Berichtsblatt

1. ISBN oder ISSN entfällt	2. Berichtsart Abschlussbericht

3a. Titel des Berichts

Abschlussbericht Netzwerk "Integrated genomic investigation of colorectal carcinoma"
Teilprojekt Statistics and Genetic Epidemiology

3b. Titel der Publikation

4a. Autoren des Berichts (Name, Vorname(n)) Nothnagel, Michael Krawczak, Michael	5. Abschlussdatum des Vorhabens 31.12.2011
4b. Autoren der Publikation (Name, Vorname(n))	6. Veröffentlichungsdatum 30.04.2012
	7. Form der Publikation Abschlussbericht
8. Durchführende Institution(en) (Name, Adresse) Institut für Medizinische Informatik und Statistik Universitätsklinikum Kiel Arnold-Heller-Str. 3, Haus 31 24105 Kiel	9. Ber. Nr. Durchführende Institution
	10. Förderkennzeichen *) PKC-01GS08182-3
	11a. Seitenzahl Bericht 9
	11b. Seitenzahl Publikation
13. Fördernde Institution (Name, Adresse) Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	12. Literaturangaben keine
	14. Tabellen 3
	15. Abbildungen 4

16. Zusätzliche Angaben

17. Vorgelegt bei (Titel, Ort, Datum)
BMBF

18. Kurzfassung

Das Teilprojekt hat zwei neue statistisch-methodische Ansätze zur Analysen von Daten aus dem Next-Generation Sequencing (NGS) entwickelt. Durch die wachsende Masse und Bedeutung von Daten, die durch NGS generiert werden, z.B. Transkriptom-, Methylierungs- und Sequenzdaten, ergeben sich eine Vielzahl möglicher Anwendungen. Der statistische Support hat den Erfolg anderer Teilprojekte befördert, u.a. bei der Identifizierung bisher unbekannter humaner uORFs und N-terminaler Transkriptionserweiterungen mit Hilfe Neuronaler Netze und bei der Detektion genomischer Alterationen und aberranter DNA-Methylierungsmuster. Methodische Entwicklungen und statistischer Support wurden, wo notwendig, in einer effizienten Weise implementiert, um die großen Datenmengen aus dem NGS verarbeiten zu können und die Anwendbarkeit in künftigen Projekten sicherzustellen.

19. Schlagwörter

20. Verlag	21. Preis

*) Auf das Förderkennzeichen des BMBF soll auch in der Veröffentlichung hingewiesen werden.

Document Control Sheet

1. ISBN or ISSN	2. Type of Report
3a. Report Title Final Report "Integrated genomic investigation of colorectal carcinoma" Subproject Statistics and Genetic Epidemiology	
3b. Title of Publication	
4a. Author(s) of the Report (Family Name, First Name(s)) Brosch, Mario Hampe, Jochen	5.End of Project 31.12.2011
	6. Publication Date
4b. Author(s) of the Publication (Family Name, First Name(s))	7. Form of Publication Final Report
8. Performing Organization(s) (Name, Address) Institut für Medizinische Informatik und Statistik Universitätsklinikum Kiel Arnold-Heller-Str. 3, Haus 31 24105 Kiel	9. Originator's Report No.
	10. Reference No. PKC-01GS08182-3
	11a. No. of Pages Report 9
	11b. No. of Pages Publication
13. Sponsoring Agency (Name, Address) Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	12. No. of References 7
	14. No. of Tables 3
	15. No. of Figures 4
16. Supplementary Notes	
17. Presented at (Title, Place, Date)	
18. Abstract The project has developed two new statistical-methodological approaches for the analysis of data generated by Next-Generation Sequencing (NGS). Due to the growing importance of data generated by NGS, for example transcriptome, methylation and genomic sequencing, a variety of possible applications arise. The statistical support has promoted the success of other parts of the project, including the identification of previously unknown human uORFs and N-terminal extensions transcription using neural networks and in the detection of genomic alterations and aberrant DNA methylation patterns. Methodological developments and statistical support were, when necessary, implemented in an efficient way to process the large amounts of data from NGS.	
19. Keywords	
20. Publisher	21. Price

SCHLUSSBERICHT:

VERBUND: Nationales Genomforschungsnetzwerk Integrated genomic investigations of colorectal carcinoma

Teilprojekt: Statistics and Genetic Epidemiology

Zuwendungsempfänger	Universitätsklinikum Schleswig-Holstein, Campus Kiel
Projektleiter	Prof. Dr. rer. nat. Michael Krawczak, PD Dr. rer. medic. Michael Nothnagel
Förderkennzeichen	PKC-01GS08182-3
Vorhabensbezeichnung	Statistics and Genetic Epidemiology
Laufzeit des Vorhabens	01.06.2008 bis 31.12.2011
Institution	Universitätsklinikum Schleswig-Holstein / Christian-Albrechts-Universität
Abteilung	Institut für Medizinische Informatik und Statistik
Anschrift	Arnold-Heller-Straße 3, Haus 10, 24105 Kiel Tel: ++49 (0)431 597 3200 Fax: ++49 (0)431 597 3193

1 Ziele und Struktur

1.1 Projektziele

Ziel des Teilprojekts war die statistische, genetisch-epidemiologische und bioinformatische Unterstützung der anderen Teilprojekte innerhalb des NGFN-Konsortiums. Insbesondere beinhaltete dies eine konsistente Datenhaltung, eine umfassende Qualitätskontrolle der Daten, die innerhalb des NGFN-Konsortiums erzeugt wurden, sowie die Anwendung, Anpassung und Entwicklung statistischer Methoden für deren Analyse. Ein weiteres Ziel war die effiziente Implementation rechenintensiver Analysen.

1.2 Projektvoraussetzungen

Das Institut der Antragsteller war mit einer breiten, langjährigen Erfahrung in der Statistischen Genetik, Genetischen Epidemiologie und Bioinformatik ausgewiesen. Häufige Interaktionen mit klinischen und biologischen Kooperationspartnern in der Vergangenheit waren eine exzellente Voraussetzung für das frühzeitige Einbringen statistischer Expertise in das NGFN-Konsortium.

1.3 Planung und Ablauf des Vorhabens

Das Projekt wurde als Verbundprojekt unter Koordination durch Prof. Kari Hemminki aus Heidelberg durchgeführt. Am Kieler Standort wurde in vier Teilprojekten an einer integrierten somatischen und genomischen Analyse des Kolonkarzinoms gearbeitet.

1.4 Wissenschaftlicher Stand bei Projektstart / Vorarbeiten

Die Teilprojektleiter hatten bei Projektstart umfassende Erfahrungen in der Analyse, Qualitätskontrolle und Prozessierung großer genomweiter Genotyp-Datensätze. Zum Zeitpunkt des Projektstarts wurden, bedingt durch die Entwicklung neuartiger Technologien, zunehmend auch neue Arten genomweiter Daten verfügbar. Dies beinhaltete z.B. somatische Mutationsspektren, Methylierungsmuster zur epigenetischen Regulation und allelische Spektren der Transkription. Die meisten dieser Daten basieren in mindestens einem Prozessschritt auf dem Next-Generation-Sequencing, für das noch kein Standard sowohl hinsichtlich der Qualitätskontrolle als auch der Analyse existierte. Ebenso wenig waren bioinformatische Werkzeuge für die effiziente Durchführung von Analyse und Qualitätskontrolle dieser Daten oder für die Integration zusätzlicher Information verfügbar.

1.5 Zusammenarbeit mit anderen Stellen

Primäre Kooperationspartner im Rahmen des Projektes waren die Kieler Mitglieder des Verbundes.

2 Wissenschaftliche Ergebnisse

2.1 Zusammenfassung der Ergebnisse

1. Aus dem Next-Generation-Sequencing (NGS) gewonnene Rohdaten werden häufig unkritisch und nach bestenfalls eingeschränkter Qualitätskontrolle analysiert. Wir haben einen neuen statistischen Ansatz entwickelt, um den Anteil falsch-positiver Varianten in NGS-Daten zu schätzen. Durch Anwendung auf die Daten des 1000-Genome-Projekts, das als Referenz für eine Vielzahl von Anwendungen dient und dienen soll, konnten wir substantielle Fehlerraten nachweisen. Dies geschah in Kollaboration mit den AG Hampe und Siebert.
2. In Zusammenarbeit mit den AGs Hampe, Brosch und Siebert wurde ein statistisch-methodischer Ansatz zur Detektion allelischer Imbalancen in Transkriptom-Daten und zur simultanen SNP-Genotyp-Inferierung entwickelt. Eine natürliche Anwendung dieses Ansatzes ergibt sich bei der Analyse somatischer Mutationsspektren, etwa in Tumorzellen.
3. Wir haben unter Nutzung Neuronaler Netze einen Ansatz für die statistische Identifizierung bisher unbekannter humaner uORFs und N-terminaler Transkriptionserweiterungen entwickelt, in Zusammenarbeit mit den AGs Brosch und Hampe. Dieser Ansatz eignet sich durch seine effiziente Implementierung insbesondere für die genomweite Analyse.
4. Durch statistischen Support für die AGs Hampe und Siebert wurde die Analyse von CRC-Daten aus der RainDance Technologie sowie aus dem NGS zur Detektion genomischer Alterationen und aberranter DNA-Methylierungsmuster ermöglicht. Dies beinhaltete die Entwicklung eines neuartigen Hepitypingansatzes, u.a. zur Untersuchungen der Tumorevolution.
5. Wir haben zusätzlich statistischen Support für eine Anzahl weiterer Projekte im NGFN-Konsortium geleistet.

2.2 Fehlerschätzung in Next-Generation-Sequencing (NGS)-Daten

Das Next-Generation-Sequencing (NGS) ist eine vergleichsweise junge Technologie, die die Grundlage sehr vieler Daten in der genetischen Forschung bildet bzw. bilden wird. Allerdings existiert noch kein Standard sowohl hinsichtlich der Qualitätskontrolle als auch der Analyse der Daten. NGS-basierte Rohdaten werden häufig unkritisch und nach bestenfalls eingeschränkter Qualitätskontrolle analysiert. Im Ergebnis basiert eine Anzahl von z.T. hochrangig publizierten Studien auf fehlerhaften Daten. Die Schätzung der Fehlerrate solcher NGS-Daten und die Integration von Fehlern in das statistische Modell sind daher wichtige Bausteine für die korrekte Analyse.

Wir haben einen neuen statistischen Ansatz entwickelt, um den Anteil falsch-positiver Varianten in NGS-Daten zu schätzen (Nothnagel et al., 2011b). Dieser Ansatz basiert auf der Verteilung variantenspezifischer Allel-Read-Verteilungen. Er vergleicht die Verteilungen von Varianten, die in der HapMap-Datenbank vorkommen und daher als gesichert gelten können (Gold-Standard) mit solchen, die nicht in HapMap geführt werden und daher zunächst ungesichert sind (siehe Abbildung 1). Im Rahmen des 1000-Genome-Projekts wurden für zwei Personen parallel auf drei verschiedenen Sequenzierungsplattformen NGS-Daten erhoben; die somit den direkten Vergleich verschiedener Plattformen an derselben Probe erlauben.

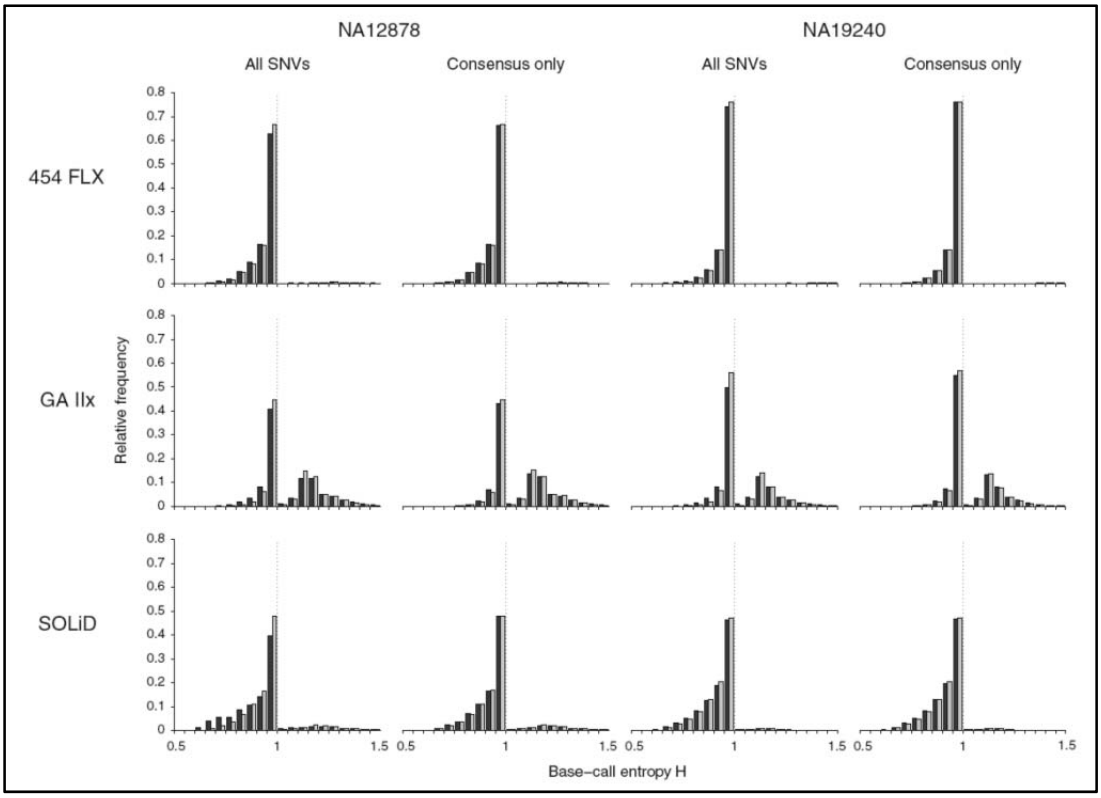

Abbildung 1: Genomweite Read-Entropieverteilung detektierter Einzel-Nukleotid-Varianten, stratifiziert nach der datengenerierender NGS-Plattform, nach Anwesenheit im HapMap-Katalog und nach Klassifizierung in Single- vs. Konsens-Call (d.h. detektiert mit allen drei Plattformen), für die zwei analysierten Individuen aus dem 1000-Genome-Projekt (aus: Nothnagel et al., 2011; http://dx.doi.org/10.1007/s00439-011-0971-3).

Durch Anwendung unseres neuartigen Ansatzes zur Fehlerschätzung konnten wir zeigen, dass alle Plattformen Daten liefern, die fehlerhaft sind; allerdings unterscheidet sich der Anteil zwischen den Plattformen erheblich (2.9-17.1%; siehe Tabelle 1). Dabei zeigte 454 FLX die besten Ergebnisse, vermutlich auch aufgrund längerer Read-Längen im Vergleich zu den anderen Plattformen. Wir konnten außerdem zeigen, dass Konsens-Calling aufgrund der Daten von zwei oder drei Plattformen die Falsch-positiv-Rate immer senkt und auf plattform-spezifische Fehler hinweist.

	NA12878			NA19240		
	All SNVs	Consensus only	*P* value	All SNVs	Consensus only	*P* value
454 FLX™	6.3 (6.1–6.5)	0.7 (0.5–3.6)	$<10^{-4}$	2.9 (2.7–3.2)	2.6 (1.2–4.7)	0.08
GA IIx™	8.4 (8.0–8.7)	3.5 (3.1–3.9)	$<10^{-4}$	11.1 (10.9–11.3)	3.9 (3.5–4.3)	$<10^{-4}$
SOLiD™	17.1 (16.9–17.4)	0.8 (0.1–2.6)	$<10^{-4}$	7.3 (6.8–7.8)	4.0 (3.1–4.8)	$<10^{-4}$

Tabelle 1: Schätzung des Anteils falsch-positiver Einzel-Nukleotid-Varianten-Detektionen für drei NGS-Plattformen, unterteilt nach Single- vs. Konsens-Call, für die zwei analysierten Individuen aus dem 1000-Genome-Projekt (aus: Nothnagel et al., 2011b; http://dx.doi.org/10.1007/s00439-011-0971-3).

Diese Ergebnisse sind nicht allein aufgrund des verwendeten neuen Ansatzes zur Fehlerschätzung von Interesse. Das 1000-Genome-Projekt ist als Referenzprojekt für eine Vielzahl von Anwendungen geplant, etwa zur Genotyp-Imputation, für populationsgenetische Analysen u.ä. Unsere Ergebnisse

zeigen, dass die nachgeordnete Qualitätskontrolle dieser Daten bisher nicht ausreichend war. Sie demonstrieren auch, dass die bioinformatische Analyse häufig noch sehr Erfahrung-getrieben ist und dass bestimmte Fehler nicht auf die NGS-Plattform, sondern auf die nachfolgende bioinformatische Pipeline zurückzuführen sind.

2.3 Genomweite Analyse allelischer Imbalancen in der Transkription

Die Realisierung genetischer Information unterliegt vielen Regulationsschritten. Ein mögliches Phänomen dieser Regulation ist das der Allelischen Imbalance (AI), bei der die beiden Transkriptformen einer autosomal-exonischen hetorozygoten Variante in einem sehr ungleichen Verhältnis beobachtet werden. Wir haben einen auf dem Maximum-Likelihood-Prinzip basierenden statistischen Ansatz zur Inferierung des nukleären Genotyps und zum statistischen Testen hinsichtlich des Vorhandenseins von AI entwickelt (Nothnagel et al., 2011a). Das NGS-Daten fehlerbehaftet sind, zeichnet sich unser Ansatz insbesondere durch die Berücksichtigung möglicher Fehler aus. Die Fehlermodellierung erlaubt die Adaptation an verschiedene NGS-Plattformen, verschiedene Spezies u.ä.

		Allele Call (X)			
		A	C	G	T
Nuclear allele (Y)	A	–	0.40210	0.48006	0.11784
	C	0.22214	–	0.25456	0.52330
	G	0.53012	0.26919	–	0.20070
	T	0.13354	0.44217	0.42429	–

Tabelle 2: Schätzung der Transitionswahrscheinlichkeiten zwischen den Basen bei Sequenzierungsfehlern für die GAIIx-Plattform (aus: Nothnagel et al., 2011a; http://dx.doi.org/10.1002/humu.21396).

Auf der Grundlage genomweiter Genotyp- und Transkript-Daten für verschiedene Zelllinien des Coriell-Instituts konnten wir so unter anderem die Wahrscheinlichkeiten für die Richtung von Sequenzierungsfehlern für die Illumina GAIIx-Plattform schätzen (siehe Tabelle 2). Diese Schätzung ist in sehr guter Übereinstimmung mit Schätzungen aus der Literatur.

	Sequencing coverage				
Allelic imbalance	5 ×	10 ×	20 ×	50 ×	100 ×
50:50	93.5	98.9	100.0	100.0	100.0
60:40	90.9	97.9	99.8	100.0	100.0
70:30	82.8	93.0	98.3	100.0	100.0
80:20	67.2	79.7	88.3	97.3	99.6
90:10	40.9	51.7	51.4	52.6	52.8
95:5	22.8	29.5	20.1	9.9	3.6

Tabelle 3: Anteil korrekt aus Transkriptionsdaten inferierter nukleärer Genotypen für verschiedene Grade allelischer Imbalance und Sequenzierungstiefe (aus: Nothnagel et al., 2011a; http://dx.doi.org/10.1002/humu.21396).

Wir haben unseren Ansatz einer gründlichen Evaluation durch Simulation unterzigen. Die Simulationen erfolgten dergestalt, dass die Eigenschaften des menschlichen Genoms bezüglich

Triplet-Häufigkeit und Triplet-spezifischen Mutationshäufigkeiten bestmöglich abgebildet wurden. Im Ergebnis zeigten die Simulationen eine sehr hohe Power unseres Ansatzes, bei geringer bis moderater AI und schon bei sehr geringer Sequenztiefe (Coverage) den zugrundeliegenden nukleären Genotyp allein aus den vorliegenden Transkriptdaten zu inferieren (siehe Tabelle 3). Bei geringer Sequenzierungstiefe ergab sich weiterhin ein substantieller Power-Gewinn durch den Einschluss fehlerhafter allelischer Reads im Vergleich zu einem naiven Test, der diese Information verwirft (siehe Tabelle 4).

	Sequencing Coverage									
	5×		10×		20×		50×		100×	
Allelic imbalance	LRT	Chi-square test	LRT	Chi-square test	LRT	Chi-square test	LRT	Chi-square test	LRT	Chi-square test
50:50	0.000 (0.000)	0.000 (0.000)	0.099 (0.000)	0.011 (0.000)	0.042 (0.000)	0.042 (0.000)	0.063 (0.000)	0.063 (0.000)	0.055 (0.000)	0.055 (0.000)
60:40	0.000 (0.000)	0.000 (0.000)	0.162 (0.000)	0.027 (0.000)	0.125 (0.000)	0.125 (0.000)	0.329 (0.000)	0.329 (0.000)	0.538 (0.001)	0.537 (0.001)
70:30	0.000 (0.000)	0.000 (0.000)	0.333 (0.000)	0.084 (0.000)	0.407 (0.000)	0.407 (0.000)	0.856 (0.007)	0.856 (0.007)	0.987 (0.156)	0.987 (0.113)
80:20	0.000 (0.000)	0.000 (0.000)	0.590 (0.000)	0.216 (0.000)	0.777 (0.000)	0.777 (0.000)	0.997 (0.167)	0.997 (0.160)	1.000 (0.907)	1.000 (0.866)
90:10	0.000 (0.000)	0.000 (0.000)	0.859 (0.000)	0.485 (0.000)	0.977 (0.000)	0.977 (0.000)	1.000 (0.882)	1.000 (0.760)	1.000 (1.000)	1.000 (1.000)
95:5	0.000 (0.000)	0.000 (0.000)	0.957 (0.000)	0.697 (0.000)	0.998 (0.000)	0.998 (0.000)	1.000 (0.998)	1.000 (0.992)	1.000 (1.000)	1.000 (1.000)

Tabelle 4: Power für die Detektion Allelischer Imbalance (AI) mithilfe eines Tests, der fehlerhafte Allel-Reads mit einbezieht (LRT), im Vergleich zu einem, der diese Information verwirft (Chi-square test), für verschiedene Sequenzierungstiefen und Grade von AI in simulierten Daten (aus: Nothnagel et al., 2011a; http://dx.doi.org/10.1002/humu.21396).

Bei Anwendung unseres Ansatzes auf einen publizierten Transkriptomdatensatz, der Informationen über Genotypen und Transkripte enthielt, zeigte sich erneut ein substantieller Powergewinn bei niedriger Sequenztiefe durch den Einschluss fehlerhafter allelischer Reads in die statistische Betrachtung (siehe Abbildung 2).

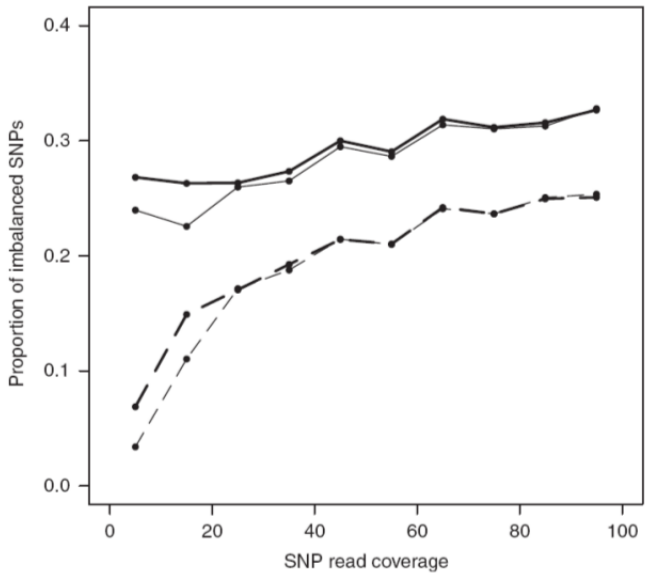

Abbildung 2: Anteil genetischer Varianten, der durch einen statistischen Test mit (schwarz) bzw. ohne (grau) Einbeziehung fehlerhafte Allel-Reads als signifikant imbalanciert detektiert wurde; bei zusätzlich vorgeschalteter Inferierung des nukleären Genotyps aus den Transkripten (gestrichelte Linie) sinkt die Power beider Verfahren aufgrund der Nichtunterscheidbarkeit von Homozygotie und extremer AI (aus: Nothnagel et al., 2011a; http://dx.doi.org/10.1002/humu.21396).

Die Bedeutung dieses Powergewinns wird deutlich, wenn man die Verteilung der Anzahl von Reads pro genetische Variante in dem realen Datensatz betrachtet. Die meisten Varianten (~75%) zeigten eine Sequenztiefe von 20 oder weniger Reads und ca. 50% hatten 10 Reads oder weniger (siehe Abbildung 3). Damit fallen die meisten Varianten genau in den Bereich, in dem ein Einschluss fehlerhafter Reads in das statistische Modell einen Powergewinn liefert. Aufgrund der beträchtlichen Kosten des NGS ist in den kommenden Jahren mit einem gehäuften Vorliegen dieses Umstands zu rechnen.

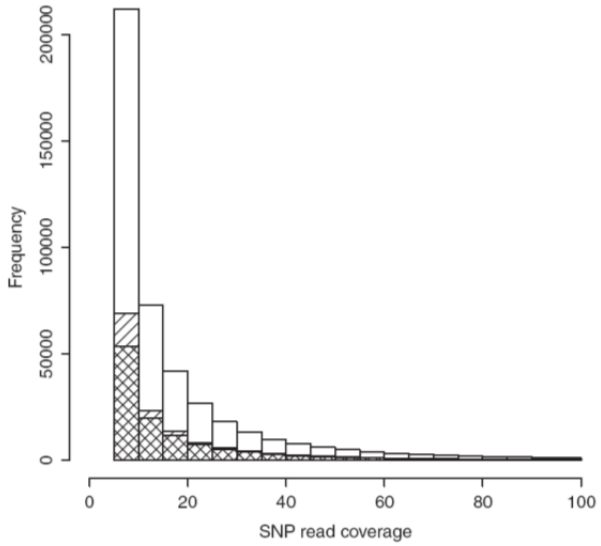

Abbildung 3: Verteilung der Anzahl von Reads pro genetischer Varianten in einem publizierten realen Transkriptomdatensatz (aus: Nothnagel et al., 2011a; http://dx.doi.org/10.1002/humu.21396).

2.4 Statistische Identifizierung bisher unbekannter humaner uORFs und N-terminaler Transkriptionserweiterungen

Ziel dieser Kooperation war die Annotation translationaler Start-Sites mithilfe einer bioinformatischen Sequenzanalyse anstelle eines direkten experimentellen Nachweises. Zu diesem Zweck wurde das so-genannte „ribosomale Footprint-Sequencing" an Puromycin-behandelte Zellen adaptiert, um die Daten mit tatsächlichen Initiationsereignisse anzureichen. Auf diesem Weg wurde eine Transkriptom-weite Karte von ribosomalen Initiationsstellen in einer humanen monozytischen Zelllinie erstellt. Das Teilprojekt hat diese Kooperation durch die Adaptation Neuronaler Netze an die konkrete Aufgabenstellung unterstützt. Insbesondere wurde das Netzwerk auf die Erkennung von Initiationssignaturen von annotierten AUGs trainiert und dann für die Prädiktion neuer, nicht annotierter Initiationsereignisse eingesetzt. Wir konnten in dieser Kollaboration zeigen, dass der gewählte statistische Ansatz gut uORFs vorhersagt, u.a. im Vergleich mit uORFs aus der Literatur (Fritsch et al., under revision). Zusätzlich wurde eine Vielzahl neuer, bisher unbekannter uORFs vorhergesagt. Diese uORFs zeigen eine Konservierung in Primaten sowohl für AUG als auch für „near-cognate"-Initiationen. Die Erstellung einer Transkriptom-weiten Karte von humanen Initiationsstellen hat das Potential, zu einem besseren Verständnis der Proteom-Diversität aufgrund alternativer Initiationen und des Einflusses der uORF-Regulierung beizutragen.

2.5 Statistischer Support für die Detektion genomischer Alterationen und aberranter DNA-Methylierungsmuster aus RainDance- und NGS- Daten

In dieser Kooperation wurden 50 Gene untersucht auf Mutationen, SNPs, LOH, Imbalancen sowie nach Bisulfit-Konvertierung auf Veränderungen im DNA-Methylierungsmuster, basierend auf dem Next-Generation-Sequencing (NGS). Hierfür leistete das Teilprojekt Support für statistische und phylogenetische Analysen. Die untersuchten Gene zeigten in vorangegangenen eigenen Bisulfit-Pyrosequenzierungs- sowie arraybasierten Studien eine Hypermethylierung in Tumoren. Im Rahmen der Analysen konnte ein neuer "hepityping"- Ansatz zur Analyse der Tumorevolution im CRC entwickelt und erstmalig angewandt werden (Abbildung 4). Das Verfahren ist inzwischen publiziert (Herrmann et al., 2011), eine umfangreichere Analyse zahlreicher Tumorproben mit diesem Verfahren scheiterte jedoch an der Nicht-Fortführung des Projektes.

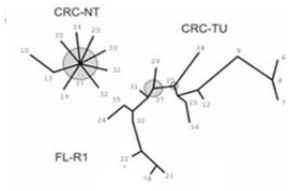

Abbildung 4: Maximum-Parsimony Bäume, erstellt mit Hepitypen mit einer minimalen Frequenz von 1% (aus: Herrmann et al., 2011; http://dx.plos.org/10.1371/journal.pone.0021332).

Des Weiteren wurde im Tumor- und Normalgewebe eines CRC-Patienten das putativ tumorrelevante Gen SMAD7 nach Anreicherung mittels Raindance-Technologie vollständig sequenziert (genomisch und Bisulfit-Sequenzierung). Dabei zeigte eine dem Risikoallel "Novel1" (Pittman et al., 2009) benachbarte Region des SMAD7-Gens in der Tumorprobe eine Hypomethylierung. Allerdings konnte in 100 weiteren Tumorgeweben keine rekurrente Hypomethylierung dieses Risiko-SNPs im Vergleich zu Normalgeweben mittels Bisulfit-Pyrosequenzierung nachgewiesen werden. Infolge dessen wurde auf die Analyse weiterer Proben verzichtet.

2.6 Statistischer Support für weitere Projekte

Weitere Projekte wurden in unterschiedlichem Maße bei der statistischen Analyse ihrer Daten unterstützt und beraten (Schafmayer et al. 2008, Buch et al. 2010, Helbig et al. 2012, Brosch et al. 2012).

3 Vorraussichtlicher Nutzen / Verwertungsplan

Die Aussichten einer wirtschaftlichen, wissenschaftlichen oder technischen Nutzung bzw. Verwertung sind sehr gut. Wir erwarten eine weitere Aufklärung der molekularen Mechanismen des kolorektalen Karzinoms in den nächsten 2-5 Jahren. Die statistischen Verfahren werden auch auf andere Fragestellungen anwendbar sein.

4 Publikationen:

Schafmayer C, Buch S, Völzke H, von Schönfels W, Egberts J, Schniewind B, Brosch M, Ruether A, Franke A, Mathiak M, Sipos B, Henopp T, Catalcali J, Hellmig S, ElSharawy A, Katalinic A, Lerch MM, John U, Fölsch UR, Fändrich F, Kalthoff H, Schreiber S, **Krawczak M**, Tepel J, Hampe J (2008). Investigation of the colorectal cancer susceptibility region on chromosome 8q24.21 in a large German patient sample. *Int J Cancer*, 124: 75-80.

Buch S, Schafmayer C, Völzke H, Seeger M, Miquel JF, Sookoian SC, Egberts JH, Arlt A, Pirola CJ, Lerch MM, John U, Franke A, von Kampen O, Brosch M, **Nothnagel M**, Kratzer W, Boehm BO, Bröring DC, Schreiber S, **Krawczak M**, Hampe J (2010). Loci from a genome-wide analysis of bilirubin levels are associated with gallstone risk and composition. *Gastroenterology*, 139(6):1942-1951.e2.

Herrmann A, Haake A, Ammerpohl O, Martin-Guerrero I, Szafranski K, Stemshorn K, **Nothnagel M**, Kotsopoulos SK, Richter J, Warner J, Olson J, Link DR, Schreiber S, **Krawczak M**, Platzer M, Nürnberg P, Siebert R, Hampe J (2011). Pipeline for large-scale microdroplet bisulfite PCR-based sequencing allows the tracking of hepitype evolution in tumors. *PLoS One*, 6(7):e21332.

Nothnagel M, **Wolf A**, Herrmann A, Szafranski K, Vater I, Brosch M, Huse K, Siebert R, Platzer M, Hampe J, **Krawczak M** (2011). Statistical inference of allelic imbalance from transcriptome data. *Hum Mutat*, 32(1):98-106.

Helbig KL, **Nothnagel M**, Hampe J, Balschun T, Nikolaus S, Schreiber S, Franke A, Nöthlings U (2012). A case-only study of gene-environment interaction between genetic susceptibility variants in NOD2 and cigarette smoking in Crohn's disease aetiology. *BMC Med Genet*, 14;13:14.

Brosch M, von Schönfels W, Ahrens M, **Nothnagel M**, **Krawczak M**, Laudes M, Sipos B, Becker T, Schreiber S, Röcken C, Schafmayer C, Hampe J (2012). SFRS10--a splicing factor gene reduced in human obesity? *Cell Metab*, 7;15(3):265-6; author reply 267-9.

Nothnagel M, Herrmann A, **Wolf A**, Schreiber S, Platzer M, Siebert R, **Krawczak M**, Hampe J (2011). Technology-specific error signatures in the 1000 Genomes Project data. *Hum Genet*, 130(4):505-16.

Fritsch C, Herrmann A, **Nothnagel M**, Szafranski K, Huse K, Schreiber S, Platzer M, **Krawczak M**, Hampe J, Brosch M. Genome-wide search for novel human uORFs and N-terminal transcript extensions using footprint analysis of initiating ribosomes. (under revision)