 669–680. doi:10.1038/nrg2641.

Phizicky, Eric M, und Anita K Hopper. 2010. „tRNA biology charges to the front“. *Genes & Development* 24 (17) (September 1): 1832–1860. doi:10.1101/gad.1956510.

Rakoff-Nahoum, Seth, Justin Paglino, Fatima Eslami-Varzaneh, Stephen Edberg, und Ruslan Medzhitov. 2004. „Recognition of commensal microflora by toll-like receptors is required for intestinal homeostasis“. *Cell* 118 (2) (Juli 23): 229–41. doi:15260992.

Rosenkranz, Ruben, Tatiana Borodina, Hans Lehrach, und Heinz Himmelbauer. 2008. „Characterizing the mouse ES cell transcriptome with Illumina sequencing“. *Genomics* 92 (4) (Oktober): 187–194. doi:10.1016/j.ygeno.2008.05.011.

Rudolf, Matthias, und Wiltrud Kuhlisch. 2008. *Biostatistik: Eine Einführung für Biowissenschaftler*. 1. Aufl. Pearson Studium.

Ruß, Andreas. 2009. *Arzneimittel pocket 2010*. 15. Aufl. Börm Bruckmeier.

Saini, Harpreet Kaur, Sam Griffiths-Jones, und Anton James Enright. 2007. „Genomic analysis of human microRNA transcripts“. *Proceedings of the National Academy of Sciences of the United States of America* 104 (45) (November 6): 17719–17724. doi:10.1073/pnas.0703890104.

Sanger, F, G M Air, B G Barrell, N L Brown, A R Coulson, C A Fiddes, C A Hutchison, P M Slocombe, und M Smith. 1977. „Nucleotide sequence of bacteriophage phi X174 DNA“. *Nature* 265 (5596) (Februar 24): 687–95. doi:870828.

Schadt, Eric E, Steve Turner, und Andrew Kasarskis. 2010. „A window into third-generation sequencing“. *Human molecular genetics* 19 (R2) (Oktober 15): R227–240. doi:10.1093/hmg/ddq416.

Schmid, Marc W, Anja Schmidt, Ulrich C Klostermeier, Matthias Barann, Philip Rosenstiel, und Ueli Grossniklaus. 2012. „A powerful method for transcriptional profiling of specific cell types in eukaryotes: laser-assisted microdissection and RNA sequencing“. *PloS one* 7 (1): e29685. doi:10.1371/journal.pone.0029685.

Schreiber, Stefan, Philip Rosenstiel, Mario Albrecht, Jochen Hampe, und Michael Krawczak. 2005. „Genetics of Crohn disease, an archetypal inflammatory barrier disease“. *Nature Reviews. Genetics* 6 (5) (Mai): 376–388. doi:10.1038/nrg1607.

Schröder, Natalia, Aravind Sekhar, Insa Geffers, Julia Müller, Oliver Dittrich-Breiholz, Michael Kracht, Jochen Wedemeyer, und Achim Gossler. 2006. „Identification of mouse genes with highly specific expression patterns in differentiated intestinal epithelium“. *Gastroenterology* 130 (3) (März): 902–907. doi:10.1053/j.gastro.2005.12.025.

Sekirov, Inna, Shannon L Russell, L Caetano M Antunes, und B Brett Finlay. 2010. „Gut microbiota in health and disease“. *Physiological Reviews* 90 (3) (Juli): 859–904. doi:10.1152/physrev.00045.2009.

Shell, Scott A, Candice Hesse, Sidney M Morris, und Christine Milcarek. 2005. „Elevated levels of the 64-kDa cleavage stimulatory factor (CstF-64) in lipopolysaccharide-stimulated macrophages influence gene expression and induce alternative poly(A) site selection“. *The Journal of Biological Chemistry* 280 (48) (Dezember 2): 39950–39961. doi:10.1074/jbc.M508848200.

Shi, H, und P B Moore. 2000. „The crystal structure of yeast phenylalanine tRNA at 1.93 A resolution: a classic structure revisited“. *RNA (New York, N.Y.)* 6 (8) (August): 1091–1105.

Shiraki, Shinji Kondo, Shintaro Katayama, Kazunori Waki, Takeya Kasukawa, Hideya Kawaji, Rimantas Kodzius, u. a. 2003. „Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage“. *Proceedings of the National Academy of Sciences of the United States of America* 100 (26) (Dezember 23): 15776–15781. doi:10.1073/pnas.2136655100.

Sina, Christian, Alexander Arlt, Olga Gavrilova, Emilie Midtling, Marie-Luise Kruse, Susanne Sebens Müerköster, Rajiv Kumar, u. a. 2010. „Ablation of gly96/immediate early gene-X1 (gly96/iex-1) aggravates DSS-induced colitis in mice: role for gly96/iex-1 in the regulation of NF-kappaB“. *Inflammatory Bowel Diseases* 16 (2) (Februar): 320–331. doi:10.1002/ibd.21066.

Soller, M. 2006. „Pre-messenger RNA processing and its regulation: a genomic perspective“. *Cellular and Molecular Life Sciences: CMLS* 63 (7-8) (April): 796–819. doi:10.1007/s00018-005-5391-x.

Stanley, Samuel L. 2003. „Amoebiasis“. *Lancet* 361 (9362) (März 22): 1025–1034. doi:10.1016/S0140-6736(03)12830-9.

Stuart, Kenneth D, Achim Schnaufer, Nancy Lewis Ernst, und Aswini K Panigrahi. 2005. „Complex management: RNA editing in trypanosomes“. *Trends in Biochemical Sciences* 30 (2) (Februar): 97–105. doi:10.1016/j.tibs.2004.12.006.

Sultan, Marc, Marcel H Schulz, Hugues Richard, Alon Magen, Andreas Klingenhoff, Matthias Scherf, Martin Seifert, u. a. 2008. „A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome“. *Science (New York, N.Y.)* 321 (5891) (August 15): 956–60. doi:1160342.

Sutherland, Greg T, Michal Janitz, und Jillian J Kril. 2011. „Understanding the pathogenesis of Alzheimer's disease: will RNA-Seq realize the promise of transcriptomics?" *Journal of neurochemistry* 116 (6) (März): 937–946. doi:10.1111/j.1471-4159.2010.07157.x.

Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, u. a. 2009. „mRNA-Seq whole-transcriptome analysis of a single cell". *Nature Methods* (April 6). doi:10.1038/nmeth.1315. http://www.ncbi.nlm.nih.gov/pubmed/19349980.

Tavazoie, S, J D Hughes, M J Campbell, R J Cho, und G M Church. 1999. „Systematic determination of genetic network architecture". *Nature Genetics* 22 (3) (Juli): 281–285. doi:10.1038/10343.

Toledo-Arana, Alejandro, und Cristina Solano. 2010. „Deciphering the physiological blueprint of a bacterial cell: revelations of unanticipated complexity in transcriptome and proteome". *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 32 (6) (Juni): 461–467. doi:10.1002/bies.201000020.

Torres, Tatiana Teixeira, Muralidhar Metta, Birgit Ottenwälder, und Christian Schlötterer. 2008. „Gene expression profiling by massively parallel sequencing". *Genome Research* 18 (1) (Januar): 172–177. doi:10.1101/gr.6984908.

Trapnell, Cole, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, und Lior Pachter. 2010. „Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation". *Nature Biotechnology* (Mai 2). doi:10.1038/nbt.1621. http://www.ncbi.nlm.nih.gov/pubmed/20436464.

Velculescu, L Zhang, B Vogelstein, und K W Kinzler. 1995. „Serial analysis of gene expression". *Science (New York, N.Y.)* 270 (5235) (Oktober 20): 484–7. doi:7570003.

Velculescu, Lin Zhang, Wei Zhou, Jacob Vogelstein, Munira A. Basrai, Douglas E. Bassett, Phil Hieter, Bert Vogelstein, und Kenneth W. Kinzler. 1997. „Characterization of the Yeast Transcriptome". *Cell* 88 (2) (Januar 24): 243–251. doi:10.1016/S0092-8674(00)81845-0.

Wang, Mark Gerstein, und Michael Snyder. 2009. „RNA-Seq: a revolutionary tool for transcriptomics". *Nature Reviews. Genetics* 10 (1) (Januar): 57–63. doi:10.1038/nrg2484.

Wilhelm, Brian T, Samuel Marguerat, Stephen Watt, Falk Schubert, Valerie Wood, Ian Goodhead, Christopher J Penkett, Jane Rogers, und Jürg Bähler. 2008. „Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution". *Nature* 453 (7199) (Juni 26): 1239–43. doi:10.1038/nature07002.

Winkler, Hans. 1920. *Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreiche*. Jena :: G. Fischer,. http://www.biodiversitylibrary.org/item/16372.

Wostmann, B. S., J. R. Pleasants, Patricia Bealmear, und P. W. Kincade. 1970. „Serum proteins and lymphoid tissues in germ-free mice fed a chemically defined, water soluble, low molecular weight diet". *Immunology* 19 (3) (September): 443–448.

Young, Matthew D, Matthew J Wakefield, Gordon K Smyth, und Alicia Oshlack. 2010. „Gene ontology analysis for RNA-seq: accounting for selection bias“. *Genome Biology* 11 (2) (Februar 4): R14. doi:10.1186/gb-2010-11-2-r14.

Zusammenfassung

Die Schleimhaut des Darmtraktes ist charakterisiert durch komplexe metabolische und immunologische Prozesse und wird gesteuert durch hochdynamische Genexpressions-Programme. Mit der Verfügbarkeit von *next generation sequencing* und ihrer Nutzbarkeit für die Analyse von RNA-Sequenzen wurde die Genauigkeit über die globale Architektur des Transkriptoms gegenüber anderen Methoden wie *microarrays* auf eine neue Ebene befördert. Basierend auf dem 3‘ Oligo-dT *priming* von polyadenylierten Boten-RNA gefolgt von reverser Transkription und einem sogenannten *template switch* wurde eine Methode für RNA-Seq auf der Life Technologies SOLiD Plattform etabliert. Dieser Ansatz zeigte zuverlässige Informationen über das untersuchte Gewebe, z. B. in Bezug auf Genexpression und der Beobachtung von nicht-annotierten, transkriptionell aktiven Bereichen im Genom.

In dieser Arbeit wird über die Tiefencharakterisierung des polyadenylierten Transkriptoms in zwei nah verwandten, dennoch unterschiedlichen Geweben des murinen Intestinaltrakts (Dünn- und Dickdarm) berichtet. Eine gewebespezifische Architektur des Transkriptoms und die Präsenz von zuvor nicht bekannten Bereichen transkriptioneller Aktivität wurden gefunden. Im ersten Schritt wurden Signaturen von 20.541 NCBI-*RefSeq*-Transkripten im Darm identifiziert (74,1 % der annotierten Gene), davon fanden sich 16.742 in beiden untersuchten Geweben. Obwohl die Mehrheit der Sequenzen annotierten Genen zugeordnet werden konnten, fanden sich 27.543 nicht-annotierte, transkriptionell aktive Regionen im Genom im Widerspruch zur aktuellen Genannotationen von *RefSeq* oder *Ensembl*. Unter Nutzung eines zweiten, unabhängigen, strangspezifischen Protokolls konnten 20.966 dieser Befunde bestätigt werden, die Mehrheit davon in direkter Nähe zu bekannten Genen.

In der Folge wurden die Befunde bezüglich ihrer Nähe zu beschriebenen Exon-Elementen kategorisiert. Regionale Unterschiede zwischen Dünn- und Dickdarm dieser transkriptionell-aktiven, aber nicht annotierten Elemente wurden untersucht.

Die vorliegende Arbeit demonstriert die Komplexität eines typischen intestinalen mRNA-Transkriptoms von Säugetieren anhand einer strangspezifischen Auflösung bis hin zur Einzelbase. Die Analysen zeigten zum ersten Mal ein strangspezifisches Bild von nicht-annotierten, transkriptionell aktiven Bereichen in zwei Geweben und repräsentieren eine Ressource für weitere Untersuchungen transkriptioneller Prozesse, welche zur molekularen Gewebeidentität beitragen. Die mittels RNA-Seq in dieser Arbeit erhobenen Daten waren von hoher, zuvor nicht zugänglicher Qualität, so dass RNA-Seq in der Zukunft vermutlich die neue Standardmethode für die genomweite Untersuchung des Transkriptoms darstellen wird.

Summary

The intestinal mucosa is characterized by complex metabolic and immunological processes driven highly dynamic gene expression programs. With the advent of next generation sequencing and its utilization for the analysis of the RNA sequence space, the level of detail on the global architecture of the transcriptome reached a new order of magnitude compared to other methods like microarrays.

A method for RNA-Seq based on 3' oligo-dT priming of polyadenylated messenger RNA followed by reverse transcription and a template switch was established on the Life Technologies SOLiD platform. This approach showed robust information about the characterized tissue, for example in terms of gene expression and the observation of non-annotated transcriptionally active regions of the genome.

This thesis reports the ultra-deep characterization of the polyadenylated transcriptome in two closely related, yet distinct regions of the mouse intestinal tract (small intestine and colon). The tissue-specific transcriptomal architecture and the presence of novel transcriptionally active regions were assessed. In the first step, signatures of 20,541 NCBI *RefSeq* transcripts could be identified in the intestine (74.1% of annotated genes), thereof 16,742 are common in both tissues. Although the majority of reads could be linked to annotated genes, 27,543 non-annotated transcriptionally active regions not consistent with current gene annotations in *RefSeq* or *Ensembl* were identified. By use of a second independent strand-specific RNA-Seq protocol, 20,966 of these nTARs were confirmed, most of them in vicinity of known genes.

This thesis further categorized the findings by their relative adjacency to described exonic elements and investigated regional differences of novel transcribed elements in small intestine and colon.

The current study demonstrates the complexity of an archetypal mammalian intestinal mRNA transcriptome in high resolution and identifies novel transcriptionally active regions at strand-specific, single base resolution. The analysis for the first time shows a strand-specific comparative picture of non-annotated transcriptionally active regions in two tissues and represents a resource for further investigation of the transcriptional processes that contribute to molecular tissue identity. RNA-Seq generated data showed high, to date unseen quality. This suggests that RNA-Seq will be the new standard method for genome-wide analysis of the transcriptome.

Anhang

A. Genutzte Chemikalien und Reagenzien

3 M Natriumacetat-Lösung (pH 5,2) - Sigma-Aldrich Chemie GmbH, München, Deutschland

Advantage 2 PCR Kit – BD, Heidelberg, Deutschland

Affymetrix Mouse 430 2.0-Microarrays – Affymetrix UK Ltd., High Wycombe, Großbritannien

Agarosegel – Biozym Scientific GmbH, Hessisch Oldendorf, Deutschland

Agencourt AMPure XP Reagent – Beckman Coulter GmbH, Krefeld, Deutschland

Biotin-SMART-Primer - Metabion, Martinsried, Deutschland

Desoxyribonuklease I (DNase I) – Fermentas, St. Leon-Rot, Deutschland

Dithiothreitol (DTT, 20mM) - Fermentas, St. Leon-Rot, Deutschland

dNTP (je 10mM) - Fermentas, St. Leon-Rot, Deutschland

Dynabeads Streptavidin-Magnetkügelchen – Life Technologies GmbH, Darmstadt, Deutschland

E. coli DNA-Polymerase I - Fermentas, St. Leon-Rot, Deutschland

E. coli Ribonucelase H (5U/µl) - Fermentas, St. Leon-Rot, Deutschland

Ethanol – Merck, Darmstadt, Deutschland

Ethylendiamintetraessigsäure (EDTA, 500mM, pH 8) - Sigma-Aldrich Chemie GmbH, München, Deutschland

GAPDH-Primer - Beckman Coulter GmbH, Krefeld, Deutschland

GoTaq 5x Puffer – Promega, Mannheim, Deutschland

GoTaq-Polymerase – Promega, Mannheim, Deutschland

MicroPoly(A)Purist - Life Technologies GmbH, Darmstadt, Deutschland

mirVana miRNA Isolation - Life Technologies GmbH, Darmstadt, Deutschland

MMLV Reverse Transkriptase - Promega, Mannheim, Deutschland

Modifizierter Oligo-dT-Primer (mit Roche FLX B-Adaptor) - Metabion, Martinsried, Deutschland

Oligotex Direct mRNA Mini Kit - Qiagen, Hilden, Deutschland

Phenol-Chloroform-Isoamylalkohol-Lösung - Life Technologies GmbH, Darmstadt, Deutschland

Purelink RNA Micro Kit - Life Technologies GmbH, Darmstadt, Deutschland

Qiashredder - Qiagen, Hilden, Deutschland

RevertAid H Minus First Strand cDNA Synthesis-Kits - Fermentas, St. Leon-Rot, Deutschland

RNA 6000 Nano Kit – Agilent Technologies GmbH, Waldbronn, Deutschland

RNase-Inhibitor - Fermentas, St. Leon-Rot, Deutschland

RNeasy Mini Kit - Qiagen, Hilden, Deutschland

SMART PCR cDNA Synthesis Kit - BD, Heidelberg, Deutschland

SOLiD Total RNA-Seq Kit - Life Technologies GmbH, Darmstadt, Deutschland

ß-Mercaptoethanol - Sigma-Aldrich Chemie GmbH, München, Deutschland

Superscript II Reverse Transkriptase - Life Technologies GmbH, Darmstadt, Deutschland

T4 DNA-Polymerase (5U/µl) - Fermentas, St. Leon-Rot, Deutschland

V2 35 bp fragment library Kit - Life Technologies GmbH, Darmstadt, Deutschland

B. Publikationen

Veröffentlichte Forschungsartikel:

Mol Biol Evol. 2011 May 23. Defining the origins of the NOD-like receptor system at the base of animal evolution. Lange C, Hemmrich G, Klostermeier UC, López-Quintero JA, Miller DJ, Rahn T, Weiss Y, Bosch TC, Rosenstiel P.

Cell. 2011 May 27. Maternal epigenetic pathways control parental contributions to Arabidopsis early embryogenesis. Autran D, Baroux C, Raissig MT, Lenormand T, Wittig M, Grob S, Steimer A, Barann M, Klostermeier UC, Leblanc O, Vielle-Calzada JP, Rosenstiel P, Grimanelli D, Grossniklaus U.

BMC Genomics. 2011 Jun 10. A tissue-specific landscape of sense/antisense transcription in the mouse intestine. Klostermeier UC, Barann M, Wittig M, Häsler R, Franke A, Gavrilova O, Kreck B, Sina C, Schilhabel MB, Schreiber S, Rosenstiel P.

Proc Natl Acad Sci U S A. 2011 Nov 29. Transcriptomic resilience to global warming in the seagrass Zostera marina, a marine foundation species. Franssen SU, Gu J, Bergmann N, Winters G, Klostermeier UC, Rosenstiel P, Bornberg-Bauer E, Reusch TB.

PLoS One. 2012 Jan 26. A powerful method for transcriptional profiling of specific cell types in eukaryotes: laser-assisted microdissection and RNA sequencing.Schmid MW, Schmidt A, Klostermeier UC, Barann M, Rosenstiel P, Grossniklaus U.

Mol Biol Evol. 2012 May16. Molecular signatures of the three stem cell lineages in Hydra and the emergence of stem cell function at the base of multicellularity. Hemmrich G, Khalturin K, Boehm AM, Puchert M, Anton-Erxleben F, Wittlieb J, Klostermeier UC, Rosenstiel P, Oberg HH, Domazet-Loso T, Sugimoto T, Niwa H, Bosch TC.

Weitere Forschungsartikel befinden sich in Vorbereitung oder sind zur Begutachtung eingereicht. Zusätzlich wurden Inhalte der Doktorarbeit in Vorträgen oder Posterpräsentationen im Rahmen von Symposien der Exzellenzcluster „Inflammation at Interfaces“ und „The Future Ocean“, des Nationalen Genomforschungsnetzes und der 1. Studententagung Schleswig-Holstein vorgestellt.

C. Erklärung

Ich erkläre, dass ich die vorliegende Dissertation eigenständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Daneben versichere ich, dass die Arbeit weder ganz oder zum Teil im Rahmen eines Dissertationsverfahrens vorgelegen hat, veröffentlicht worden ist oder zur Veröffentlichung eingereicht wurde.

Die Arbeit ist unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft entstanden.

Kiel, den 12. Juli 2012

_____

(Ulrich C. Klostermeier)

D. Lebenslauf

Name	Klostermeier, Ulrich Christopher
Geburtstag	23.02.1979
Geburtsort	Eckernförde
Familienstand	ledig

Schulausbildung

1986-1990	Besuch der Grund- und Hauptschule Osdorf
1990-1999	Besuch der Jungmannschule Eckernförde
Juli 1999	Erhalt der Hochschulreife
1999-2000	Wehrdienst, Kaserne auf der Freiheit, Schleswig

Hochschulausbildung

2000-2005	Studium der Biologie an der Christian-Albrechts-Universität zu Kiel Hauptfach: Zellbiologie Nebenfächer: Biochemie, Mikrobiologie Zusätzliche Nebenfächer: Informatik, Zoologie
2004-2005	Diplomarbeit am Biochemischen Institut der Christian-Albrechts-Universität zu Kiel unter der Betreuung von Dr. rer. nat. Andreas Ludwig. Titel der Diplomarbeit: Wechselwirkungen zwischen Interleukin-6 und Chemokinen bei der Leukozytenaktivierung.
Juli 2005	Abschluss des Studiums als Diplom-Biologe
2005-2012	Studium der Humanmedizin an der Christian-Albrechts-Universität zu Kiel
2005-2007	Studentischer Mitarbeiter in der Arbeitsgruppe von Prof. Dr. Paul Saftig, Biochemisches Institut
Seit 2007	Wissenschaftlicher Mitarbeiter in der Arbeitsgruppe Zellbiologie von Prof. Dr. Philip Rosenstiel, Instituts für Klinische Molekularbiologie, mit dem Ziel der Promotion
Mai 2012	Abschluss des Studiums der Humanmedizin

E. Danksagung

Ich bedanke mich bei Prof. Dr. Philip Rosenstiel für die Vergabe des Themas der Dissertation, vielen Freiräumen in der Ausgestaltung, immer aber eine enge Betreuung oder guter Rat in freundschaftlicher Atmosphäre. Auch großen Dank ist ihm für die Nachsicht geschuldet, mit der er meinem Studium der Humanmedizin begleitete – nur so konnte ich der Starrheit des Studiums mit viel Flexibilität im Labor begegnen.

Herrn Prof. Dr. Dr. hc. Thomas C. G. Bosch möchte ich für die Übernahme der Betreuung seitens der mathematisch-naturwissenschaftlichen Fakultät danken. Aus meiner Sicht setzt sich Prof. Bosch wie kein anderer an der CAU für eine interdisziplinäre Forschung ein, um Köpfe verschiedenster Fachrichtungen zusammenzubringen. Ich bin ihm dankbar, dass er auch meine Exkursionen in die Tiefen der medizinischen Forschung wohlwollend unterstützt hat.

Herr Prof. Dr. Stefan Schreiber hat zu Beginn der Arbeit, als ich mich mit dem Wunsch an ihn gewandt habe, trotz eines parallelen Medizinstudiums eine naturwissenschaftliche Doktorarbeit anzufertigen, keine Sekunde gezögert und den Kontakt zu Prof. Rosenstiel hergestellt. Diese Freude daran, etwas zu bewegen, lässt sich auch an der exzellenten Ausstattung und den Möglichkeiten ablesen, die Wissenschaftler am Institut für Klinische Molekularbiologie haben. Dafür möchte ich mich bedanken.

Großer Dank gilt auch den vielen Mitarbeitern am Institut für Klinische Molekularbiologie, insbesondere in den Arbeitsgruppen *Next Generation Sequencing*, Bioinformatik sowie in der Zellbiologie. Ich bin vielen bewundernswerten Menschen begegnet, die mit viel Hingabe ihre Arbeit verrichten, dabei aber immer auch Zeit für ein nettes Wort fanden.