Aus dem Max-Delbrück-Centrum für Molekulare Medizin
in Zusammenarbeit mit der
Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Modellorganismen und Genomik in der Krebsforschung

zur Erlangung des akademischen Grades
Doctor rerum medicarum (Dr. rer. medic.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Sebastian Hörsch

aus Stuttgart

Gutachter: 1. **Prof. em. Dr. med. Jens G. Reich**

2. Priv.-Doz. Dr. rer. nat. Christine Sers

3. Priv.-Doz. Dr. rer. medic. Robert Preißner

Datum der Promotion: 3. Juni 2012

Inhaltsverzeichnis / Table of Contents

Zusammenfassung

Titel und Autoren

Modellorganismen und Genomik in der Krebsforschung

vorgelegt von Sebastian Hörsch

bestehend aus den folgenden Veröffentlichungen (in zeitlicher Reihenfolge):

RNA interference and retinoblastoma-related genes are required for repression of endogenous siRNA targets in *Caenorhabditis elegans*.
Proceedings of the National Academy of Sciences of the United States of America 105(51): 20386-20391 (2008).
Alla Grishok, Sebastian Hoersch, and Phillip A Sharp.

Periostin shows increased evolutionary plasticity in its alternatively spliced region.
BMC Evolutionary Biology 10: 30 (2010).
Sebastian Hoersch and Miguel A Andrade-Navarro.

Highly aneuploid zebrafish malignant peripheral nerve sheath tumors have genetic alterations similar to human cancers.
Proceedings of the National Academy of Sciences of the United States of America 107(39): 16940-16945 (2010).
GuangJun Zhang, Sebastian Hoersch, Adam Amsterdam, Charles A. Whittaker, Jacqueline A Lees, and Nancy Hopkins.

Kurzbeschreibung

In dieser Dissertation werden neue Wege in der Krebsforschung beleuchtet, die durch die Analyse genomischer Daten von Modellorganismen ermöglicht werden. Wie am Beispiel der drei Publikationen dieser Dissertation gezeigt, hat diese Herangehensweise ein beträchtliches und nicht erschöpfend genutztes Potential, auf ganz verschiedenen Ebenen neue Einsichten in die Krebsbiologie zu gewinnen.

Wie in einer dieser Veröffentlichungen beschrieben (1), können genomische Sequenzdaten auf zwischenartlich-vergleichender Basis zu interessanten Einsichten auch im Rahmen einer rein bioinformatische Studie führen. In dieser Studie untersuchen wir ein bestimmtes Gen, welches das extrazelluläre Matrixprotein Periostin kodiert. Periostin ist auch als Krebsmarker bekannt, über seine Funktion jedoch weiss man nur recht wenig. Im Zuge dieser Arbeit wurde das Verständnis von Periostins C-terminaler Region verbessert, indem Periostin-Sequenzen verschiedener Vertebraten verglichen wurden. Die Ergebnisse machten das Fehlen bekannter Proteindomänen oder Sequenzähnlichkeiten zu Nicht-Periostin-Proteinen verständlich, zeigten seine bemerkenswerte genomische und trankriptomische Variabilität auf und legten die Möglichkeit für eine Sekundärstruktur und einen funktionellen Mechanismus nahe.

Die anderen beiden Studien dieser Dissertation beruhen auf nicht-murinen Modellorganismen – in der Krebsforschung eher unüblich – und enthalten wichtige bioinformatische Komponenten, die mit den experimentellen, von Forscherkollegen beigebrachten Teilen integriert sind. In diesen Fällen stellten genomische Daten eine notwendige Grundlage dar, welche sowohl die unmittelbare Forschung im Modellorganismus als auch deren Projektion auf den Menschen ermöglichte.

In der einen Studie (2), am Fadenwurm *Caenorhabditis elegans* erarbeitet, werden Daten zu der Funktion bestimmter Gene (*lin-35, zfp-1, rde-4, alg-1*) vorgestellt und interpretiert, und wir beschreiben, dass diese in kooperativer Weise mit endogenen kurzen interferierenden RNAs („small interfering RNAs", siRNAs) agieren. mRNA-Expressionsdaten von *C. elegans*-Mutanten wurden in einem bioinformatischen Kontext analysiert, welcher genomweite funktionelle Interpretationen unter Berücksichtigung von relevanten genomischen Daten aus anderen Publikationen erlaubte. Diese Studie ist onkologisch von Bedeutung, weil die untersuchten Gene (mit einer Ausnahme) homolog zu bekannten Krebsgenen im Menschen sind, und weil die Ergebnisse auf das Konzept von Krebs als einem Zustand verweisen, bei welchem Keimbahn-Gene in somatischen Zellen pathologisch dereprimiert sind. Demgemäß handelt es sich bei dieser Studie um Grundlagenforschung mit einer wegen der untersuchten Gene und Mechanismen intrinsischen Bedeutung für die Krebsbiologie.

Die Ergebnisse der auf Zebrafisch-Tumoren beruhenden Studie (3) sind unmittelbar krebsspezifisch. Hier wurde ein genomweites DNA-Mikroarray konzipiert, um vergleichende genomische Hybridisierungsdaten („comparative genomic hybridization", CGH) zu erzeugen, und zwar von malignen peripheren Nervenscheidentumoren (MPNSTs) des Zebrafischs (*Danio rerio*). Die Datenanalyse zeigte, dass Zebrafisch-MPNSTs chromosomale und subchromosomale Veränderungen aufweisen, ähnlich solchen, die auch von menschlichen Tumoren (einschließlich MPNSTs) einschlägig dokumentiert sind. Manche der betroffenen Gene sind bei menschlichen Tumoren als häufig amplifiziert bekannt (z.B. *met, ccnd2a, cdk6*). Diese Erkenntnis führt zu einem möglichen Modellsystem für die Aneuploidie, einem Krebs-Merkmal, das in Mausmodellen generell nicht gut darstellbar ist. Des weiteren, und von unmittelbarer praktischer Bedeutung, stellen die

Ergebnisse eine Strategie in Aussicht, mit der Kandidaten für krebsfördernde Gene identifiziert werden könnten, indem man die Chromosomenprofile mit ihren Änderungen von Mensch und Zebrafisch vergleichend überlagert.

Zusammengenommen stellen diese drei Abhandlungen Beiträge zur onkologischen Forschung dar, die sich in kritischer Weise auf Modellorganismen – und ihre Gegenstücke *in silico*, Modellgenome – stützen. Auf dieser Basis wurden hier bioinformatische Untersuchungen möglich, die entweder eng mit experimenteller Forschung verschränkt waren oder die in eigenständiger Weise nützliche Einsichten und Hypothesen hervorbrachten, und die so wiederum Fortschritte in unserem Verständnis von Krebs im Menschen erlaubten.

Einführung

Bis zum heutigen Tag stellt Krebs als Krankheit und biologisches Phänomen eine grundsätzliche Herausforderung dar. Dies gilt auf verschiedenen Ebenen – vom fundamentalen Verständnis der Gründe bis hin zu Diagnose und Therapie. Es bleibt auch der Fall angesichts der großen Datenmengen, die im Laufe vieler Jahrzehnte der Krebsforschung angesammelt wurden.

Das Konzept von Krebs als einer Krankheit ohne Parallele (4) ist gut begründet. Keine andere Krankheit verfügt über ihre eigene Biologie in vergleichbarem Ausmaß, wie eine wegweisende Veröffentlichung vor elf Jahren darlegte (5). Seitdem haben die definierenden Kennzeichen von Krebs („the hallmarks of cancer") stetig an Profil gewonnen (6), seit genomische Technologien die Charakterisierung zugrunde liegender molekularer Vorgänge in immer feinerer Auflösung erlaubten. Zur gleichen Zeit erstarkte die Hoffnung, dass letztendlich ein funktionelles Verständnis der krebstypischen Veränderungen möglich ist und diese somit therapeutisch zugänglich werden.

Wie die Verfügbarkeit von Sequenzdaten des gesamten Genoms sowohl des Menschen als auch wichtiger Modellorganismen deutlich macht, haben sich die definierenden und beschränkenden Parameter in der Krebsforschung grundlegend verschoben. Dank genomweiter DNA-Mikroarrays oder, in jüngster Vergangenheit, massiv paralleler Sequenzierungstechniken, können Krebsproben hinsichtlich ihrer Genexpression, chromosomaler Abweichungen und ihres Mutationsstatus untersucht werden.

Modellorganismen – für die onkologische Forschung unverzichtbar, wie stetig verfeinerte Mausmodelle belegen – können in ähnlicher Weise untersucht werden und die Ergebnisse mit menschlichen Daten integriert werden. Außerdem können genomische Sequenzen vom Menschen und anderen Organismen direkt verglichen werden, um Fragen beispielsweise nach der Größe von Genfamilien oder der Konservierung regulatorischer Elemente nachzugehen. Dies ist relevant für das Problem der grundsätzlichen Kompatibilität zwischen Mensch und Modellorganismus in einer gegebenen Fragestellung.

Diese Dissertation umfasst für die Krebsbiologie relevante Ergebnisse, welche zum einen auf zwei Modellorganismen beruhen (dem Fadenwurm *Caenorhabditis elegans* und dem Zebrafisch *Danio rerio*) und zum anderen auf einer Studie, die eine breitere Auswahl von Vertebraten-Modellgenomen umfasst. In diesem Zusammenhang kann die Bedeutung genomischer Daten als ein Element, das die Studien überhaupt erst ermöglichte, gar nicht unterschätzt werden. Es ist unwahrscheinlich, dass nicht-murine Modellorganismen ohne umfassende genomische Sequenzinformationen in der onkologischen Forschung heute eine gewichtige Rolle spielen

könnten, und schon allein die Verfügbarkeit genomischer Sequenzdaten öffnet innovativer Forschung die Tür.

Zielsetzung

Die Zielsetzung dieser Dissertation umfasst unkonventionelle bioinformatische Strategien, genomische Daten für die Krebsforschung zu nutzbar zu machen. Die Umsetzung dieser Strategien ist in drei unabhängigen Publikationen (1-3) beschrieben, in welchen meine bioinformatischen Beiträge die folgenden Ziele erfüllen:

(1): Periostin ist ein Protein der extrazellulären Matrix und ist in vielen Krebsformen epithelialer Herkunft überexprimiert. Periostins wenig verstandene C-terminale Region soll mit Hilfe eines Ansatzes der vergleichenden Genomik charakterisiert werden. Besonderes Gewicht sollen dabei dem offenbaren Fehlen jeglicher funktioneller Merkmale wie Proteindomänen oder Homologien zu anderen Proteinen sowie dem wiederholt für diese Region beschriebenen alternativen Spleißen zukommen.

(2): Im Fadenwurm *Caenorhabditis elegans* wurden vier Gene mit Homologien zu menschlichen Krebsgenen durch Mutationen unabhängig voneinander ausgeschaltet. Von den vier mutierten Linien sowie einer Referenzlinie wurden dann genomweit DNA-Mikroarray-basierte Gen-Expressionsdaten gewonnen. Dieser Datensatz soll hinsichtlich differentieller Expressionsmuster analysiert werden, und die Ergebnisse sollen in einem geeigneten System umfassend funktionell zugänglich gemacht werden. Dieses System soll außerdem eine statistisch gestützte Interpretation dieser Daten im Zusammenhang mit Datensätzen aus der wissenschaftlichen Literatur ermöglichen.

(3): Für das Genom des Zebrafisch (*Danio rerio*) soll ein DNA-Mikroarray zur Erfassung vergleichender genomischer Hybridisierungsdaten ausgelegt werden. Die mit diesem Array gewonnenen Datensätze sollen dann bezüglich chromosomaler Veränderungen in Zebrafisch-Tumoren analysiert werden, und diese Veränderungen mit von menschlichen Tumoren bekannten verglichen werden. Darüber hinaus sollen, als unabhängiger Ansatz, massiv parallele Sequenzierungsdaten dieser Tumoren ausgewertet werden, um die Mikroarray-gestützten Erkenntnisse zu bestätigen.

Bioinformatische Methoden

Allgemeine Prinzipien

Einleitend darf festgestellt werden, dass die bioinformatische Arbeit für ein Projekt vor dem Schreiben von Programmcode, vor der Anwendung von Software und Algorithmen beginnt. Nicht nur müssen die zu lösenden Probleme verstanden werden, sie müssen auch gepaart werden mit verfügbaren und praktikablen Optionen für ihre Lösung. Vom machmal komplexen und oft pragmatischen Entscheidungsprozess, der zu der Wahl der letztlich benutzten Methodik führt, ist in Fachartikeln häufig nichts zu sehen, und auch im Rahmen dieser Dissertation forderten diese Arbeitsphasen „hinter den Kulissen" meinerseits erhebliche Beachtung.

Grundsätzlich und konzeptionell fußen die drei Teilprojekte dieser Dissertation auf einem gemeinsamen Kern bioinformatischer Methodik hinsichtlich der Analyse und Visualisierung biologischer Sequenz- und Mikroarray-Daten.

Suche und Vergleich von biologischen Sequenzen sind von bestimmender Bedeutung in allen drei Projekten und umfassen weit verbreitete Methoden wie BLAST und BLAT, aber auch die neueren Programme (z.B. BWA) zum besonders effizienten „Alignment" kurzer Sequenzen. Die Analyse von DNA-Mikroarray-Daten ist kennzeichnend für zwei der drei Studien (2; 3), und umfasst geeignete Strategien für die Normalisierung der Daten, deren Organisation (z.B. durch „Clustering"), und der Sequenzanalyse von Oligonukleotid-Sonden.

Projekt-spezifische Methoden

Die Details dieser beiden Themenkomplexe in Bezug auf ihre praktische Anwendung variieren mit den verschiedenen Schwerpunkten und experimentellen Ansätzen der drei Studien. Es folgt eine projektbezogene Kurzbeschreibung der relevanten bioinformatischen Methodik.

Die Periostin-Studie (1) ist rein bioinformatischer Natur, mit Sequenzanalyse und Phylogenetik als expliziten Schwerpunkten. Zu den verwendeten Methoden gehören verschiedene Versionen des Sequenz-Suchalgorithmus BLAST (z.B. TBLASTN für die Identifizierung von Periostin-Exons in wenig annotierten Genomsequenzen und PSI-BLAST, um Homologie zwischen Sequenzen nachzuweisen, welche von geringer Komplexität und Sequenzmotiv-Wiederholungen geprägt sind). „Multiple Sequence Alignments" und davon abgeleitete phylogenetische Stammbäume wurden mit ClustalW erstellt. Um das repetitive Grundmuster von der C-terminalen Region von Periostin aufzuzeigen, wurden „Dot-Matrix Plots" verwendet. Periostins genomische Sequenzen verschiedener Organismen wurden mittels VISTA verglichen. PsiPred wurde für die Vorhersage der Protein-Sekundärstruktur der C-terminalen Region verwendet, welche dann die Grundlage für eine Hypothese bezüglich der Funktion des C-Terminus bildete.

Die bioinformatische Methodik für die Publikation zu den *C. elegans*-Mutanten (2) umfasst die Prozessierung eines auf einem genomischen Mikroarray beruhenden Genexpressions-Datensatzes, gefolgt von einer Analyse der differentiell exprimierten Gene zwischen Mutantenstämmen und dem Kontrollstamm. Von zentraler Bedeutung ist die Adaption von TOPOMAP (7), einer mathematischen Projektion von Gen-Expressionsassoziationen in ein zweidimensionales Koordinatensystem, um eine schnelle, umfassende und genomweite funktionelle Klassifizierung von Gen-Gruppen zu erzielen, die sowohl von unseren Experimenten als auch von maßgeblichen publizierten Datensätzen stammten. Vorausgegangen war eine Evaluierung von „Gene Ontology"-basierten

Methoden, die jedoch für *C. elegans* aufgrund der ungleichmäßigen Repräsentation unterschiedlicher Gen-Funktionsgruppen als unzureichend befunden wurden. Die TOPOMAP-Daten wurden in einer grossen Tabelle erfasst und in einem entsprechenden Programm (MS Excel) den Projekt-Teilnehmern verfügbar gemacht, so dass die Schnittmengen zwischen beliebigen Gengruppen und TOPOMAP-Expressionsgruppen („Mounts") mühelos etabliert und zusätzliche Gengruppen aus der Literatur nach und nach hinzugefügt werden konnten. Um die statistische Signifikanz solcher Schnittmengen zu bestimmen, wurde „Fischer's Exact Test" verwendet. siRNA-Sequenzen aus Literatur-Daten wurden mittels BLASTN dem Transkriptom von *C. elegans* zugeordnet, so dass eine Evaluierung vermittels TOPOMAP möglich wurde.

In Ermangelung einer kommerziell verfügbaren Mikroarray-Plattform für vergleichende Genom-Hybridisierungsexperimente („array-based comparative genomic hybridization", aCGH) wurde im Rahmen der Studie zur Aneuploidie in Zebrafisch-Tumoren (3) eine solches Mikroarray von Grund auf konzipiert und schließlich bei der Firma Agilent in Auftrag gegeben. Aus fünf Millionen Oligonukleotid-Sequenzen, die uns die Firma Agilent zur Verfügung stellte, wurden 15000 ausgewählt, indem bestimmte Kriterien sukzessive kombiniert wurden, nämlich Kennzahlen zur Hybridisierung, durch BLAT erzielte Sequenz-Suchergebnisse (um Einzigartigkeit innerhalb des Genoms sicherzustellen) und schließlich eine Heuristik, um einen gleichmäßigen genomischen Abstand der Oligonukleotid-Sonden zu erreichen. BLAT wurde in ähnlicher Weise verwendet, um die Oligonukleotid-Sonden des Arrays auf eine neue Zebrafisch-Genomversion (Zv8) zu übertragen. Die vermittels Agilents „Feature Extraction Software" prozessierten und normalisierten Daten wurden mit dem Programm DNAcopy segmentiert und in Bezug auf chromosomale und subchromosomale Änderungen mit dem Programm STAC analysiert. Schließlich führten Vergleiche der Syntenie zwischen Mensch und Maus einerseits und Mensch und Zebrafisch andererseits zur Entwicklung der in der Publikation diskutierten Hypothese, dass aktive Krebsgene („cancer driver genes") über Schnittmengen der in den Tumoren beider Organismen veränderten Chromosomen(abschnitte) identifiziert werden können.

Ergebnisse und Diskussion

Genomische Technologie und Modellorganismen in der onkologischen Forschung

Diese Dissertation beruht auf der effektiven Kombination von Modellorganismen und genomischen Technologien.

Genomische Technologien sind darauf ausgelegt, Biomoleküle wie DNA, RNA, Proteine oder Metaboliten in ihrer Gesamtheit zu erfassen. Die Analyse von Nukleinsäuren ist dabei am weitesten verbreitet. Ursprünglich ausschließlich auf Sequenzierungsmethoden (Sanger-Sequenzierung) beruhend, rückten später DNA-Mikroarrays mit immer höherer Dichte in den Vordergrund, bis schließlich seit einigen Jahren mit der raschen Entwicklung massiv-paralleler Technologien die Sequenz-Analyse wieder dominiert.

Unter Modellorganismen versteht man im Allgemeinen eine kleine und recht scharf umrissene Gruppe von repräsentativen Organismen verschiedener phylogenetischer Gruppen. Historisch wurden diese in der Regel um bestimmter Eigenschaften willen ausgewählt, die das Studium ausgewählter, oft eng begrenzter biologischer Fragestellungen erleichterten oder erst ermöglichten. Bekannte Beispiele sind die Bäckerhefe (*Saccharomyces cerevisiae*), die Fruchtfliege (*Drosophila melanogaster*), die Ratte (*Rattus norvegicus*) und die Maus (*Mus musculus*).

Die Maus ist auch mit Abstand der wichtigste Modellorganismus in der onkologischen Forschung (Abbildung 1). Aufbauend auf einer 100-jährigen Geschichte in der biomedizinischen Forschung können murine Krebsmodelle heutzutage im Hinblick auf bestimmte Veränderungen in bestimmten Krebsgenen präzise erzeugt werden (siehe z.B. (8)). Oft ahmen sie ihre menschlichen „Vorbilder" bis in die molekularen Details nach (9; 10).

Abgesehen vom Kontext präklinischer Tierversuche, stehen der Maus in der Krebsforschung nicht viele Alternativen gegenüber. Auf einem Spektrum der Nähe zum Menschen in Bezug auf Phylogenetik, Physiologie und – vermutlich – Ätiologie von Krebserkrankungen findet sich einerseits der Haushund (*Canis lupus familiaris*), dem in der genomisch fundierten Krebsforschung möglicherweise eine zunehmend wichtige Rolle zukommen wird (11).

Auf der anderen Seite dieses Spektrums befinden sich zwei Modellorganismen, die in dieser Dissertation eine definierende Rolle spielen: der Fadenwurm oder Nematode *Caenorhabditis elegans* und der Knochenfisch oder Teleostier *Danio rerio* (2; 3). Ihre Bedeutung für die Krebsforschung mag nicht offensichtlich sein. Beide sind – besonders auf dem Gebiet der Onkologie (Abbildung 1) – relativ junge Modellorganismen, die diesen Status im wesentlichen dem Werk einzelner Wissenschaftler zu verdanken haben.

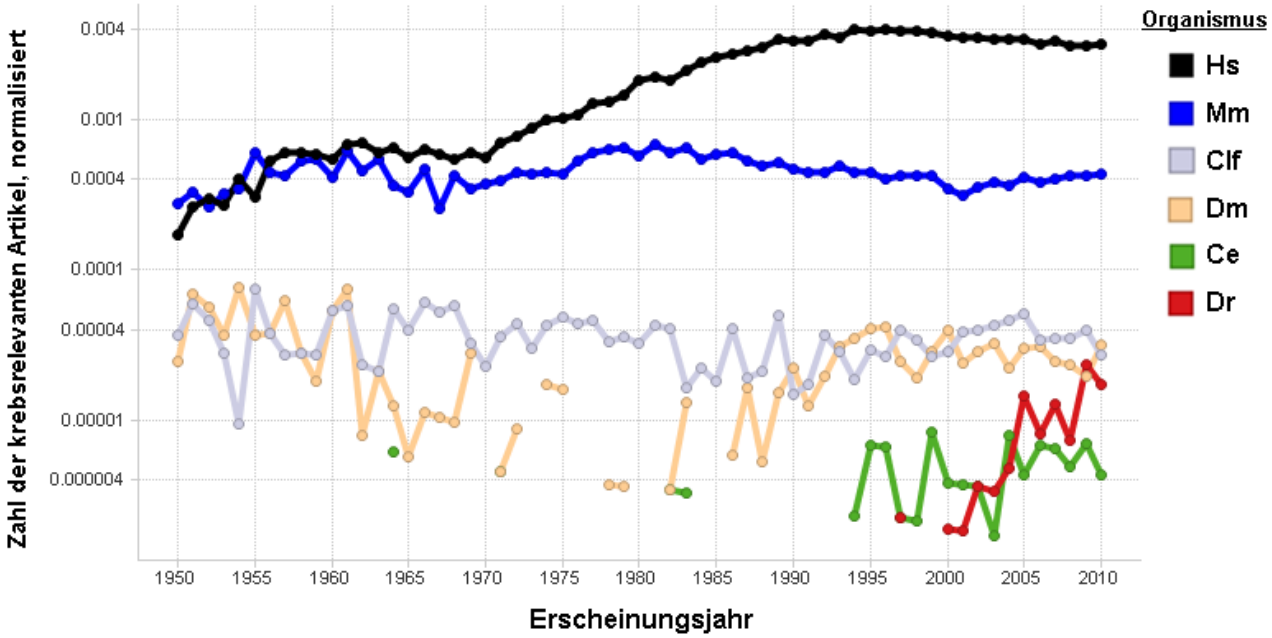


Abbildung 1: Titel der in der Medline-Datenbank zugänglichen Literatur von 1950 – 2010, die Krebs und ausgewählte Modellorganismen oder den Menschen (Hs) zusammen erwähnen, dargestellt nach „MLTrends" (12)**.** Die logarithmische Y-Achse stellt die Zahl der Artikel dar, die die Suchbegriffen enthalten, geteilt durch die Gesamtzahl der Publikationen per Jahr. Auf diese Weise entspricht eine dargestellte Zunahme einem zunehmenden Anteil an der Gesamtliteratur. Verglichen mit Maus (Mm), Hund (Clf) und Fruchtfliege (Dm) erscheinen Fadenwurm (Ce) und Zebrafisch (Dr) erst spät, und Zebrafisch zeigt gegenwärtig als einziger Organismus eine exponentielle Zunahme der relevanten Artikel.
Titel mussten mindestens einen der Suchbegriffe „cancer(s)", „tumor(s)" oder „tumour(s)" enthalten und zusätzlich wenigstens einen Bestandteil des wissenschaftlichen Artnamens oder den allgemeinen Namen („human", „mouse", „dog", „fruit fly", „nematode", „zebrafish").

Heutzutage sind die Genomsequenzen fast aller Modellorganismen in öffentlichen Datenbanken verfügbar. Dort findet sich auch eine rasch wachsende Anzahl von Genomsequenzen anderer Organismen, welche zwar nicht an sich Modellorganismen sind, die aber für die wissenschaftliche Forschung trotzdem von Interesse sind, z.B. für Fragestellungen der Evolutionsbiologie, der Medizin, der Ökologie oder der Landwirtschaft. Wie in dieser Dissertation im Detail gezeigt (1), stellen diese zusammen genommen eine mächtige Ressource für bioinformatische Studien dar.

Ergebnisse, nach Publikationen geordnet

Periostin shows increased evolutionary plasticity in its alternatively spliced region (1) (*Periostin zeigt gesteigerte evolutionäre Plastizität in seiner alternativ gespleißten Region*)

Periostin (POSTN), ein sekretiertes Protein der extrazellulären Matrix mit Bedeutung für die Zelladhäsion, ist für seine Überexpression in epithelialen Krebsformen bekannt; seine gesteigerte Expression ist mit den für Krebs bedeutenden Phänomenen der Angiogenese und der Metastasenbildung in Zusammenhang gebracht worden. Periostin im Menschen wird von 23 Exons kodiert, was zu einem aus 836 Aminosäuren bestehenden Protein führt. Eine mit rund 180 Aminosäuren recht umfangreiche C-terminale Region ist für ausgeprägtes alternatives Spleißen bekannt, andererseits jedoch bar jeglicher Signatur definierter Protein-Domänen.

Unsere Hypothese war, dass detaillierte Vergleiche mit Periostinsequenzen anderer Organismen zu weiteren Einsichten hinsichtlich der C-terminalen Region von Periostin führen könnten. Wir zeigten dann die Homologie zwischen dieser Region in Tetrapoden und einem stark konservierten, wiederholten Sequenzmotiv aus 13 Aminosäuren in der C-terminalen Region von Periostin in Knochenfischen auf. Die Analyse verfügbarer genomischer und transkriptomischer Sequenzen ermöglichte es uns, Periostin-Sequenzen – oft in ihrer Gesamtlänge – für mehrere Vertebraten zu rekonstruieren und alternatives Spleißen der C-terminalen Region von Periostin in all jenen Tetrapoden und Teleostiern (zusammen *Euteleostomi*) zu beobachten, für die es genügend Transkriptsequenzen gab.

Dieser Sequenzvergleich ermöglichte auch die Identifikation eines vierundzwanzigsten Periostin-Exons, das nur in bestimmten Tetrapoden-Linien exprimiert wird, und eines Clusters von genomisch kodierten Kopien von Exon 19, welches nur im Krallenfrosch (*Xenopus tropicalis* und *laevis*) beobachtet wurde und damit möglicherweise eine spezielle Entwicklung in den Amphibien darstellt. In beiden Fällen sind diese „zusätzlichen" Exons Teil der C-terminalen Region und werden unseren Beobachtungen zufolge alternativ gespleißt.

Des weiteren stellten wir aufgrund von Sequenzvergleichen in fünf Knochenfisch-Genomen fest, dass diese zwei Periostin-Kopien enthalten, offenbar in Folge der Genomduplikation, die der Radiation der Teleostier vorausging. Die beiden Periostin-Kopien zeigen zum Teil erhebliche Divergenz im C-Terminus. Im Gegensatz dazu scheint Periostins Paralog *TGFBI* (transforming growth factor, beta-induced) in Knochenfischen universell nur einfach vorhanden zu sein.

So präsentierte sich Periostin schließlich in seiner C-terminalen Region als außergewöhnlich variabel, sowohl im Hinblick auf die Transkription (alternatives Spleißen) als auch auf genomischer und evolutionärer Ebene (hoher Grad an Sequenzdivergenz, Unterschiede von Zahl und Länge der Exons). Unserer Interpretation nach könnte dies eine Folge von aktiver Sub- oder Neofunktionalisierung nach der Aufspaltung zwischen Periostin und seinem Paralog *TGFBI* darstellen. Interessanterweise scheint diese Dynamik innerhalb der Knochenfische zwischen den beiden Periostin-Kopien in ähnlicher Weise erneut zum Tragen zu kommen

Schließlich führten wir Sekundärstrukturvorhersagen durch, gemäß derer Periostins C-terminale Region in phylogenetisch konservierter Weise von aufeinanderfolgenden beta-Strängen gekennzeichnet ist. Dies erlaubte es uns, als erste eine biologische Funktion für diese Region vorzuschlagen: Unter den für Periostin bekannten Bindungspartnern in der extrazellulären Matrix ist unter anderem Fibronektin. Unserer Hypothese zufolge könnten Periostins beta-Stränge eine Bindung mit anderen Proteinen (z.B. Fibronektin) über einen ausgestreckte „beta zipper"-Struktur vermitteln. Dies könnte in ähnlicher Weise geschehen wie für wiederholte beta-Strang-Einheiten in bakteriellen Zellwand-Proteinen beschrieben, welche menschliches Fibronektin binden können.

Beiträge der einzelnen Autoren

Die Idee für diese Studie stammt von mir; das Konzept für ihe Durchführung stammt ebenfalls von mir und wurde durch Beiträge von Koautor Miguel A. Andrade-Navarro verfeinert. Alle Analysen wurden von mir durchgeführt, mit der Ausnahme der PSI-BLAST-Analyse, die zusammen mit M.A.A. durchgeführt wurde. Die Arbeit wurde mit M.A.A. auf kontinuierlicher Basis diskutiert. Das Manuskript mit allen Abbildungen und zusätzlichen Materialien wurde von mir verfasst und unter Berücksichtigung der Kommentare von M.A.A. in seine endgültige Form gebracht.

RNA interference and retinoblastoma-related genes are required for repression of endogenous siRNA targets in *Caenorhabditis elegans* (2)
(*RNA Interferenz und „Retinoblastoma"-assoziierte Gene sind notwendig für die Repression von endogenen siRNA-Zielgenen in Caenorhabditis elegans*)

Die Erkenntnis, dass RNA unabhängig von seiner klassischen Rolle in der zellulären Proteinsynthese auch als regulatorisches Molekül funktioniert, beflügelte in jüngster Vergangenheit ein rapide wachsendes Forschungsfeld, mit deutlichem Einfluss auch auf die Krebsforschung (13).

Der Fadenwurm *Caenorhabditis elegans* war für die Arbeit an RNA-basierter Genregulation von Beginn an von großer Bedeutung. Sein Genom kodiert die größte für einen Organismus bekannte Zahl von mit kurzen RNAs interagierenden Argonaut-Proteinen sowie tausende von Genen für endogene kurze interferierende RNAs („endogenous short interfering RNAs"; endo-siRNAs).

Wir verwendeten DNA-Mikroarrays für eine differentielle Genexpressionsstudie in *C. elegans* zwischen einem Wildtyp-Stamm und vier Mutanten-Stämmen. Diese waren entweder im Gen *lin-35*, einem Homolog des menschlichen Tumorsuppressor-Gens *RB1* („retinoblastoma"), mutiert oder in jeweils einem von drei Genen, die für RNA-Interferenz bedeutend sind (*zfp-1, rde-4* oder *alg-1*). Um die Mikroarray-basierten Daten in Bezug auf biologische Funktionen interpretieren zu können, adaptierten wir Daten für eine virtuelle Genexpressions-Landschaft („gene expression terrain map"; TOPOMAP) (7). Dies ist eine mathematische Projektion koexprimierter Gen-Muster in Gruppen (oder, nach (7), Bergen („mounts")) auf eine zweidimensionale Ebene. Die TOPOMAP-Analyse in (7) beruht auf einem großen, viele verschiedene Bedingungen umfassenden Genexpressionsdatensatz, der eine große Mehrheit aller *C. elegans*-Gene enthält. In der Tat ergaben unsere Untersuchungen, dass der Anteil in TOPOMAP vertretener *C. elegans*-Gene mit 77% sehr hoch war – über „Gene Ontology"-Annotationen fanden wir nur 46% aller Gene annotiert. Zudem ist die Repräsentation in TOPOMAP vom Bekanntheitsgrad eines Gens weitgehend unabhängig, so dass wir hiermit eine ausgewogenere Funktionalisierung des *C. elegans* Transkriptoms erreichten, als dies über „Gene Ontology"-Annotationen möglich gewesen wäre.

Wir beobachteten, dass sich die Genexpressionsmuster zweier Mutanten-Stämme – verglichen mit Wildtyp-Würmern – ähnelten. Betroffen waren hier die Mutante für das Gen *zfp-1* (ein Chromatin-Faktor und Homolog des menschlichen Gens *MLLT10* oder auch *AF10* („acute lymphoblastic leukemia-1 (ALL-1)-fused gene from chromosome 10") und die Mutante für das Gen *rde-4*, das ein Bindungsprotein für doppelsträngige RNA kodiert. Auf der Ebene funktioneller Annotationen (d.h. TOPOMAP „mounts") betrachtet, verstärkte sich dies noch und verwies klar auf einen gemeinsamen Signalweg dieser beiden Gene.

Wir verwendeten TOPOMAP zur Integration unserer eigenen Mikroarray-Daten mit relevanten publizierten Daten. Unsere Untersuchung von drei eigenständigen Datensätzen mit endo-siRNA Zielgenen ergab statistisch signifikant vergrößerte Schnittmengen mit in den Mutanten von *zfp-1*, *rde-4* und *lin-35* überexprimierten Genen, und zwar sowohl auf der Ebene einzelner Gene als auch funktioneller Gruppen. Dies legte nahe, dass es sich bei in den Mutantenstämmen überexprimierten Genen um direkte Zielgene für RNA-Interferenz handelt.

Zusammenfassend legten wir Nachweise vor für eine in großem Maßstab erfolgende Kooperation zwischen endo-siRNAs und Chromatinfaktoren zur Regulation sich überschneidender Gengruppen. Wir sehen eine bedeutende Rolle voraus für über RNA-Interferenz herbeigeführte, Chromatin-basierte negative Regulierung der Genexpression in *C. elegans.*

Beiträge der einzelnen Autoren

Idee und Gesamtkonzept für diese Arbeit stammen von Erstautorin Alla Grishok und von Professor Philip A. Sharp; das Detailkonzept wurde unter meiner Mithilfe entwickelt.

Alle Experimente und die Analyse der PCR-Experimente wurden von A.G. durchgeführt.

Die gesamte bioinformatische Arbeit wurde von mir durchgeführt, im Detail: Datenprozessierung und Normalisierung der DNA-Mikroarray-Daten (mit Unterstützung durch Charles A. Whittaker); die differentielle Expressionsanalyse der Mikroarray-Daten; die Bewertung einer Eignung von „Gene Ontology"-basierten funktionellen Annotationen für dieses Projekt; die Anpassung von TOPOMAP als Plattform für funktionelle Annotationen und die Veröffentlichung entsprechender allgemein nutzbarer Dateien; der Einsatz von Methodik, um die Überschneidung von Gengruppen innerhalb von TOPOMAP statistisch zu bewerten; die Adoption von publizierten Datensätzen, insbesondere von endo-siRNAs, aber auch von anderen, welche nur teilweise Bestandteil der Publikation sind.

Zusätzlich zu Projekt-Besprechungen zwischen den Koautoren P.A.S. und A.G. (mit meiner gelegentlichen Teilnahme) trafen sich A.G. und ich regelmäßig für detaillierten Diskussionen, bei denen – oft unter meiner Leitung – Zwischenergebnisse bewertet wurden und über Untersuchungsmethodik oder zusätzliche publizierte Datensätze entschieden wurde.

Das Manuskript wurde von A.G. aufgesetzt und mit Unterstützung von P.A.S. und mir überarbeitet. Die Abschnitte bezüglich der bioinformatischen Methodik wurden von mir verfasst. Die Abbildungen 1, S1, S2 sowie alle drei ergänzenden Datentabellen wurden von mir erstellt.

Die Ablage der Mikroarray-Genexpressionsdaten in der öffentlichen Datenbank (Gene Expression Omnibus, GSE13258) wurde von mir durchgeführt.

Highly aneuploid zebrafish malignant peripheral nerve sheath tumors have genetic alterations similar to human cancers (3)
(*Hochgradig aneuploide maligne periphere Nervenscheidentumoren im Zebrafisch zeigen menschlichem Krebs ähnliche genetische Veränderungen*)

Chromosomale Instabilität, ein wichtiges Kennzeichen in menschlichen Krebserkrankungen, führt zu Aneuploidie und zu subchromosomalen Veränderungen wie Translokationen, Inversionen, Deletionen, und Amplifzierungen, und es ist schwierig, die vielen nicht-spezifischen Abweichungen („passengers") von solchen zu unterscheiden, die aktiv krebsfördernd wirken („drivers"). Mausmodelle sind in dieser Hinsicht nicht sehr hilfreich, weil chromosomale Instabilität in diesen typischerweise eine untergeordnete Rolle spielt.

In (3) untersuchten wir mit Hilfe von extra konzipierten Mikroarrays für die vergleichende genomische Hybridisierung (CGH) und von massiv paralleler Sequenzierung die chromosomalen Veränderungen in malignen peripheren Nervenscheidentumoren (MPNSTs) des Zebrafischs. Diese Tumoren entstanden durch Mutationen in ribosomalen Proteingenen oder im Gen des Tumorsuppressors *p53* nach einer beträchtlichen Latenzzeit von 9 – 24 Monaten.

Unsere Untersuchungen ergaben, dass Zebrafisch-MPNSTs – wie viele menschliche Tumoren – hochgradig aneuploid sind, im Allgemeinen mit einer durchschnittlichen Ploidie von 3N. Für bestimmte Chromosomen zeigten sich dabei in den 36 unabhängigen MPNST-Proben deutliche Tendenzen: So waren z.B. die Chromosomen 25, 11 und 10 vorwiegend überrepräsentiert, während die Chromosomen 15, 8 und 5 vor allem unterrepräsentiert waren. Andere

Chromosomen, z.B. 16, 13 oder 3 zeigten hingegen keine eindeutige Tendenz in die eine oder andere Richtung.

Wir beobachteten auch subchromosomale Amplifikationen, am deutlichsten auf Chromosom 25. In diesem Fall fanden sich in der amplifizierten Region z.B. die Gene *slc45a3*, *ccnd2a* und *met* – all dies Gene, deren menschliche Gegenstücke schon wiederholt in chromosomalen Veränderungen im Menschen beobachtet worden sind und die demgemäß als potentiell krebsfördernd gelten.

Wie fanden auch subchromosomale Veränderungen, die nur wenige hunderttausend Basenpaare betrafen („narrow focal changes"). Allerdings stellte sich ihre Auswertung als problematisch heraus, und zwar wegen des vorläufigen Charakters der für den Zebrafisch verfügbaren Genomsequenzen. So bemerkten wir, dass viele dieser begrenzten Veränderungen in verschiedenen Genomsequenz-Versionen zu unterschiedlichen Chromosomen gehörten. Wir schlossen daraus, dass diese Art von Veränderung einer Validierung bedarf, die von heute verfügbaren Genomsequenzen unabhängig ist.

Schließlich führten wir eine vorläufige und indirekte Validierung eines Gens (*fgf6a*) in einer subchromosomal amplifizierten Region auf Chromosom 25 durch. Von Säugetieren ist bekannt, dass verschiedene Mitglieder der Familie der Fibroblasten-Wachstumsfaktoren (FGF) übergreifend an vier FGF-Rezeptoren binden können, woraufhin sie über gemeinsame MAP-Kinase Signalwege wirken. Wir zeigten, dass die Überexprimierung eines anderen Fibroblasten-Wachstumsfaktoren (*fgf8a*) zu einem beschleunigten Auftreten von MPNSTs in den p53-Mutantenstämmen führte, und stellten die Hypothese auf, das *fgf6a* in ähnlicher Weise fördernd auf MPNST wirken könnte.

Zusammenfassend fanden wir in MPNSTs des Zebrafischs chromosomale Veränderungen, die den von menschlichem Krebs her bekannten ähneln. Auf dieser Basis erscheint der Zebrafisch als ein wertvoller Modellorganismus für die Untersuchung von Aneuploidie, einem wichtigen onkologischen Merkmal, das Mausmodellen nicht in direkter Weise zugänglich ist.

Beiträge der einzelnen Autoren

Idee und Gesamtkonzept für diese Studie stammen von Erstautor GuangJun Zhang und Koautoren Adam Amsterdam, Professor Jacqueline A. Lees und Professor Nancy Hopkins; das Detailkonzept wurde mit meiner Mithilfe entwickelt. Alle experimentellen Arbeiten wurden von G.Z. und A.A. durchgeführt, die auch die Datenanalyse für die Durchflusszytometrie, das „Southern Blotting" und Chromosomezählungen durchführten.

Die gesamte bioinformatische Arbeit wurde von mir durchgeführt, mit Ausnahme der Prozessierung der massiv parallelen Sequenzierungsdaten, die von Charles A. Whittaker vorgenommen wurde. Im Detail führte ich die folgenden Untersuchungen durch: die Evaluierung verschiedener Designstrategien für das eigens für diese Studie konzipierte CGH-Mikroarray für Zebrafisch, Auslegung des tatsächlich hergestellten und verwendeten Arrays auf der Basis von etwa fünf Millionen Oligonukleotid-Sequenzen (von der Firma Agilent zur Verfügung gestellt), einschließlich der nötigen Sequenzanalyse, um Qualität, Einzigartigkeit, und Kompatibilität der Sondensequenzen mit dem verwendeten experimentellen Protokoll sicherzustellen; die erneute Analyse der Array-Sondensequenzen im Kontext einer neuen Version der Zebrafisch-Genomsequenz; Prozessierung und Normalisierung der Arraydaten (Agilent Feature Extraction); und die Analyse der Array- und der Sequenzierungs-Daten im Hinblick auf chromosomale Abweichungen vor und nach erfolgter Segmentierung (DNAcopy, STAC).

Bestimmte Details des experimentellen Designs dieser Studie wurden unter meiner Leitung erarbeitet, insbesondere die kontinuierliche Verwendung von gepaarten Kontrollen (zusätzlich zu Tumor-Kontroll-Paaren), um zu einer verbesserten Einschätzung des Grundrauschens in den Daten zu gelangen. Diese gepaarten Kontrollen werden im Artikel zwar nicht erwähnt, ihre Datensätze wurden aber – zusammen mit denen der Tumor-Kontroll-Paaren – in der öffentlichen Datenbank abgelegt. Auch erfolgte die Wahl massiv paralleler Sequenzierung als Validierungsmethode für die Arraydaten auf meine (und C.A.W.s) Initiative hin.

Ich nahm regelmäßig an Projektbesprechungen teil, die üblicherweise die Koautoren G.Z., A.A., N.H. und mich einbezogen. Während dieser brachte ich häufig meine bioinformatische Erfahrung ein, auch bezüglich der Machbarkeit und vorläufiger Ergebnisse für einen potentiellen zukünftigen Vergleich zwischen den chromosomalen Veränderungen im Menschen und Zebrafisch.

Das Manuskript wurde zusammen von G.Z., A.A., N.H. und mir geschrieben, die bioinformatische Methoden betreffenden Abschnitte wurden von mir verfasst (unter Berücksichtigung von C.A.W.s Kommentaren). Abbildungen 2 und 3A & B wurden von mir erstellt.

Schließlich wurden die von den Mikroarrays und der massiv-parallelen Sequenzierung stammenden Datensätze von mir in der öffentlichen Datenbank abgelegt (Gene Expression Omnibus, GSE23666).

Weiterführende Untersuchungen

Alle drei Publikationen dieser Dissertation repräsentieren nicht so sehr Endpunkte als viel mehr Momentaufnahmen aktiver Forschungsanstrengungen.

Die Periostin-Studie (1) liefert eine sequenzanalytische Grundlage für weitere Ergebnisse, zu denen sich ein Manuskript in Vorbereitung befindet (14) und die teilweise schon in einer Patentschrift offengelegt wurden, die von Millennium Predictive Medicine (jetzt Takeda, the Millennium Oncology Company, Cambridge, USA) eingereicht wurde (15). So ermöglichen unsere Einsichten in die transkriptionelle Variabilität und in einen möglichen funktionellen Mechanismus eine verbesserte Interpretation unserer auf klinischen Proben beruhenden Beobachtung Brustkrebs-assoziierten alternativen Spleißens in Periostins C-terminaler Region.

Während der Veröffentlichung von (2) übernahm die Erstautorin ihre eigene Forschungsgruppe an der Columbia Universität in New York, wo sie ihre Studien der Biologie von *C. elegans* mit einem Schwerpunkt RNA-Interferenz fortsetzt. Sie unterhält weiterhin eine informelle Zusammenarbeit mit mir im Zusammenhang mit verschiedenen bioinformatischen Fragestellungen. Zu diesen gehört auch die fortgesetzte Verwendung von TOPOMAP, das für diese Studie adaptiert worden war, als einer Plattform für die funktionelle Annotation und die Integration von Gengruppen aus diversen Quellen. Wir haben festgestellt, dass diese Plattform gut erweiterungsfähig ist, so dass neue Daten einfach hinzugefügt und mit älteren Daten verglichen werden können. Ein Fachartikel, das diese fortgesetzte Nutzung widerspiegelt ist bereits eingereicht und in der Begutachtungsphase (16), ein weiterer ist in Vorbereitung.

Die Untersuchungen am Zebrafisch-Krebsmodell laufen nach (3) in direkter Weise und unter meiner Mitarbeit weiter, und wichtige im Diskussionsteil dieses Artikels angesprochene Aspekte werden nun verfolgt. Statt des eigens entwickelten Mikroarrays kommt in dieser Phase ausschließlich massiv-parallele Sequenzierungstechnologie zum Einsatz, welche ja in (3) nur zu Validierungszwecken benutzt worden war – ein Zeugnis der geradezu explosiven Fortentwicklung

dieser Technologie. Etliche hundert Zebrafisch-Krebsproben – zum größten Teil wieder MPNSTs, aber auch zwei andere Krebsarten – wurden bereits sequenziert und werden zurzeit analysiert.

Des Weiteren folgen wir der Idee, dass der beträchtliche evolutionäre Abstand zwischen Mensch und Zebrafisch zur Identifikation krebsfördernder Gene benutzt werden kann, indem man aus den chromosomalen Veränderungen der menschlichen und Zebrafisch-Tumoren auf Gen-Ebene gewissermaßen eine Schnittmenge bildet. Verglichen mit Gengruppen, die den chromosomalen Veränderungen eines einzelnen Organismus entsprechen, sind die aus dieser Herangehensweise resultierenden Gengruppen deutlich kleiner und – so die Erwartung – angereichert mit Krebs-relevanten Genen. Diese könnten dann innerhalb dieser relativ kleinen Gengruppen mittels RNA-Interferenz-Screening identifiziert werden (siehe z.B. (17; 18)).

Die folgenden Abschnitte stellen einen Versuch dar, die Projekte meiner Dissertation sowohl in historischen als auch aktuellen Kontext einzuordnen, so gut dies in der gebotenen Kürze möglich ist. Diese Überlegungen werden von den für diese Dissertation so zentralen Motiven „Genomik" und „Modellorganismen" geleitet und bleiben schwerpunktmäßig der Krebsforschung und ihrem Fortschritt verhaftet.

Caenorhabditis elegans als Modellorganismus in der Krebsforschung

Der Fadenwurm *Caenorhabditis elegans* erfuhr seine systematische Einführung als Modellorganismus für genetische Studien durch Sydney Brenner in 1974 (19). Seitdem spielt er eine bedeutende Rolle in der biologischen Forschung, wie zum Beispiel bei der Entschlüsselung der Mechanismen der RNA-Interferenz (RNAi) in 1998 (20).

Krebserkrankungen an sich sind in Fadenwürmern unbekannt, wobei allerdings bestimmte Genmutationen eine stark erhöhte Proliferation in manchen Zelllinien bewirken können. Dennoch ist *C. elegans* für die onkologische Forschung von erheblicher Bedeutung. Zum einen führte die Entdeckung des biologischen Phänomens RNAi bald zu einer RNAi-basierten Labortechnik, welche auch der Krebsforschung in bedeutender Weise zugute kommt. So hat zum Beispiel die durch RNAi eröffnete Möglichkeit, die Auswirkungen von Genverlusten mit hohem Durchsatz zu eruieren („loss-of-function screening") genomweite Studien ermöglicht, in denen vergleichende Analysen der RNAi-Ergebnisse aus verschiedenartigen Zelllinien zur Entschlüsselung onkogenischer Signalwege führten (17; 18).

Zum anderen erweisen sich zentrale biologische Prozesse und ihre Gene im Allgemeinen als konserviert zwischen Fadenwürmern und Vertebraten (13). Insofern sind in *C. elegans* gewonnene Einsichten für die menschliche Krebsbiologie relevant, in der die häufige Deregulierung von Schlüsselprozessen wie Zellzyklus, Wachstumsfaktor-Signalwegen oder Apoptose gut bekannt ist.

In unserer Studie (2) sind drei der vier für die Mikroarray-Analyse mutierten Gene Homologe von bekannten Krebsgenen. So ist zum Beispiel *lin-35* ein Homolog von *RB1* (retinoblastoma 1), einem Tumorsuppressor und den Zellzykus regulierenden Gens, dessen Mutation im Menschen zu Retinoblastom im Kindesalter, Blasenkrebs und Osteosarkom führen kann. Ein zweites Gen, *zfp-1*, ist ein Homolog von *MLLT10* („myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 10"), auch *AF10* genannt. *MLLT10* kodiert einen Transkriptionsfaktor und ist am besten bekannt für seine Rolle in chromosomalen Translokationen, die über die Erzeugung von Fusionstranskripten und –proteinen zu Leukämien führen. Es ist allerdings

bemerkenswert, dass die eigentliche Funktion von MLLT10 weitgehend unbeschrieben ist. Vor diesem Hintergrund wird das aus dieser Studie und nachfolgenden Arbeiten hervorgehende Bild von ZFP-1 als einem negativen Modulator seiner Zielgene auch für das Verstehen der Funktion von MLLT10 von Bedeutung sein. Die Beobachtung, dass *zfp-1* und *MLLT10* sich im Bezug auf die Komplexität ihrer Transkriptvarianten stark ähneln, ist dabei ebenfalls interessant.

Krebszellen können überlebenswichtige Eigenschaften gewinnen, indem sie normalerweise auf Keimbahnzellen beschränkte Genexpressionsmuster aktivieren. Im Menschen liefert die Klasse der sogenannten „Cancer-Testis Antigene" ein Beispiel (21). Eine neuere Untersuchung in der Fruchtfliege *Drosophila melanogaster* (22) verweist auf eine Rolle von RB1-homologen Proteinen, die Expression von Keimbahn-Genen in somatischen Zellen auf Transkriptionsebene zu unterdrücken. Während der Nachweis für diese Funktion von RB1 in Säugetieren noch aussteht, spiegelt unsere Untersuchung diese Ergebnisse wider: Wir zeigen eine Überexpression gerade von Keimbahn-Genen in dem in *lin-35* mutierten *C. elegans*-Stamm.

All dies unterstreicht, dass biologische Erkenntnisse in *C. elegans* für die Krebsforschung sowohl auf der Ebene individueller Gene als auch auf der Ebene grundsätzlicher Prinzipien von Bedeutung sind.

Danio rerio als Modellorganismus in der Krebsforschung

1981 stellten George Streisinger und seine Kollegen den Knochenfisch *Danio rerio*, den Zebrafisch, als für genetische Screening-Verfahren geeignet vor (23). Seitdem entwickelten sich die Methoden erheblich weiter, und wichtige Einsichten, vor allem im Bereich der Entwicklungsbiologie, wurden anhand des Modellorganismus Zebrafisch gewonnen. Erst zwei Jahrzehnte später wurde der Zebrafisch als ein Modellsystem für die onkologische Forschung vorgeschlagen, obgleich dieser Vorstoß zunächst hauptsächlich auf theoretischer Ebene erfolgte (24). In den folgenden Jahren wurden etliche Zebrafisch-Krebsmodelle in der Literatur vorgestellt (z.B. für Bauchspeicheldrüsenkrebs, Melanom, Leukämie), darunter eine Studie über Mutationen in Genen für ribosomale Proteine, welche zur Krebsbildung führen können, vor allem zu malignen peripheren Nervenscheidentumoren (MPNSTs) (25). Die Zebrafisch-Studie dieser Dissertation baut in direkter Weise auf diese frühere Arbeiten auf.

Es gibt übrigens ältere, von der Entwicklung genetischer Screening-Verfahren unabhängige Arbeiten, die bereits auf eine Rolle des Zebrafischs in onkologischen Studien verweisen: Schon 1965 wurde beschrieben, dass Zebrafische in reproduzierbarer Weise Krebs entwickeln, wenn sie karzinogenen Substanzen ausgesetzt werden (26), und ähnliche Studien in anderen Fischarten gehen sogar noch weiter zurück. Möglicherweise könnten diese Arbeiten wieder aufgegriffen und die Ergebnisse mit jenen, die von genetischen Krebsmodellen und deren genomischer Charakterisierung stammen, verglichen und integriert werden.

Durch die Tatsache, dass Zebrafische an Krebs erkranken, insbesondere an Krebs mit direkter histopathologischer Entsprechung im Menschen, kommt diesem Organismus eine direktere Rolle in der Krebsforschung zu als dem Fadenwurm. Wie auch bei Mausmodellen ist eine exakte funktionelle Äquivalenz zwischen einem menschlichen Krebstyp und dem Modell-Gegenstück nur schwer nachzuweisen und deshalb bei der Interpretation Vorsicht geboten. So ist eine direkte Entsprechung zu der im Zebrafisch onkogenen Wirkung von Mutationen in den ribosomalen Protein-Genen im Menschen bisher nicht gezeigt worden.

Vor diesem Hintergrund ist unsere Studie bedeutsam, weil sie die in Zebrafisch MPNSTs häufigen chromosomalen Veränderungen dokumentiert, welche wiederum auf zum Teil wohlbekannte Krebsgene verweisen (z.B. *met, cdk6*, *slc45a3*, *ccnd2a*). Gemessen am Ziel einer möglichst genauen Krebs-Modellierung stellt dies – verglichen mit Mausmodellen – eine Verbesserung hinsichtlich des übergreifenden Phänomens der Aneuploidie dar, ohne dass dafür der Zusammenhang mit bekannten Krebsgenen aufgegeben würde.

Die Genome von *C. elegans* und *D. rerio* im Vergleich

Ein Vergleich zwischen *C. elegans* und *D. rerio* aus genomischer und bioinformatischer Perspektive ist im Kontext dieser Dissertation von Interesse, und zwar sowohl in Bezug auf die Genomstruktur als auch auf den Stand der Genomsequenzen, die ja die Basis für den Hauptanteil der analytischen Arbeit darstellen.

Die Genomsequenz von *C. elegans* wurde 1998 veröffentlicht (27), als erste Genomsequenz eines Vielzellers. Heute ist sie so gut etabliert, dass Unterschiede zwischen aufeinanderfolgenden Versionen minimal sind. Interessanterweise wurden Schätzungen der Gesamtzahl proteinkodierender Gene in *C. elegans* über die Jahre kontinuierlich nach oben korrigiert, während sie für das menschliche Genom dramatisch gefallen sind, so dass sie heute für beide Organismen bei ungefähr 20000 Genen liegen. Diese Tatsache sollte beim Studium von Krebsgenen in *C. elegans* berücksichtigt werden. Zum Beispiel wird das Argument vorgebracht, dass genetische Netzwerke in *C. elegans* im Vergleich zu Säugetieren weniger redundant seien (13), was sich an Beispielen durchaus illustrieren lässt: So entsprechen dem *C. elegans*-Gen *lin-35* im Menschen nicht nur *RB1*, sondern auch zwei Paraloge, *RBL1* (p107) und *RBL2* (p130). Ganz ähnlich ist die Situation mit *cep-1*, dem auf menschlicher Seite drei Homologe, *TP53*, *TP63* und *TP73* gegenüberstehen. Andererseits verweist die Tatsache der in beiden Linien so ähnlichen Genzahl jedoch zwingend auf Fälle, wo eine derart reduzierte Redundanz nicht auftritt oder sich gar in umgekehrter Weise manifestiert. Auch haben manche *C. elegans*-Gene, wie zum Beispiel *rde-4* (2), keine erkennbaren Homologe in Vertebraten, was die Übertragung genetischer Modelle vom Fadenwurm auf den Menschen natürlich erschwert.

Für den Zebrafisch ist die Situation eine deutlich andere. Das gesamte Genom hat in evolutionsgeschichtlich jüngerer Vergangenheit eine Duplikation durchgemacht, die sich vor der Radiation der Teleostier ereignete. Wie anhand der Gene *postn* und *tgfbi* diskutiert (1), resultiert dies für manche Gene in einer von Neo- oder Subfunktionalisierung begleiteten Beibehaltung beider Kopien (*postn*) und für andere im Verlust einer Kopie (*tgfbi*). Diesen Sachverhalt spiegelt die Nomenklatur von Zebrafisch-Genen wider, deren Kurznamen oft das Suffix ‚a' oder ‚b' tragen, um die paralogen Kopien zu bezeichnen. Leider werden Paraloge nicht immer zuverlässig erkannt und annotiert, wie das Beispiel von *postn* und seinem unbezeichneten Paralog verdeutlicht. Die biologische Bedeutung dieser Situation liegt in der resultierenden zusätzlichen Hürde, das „funktionelle Ortholog" für ein bestimmtes Säugetier-Gen zu bestimmen.

Das Projekt für die Genomsequenzierung von *D. rerio* wurde im Jahre 2001 begonnen, und eine erste Version wurde 2003 veröffentlicht (28). Die Genomsequenz ist bis heute durch die Häufigkeit von Polymorphismen gekennzeichnet, welche von der großen Zahl der für die ursprüngliche DNA-Gewinnung genutzten diploiden Embryonen herrühren. Im Unterschied zur Genomsequenz des Menschen oder des Wurms sind die Unterschiede zwischen aufeinanderfolgenden Versionen der Genomsequenz beim Zebrafisch beträchtlich, und bedingen in der Praxis oft erheblichen

bioinformatischen Aufwand (siehe oben und (3)). Auch gibt es bis heute bedeutende (und zwischen Versionen wechselnde) Teile der Genomsequenz ohne Zuweisung zu einem der 25 Chromosomen, was ebenfalls Schwierigkeiten für bioinformatische Studien und Datenanalyse bedingt.

Von Modellorganismen zu Modellgenomen

Zwei Studien (2; 3) dieser Dissertation beschreiben an jeweils einem Modellorganismus ausgeführte Forschungsprojekte mit dem Potential der Übertragbarkeit auf menschliche Biologie. Im Gegensatz dazu verkörpert (1) eine dem genomischen Zeitalter vorbehaltene Art und Weise, mit Modellorganismen zu arbeiten. Hier wurden die Sequenzen eines bestimmten Gens, *POSTN*, innerhalb der Genomsequenzen verschiedener Organismen, von Säugetieren bis zu Knochenfischen (und ansatzweise darüber hinaus), identifiziert, wenn nötig rekonstruiert, und schließlich verglichen.

Dieser gewissermaßen vertikale Ansatz, der im Prinzip für jedes beliebige Gen zur Verfügung steht, wäre ohne die große und schell wachsende Zahl öffentlich verfügbarer Genomsequenzen verschiedener Organismen nicht möglich gewesen und steht damit für eine neue, genomisch-bioinformatisch geprägte Perspektive der Arbeit mit Modellorganismen.

Der Aufstieg eines Organismus zur „Prominenz" des Modell-Status ist traditionell von wissenschaftlichen und auch praktischen Überlegungen bestimmt. Dazu zählen charakteristische oder exemplarische biologische Merkmale, kurze Generationszeiten, Einfachheit der Haltung und schließlich Konsens innerhalb der Wissenschaftsgemeinde. Heute jedoch stehen Organismen als Modelle für genomische, phylogenetische oder bioinformatische Studien allein durch ihre genomischen Information zur Verfügung. Die folgenden zwei Beispiele verdeutlichen diesen Übergang.

(i) Die Ackerschmalwand (*Arabidopsis thaliana*) erfuhr erst später als Fadenwurm und der Zebrafisch allgemeine Akzeptanz als Modellorganismus. Als der Wissenschaftsgemeinde im Jahre 1985 die konzeptionellen Vorteile dieses Modells dargelegt wurden (29), wurde auch ihr Genom im Detail diskutiert. Während die Sequenzierung ganzer Genome noch in weiter Ferne zu liegen schien, wurden – neben klassischen Vorzügen wie der kurzen Generationszeit, der hohen Zahl von Samen, und der Einfachheit, durch Selbstbefruchtung homozygote Pflanzen zu erzeugen – bereits Eigenschaften des *Arabidopsis*-Genoms vorgebracht, insbesondere die geringe Größe (das bedeutete wenig benötigte Klone für eine umfassende DNA-Bibliothek), die geringe Chromosomenzahl und der niedrige Anteil repetitiver Sequenzen.

(ii) Das wahrscheinlich erste Beispiel eines Modellorganismus, der allein auf genomischer Basis vorgeschlagen wurde, stammt aus dem Jahre 1993, als Sydney Brenner – einmal mehr – und seine Kollegen für die Charakterisierung des Genoms eines Kugelfischs (*Takifugu rubripes*) plädierten (30). Sie prägten die Bezeichnung „Vertebraten-Modellgenom" („vertebrate model genome") und argumentierten, dass das mit 400 Mb äußerst kompakte Genom dieses Kugelfischs für die damals existierende oder wenigstens vorstellbare Sequenzierungstechnologie ein erreichbares Ziel darstelle. Auch sei es durch seine geringe Komplexität und sein hohes Verhältnis von kodierenden zu nicht kodierenden Sequenzabschnitten ein ideales Werkzeug für die Gen-Identifikation im Menschen. Doch während das Genom von *Takifugu rubripes* fast viermal kleiner ist als das des Zebrafisches, sind Kugelfische aus praktischen Gründen als allgemeine Modellorganismen nicht gut geeignet.

Nachdem die Sequenz des menschlichen Genoms im Jahre 2001 vorgestellt worden war, erfolgte die Publikation neuer Genomsequenzen immer schneller, was durch die rasche Optimierung der Sanger-Sequenzierungstechnologie und durch die dadurch bedingten Kapazitätsgewinne möglich wurde. Zu Genomprojekten für etablierte Modellorganismen mit großen Genomen (Maus, Hund, Krallenfrosch, Zebrafisch) gesellten sich bald solche, die vor allem aufgrund evolutionsbiologischer und phylogenetischer Bedeutung (Opossum, Platypus, Rotkehlanolis (*Anolis carolinensis*)) oder wegen landwirtschaftlicher Anwendungsmöglichkeiten (Rind, Reis, Weintraube) unternommen wurden.

Krebsfoschung und die sich beschleunigende genomische Revolution

Genomische Forschung ist heute gekennzeichnet durch ausgereifte DNA-Mikroarray-Technologie und durch sich rasant entwickelnde massiv parallele Sequenzierungstechnologien. Letztere ersetzen zunehmend ältere Technologien (Sanger-Sequenzierung), auch für genomweite Analysen, wie dies die Sequenzierungsstatistiken eindrucksvoll zeigen, die auf der Webseite der Genomprojekte des U.S. Energieministeriums veröffentlicht werden (31). Auch verdrängen sie Mikroarrays bei Anwendungen, für die diese viele Jahre lang Standard waren (Genexpression, Chromosomenzahl, Mutationen). Dieser Trend beschleunigt sich durch die Aussicht, gleichsam „nebenbei" Einsichten in zusätzliche und für Mikroarrays bisher nicht oder nur unter Schwierigkeiten zugängliche Fragestellungen (z.B. Translokationen, Spleißvarianten) gewinnen zu können.

Unsere Folgearbeiten im Anschluss an (3) an hunderten von Zebrafisch-Tumoren beruhen nicht länger auf Mikroarrays, sondern ausschließlich auf voller Sequenzierung. Auch in unserer fortgesetzten Forschung am Fadenwurm stammen die zu Vergleichen herangezogenen publizierten Literatur-Datensätze immer öfter von Sequenzierungsprojekten und können dementsprechend extrem groß sein.

Für die Untersuchung von Modellgenomen hat diese Entwicklung die Möglichkeit eröffnet, die Variation innerhalb einer Art in beispielloser Breite und Auflösung zu studieren. In der Tat werden menschliche Genome schon weithin sequenziert, und das Ausmaß natürlicher Variation im Menschen (das „Variom") beginnt deutlich zu werden – das Ende des vereinfachten (und bioinformatisch so bequemen) Konzepts „ein Organismus, ein Genom" zeichnet sich ab.

Im Bereich der Krebsbiologie spiegelt sich diese Entwicklung in onkogenomischen Ansätzen wider, die dem erklärten Ziel eines erschöpfend beschriebenen, Krebsform-spezifischen Modellgenoms (mit Transkriptom und Variom) immer näher kommen. Es ist faszinierend, dies als Bestätigung der Idee des Evolutionsbiologen Leigh van Valen zu sehen, nach der Krebszelllinien (in diesem Fall HeLa) als eine eigene biologische Art anzusehen seien (32). Wie ernst es van Valen mit diesem Konzept damals war, ist nicht klar – es findet sich jedoch heute in Studien wieder, die Krebs mit den Parametern der Evolutionsbiologie ergründen (33), und in Beispielen klinischer Forschung, bei denen das gesamte Krebsgenom eines Patienten sequenziert und analysiert wird (34; 35), mit dem Endziel aus den Ergebnissen individualisierte und damit optimierte Therapieoptionen abzuleiten. Eine solcherart personalisierte Medizin ist schon vor mehr als 35 Jahren angedacht worden (36), wird jedoch erst heute umfassend möglich.

Die hohe Rate der Neusequenzierung vor allem menschlicher Genome kann leicht den Blick darauf verstellen, dass die Situation nicht-menschlicher Modellgenome, einschließlich solcher von weniger etablierten Modellorganismen weit weniger fortgeschritten ist, ohne dass hier Besserung absehbar

wäre. Die Situation für das Zebrafisch-Genom ist weiter oben schon erläutert worden. Schlechter steht es um andere Genome, deren Sequenzen zum Teil ohne Zuweisung zu Chromosomen verbleiben (z.B. Krallenfrosch) und die mitunter so fragmentiert sind, dass sogar einfache Versuche, einzelne Gene zu identifizieren zu einer Herausforderung werden (Pflugnasenchimäre (*Callorhinchus milii*), Neunauge (*Petromyzon marinus*)). Ohne finanzielle Förderung der entsprechenden Projekte verkümmern diese Genomsequenzen in jetzigen Zustand ihrer teils extrem eingeschränkten Nutzbarkeit. Diese Situation hat zum Beispiel schlüssige Ergebnisse zu Periostin verhindert, als wir versuchten, die C-terminalen Region in phylogenetischen Linien jenseits der Knochenfische zu identifizieren (siehe zusätzliche Datei 6 in (1)), und behindert uns durch die weiter oben beschriebenen Unzulänglichkeiten in der Zebrafisch-Genomsequenz weiterhin bei der Fortsetzung der Arbeiten im Anschluss an (3).

Wir dürfen vielleicht hoffen, dass die neuen Sequenzierungstechnologien schließlich auch hier Abhilfe schaffen könnten, wenn ihr Einsatz weithin zur Routine geworden ist und die Kosten niedrig genug sind. Die Erkenntnis, dass eine phylogenetisch breite Sammlung von Genomsequenzen, die alle bestimmte minimale Gütekriterien erfüllen, im Interesse der wissenschaftlichen Gemeinschaft ist, sollte es ermöglichen, derartige Genomsequenzen geringer Qualität durch neu erzeugte Sequenzen „aufzupolieren". Dieses Konzept ist bereits bei der neuesten Version des Zebrafisch-Genoms eingesetzt worden – vielleicht lässt es sich ja auch auf „verwaiste" Genomprojekte übertragen.

Fortschritte im Verständnis der Krebsbiologie werden heute in bedeutender und systematischer Weise durch das Studium von Modellorganismen und genomische Forschung ermöglicht. Die vorliegende Dissertation zeigt einige der unterschiedlichen Modalitäten auf, die diesem Konzept folgen, und die bestimmte exemplarische Koordinaten in einem komplexen Raum einnehmen, der durch Achsen wie „bioinformatische oder experimentelle Ansätze", „Zahl der berücksichtigten Organismen", „Zahl der untersuchten Gene", „Grundlagenforschung oder angewandte Forschung" und vielen anderen definiert ist. Und sie nährt die Hoffnung, dass letztendlich die Herausforderung gemeistert werden kann, all die wachsenden Inseln der Einsicht zu einem detaillierten und übergreifenden Verständnis zu verschmelzen, aus dem heraus Krebs tatsächlich heilbar wird.

Thesis Summary in English

Title and Authors

Model Organisms and Genomics in Cancer Research

by Sebastian Hoersch

constituent of the following publications (in chronological order):

RNA interference and retinoblastoma-related genes are required for repression of endogenous siRNA targets in *Caenorhabditis elegans*.
Proceedings of the National Academy of Sciences of the United States of America 105(51): 20386-20391 (2008).
Alla Grishok, Sebastian Hoersch, and Phillip A Sharp.

Periostin shows increased evolutionary plasticity in its alternatively spliced region.
BMC Evolutionary Biology 10: 30 (2010).
Sebastian Hoersch and Miguel A Andrade-Navarro.

Highly aneuploid zebrafish malignant peripheral nerve sheath tumors have genetic alterations similar to human cancers.
Proceedings of the National Academy of Sciences of the United States of America 107(39): 16940-16945 (2010).
GuangJun Zhang, Sebastian Hoersch, Adam Amsterdam, Charles A. Whittaker, Jacqueline A Lees, and Nancy Hopkins.

Abstract

This thesis explores new territory in cancer research, enabled by genomic data originating from model organisms. As exemplified in three publications forming the core of this thesis, this approach has a considerable and generally under-used potential for gaining insights into cancer biology on many distinct levels.

As described in one of these publications (1), genomic sequence data can be used in comparative mode across species with insightful results even in the context of a purely computational project. Here, we study one particular gene of interest, which encodes the extracellular matrix protein periostin. Known also as a cancer marker, periostin function remains poorly characterized overall, but as part of this project, an improved understanding of periostin's C-terminal region was achieved by comparing periostin sequences across a range of vertebrate genomes. We explained the absence of known domains or sequence similarities to any non-periostin proteins for this region, demonstrated its remarkable genomic and transcriptomic variability, and suggested a possible secondary structure and functional mechanism.

The other two studies reported in this thesis are based on non-murine model organisms – the use of which is generally not common in cancer research – and contain key bioinformatic components in tight integration with experimental results that were contributed by other members of the research teams. Here, genomic data was a necessary foundation enabling both the immediate, model organism-centric research and its projection into the human realm.

The study on the nematode *Caenorhabditis elegans* (2) presents and interprets data on the functional roles of select genes (*lin-35, zfp-1, rde-4, alg-1*), finding that they act cooperatively with endogenous small interfering RNAs (siRNAs). Here, mRNA expression profiling data of mutant *C. elegans* strains were evaluated in a bioinformatic framework that allowed genome-wide functional interpretations in conjunction with relevant genomic data from the public domain. This study is significant because the genes investigated are, with one exception, homologs of human genes with known importance in cancer, and because the findings point to the notion of cancer as a condition where germline genes are pathologically de-repressed in the soma. Thus, this study constitutes basic research with an intrinsic relevance for cancer due to the identity of the genes and mechanisms examined.

The third study focuses on tumors occurring in zebrafish (*Danio rerio*), with results that are directly cancer-specific (3). A genome-wide DNA microarray was custom-designed to generate comparative genomic hybridization (aCGH) data from zebrafish malignant peripheral nerve sheath tumors (MPNSTs). Analysis of this data showed that zebrafish MPNSTs have chromosomal and subchromosomal alterations akin to those widely documented in human cancers (including human MPNSTs), some of them comprising genes known to be amplified in human cancers (*met, ccnd2a, cdk6*). The results suggest a model system for aneuploidy, an aspect of cancer not generally well replicated in mouse models. Furthermore, and practically forward-looking, they also evoke the promise of a strategy to effectively screen for cancer driver gene candidates by "intersecting" human and zebrafish copy number alteration profiles.

Together, these three studies exemplify contributions to oncology research that rely critically on model organisms and their *in silico* counterparts, model genomes. On this basis, bioinformatic analyses became possible – either closely integrated with experimental research or generating useful insights and hypotheses on their own – that enabled advances in our understanding of human cancer.

Introduction

Cancer remains a profoundly challenging disease today. This is true on many different levels, from a fundamental understanding of its causes all the way to diagnosis and therapy. It remains true also in the face of the large amounts of data collected and all the insights gained over many decades of cancer research.

The notion of cancer as a disease in its own league (4) is well-founded: No other disease commands its own biology to a remotely similar extent, as documented in a seminal paper published eleven years ago (5). Since then, "the hallmarks of cancer" have come into even sharper focus (6), as genomic technologies allowed characterizations of underlying molecular events at ever higher granularity, while at the same time raising hopes that, ultimately, the aggregate of these alterations can be functionally parsed, understood, and targeted with medical therapies.

Most prominently epitomized by the availability of entire genome sequences for human as well as important model organisms, the advent of the "genomic age" has profoundly altered the defining and limiting parameters in basic and clinical cancer research. Thanks to technologies such as genome-wide DNA microarrays or, most recently, high-throughput sequencing, clinical cancer specimens are amenable to molecular analyses of their gene expression, chromosomal alterations, and mutational status.

Model organisms – indispensable accessories to cancer research as attested by ever more refined mouse models of cancer – can be similarly investigated and their data integrated with human data. Also, and of particular interest regarding the compatibility of a model organism with human for a given question, genomic sequences of humans and other organisms can be compared directly to address diverse questions concerning for example gene content, gene family size, or conservation of regulatory elements.

This thesis comprises results relevant to the biology of cancer obtained from two non-murine model organisms (the nematode *Caenorhabditis elegans* and the zebrafish *Danio rerio*) and from a wider analysis using a collection of vertebrate model genomes. In this context, the importance of genomic data as an enabling factor cannot be understated: It is likely that without comprehensive genomic sequence data, model organisms other than mouse could not play a significant role in cancer research today, and the availability of genome sequences alone opens up innovative avenues of investigation.

Goals

The goal of this thesis is to explore unconventional computational approaches to utilizing genomic data in cancer research. These approaches are exemplified in three separate publications (1-3), where my bioinformatic contributions address the following specific core objectives:

(1): Utilize a comparative genomics approach for the characterization of the poorly understood C-terminal region of the extracellular matrix protein periostin, which is over-expressed in many cancers of epithelial origin. In particular, focus on the apparent absence of functional clues such as functional domains or homology to other known proteins, and its prevalence of alternative splicing.

(2): Perform a comparative analysis of genomic microarray data profiling the gene expression of *Caenorhabditis elegans* wild type and four strains mutated in genes homologous to human cancer genes. Adapt an appropriate functional genomics system to enable comprehensive functional interpretation and statistical inferences of this data in conjunction with relevant publicly available data.

(3): Use genome sequence data for the design of a custom microarray-based comparative genomic hybridization platform for the zebrafish (*Danio rerio*) in order to measure and evaluate evidence for copy number alterations in zebrafish tumors. Compare the resulting chromosomal gains and losses with those observed in human cancer. Analyze high-throughput sequencing data as an orthogonal methodology to confirm the microarray-based findings.

Bioinformatics Methods

General principles

It should be stated that bioinformatic work for any project starts prior to writing program code and prior to the application of software tools and algorithms. Not only need the problems at hand understanding, they also need to be paired with available and practical options for addressing them. Journal articles often betray little of the sometimes complex and often pragmatic decision-making process leading to the final choice of methods, and also in the context of this thesis, these phases "behind the scenes" demanded my considerable attention.

Fundamentally and conceptually, the three projects of this thesis build on a common core of bioinformatic methodology for analysis and visualization of biological sequences and microarray data.

Sequence search and comparison are prevalent in all three studies, comprising successive generations of widely-used methods such as BLAST and BLAT, but also the more recently developed high-throughput, short read alignment programs like BWA. DNA microarray analysis is integral to two of the three publications (2; 3) and requires appropriate strategies for data normalization, organization (for example via clustering), and probe sequence analysis.

Project-specific methods

Given these two themes, the details of their actual application vary across the three studies, mirroring the diverse study goals and the different approaches chosen in their pursuit. A brief, project-specific characterization of the relevant bioinformatic methods follows.

The periostin publication (1) summarizes a pure bioinformatics project, with a stated, exclusive focus on sequence analysis and phylogenetic aspects. Methods applied include various versions of the sequence similarity search algorithm BLAST, such as TBLASTN (to identify periostin exons in genomic sequences of poorly annotated organisms) and PSI-BLAST (to establish homology that is obscured by low complexity and repeats). Multiple sequence alignments and derived phylogenetic trees were generated with ClustalW. Less commonly used sequence analysis methods were dot-matrix plots to visualize the repetitive nature of the periostin C-terminal region and VISTA for a multi-species alignment of the periostin genomic region. PsiPred was used for secondary structure prediction, which provided a basis for a speculative proposal regarding periostin C-terminus function.

Bioinformatic methods for the paper on *C. elegans* mutants (2) comprised processing and normalization of a genome-wide gene expression microarray dataset, followed by determination of gene sets differentially expressed between wild type and mutant strains. The central bioinformatic contribution is the adaptation of the expression cluster terrain map TOPOMAP (7) for rapid, universal, genome-scale functional classification of gene sets, derived both from our own data as well as from public data. Gene Ontology-based strategies were evaluated and found wanting due to the uneven coverage of genes from different functional groups. The TOPOMAP data were organized in a large matrix and deployed in a spreadsheet program (MS Excel), so that overlaps between gene sets and TOPOMAP expression clusters ("mounts") could be easily ascertained and additional literature-based gene sets of interest could be incrementally added for going forward. Fisher's Exact Test was adopted for assessing statistical significance of such overlaps. BLASTN was used to map siRNA sequences from other studies to the *C. elegans* transcriptome for adaption into the TOPOMAP framework.

The lack of a commercial platform to perform array-based comparative genomic hybridization (aCGH) experiments in zebrafish necessitated a custom design for such a microarray in the context of the aneuploidy study in zebrafish tumors (3). 15,000 array probes were selected from a collection of five million probes provided to us by Agilent, using a successive combination of criteria: hybridization scores, BLAT search results (to determine uniqueness in the genome sequence), and a heuristic to achieve an approximation of equidistant spacing of probes along the genome. BLAT was similarly used to remap and re-evaluate the array content when a newer reference zebrafish genome assembly (Zv8) was chosen. Processed and normalized array data (via Agilent Feature Extraction) were subjected to the program DNAcopy for segmentation. Segmented data were analyzed for sub-chromosomal copy-number changes using STAC. Finally, bioinformatic efforts comparing human-zebrafish and human-mouse synteny centrally contributed to forming and supporting the hypothesis laid out in the paper's discussion section that candidate cancer driver genes can be effectively screened for by "intersecting" chromosomal alterations found in human and zebrafish tumors.

Results and Discussion

Genomic technologies and model organisms in oncology research

This thesis takes advantage of the combined power of model organisms and genomic technologies.

Genomic technologies are those that in aim and scope allow measurement of near-complete complements of DNA, RNA, protein or other biomolecules in a biological sample. Among these analytes, nucleic acids (usually genomic DNA or messenger RNA) have been most widely accessible to a range of methods. Exclusively sequencing-based in their early incarnations (Sanger sequencing), their primary focus shifted later to DNA microarrays of ever increasing density, only to recently return to sequencing with the advent of massively parallel technologies.

Model organisms are generally understood to be a small set of representatives from different phylogenetic groups, selected originally for properties that facilitate the study of select, often narrowly defined biological questions. Iconic examples include baker's yeast (*Saccharomyces cerevisiae*), the fruit fly (*Drosophila melanogaster*), the rat (*Rattus norvegicus*), and the mouse (*Mus musculus*).

The mouse is by far the most prominent model organism in oncology research (Figure 1). With a background of a 100-year history in biomedical research, today's mouse cancer models can be precisely engineered with specific changes to specific cancer genes (for example, see (8)), and often mimic their human counterparts down to molecular levels of detail (9; 10).

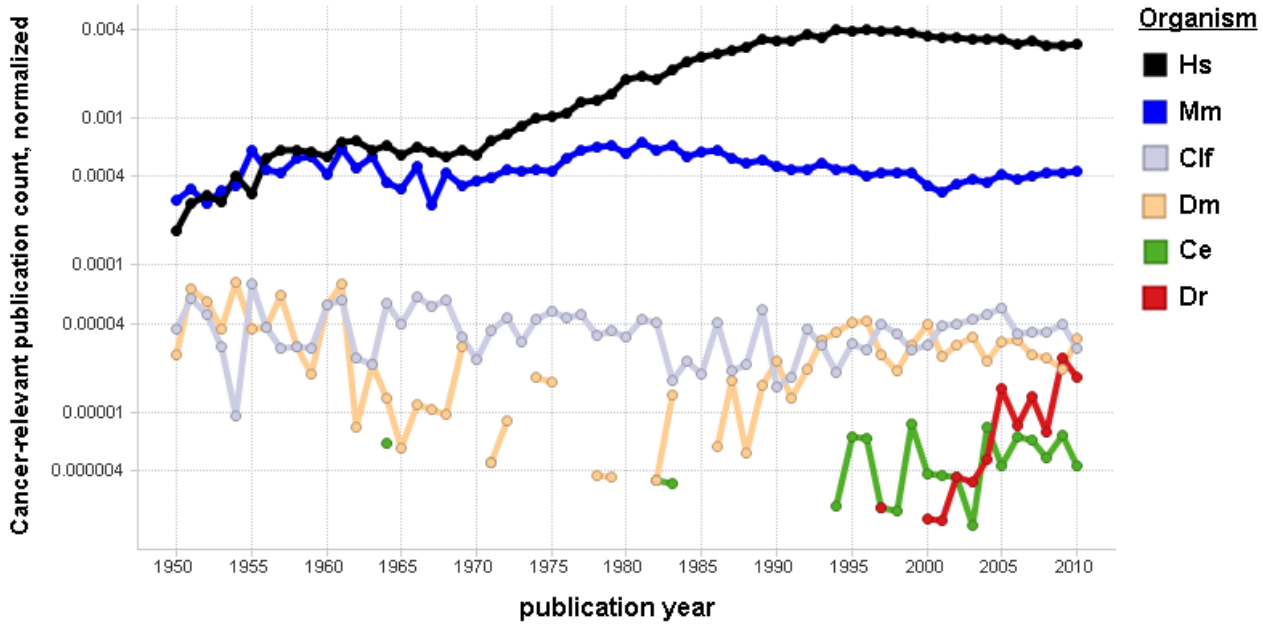


Figure 1: Titles of the scientific literature accessible via the Medline database linking cancer and select model organisms or human (Hs) from 1950 to 2010, graphed according to MLTrends (12).
The logarithmic Y-axis represents the number of entries matching the keywords, divided by the total number of publications per given year, so that an increase over time indicates an actually growing proportion of the total literature. Compared to mouse (Mm), dog (Clf), and fruit fly (Dm), nematode (Ce) and zebrafish (Dr) appear late in the cancer field, the latter being the only model organism to presently show explosive growth in its share of relevant publications.
Any title had to match at least one of the keywords "cancer(s)", "tumor(s)", or "tumour(s)" and one component of the scientific species name or the common name, as mentioned above.

Leaving aside the context of pre-clinical testing, the list of alternative metazoan model organisms in cancer research is quite short indeed. On the spectrum of proximity to humans with respect to phylogenetics, physiology, and presumably cancer etiology, we find on one hand the dog, with a potentially increasingly prominent role in genomics-based cancer research going forward (11).

On the other end of the spectrum are two model organisms that are prominent subjects of this thesis: the nematode *Caenorhabditis elegans* and the teleost fish *Danio rerio* (2; 3). They are phylogenetically more distant from human than mammals, and their relevance to cancer research may not be obvious. Both are relatively recent model organisms, especially in the oncology field (Figure 1), elevated to this status essentially single-handedly by dedicated scientists.

Today, genomic sequences are available for almost all model organisms in public databases, where they reside alongside a rapidly expanding collection of genome sequences from organisms that – while not model organisms *per se* – are of interest to the scientific community in the context of domains as diverse as evolutionary biology, disease biology, ecology, or agriculture. Together, these form a formidable repository amenable to bioinformatic analyses, of which this thesis provides a detailed example (1).

Results by publication

Periostin shows increased evolutionary plasticity in its alternatively spliced region (1)

Periostin (POSTN), a secreted, extracellular matrix protein with a role in cell adhesion, has been found to be over-expressed in numerous cancers of epithelial origin and its increased expression has been associated with angiogenesis and metastasis. Human periostin is encoded by up to 23 exons, giving rise to an 836-amino acid protein. It has a C-terminal region that constitutes a substantial part of the periostin protein (~180 amino acids) and is known to be subject to alternative splicing, but is currently devoid of defined protein domain annotations.

We hypothesized that an in-depth sequence comparison to other organisms would shed light on the C-terminal region. We established homology between this region in tetrapod periostin and a strongly conserved C-terminal 13 amino acid repeat in periostin of teleost fish. Evaluating existing genomic and transcript sequence data enabled us to infer the full-length periostin sequence for a range of vertebrate species and to observe alternative splicing of the periostin C-terminal region for all euteleostome lines with sufficient transcript evidence, including teleost fish.

This sequence comparison exercise allowed us to identify a 24th periostin exon that is expressed in only a subset of tetrapod lines, and a cluster of 8 additional genomically encoded copies of exon 19 unique to the claw frog *Xenopus tropicalis*. In both these cases, these non-canonical exons are part of the periostin C-terminal region and were also found to be alternatively spliced.

Moreover, comparing five teleost genome sequences, we found that teleost fish have – presumably as a consequence of the whole genome duplication event occurring prior to teleost radiation – two copies of periostin that show considerable divergence in their C-terminus. By contrast, the periostin paralog transforming growth factor, beta-induced (*TGFBI*), which lacks periostin's extended, alternatively spliced C-terminus, is uniformly not duplicated in teleosts.

Thus emerged a picture of the periostin C-terminal region as remarkably variable both on a transcriptional level (alternative splicing) as well as on a genomic and evolutionary level (having a high degree of sequence divergence and a variable number and length of exons). In our

interpretation, this may be a consequence of active sub- or neo-functionalization following the split between periostin and its paralog *TGFBI*. Interestingly, this situation is mirrored in the teleosts, where the two periostin copies show considerable variation in their C-terminal region.

Finally, we found a predicted structure for the C-terminal region of consecutive beta strands separated by turns to be conserved across all vertebrate species, regardless of exon structure or the degree of conservation of the 13 amino acid repeats. This allowed us to propose, for the first time, a biological function for the periostin C-terminal region: Periostin is known to bind to, among other extracellular matrix proteins, fibronectin, and we hypothesized that the beta strands may mediate binding with other proteins through an extended beta-zipper, reminiscent of the way beta-strand repeat units in bacterial cell wall proteins have been described to bind human fibronectin.

Author contributions

This study and its design was conceived of by me and refined with input from co-author Miguel A. Andrade-Navarro. All analyses were performed by me, with the exception of the PSI-BLAST analysis conducted jointly with M.A.A. The work was discussed with M.A.A. on an ongoing basis. The manuscript, including all Figures and supplemental material, was drafted by me and revised and finalized with input from M.A.A.

RNA interference and retinoblastoma-related genes are required for repression of endogenous siRNA targets in *Caenorhabditis elegans* (2)

The realization that RNA serves as a functional regulatory molecule independent of its classical role in cellular protein synthesis stimulated an explosive field of research recently, with significant implications in cancer biology (13).

The nematode *Caenorhabditis elegans* has been at the forefront of work on RNA-based gene regulation. Its genome encodes the largest number of short RNA-interacting Argonaut proteins known in any organism, along with thousands of genes coding for endogenous short interfering RNAs (endo-siRNAs).

We used microarray analysis to concurrently study the gene expression in mutants for three RNA interference related genes, *zfp-1, rde-4*, and *alg-1,* and for *lin-35*, a nuclear factor and homolog to the human tumor suppressor gene retinoblastoma (RB1) in comparison to a wild-type strain. For a functional interpretation of the microarray-based results, we adapted the framework provided by the gene expression terrain map ("TOPOMAP") (7). This is a mathematical projection of co-expressed gene signatures into clusters ("mounts") on a two-dimensional plane. The TOPOMAP analysis in (7) is based on a large collection of gene expression data obtained under a wide variety of conditions, covering the majority of *C. elegans* genes. Indeed, we found the fraction of *C. elegans* genes represented in TOPOMAP to be very high (77%), whereas only 46% of genes had Gene Ontology annotations. Also, the TOPOMAP representation is largely independent of the degree to which a gene is known or studied, affording us a more balanced functionalization of the *C. elegans* transcriptome than would have been possible using Gene Ontology annotations.

We found that mutants of *zfp-1* (a chromatin factor and homolog of human acute lymphoblastic leukemia-1 (ALL-1)-fused gene from chromosome 10 (AF10), *MLLT10*) and *rde-4* (encoding a Dicer-interacting, double-stranded RNA-binding protein) exhibit similar expression changes

compared to wild-type worms. This notion emerged even more forcefully on the level of functional annotations (TOPOMAP mounts), clearly indicating involvement in a common pathway (RNAi-induced transcriptional gene silencing) for these two genes.

We used TOPOMAP as an integrating platform for comparing our own microarray data with relevant data in the public domain. Analyzing three separate datasets of endo-siRNA target genes in this fashion, we found statistically significant enrichment with overexpressed genes in *zfp-1*, *rde-4*, and *lin-35* mutants both on the level of individual genes and of functional groups. This suggested that overexpressed genes in the mutants represented direct RNAi targets.

Together, we provided evidence for a large-scale cooperation between endo-siRNAs and chromatin factors in regulating overlapping gene sets, and we anticipate a significant role for RNAi-mediated chromatin silencing in *C. elegans* gene expression regulation.

Author contributions

This study and its overall design was conceived of by first author Alla Grishok and Professor Philip A. Sharp, and the detailed study design was developed with my input.

All experimental work and data analysis of PCR experiments was conducted by A.G.

All bioinformatic analyses were conducted by myself, specifically: data processing and normalization of DNA microarray data (with input from Charles A. Whittaker); differential expression analysis of microarray data; evaluation of the suitability of Gene Ontology functional annotations for this study; adaptation and deployment of TOPOMAP as a platform for functional annotation; implementation of a statistical framework to evaluate gene set overlaps within TOPOMAP; mapping of 3rd-party datasets, notably endogenous siRNA collections, but also others only partially represented in the final publication.

In addition to project discussions between co-authors P.A.S. and A.G. (with my occasional participation), A.G. and I met for project discussions on a regular basis, reviewing results and evaluating and deciding on analysis strategies or feasibility of public dataset analyses in the context of our data, often with my guidance.

The manuscript was drafted by A.G. and revised with input from P.S. and myself. I wrote the Bioinformatics method sections and created Figures 1, S1, S2 and all three supporting data tables.

In addition, I handled the submission of the gene expression microarray data to the public data repository (Gene Expression Omnibus, accession no. GSE13258).

Highly aneuploid zebrafish malignant peripheral nerve sheath tumors have genetic alterations similar to human cancers (3)

Chromosomal instability, a key characteristic of human cancer, leads to aneuploidy as well as subchromosomal abnormalities such as translocations, inversions, deletions, and amplifications. It is difficult to distinguish the many non-specific abnormalities ("passengers") from those actively promoting disease progression ("drivers"). Mouse models are not very helpful at studying chromosomal instability, because they typically exhibit a substantially lower level of chromosomal instability than their human counterparts.

In (3), using custom-designed microarrays for comparative genomic hybridization aided by confirmatory high-throughput sequencing, we investigated chromosomal changes in malignant

peripheral nerve sheath tumors (MPNSTs) arising in *rp* or *p53* mutant zebrafish (*Danio rerio*) after a considerable latency period of 9 – 24 months.

We found that zebrafish MPNSTs are highly aneuploid, generally with an average ploidy of 3N, suggesting that they represent a good model for aneuploidy in human cancer. Among the 36 independent MPNST samples, we observed whole chromosome changes that were non-random; examples include over-representation of chromosomes 25, 11, and 10 and under-representation of chromosomes 15, 8, and 5. On the other hand, certain other chromosomes (e.g. 16, 13, and 3) showed no consistent trend for either over- or under-representation.

We also observed sub-chromosomal amplifications, most notably on chromosome 25. Genes in these amplified regions included *slc45a3*, *ccnd2a*, and *met* – all genes whose human counterparts have been observed repeatedly in the context of chromosomal changes in human cancers and implicated as potential cancer drivers.

Narrower focal changes (only several hundred thousand bases wide) were also found, although their evaluation proved problematic because of the preliminary status of the zebrafish genome assembly. Indeed, we observed that many of these narrow changes were mapped to different chromosomes in different genome assembly versions. In conclusion, these focal amplifications require validation independent of the genomic references currently available.

Finally, we performed a preliminary, indirect validation of a candidate gene (*fgf6a*) from a sub-chromosomal amplification on chromosome 25. In mammals, different members of the fibroblast growth factor (FGF) family are known to bind across four FGF receptors, effecting signaling through common MAP-kinase pathways. We showed that over-expression of a different fibroblast growth factor (*fgf8a*) led to accelerated onset of MPNSTs in p53 mutant background, and hypothesized that *fgf6a* may similarly be a driver gene for MPNST.

Overall, we found that zebrafish MPNSTs display chromosomal changes resembling those observed in human cancer, suggesting zebrafish as a valuable model organism for the study of aneuploidy, a key aspect of cancer not readily accessible with mouse models.

Author contributions

This study, including overall research design, was conceived of by first author GuangJun Zhang with co-authors Adam Amsterdam, Professor Jacqueline A. Lees and Professor Nancy Hopkins. The detailed study design was developed with my input. All experimental work was conducted by G.Z. and A.A, who also performed data analysis of flow cytometry, Southern Blot, and chromosome counting data.

I conducted all bioinformatic work – except for processing of the high-throughput sequencing data performed by co-author Charles A. Whittaker. Specifically, I carried out the following: evaluation of different design strategies for a custom CGH array for zebrafish; design of the array actually manufactured and used based on a collection of about 5 million probes provided by Agilent, including sequence analysis steps to ensure probe quality, uniqueness, and compatibility with the experimental protocol used; re-mapping (including quality assessment) of the existing array probes to a newly released zebrafish genome assembly; array data processing and normalization (Agilent Feature Extraction) and analysis of whole-chromosome as well as segmented copy number data (DNAcopy, STAC) in both array and sequencing data.

Certain detailed aspects of experimental design for this study were guided by me, especially the ongoing use of two-channel control/control samples to gauge noise levels in the data. These samples are not commented on in the publication, but were submitted to the public data repository for this study together with the tumor/normal samples. Also, the choice of high-throughput sequencing for array validation was advocated by me (with C.A.W.).

I participated in project discussions on a regular basis, usually involving co-authors G.Z., A.A. and N.H. During these, I frequently provided guidance on bioinformatic analysis, also on the feasibility and preliminary data for potential future zebrafish/human comparisons outlined in the Discussion section.

The manuscript was jointly written by G.Z., A.A., N.H., and myself; I wrote the bioinformatics methods section with input by C.A.W. Figures 2 and 3A & B were created by me.

Finally, I handled submission of the microarray and the high-throughput sequencing CGH data to the public data repository (Gene Expression Omnibus, accession no. GSE23666).

Continuing research

All three publications constituting this thesis reflect snapshots from active areas of research rather than final end points.

The study on periostin (1) provides a foundation for further work, for which a manuscript is in preparation (14) and certain core aspects of which have already been disclosed in a patent application filed by Millennium Predictive Medicine (now Takeda, the Millennium Oncology Company, Cambridge, US) (15). Our insights into periostin's C-terminal variability, including a potential functional mechanism, enable an improved interpretation of our findings of a cancer-associated alternative splicing pattern in this region in clinical breast cancer samples.

During publication of (2), the first author started her own laboratory at Columbia University, New York, where she continues research in *C. elegans* biology with a particular focus on RNA interference. She maintains an informal collaboration with me on a range of bioinformatics problems and specifically on the continued utilization of the TOPOMAP platform that was adapted for this publication as a tool for functionally annotating and integrating gene sets from a diverse spectrum of sources. We have found that this capability scales well, thus allowing additional data to be incorporated and easily compared to earlier gene sets. One manuscript reflecting this has been submitted for publication and is under review (16), another one is in preparation.

Investigations into zebrafish cancer following (3) continue in a very direct fashion with my participation, as major points outlined in the Discussion section of this paper are now being pursued. Instead of the custom-designed microarrays used in (3), we are now using high-throughput sequencing exclusively – a testament to the extremely rapid advancement of this technology. Several hundred zebrafish tumor samples – mostly from MPNSTs, but also from two other tumor types – have already been sequenced and data analysis is ongoing.

Furthermore, efforts are underway to explore the idea that the considerable evolutionary distance between humans and zebrafish can be used to hone in on cancer driver genes by "intersecting" the genes affected by chromosomal gains or losses observed in human and zebrafish cancer. Compared to considering all genes affected by chromosomal alterations in a single organism, the resulting collection of candidate genes is considerably reduced and expected to be enriched with

cancer-relevant genes. These, then, could be identified within this relatively small gene set using RNAi-screening (see e.g. (17; 18).

The following sections are an attempt to provide both historical and current context for the projects of my thesis, space constraints notwithstanding. These considerations are guided by the two areas so central to this thesis, that is, genomics and model organisms. They continue to emphasize in their outlook the biomedical research into cancer, and its progress.

Caenorhabditis elegans as a cancer model organism

The nematode *Caenorhabditis elegans* was systematically introduced as a model organism for genetic research by Sydney Brenner in 1974 (19). Since then, it has been at the forefront of groundbreaking biological research, including for example the unraveling of the mechanisms of RNA interference (RNAi) in 1998 (20).

Nematodes do not develop cancerous tumors, although hyper-proliferation of some cell lineages can occur as a consequence of mutations in certain genes. Nevertheless, the impact of *C. elegans* on oncology research is very considerable. Firstly, the discovery of RNAi as a biological phenomenon led to the development of RNAi as a laboratory technique, with a significant impact in cancer research. For example, the resulting RNAi-based ability to perform loss-of-function screens has enabled genome-wide studies elucidating oncogenic pathways by virtue of profiling and comparing multiple cancer cell lines (17; 18).

Secondly, central biological processes and their genes are generally conserved between nematodes and vertebrates (13). Thus, insights obtained in *C. elegans* into the biological workings of these genes and pathways are relevant for vertebrate cancer biology, which routinely involves the deregulation of such conserved processes as cell cycle progression, growth factor signaling, or apoptosis.

In our work (2), three of the four genes mutated in the microarray profiling panel are homologs of prominent cancer genes: For example, *lin-35* is a homolog of the tumor suppressor and negative cell cycle regulator gene *RB1* (retinoblastoma 1), that causes childhood retinoblastoma, bladder cancer, and osteosarcoma, when mutated. A second gene, *zfp-1* is a homolog of *MLLT10* (myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 10), also called *AF10*. MLLT10 is a transcription factor best known for its involvement in chromosomal translocation events leading to leukemias via the generation of fusion transcripts and proteins. It is notable, however, that MLLT10 function in its native configuration remains largely uncharacterized. Consequently, the picture emerging from this and follow-up work of ZFP-1 as a negative modulator of its targets will be important in elucidating native MLLT10 function. It is interesting to note that *zfp-1* appears to echo *MLLT10* in terms of transcript isoform complexity.

Cancer cells can gain properties important for survival by expressing genes whose expression is normally restricted to germline cells. In humans, a class of genes illustratively named cancer-testis antigens provides an example (21). A more recent study in the fruit fly *Drosophila melanogaster* (22) implicates fly retinoblastoma protein homologs in a role of keeping germline genes transcriptionally repressed in somatic cells. While this function of RB1 still remains to be

demonstrated in mammals, our study mirrors these findings: we find over-expression of germline genes in the *lin-35* mutant *C. elegans* strain. Thus we show that *C.elegans* biology informs cancer research for both individual gene function as well as broad trends in cancer.

Danio rerio as a cancer model organism

George Streisinger and colleagues introduced the teleost *Danio rerio*, the zebrafish, as a model organism for conducting genetic screens in 1981 (23). Over the next two decades, the screening methodology evolved considerably, leading to important insights mainly in the field of developmental biology. Only then, a forceful, but largely theoretical case was made for using zebrafish as a model system for cancer (24). Over the following years, several zebrafish cancer models (pancreatic cancer, melanoma, leukemia) were introduced in the literature. One study demonstrated that mutations in ribosomal protein genes can lead to tumors, mostly malignant peripheral nerve sheath tumors (MPNSTs) (25). The zebrafish study reported here continues and builds on that earlier work.

As an aside, there is an earlier strain of work pointing to a role for zebrafish in oncology studies that is distinct from the historical background of genetic screening: The fact that zebrafish reproducibly develop tumors in response to carcinogen exposure was documented as early as 1965 (26), with similar work in other types of fish going back even further. It is interesting to contemplate the prospect of re-opening this line of research to compare and integrate the results with those obtained from genetic cancer models and their genomic characterization.

The fact that zebrafish develop tumors, and tumors with direct histological counterparts in mammals, gives that organism a more immediate role in cancer research than the nematode. As is true for mouse models, caution is nevertheless warranted on the genetic level where an exact correspondence between the human cancer and its modeled counterpart is harder to ascertain. For example, in human cancer, there is, as of yet, no known direct correlate for ribosomal protein genes acting as tumor suppressors.

In this context, the contribution of our study is significant, because it documents the chromosomal alterations frequent in zebrafish MPNSTs, which in turn implicate some well known cancer genes (e.g. *met, cdk6*, *slc45a3*, *ccnd2a*). Compared to mouse models, this represents an advance with respect to the goal of modeling human cancer as accurately as possible. This zebrafish model captures aneuploidy, a common feature of many human cancer types, without giving up the context of known cancer genes.

Distinct genomic characteristics in *C. elegans* and *D. rerio*

There is a genomic and bioinformatic perspective from which a comparison between *C. elegans* and *D. rerio* is of interest in the context of this thesis. These considerations comprise both certain aspects of the organisms' genome structures and, more technically, the status of the genome assemblies that are the actual basis for much of the analytical and bioinformatic work.

The *C. elegans* genome sequence was published in 1998 (27) as the first genome of a multicellular organism. Today, the genome sequence is well established, so that differences between successive assembly releases are generally minuscule. Intriguingly, over the years, total estimates for protein-coding genes have steadily increased for *C. elegans*, while they have fallen dramatically for the human genome, so that they presently point to about 20,000 genes for either organism. This fact has to be taken into account when weighing the argument (see e.g. (13)) that distinct advantages

arise from studying cancer genes in *C. elegans* due to its less redundant gene networks. For example, the *C.elegans* gene *lin-35* is homologous to three human genes: *RB1* and its two paralogs *RBL1* (p107) and *RBL2* (p130). Similarly, the *C.elegans* gene *cep-1* is homologous to the human genes *TP53*, *TP63*, and *TP73*. On the other hand, the similar overall gene counts in both organisms indicate that there must be instances where the situation is actually reversed. Also, some *C. elegans* genes do not have recognizable vertebrate homologs at all, as illustrated by the case of *rde-4* (2), adding complexities to the mapping of *C. elegans*-based genetic models onto vertebrate gene networks.

The situation for the *D. rerio* genome is markedly different in several respects. The genome has undergone a relatively recent duplication event that happened at the base of the teleost radiation. As discussed and illustrated with the example genes of *postn* and *tgfbi* (1), this has resulted in the retention, followed by neo- or sub-functionalization of the duplicate copy, for some genes (*postn*) and the loss of one duplicate for others (*tgfbi*). This situation is reflected in zebrafish gene names that often carry the suffix 'a' or 'b' to denote the two paralogs. These are, however, not reliably recognized and labeled, as the example of zebrafish *postn* and its unnamed paralog illustrates. Biologically, an important consequence is that an additional layer of uncertainty impedes the determination of the teleost "functional ortholog" for a given mammalian gene.

The *D. rerio* genome sequencing project was started in 2001 and a first assembly was released in 2003 (28). Assemblies have been affected by polymorphisms originating from the large number of diploid embryos originally used for obtaining source DNA. Unlike for *C. elegans* or human, differences between successive zebrafish genome assemblies continue to be major to this day, and can necessitate major bioinformatic efforts in practice (see above and (3)). Parts of the genome sequence, varying between versions, are disjointed from any of the 25 chromosomes, adding to the challenges for bioinformatics and genomic data analysis.

From model organisms to model genomes

In this thesis, two studies (2; 3) summarize distinct research efforts carried out in one particular model organism, with the potential for applicability to human biology. By contrast, (1) represents a different mode of working with model organisms: it focuses on one particular gene, periostin (*POSTN*). Our study elucidates properties of this region by comparing periostin across vertebrate species, from mammals to teleosts, and tentatively beyond. Complete or even partial periostin gene models existed only for a fraction of the species examined, a shortcoming that was overcome with the help of genomic (and to a lesser degree, transcriptomic) sequence data.

This "vertical" approach – available in principle for any gene of interest – would not have been possible without the large and rapidly increasing number of genome sequences in the public domain. It is hence symptomatic for a new, genomically and bioinformatically informed perspective on model organisms.

Traditionally, scientific and practical considerations such as exemplary or characteristic biological features, abundance of offspring, short generation time, or ease of stock maintenance, and finally a sufficiently widespread consensus among the scientific community determined an organism's rise to "model organism prominence". But today, genomically, phylogenetically, or bioinformatically, it is possible for organisms to serve as research models based solely on characteristics of their genomic sequence. The following two examples illustrate this transition:

(i) The scientific community's convergence towards the thale cress (*Arabidopsis thaliana*), an angiosperm plant, as a model organism is even more recent than for *C. elegans* and *D. rerio*. When a high-profile case was made in 1985 (29), its genome was discussed in considerable detail. While whole-genome sequencing itself was not yet conceivable, genome features such as small size (implying a manageably small number of clones needed for a comprehensive library), low chromosome count, and low repeat content were explicitly put forward alongside "classical" advantages such as a short generation time, high seed counts, or the ease of obtaining homozygous plants due to self-fertilization.

(ii) Probably the first example of a model organism being advanced on an exclusively genomic basis is from 1993, when Sydney Brenner – once more – and his colleagues made the case for the characterization of a certain pufferfish (*Takifugu rubripes*) genome (30). Introducing the term "model vertebrate genome", they argued that this pufferfish's compact genome of 400 Mb put it within reach of existing or at least conceivable sequencing technology at the time and that its low complexity and high ratio of coding vs. non-coding sequence made it an ideal tool for the task of gene discovery in humans. Notably, contrary to zebrafish, pufferfish are not generally practical as model organisms – but the *T. rubripes* genome is almost four times smaller than the zebrafish genome.

After the announcement of the human genome sequence in 2001, the first decade of the 21st century was marked by an accelerating release of sequenced genomes, made possible by the rapidly progressing maturation of the Sanger sequencing-based technology and the concurrent ramp-up in capacity. Genome projects for established model organisms with large genomes (mouse, dog, claw frog, zebrafish) were soon complemented by those elected on the basis of phylogenetic interest (platypus, opossum, anole lizard) and, in parallel, of agricultural importance (cow, rice, grape).

Oncology research and the accelerating genomic revolution

Today, genomic research finds itself in a situation characterized by mature DNA microarray technology and by rapidly developing massively parallel sequencing technology. The latter is increasingly replacing traditional technologies, including for whole genomes, as stunningly evident from the sequence quantity statistics posted on the U.S. Department of Energy's genome sequencing project website (31). It is also making inroads for applications for which microarrays had been the gold standard for years (gene expression, copy number, mutations), a trend that is accelerated by the added promise to address questions not or only with difficulty accessible to microarray technology (e.g. translocations, splice variants).

In our case, follow-up work to (3) involves hundreds of zebrafish tumor samples being surveyed for copy number aberrations and relies on high-throughput sequencing rather than microarrays. Also, in our continued work in *C. elegans*, public datasets cross-correlated with project-generated data are increasingly sequencing-based and, accordingly, massive in size.

For model genomes and their analysis, this development has opened the door to characterizing within-species variation at unprecedented resolution. For human at least, the concept of "one organism – one genome sequence" as conveniently reflected in many popular genomic bioinformatics resources, is rapidly becoming outdated: human genomes are being sequenced widely, putting a spotlight on the extent of naturally occurring intra-species variation, the "variome".

Similarly for cancer, present oncogenomic research by far exceeds traditional piecemeal characterization efforts of individual select cancer-specific differences (mutations, over- and underexpressed genes) to approach the goal of comprehensively characterized, cancer type-specific model cancer genomes (plus their transcriptomes and variomes). It is intriguing to reflect on this trajectory as a vindication of sorts of the proposal by evolutionary biologist Leigh Van Valen – inscrutable regarding its intended seriousness – to classify cancer cell lines (specifically HeLa) as their own species (32). This idea reverberates today in studies tracking the evolutionary trajectories of cancer (33), and in instances of personal cancer genome sequencing in clinical research settings (34; 35), with the ultimate goal of deriving individualized and hence optimized therapy options from the results. Such "personalized medicine" and their rationale were presaged more than 35 years ago (36), but are becoming comprehensively feasible only today.

The rapid pace of genome sequencing of especially human genomes may obscure the fact that some non-human model genomes, including those of less established model organisms, are lagging behind to varying degrees, and that improvements are not in sight. The situation for the *D. rerio* genome has been discussed above. Much worse, other genomes remain without chromosome-based assemblies (claw frog), some of them so fragmented as to make even simple gene finding exercises challenging (elephant shark, lamprey), and their projects may well be at the end of their funded life cycles, effectively making them "orphans". This situation has frustrated distinct avenues of investigating periostin, when we attempted to identify the C-terminal region in phylogenetic lines outside of the euteleostomes (see Additional file 6 in (1)). It also continues to pose difficulties due to the problems in the genome assembly (described above), as our zebrafish studies progress (3).

It can be hoped that the new sequencing technologies may eventually offer a solution to this conundrum, once their cost is sufficiently low and their employment sufficiently routine. Recognizing that a phylogenetically broad collection of genome sequences that uniformly meet certain minimum standards is in the interest of the scientific community as a whole, it should be feasible to "brush up" low quality genome sequences by combining the existing with newly generated sequencing data, a model that has already been followed for the last zebrafish assembly. Maybe, a way will be found to do the same also for orphan genome projects.

Advances in our understanding of cancer biology today are significantly and systematically driven by model organisms and genomic research. This thesis illustrates some of the diverse research modalities following this concept, occupying distinct coordinates in a space defined by axes such as 'bioinformatics vs. experimental approaches', 'number of organisms considered', 'number of genes investigated', 'basic vs. applied research goals', and many others. And it nourishes the hope that ultimately, the challenge of fusing all the rapidly growing islands of insight into a cohesive understanding that can actually translate into a "cure for cancer" will be met.

Literatur-Referenzen / References

Die nachfolgende Literatur wird in der Zusammenfassung zitiert, aber in der Regel nicht in den für die Dissertation ausgewählten Publikationen. Ausnahmen betreffen Artikel, auf die in der Zusammenfassung ausführlicher eingegangen wird.

Einträge mit meiner Beteiligung als Koautor oder Miterfinder sind durch einen Stern * gekennzeichnet.

The following literature is cited in the thesis summary, but generally not in the publications selected for the thesis. Exceptions are made for articles that are subject to extended passages in the summary.

*References with my contribution (co-authorship or co-inventorship) are marked by an asterisk *.*

* **1. Hoersch S, Andrade-Navarro MA. Periostin shows increased evolutionary plasticity in its alternatively spliced region. BMC Evol. Biol 2010;10:30.**

* **2. Grishok A, Hoersch S, Sharp PA. RNA interference and retinoblastoma-related genes are required for repression of endogenous siRNA targets in Caenorhabditis elegans. Proc. Natl. Acad. Sci. U.S.A 2008 Dec;105(51):20386-20391.**

* **3. Zhang G, Hoersch S, Amsterdam A, Whittaker CA, Lees JA, Hopkins N. Highly aneuploid zebrafish malignant peripheral nerve sheath tumors have genetic alterations similar to human cancers. Proc Natl Acad Sci U S A 2010 Sep;107(39):16940-16945.**

4. Mukherjee S. The Emperor of All Maladies: A Biography of Cancer. 1st ed. Scribner; 2010.

5. Hanahan D, Weinberg RA. The Hallmarks of Cancer. Cell 2000 Jan;100(1):57-70.

6. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell 2011;144(5):646-674.

7. Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS. A gene expression map for Caenorhabditis elegans. Science 2001 Sep;293(5537):2087-2092.

* 8. Winslow MM, Dayton TL, Verhaak RGW, Kim-Kiselak C, Snyder EL, Feldser DM, Hubbard DD, DuPage MJ, Whittaker CA, Hoersch S, Yoon S, Crowley D, Bronson RT, Chiang DY, Meyerson M, Jacks T. Suppression of lung adenocarcinoma progression by Nkx2-1. Nature 2011 May;473(7345):101-104.

9. Tuveson DA, Jacks T. Technologically advanced cancer modeling in mice. Curr. Opin. Genet. Dev 2002 Feb;12(1):105-110.

10. Hirst GL, Balmain A. Forty years of cancer modelling in the mouse. Eur. J. Cancer 2004 Sep;40(13):1974-1980.

11. Gordon I, Paoloni M, Mazcko C, Khanna C. The Comparative Oncology Trials Consortium: using spontaneously occurring cancers in dogs to inform the cancer drug development pathway. PLoS Med 2009 Oct;6(10):e1000161.

12. Palidwor GA, Andrade-Navarro MA. MLTrends: Graphing MEDLINE term usage over time. J Biomed Discov Collab 2010;5:1-6.

13. Kirienko NV, Mani K, Fay DS. Cancer models in Caenorhabditis elegans. Dev. Dyn 2010 May;239(5):1413-1448.

* 14. Hoersch S, Anderson DL, Glatt KA, Xu Y, Wang Y, Toker A, Damokosh AI, Endege WO, Storz P, Bryant BM, Monahan JE. Alternative splicing of periostin in human breast cancer [Internet]. (manuscript in preparation).

* 15. Monahan JE, Hoersch S, Anderson DL, Endege WO, Ford D, Glatt K, Gorbatcheva BO, Kamatkar S, Xu YY, Gannavarapu M, Zhao X, Schlegel R, Mertens M, Bast, Jr. RC, Hortobagyi GN, Pusztai L. United States Patent: 7601505 - Compositions, kits, and methods for identification, assessment, prevention, and therapy of breast cancer. 2009.

* 16. Mansisidor AR, Cecere G, Hoersch S, Jensen MB, Kawli T, Kennedy LM, Chavez V, Tan M-W, Lieb JD, Grishok A. A Conserved PHD Finger Protein and Endogenous RNAi Modulate Insulin Signaling in C. elegans. PLoS Genetics, accepted for publication.

* 17. Vasudevan KM, Barbie DA, Davies MA, Rabinovsky R, McNear CJ, Kim JJ, Hennessy BT, Tseng H, Pochanard P, Kim SY, Dunn IF, Schinzel AC, Sandy P, Hoersch S, Sheng Q, Gupta PB, Boehm JS, Reiling JH, Silver S, Lu Y, Stemke-Hale K, Dutta B, Joy C, Sahin AA, Gonzalez-Angulo AM, Lluch A, Rameh LE, Jacks T, Root DE, Lander ES, Mills GB, Hahn WC, Sellers WR, Garraway LA. AKT-

independent signaling downstream of oncogenic PIK3CA mutations in human cancer. Cancer Cell 2009 Jul;16(1):21-32.

* 18. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, Schinzel AC, Sandy P, Meylan E, Scholl C, Fröhling S, Chan EM, Sos ML, Michel K, Mermel C, Silver SJ, Weir BA, Reiling JH, Sheng Q, Gupta PB, Wadlow RC, Le H, Hoersch S, Wittner BS, Ramaswamy S, Livingston DM, Sabatini DM, Meyerson M, Thomas RK, Lander ES, Mesirov JP, Root DE, Gilliland DG, Jacks T, Hahn WC. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 2009 Nov;462(7269):108-112.

19. Brenner S. The genetics of Caenorhabditis elegans. Genetics 1974 May;77(1):71-94.

20. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 1998 Feb;391(6669):806-811.

21. Simpson AJG, Caballero OL, Jungbluth A, Chen Y-T, Old LJ. Cancer/testis antigens, gametogenesis and cancer. Nat Rev Cancer 2005;5(8):615-625.

22. Janic A, Mendizabal L, Llamazares S, Rossell D, Gonzalez C. Ectopic expression of germline genes drives malignant brain tumor growth in Drosophila. Science. 2010 Dec 24;330(6012):1824-1827.

23. Streisinger G, Walker C, Dower N, Knauber D, Singer F. Production of clones of homozygous diploid zebra fish (Brachydanio rerio). Nature 1981 May;291(5813):293-296.

24. Amatruda JF, Shepard JL, Stern HM, Zon LI. Zebrafish as a cancer model system. Cancer Cell 2002 Apr;1(3):229-231.

25. Amsterdam A, Sadler KC, Lai K, Farrington S, Bronson RT, Lees JA, Hopkins N. Many ribosomal protein genes are cancer genes in zebrafish. PLoS Biol 2004 May;2(5):E139.

26. Stanton MF. Diethylnitrosamine-induced hepatic degeneration and neoplasia in the aquarium fish, Brachydanio rerio. J. Natl. Cancer Inst 1965 Jan;34:117-130.

27. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 1998 Dec;282(5396):2012-2018.

28. Jekosch K. The zebrafish genome project: sequence analysis and annotation. Methods Cell Biol 2004;77:225-239.

29. Meyerowitz EM, Pruitt RE. Arabidopsis thaliana and Plant Molecular Genetics. Science 1985 Sep;229(4719):1214-1218.

30. Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh B, Aparicio S. Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. Nature 1993 Nov;366(6452):265-268.

31. http://www.jgi.doe.gov/sequencing/statistics.html (accessed May 2011).

32. Van Valen LM, Maiorana VC. HeLa, a new microbial species. http://dl.dropbox.com/u/18310184/about-leigh-van-valen/Piglet%20Papers/1991%20HeLa.pdf 1991 (accessed May 2011).

33. Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, Andrews JM, Buck G, Chen L, Beare D, Latimer C, Widaa S, Hinton J, Fahey C, Fu B, Swamy S, Dalgliesh GL, Teh BT, Deloukas P, Yang F, Campbell PJ, Futreal PA, Stratton MR. Signatures of mutation and selection in the cancer genome. Nature 2010 Feb;463(7283):893-898.

34. Ding L, Ellis MJ, Li S, Larson DE, Chen K, Wallis JW, Harris CC, McLellan MD, Fulton RS, Fulton LL, Abbott RM, Hoog J, Dooling DJ, Koboldt DC, Schmidt H, Kalicki J, Zhang Q, Chen L, Lin L, Wendl MC, McMichael JF, Magrini VJ, Cook L, McGrath SD, Vickery TL, Appelbaum E, Deschryver K, Davies S, Guintoli T, Lin L, Crowder R, Tao Y, Snider JE, Smith SM, Dukes AF, Sanderson GE, Pohl CS, Delehaunty KD, Fronick CC, Pape KA, Reed JS, Robinson JS, Hodges JS, Schierding W, Dees ND, Shen D, Locke DP, Wiechert ME, Eldred JM, Peck JB, Oberkfell BJ, Lolofie JT, Du F, Hawkins AE, O'Laughlin MD, Bernard KE, Cunningham M, Elliott G, Mason MD, Thompson DM Jr, Ivanovich JL, Goodfellow PJ, Perou CM, Weinstock GM, Aft R, Watson M, Ley TJ, Wilson RK, Mardis ER. Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature 2010 Apr;464(7291):999-1005.

35. Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D, Ha C, Johnson S, Kennemer MI, Mohan S, Nazarenko I, Watanabe C, Sparks AB, Shames DS, Gentleman R, de Sauvage FJ, Stern H, Pandita A, Ballinger DG, Drmanac R, Modrusan Z, Seshagiri S, Zhang Z. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature 2010 May;465(7297):473-477.

36. Nowell PC. The clonal evolution of tumor cell populations. Science 1976 Oct;194(4260):23-28.

Anteilserklärung

Sebastian Hörsch hatte folgenden Anteil an den vorgelegten Publikationen:

Publikation 1: **RNA interference and retinoblastoma-related genes are required for repression of endogenous siRNA targets in *Caenorhabditis elegans*.**
Proceedings of the National Academy of Sciences of the United States of America 105(51): 20386-20391 (2008).
Alla Grishok, Sebastian Hoersch, and Phillip A Sharp.

40 Prozent

Beitrag im Einzelnen:

Idee und Gesamtkonzept für diese Arbeit stammen von Erstautorin Alla Grishok und von Professor Philip A. Sharp; das Detailkonzept wurde unter Sebastian Hörschs Mithilfe entwickelt.

Alle Experimente und die Analyse der PCR-Experimente wurden von A.G. durchgeführt.

Die gesamte bioinformatische Arbeit wurde von S.H. durchgeführt, im Detail: Datenprozessierung und Normalisierung der DNA-Mikroarray-Daten (mit Unterstützung durch Charles A. Whittaker); die differentielle Expressionsanalyse der Mikroarray-Daten; die Bewertung einer Eignung von „Gene Ontology"-basierten funktionellen Annotationen für dieses Projekt; die Anpassung von TOPOMAP als Plattform für funktionelle Annotationen und die Veröffentlichung entsprechender allgemein nutzbarer Dateien; der Einsatz von Methodik, um die Überschneidung von Gengruppen innerhalb von TOPOMAP statistisch zu bewerten; die Adoption von publizierten Datensätzen, insbesondere von endo-siRNAs, aber auch von anderen, welche nur teilweise Bestandteil der Publikation sind.

Zusätzlich zu Projekt-Besprechungen zwischen den Koautoren P.A.S. und A.G. (mit S.H.s gelegentlicher Teilnahme) trafen sich A.G. und S.H. regelmäßig für detaillierten Diskussionen, bei denen – oft unter S.H.s Leitung – Zwischenergebnisse bewertet wurden und über Untersuchungsmethodik oder zusätzliche publizierte Datensätze entschieden wurde.

Das Manuskript wurde von A.G. aufgesetzt und mit Unterstützung von P.A.S. und S.H. überarbeitet. Die Abschnitte bezüglich der bioinformatischen Methodik wurden von S.H. verfasst. Die Abbildungen 1, S1, S2 sowie alle drei ergänzenden Datentabellen wurden von S.H. erstellt.

Die Ablage der Mikroarray-Genexpressionsdaten in der öffentlichen Datenbank (Gene Expression Omnibus, GSE13258) wurde von S.H. durchgeführt.

Publikation 2: **Periostin shows increased evolutionary plasticity in its alternatively spliced region.**
BMC Evolutionary Biology 10: 30 (2010).
Sebastian Hoersch and Miguel A Andrade-Navarro.

90 Prozent

Beitrag im Einzelnen:

Die Idee für diese Studie stammt von S.H.; das Konzept für ihe Durchführung stammt ebenfalls von S.H. und wurde durch Beiträge von Koautor Miguel A. Andrade-Navarro verfeinert. Alle Analysen wurden von S.H. durchgeführt, mit der Ausnahme der PSI-BLAST-Analyse, die zusammen mit M.A.A. durchgeführt wurde. Die Arbeit wurde mit M.A.A. auf kontinuierlicher Basis diskutiert. Das Manuskript mit allen Abbildungen und zusätzlichen Materialien wurde von S.H. verfasst und unter Berücksichtigung der Kommentare von M.A.A. in seine endgültige Form gebracht.

Publikation 3: **Highly aneuploid zebrafish malignant peripheral nerve sheath tumors have genetic alterations similar to human cancers.**
Proceedings of the National Academy of Sciences of the United States of America 107(39): 16940-16945 (2010).
GuangJun Zhang, Sebastian Hoersch, Adam Amsterdam, Charles A. Whittaker, Jacqueline A Lees, and Nancy Hopkins.

30 Prozent

Beitrag im Einzelnen:

Idee und Gesamtkonzept für diese Studie stammen von Erstautor GuangJun Zhang und Koautoren Adam Amsterdam, Professor Jacqueline A. Lees und Professor Nancy Hopkins; das Detailkonzept wurde mit S.H.s Mithilfe entwickelt. Alle experimentellen Arbeiten wurden von G.Z. und A.A. durchgeführt, die auch die Datenanalyse für die Durchflusszytometrie, das „Southern Blotting" und Chromosomezählungen durchführten.

Die gesamte bioinformatische Arbeit wurde von S.H. durchgeführt, mit Ausnahme der Prozessierung der massiv parallelen Sequenzierungsdaten, die von Charles A. Whittaker vorgenommen wurde. Im Detail führte S.H. die folgenden Untersuchungen durch: die Evaluierung verschiedener Designstrategien für das eigens für diese Studie konzipierte CGH-Mikroarray für Zebrafisch, Auslegung des tatsächlich hergestellten und verwendeten Arrays auf der Basis von etwa fünf Millionen Oligonukleotid-Sequenzen (von der Firma Agilent zur Verfügung gestellt), einschließlich der nötigen Sequenzanalyse, um Qualität, Einzigartigkeit, und Kompatibilität der Sondensequenzen mit dem verwendeten experimentellen Protokoll sicherzustellen; die erneute Analyse der Array-Sondensequenzen im Kontext einer neuen Version der Zebrafisch-Genomsequenz; Prozessierung und Normalisierung der Arraydaten (Agilent Feature Extraction); und die Analyse der Array- und der Sequenzierungs-Daten im Hinblick auf chromosomale Abweichungen vor und nach erfolgter Segmentierung (DNAcopy, STAC).

Bestimmte Details des experimentellen Designs dieser Studie wurden unter S.H.s Leitung erarbeitet, insbesondere die kontinuierliche Verwendung von gepaarten Kontrollen (zusätzlich zu Tumor-Kontroll-Paaren), um zu einer verbesserten Einschätzung des Grundrauschens in den Daten zu gelangen. Diese gepaarten Kontrollen werden im Artikel zwar nicht erwähnt, ihre Datensätze wurden aber – zusammen mit denen der Tumor-Kontroll-Paaren – in der öffentlichen Datenbank abgelegt. Auch erfolgte die Wahl massiv paralleler Sequenzierung als Validierungsmethode für die Arraydaten auf S.H.s (und C.A.W.s) Initiative hin.

S.H. nahm regelmäßig an Projektbesprechungen teil, die üblicherweise die Koautoren G.Z., A.A., N.H. und S.H. einbezogen. Während dieser brachte S.H. häufig seine bioinformatische Erfahrung ein, auch bezüglich der Machbarkeit und vorläufiger Ergebnisse für einen potentiellen zukünftigen Vergleich zwischen den chromosomalen Veränderungen im Menschen und Zebrafisch.

Das Manuskript wurde zusammen von G.Z., A.A., N.H. und S.H. geschrieben, die bioinformatische Methoden betreffenden Abschnitte wurden von S.H. verfasst (unter Berücksichtigung von C.A.W.s Kommentaren). Abbildungen 2 und 3A & B wurden von S.H. erstellt.

Schließlich wurden die von den Mikroarrays und der massiv-parallelen Sequenzierung stammenden Datensätze von S.H. in der öffentlichen Datenbank abgelegt (Gene Expression Omnibus, GSE23666).

Sebastian Hörsch

Hoersch and Andrade-Navarro *BMC Evolutionary Biology* 2010, **10**:30
http://www.biomedcentral.com/1471-2148/10/30

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Periostin shows increased evolutionary plasticity in its alternatively spliced region

Sebastian Hoersch[1,2*], Miguel A Andrade-Navarro[3]

Abstract

Background: Periostin (POSTN) is a secreted extracellular matrix protein of poorly defined function that has been related to bone and heart development as well as to cancer. In human and mouse, it is known to undergo alternative splicing in its C-terminal region, which is devoid of known protein domains. Differential expression of periostin, sometimes of specific splicing isoforms, is observed in a broad range of human cancers, including breast, pancreatic, and colon cancer. Here, we combine genomic and transcriptomic sequence data from vertebrate organisms to study the evolution of periostin and particularly of its C-terminal region.

Results: We found that the C-terminal part of periostin is markedly more variable among vertebrates than the rest of periostin in terms of exon count, length, and splicing pattern, which we interpret as a consequence of neofunctionalization after the split between periostin and its paralog transforming growth factor, beta-induced (TGFBI). We also defined periostin's sequential 13-amino acid repeat units - well conserved in teleost fish, but more obscure in higher vertebrates - whose secondary structure is predicted to be consecutive beta strands. We suggest that these beta strands may mediate binding interactions with other proteins through an extended beta-zipper in a manner similar to the way repeat units in bacterial cell wall proteins have been reported to bind human fibronectin.

Conclusions: Our results, obtained with the help of the increasingly large collection of complete vertebrate genomes, document the evolutionary plasticity of periostin's C-terminal region, and for the first time suggest a basis for its functional role.

Background

Periostin (POSTN, PN, OSF-2) is a secreted extracellular matrix (ECM) glycoprotein of up to 93 kDa with a role in cell adhesion. It was originally identified in cells of mesenchymal lineage - osteoblasts, osteoblast-derived cell lines, the periosteum, and the periodontal ligament [1,2]. Its role in the development of bones, teeth, and cartilage has been documented subsequently (e.g. [3-7]). Furthermore, periostin has in recent years prominently emerged as important on two distinct fronts, both of them with notable clinical implications: One is its differential expression in a wide array of epithelial tumors compared to their respective normal tissues (for reviews, see [8,9]). For the majority of cancers investigated, periostin expression was found to be increased over normal tissue, but there are distinct exceptions where this pattern is reversed. The other relevant area is its expression in the developing and the diseased heart (for reviews, see [10-13]). In the developing heart, periostin has been found to be expressed in distinct substructures. In pathological heart conditions, periostin expression has been described both in the context of acute events (myocardial infarction) [14,15], as well as chronic pathological conditions (pressure overload) [15,16].

A comprehensive understanding of periostin's functional spectrum is still actively developing, but certain core aspects emerging from those three major areas (skeletal development, heart development and disease, and cancer) are coming increasingly into focus. Studies have associated periostin with epithelial-mesenchymal transition (EMT) in cancer [17,18] and with mesenchymal differentiation [19-22] in the developing heart.

Early characterizations of periostin as an adhesion protein on the basis of its apparent homology to insect

* Correspondence: hoersch@mit.edu
[1]Bioinformatics and Computing Core, Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

Hoersch and Andrade-Navarro *BMC Evolutionary Biology* 2010, **10**:30
http://www.biomedcentral.com/1471-2148/10/30

fasciclin [23] have subsequently been refined by expanding the collection of known ECM binding partners (see below), and secondly by illuminating functional aspects consistent with a role in contact-signaling [12]. Periostin protein, whose expression has been found to be promoted for example by TGFbeta1, 2, or 3 and BMP-2 in multiple studies [2,21,24-28], can bind to certain integrin receptors, which subsequently activate the Akt/PKB pathway via FAK and PI3K, leading to an enhanced migratory or invasive phenotype [14,29-33].

The periostin gene in human and mouse has 23 exons, with a genomic footprint covering about 36 (or 30) kilobases in human (or mouse). Both terminal exons in mouse and human are protein-coding. The periostin protein (see Figure 1) has an N-terminal signal sequence in accordance with its status as a secreted protein, an EMI domain, and four Fasciclin (FAS1) domains. The EMI domain, encoded by exons 2 and 3, is thought to be involved in protein-protein interactions or protein multimerization [34], and may be responsible for periostin dimers observed in some studies (e.g. [31,35]). A very recent report described EMI domain-mediated dimers of periostin and its paralog TGFBI (see below) [36], making it likely that the dimers observed in various periostin studies are indeed physiologically relevant.

Besides integrins, periostin has been described to bind a number of other ECM proteins, for example heparin [37], fibronectin [38], and collagen I [35,39], although the precise binding mechanisms are not defined in these cases. The four fasciclin (FAS1) domains, described as a cell adhesion module [23], are encoded by exons 3 to 14. As recently reported, each of these FAS1 domains contains an N-terminal recognition site for γ-glutamylcarboxylase, which mediates the post-translational modification of glutamate to γ-carboxyglutamate [40]. Also, integrin binding motifs are found in the second and fourth FAS1 domain, as inferred from findings of the periostin paralog transforming growth factor, beta-induced (TGFBI, also BIG-H3 or betaig-h3) [41].

Finally, exon 15 is followed by exons 16 - 23, making up 182 amino acids (in human) and thus a rather substantial part of the protein. The function of this stretch, sometimes referred to descriptively as "hydrophilic region" or, as from here onwards, "C-terminal region", is essentially unknown, although some aspects of its potential role have been investigated [42]. It is devoid of known domains and contains few known sequence motifs: (overlapping) regions of low compositional complexity and of intrinsic disorder (obviously interdependent features) and a C-terminal nuclear localization signal [43], which appears at odds with periostin's status as a secreted protein.

Interestingly, alternative splicing was described early on for human and murine periostin exclusively in this C-terminal region [1,2]. This is also suggested by EST sequence data for human or mouse periostin, where exons 17 to 21 present themselves as cassette exons that can be excluded from mature RNA message in various combinations, or individually.

The periostin paralog TGFBI is a protein with a domain structure identical to periostin (Figure 1) with the following exception: TGFBI, comprising only 17

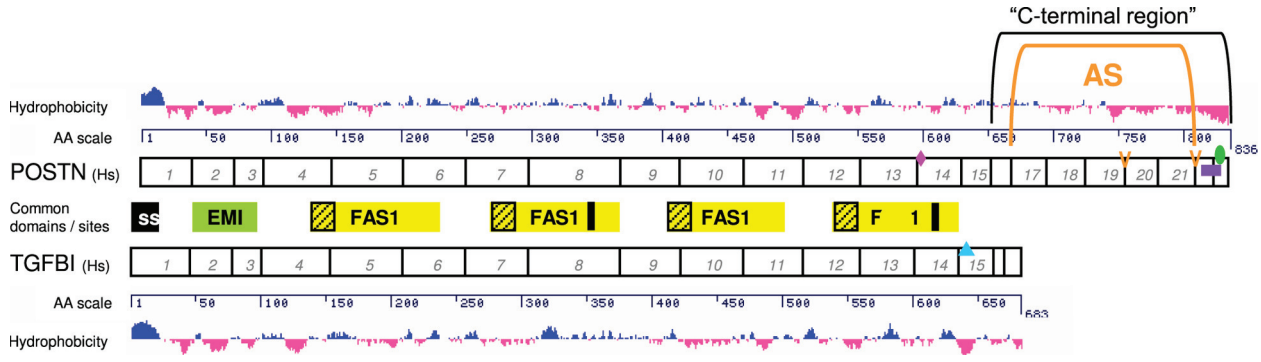


Figure 1 Periostin and TGFBI exon and domain structure. Numbered periostin (POSTN) and TGFBI exons (coding sequence only) are displayed to scale. Amino acid (AA) position scales and hydrophobicity profiles (adapted from the UCSC proteome browser [84]) for either protein are displayed above and below the exon structures, respectively. Domains and features in common are displayed in-between: a signal sequence (ss), an EMI domain, and four FAS1 domains. Vertical black bars in FAS1 domains 2 and 4 mark the integrin binding sites with conserved central DI dimers. The shaded regions in the N-terminal end of the FAS1 domains represent γ-glutamylcarboxylase recognition sites. Additional features specific to either of the two proteins are indicated by markers positioned in the respective exon structure as follows. Pink diamond: POSTN N-glycosylation site; blue triangle: TGFBI integrin RGD binding site; green oval: POSTN heparin binding motif (suspected); purple rectangle: POSTN bipartite nuclear localization signal. The two orange V's mark sites of genomic variation in periostin of other tetrapods: between exons 19 and 20, a cluster of 8 additional exon 19 copies observed in *X. tropicalis*, and an additional exon "21V22" between exons 21 and 22, observed for example in chicken. The periostin region between exons 16 and 22 is flagged as subject to widespread alternative splicing ("AS"), and the extent of the "C-terminal region" as referred to in this work is indicated. See the main text for detailed descriptions.

exons, is notably shorter and lacks completely the C-terminal region that is subject to alternative splicing in periostin. Periostin and TGFBI sequences diverge fundamentally after the fourth fasciclin domain (i.e. starting at exon 15). An Arg-Gly-Asp (RGD) binding motif for a subgroup of integrin adhesion receptors (reviewed in [44]) is found in TGFBI exon 15, which is followed by two short exons 16 and terminal exon 17. Interestingly, TGFBI $\alpha 3\beta 1$-integrin binding has been shown to be mediated not via the canonical RGD binding motif, but via two pentapeptides containing a central Asp-Ile (DI) dimer found in the second and fourth FAS1 domain [41], which were subsequently found to be conformationally similar to RGD peptides [45]. While an RGD motif is not present in periostin, the DI dimers in FAS1 domains 2 and 4 are conserved between the two proteins, strongly suggesting a mechanism for periostin integrin binding.

Much of the literature on TGFBI is dedicated to certain relatively common mutations causing a variety for corneal dystrophies (reviewed e.g. in [46]), a functional aspect for which there is no known counterpart for periostin. Like periostin, TGFBI is a TGFbeta-induced, secreted, integrin-binding ECM protein expressed during cardiac development and differentially expressed in certain epithelial cancers (e.g. [47]). But there is also mounting evidence that periostin and TGFBI can have complementary or opposite roles in cancer and development. For cancer, reports highlight a tumor-suppressive role of TGFBI (e.g. [48,49]), and complementary expression patterns of periostin and TGFBI in the developing heart were explicitly studied [50].

Here we present the results of our computational analysis of periostin's C-terminal region (exons 16 - 23), which constitutes the most visible difference between periostin and its paralog TGFBI and is hence likely to be a major determinant of the functional differences between these two genes and of periostin function per se.

Against the backdrop of unavailable functional or structural annotation and of indications for greater phylogenetic variability relative to the much better annotated rest of the gene, our interest was increased by the pervasive alternative splicing specific to this part of periostin. Periostin isoform heterogeneity is a source of significant complexity in the rapidly growing body of research on this gene, exacerbated by the fact that the precise nature of the variants observed or used experimentally is not always readily apparent. But cell-specific periostin isoform profiles have been demonstrated early on [2], and isoform-specific biological properties have been documented subsequently (e.g. [14,51,52]). It is hence not an exaggeration to expect that an improved understanding of periostin's C-terminal region in general and its alternative splicing patterns in particular may help resolve the sometimes controversial findings on periostin's biological effects.

Methods

Profile-based homology search

We compiled an alignment of two consecutive repeats from fragment 675-700 of periostin isoform 1 [*Danio rerio*] (RefSeq:NP_001071254.1) with fragments from another two sequences from *D. rerio* and one from *Xenopus tropicalis* found via BLAST [53]. The alignment was done using ClustalW [54] and manual editing. This alignment was converted into a profile and used for a search against UniRef100 with the program hmmsearch [55]. Iterative searches and addition of new positives led to a progressively refined profile with repeats from *Tetraodon nigroviridis*, *Takifugu rubripes*, and chicken, which matched mammalian periostin sequences (including human) at E-value 1.8 (3rd iteration). No false positive was observed below this E-value.

Periostin locus identification and sequence reconstruction (assembled genomes)

Periostin loci in genome assemblies of interest (detailed in Additional file 1, Table S1) were identified in the UCSC genome browser [56] (accessible at http://genome.ucsc.edu/) via a combination of annotation tracks [57], where possible. Otherwise, the BLAT [58] search function within the genome browser was used to search with known periostin sequences from other organisms.

If available, transcript data (UCSC tracks '(species) mRNAs from GenBank' and '(species) ESTs That Have Been Spliced' were used to synthesize full-length periostin sequences with a complete set of exons, commonly resulting in full-length sequences with a higher exon count than found in available RefSeq sequences (Mm, Gg, Xt, Dr).

We identified individual missing exons or refined the boundaries of individual exons with the help of BLAT-based alignments and the 'Vertebrate Multiz Alignment & Conservation track' [59], which provides sequence-level genomic alignments for many vertebrate genomes. For teleosts, we used periostin gene predictions from the Ensembl database [60] (accessible at http://www.ensembl.org), if available, in combination with transcript evidence or BLAST-based efforts to expand exon coverage.

For genome assemblies with insufficient or no transcript coverage of periostin, we performed TBLASTN [53] searches with a known full-length periostin sequence (usually human or chicken periostin for tetrapods and zebrafish locus 1 periostin for teleosts) as query against a genomic sequence fragment comprising the periostin locus as a subject sequence, changing default BLAST parameters to not mask repeats and to

report hits with E-values higher than 10.0. TBLASTN hits were then evaluated in an exon-by-exon fashion: exact exon boundaries were defined aided by the 'Vertebrate Multiz Alignment & Conservation track' and on the basis of assuming splice site conservation with canonical GT-AG intronic splice sites in all cases as common practice in the comparative genomics field [61], and the correct sequential order of the exons in the genomic sequences was ascertained. It is hence important to bear in mind the putative nature of this group of predicted periostin sequences.

Multiple sequence alignments and phylogenetic trees

All multiple sequence alignments were performed using the ClustalW algorithm as implemented in the ClustalX (version 2.0.11) software package [54], using default parameters, in particular using Neighbor Joining as the clustering algorithm. Bootstrapping values were obtained based on 1000 trials. Alignment-based phylogenetic trees were generated with gapped positions excluded and visualized with NJplot (described in [62]).

Genomic sequence alignments

In addition to the seven tetrapod and five teleost species listed in Additional file 1, Table S1, we selected 16 additional species to study periostin exon 21V22 in the context of an alignment of genomic sequences containing the periostin locus.

Five additional genomic sequences were obtained from the UCSC genome browser [56] for dog (*Canis familiaris*, assembly canFam2), cow (*Bos taurus*, assembly bosTau4), horse (*Equus caballus*, assembly equCab2), zebra finch (*Taeniopygia guttata*, assembly taeGut1), and tentatively for lamprey (*Petromyzon marinus*, assembly petMar1).

Another 11 genomic sequences were obtained with the help of the Ensembl genome browser (http://www.ensembl.org) from generally less mature genome assemblies of other organisms, after the periostin locus was identified using the 'Projected human gene' track. These organisms were: armadillo (*Dasypus novemcinctus*), dolphin (*Tursiops truncates*), European hedgehog (*Erinaceus europaeus*), hyrax (*Procavia capensis*), kangaroo rat (*Dipodomys ordii*), lesser hedgehog tenrec (*Echinops telfairi*), megabat (*Pteropus vampyrus*), microbat (*Myotis lucifugus*), squirrel (*Spermophilus tridecemlineatus*), tarsier (*Tarsius syrichta*), and tree shrew (*Tupaia belangeri*).

The genome sequence alignments were then performed using the VISTA suite of computational genomics web tools [63] (accessible at http://genome.lbl.gov/vista/), using the mVISTA option as appropriate for aligning and comparing genomic sequences from multiple species. Within mVISTA, 'Shuffle-LAGAN' [64] was chosen as the alignment program, no repeat masking was performed, and "translated anchoring" was selected for its reported potential to improve alignment of distant homologs. VISTA was run twice, with two different "reference sequences" that the alignment is anchored to: genomic periostin sequence from chicken was used as reference sequence to study exon 21V22 (Additional file 2, Figure S1 and Additional file 3, Table S2A), and genomic periostin sequence from human was used to study exon 17 (Additional file 3, Table S2B).

Sequence logos

Sequence logos were generated using the internet-based tool WebLogo (version 2.8.2) [65] (accessible at http://weblogo.berkeley.edu/), without small sample correction and with a customized color scheme.

For tetrapod exon sequence logos, exon-specific alignment blocks from seven tetrapod species (Additional file 1, Table S1) were submitted individually for sequence logo generation. The teleost repeat sequence logo was obtained by manually aligning the individual 13 amino acid repeat units within exon 18=19 from *D. rerio*, locus 1 (6 repeat units) and locus 2 (13 repeat units), *G. aculeatus*, locus 1 (7 repeat units) and locus 2 (5 repeat units), *T. nigroviridis*, locus 2 (8 repeat units), and *T. rubripes*, locus 2 (6 repeat units), totaling 45 repeat units. Alignment involved the occasional introduction of gaps and, in one case, the deletion of one amino acid from one repeat sequence to avoid the introduction of a gap position into the repeat alignment. This repeat alignment block was then duplicated (side-by-side) before being submitted for sequence logo generation.

Repeat visualization with dot-matrix plots

To visualize repetitive patterns within sequences with 2-dimensional matrix dot plots, the JAVA application JDotter [66] (accessible at http://athena.bioc.uvic.ca/tools/JDotter) was used. For genomic DNA sequences, (Additional file 4, Figure S2) we used a sliding window of size 50 (default) and a DNA scoring matrix of +5 for a match and -4 for a mismatch. Dot plots were originally generated at a resolution of 1 nucleotide per pixel and displayed with the following parameters for the "GreyMap Tool" to optimize visibility of the repeats in the display: 0 (top)/35 (bottom).

For protein sequences, we used a sliding window of size 5 and the amino acid scoring matrix BLOSUM62. The dot plot was generated at a resolution of 1 residue per pixel and with default "GreyMap Tool" parameters 0 (top)/245 (bottom) for the matrix display to obtain the maximum dynamic range of grey-values. The grey value continuum underlying this matrix display was then converted into a color continuum (white-yellow-orange-red-black) using standard image manipulation software.

Secondary structure predictions

For secondary structure predictions, we used PsiPred [67] implemented as a web-based tool [68] (accessible at http://bioinf.cs.ucl.ac.uk/psipred/). PsiPred v2.5 was used

for all predictions. Predictions were performed without any filtering options selected, in particular without the option to mask low complexity regions (selected by default).

Determination of periostin exon 21 and exon 21V22 in Xenopus laevis

For 21 EST sequences from *X. laevis* periostin, the UCSC browser indicated additional sequence without genomic match at a position coincident with an assembly gap between scaffold_505_18 and scaffold_505_19. The EST subsequences without genomic coverage were analyzed and determined to be identical, with a length of 81 nucleotides. Protein sequence comparisons with chicken periostin showed that the unmatched subsequence is periostin exon 21 and that the following exon (seen in 9 out of a total of 33 EST sequences covering this region) is actually exon 21V22 (data not shown).

Results

Exon naming conventions

Periostin is universally described as a 23-exon gene. Our findings, going beyond the scope of human and mouse as the organisms in which periostin has been investigated, expand and modify this notion on multiple counts. To keep our exon designations consistent with those in the existing literature while maintaining the capability to capture the new findings presented here, we decided on the following rules:

We consider exons 1 - 23 as described for human and murine periostin "canonical" and leave their numbering unaltered. We refer to homologous exons in other species by the same numbers, even if variations in exon count were found to exist.

Additional or modified exons without direct counterparts in human or mouse are referred to by special designations we are introducing here (see also Additional file 1, Table S1).

The tetrapod periostin C-terminal region is homologous to a repetitive part of teleost fish periostin

Our early attempts to identify homologs of known function to the C-terminal part of human or murine periostin using sequence similarity searches in protein databases produced exclusively hits to corresponding regions in tetrapod periostin sequences. An iterative search (PSI-BLAST) with this region (exons 16 - 23) alone produced an alignment (data not shown), with a sequence fragment of zebrafish (*Danio rerio*) periostin marked by a highly conserved 5-fold 13-amino acid repeat [39], covering exons 16 - 20 of the human periostin query.

In order to ascertain actual homology between the obviously repetitive part of zebrafish periostin and the C-terminal part of mammalian periostin, we conducted a profile-based homology search with repeat units of the zebrafish periostin (see Methods for details). We found that a search with only one repeat unit resulted in spurious hits, i.e., protein sequences of generically repetitive nature, e.g., myosin. However, using as few as two repeat units together for the search produced as the top hits specifically periostin C-terminal sequences from other vertebrate organisms, including human and mouse.

These results confirmed an evolutionary relationship between mammalian periostin C-terminal sequence and a zebrafish periostin repeat region, but since no significant hits to other proteins were found, it also marked the "end of the road" for attempts to identify similarities to gene sequences other than periostin for a tentative functional annotation. We hence decided to systematically compare complete periostin sequences and especially their C-terminal region across a range of phylogenetically diverse vertebrates in the hope of gaining clues to the functional role of the C-terminal region.

Full length periostin sequence, including the alternatively spliced part, can be inferred for many vertebrates with genomic data

Besides human and mouse as representatives of placental (eutherian) mammals, we selected as the basis for detailed study of periostin the complete genome sequences of the following tetrapod species: gray short-tailed opossum (*Monodelphis domestica*), platypus (*Ornithorhynchus anatinus*), chicken (*Gallus gallus*), green anole lizard (*Anolis carolensis*), and western clawed frog (*Xenopus tropicalis*). Furthermore, teleost fish genomes were added, namely of zebrafish (*Danio rerio*), stickleback (*Gasterosteus aculeatus*), medaka (*Oryzias latipes*), and two pufferfish, *Tetraodon nigroviridis* and *Takifugu rubripes*, adding up to 12 vertebrate - or, more narrowly, euteleostome - species.

Full-length periostin sequences comprising all exons are only readily available or annotated in public sequence databases for a subset of these vertebrate species. This was the case for human, mouse, chicken, frog, and a subset of teleost fish, although for some of these, several mRNA and/or EST sequences had to be combined to obtain complete sequences, or quality deficiencies in the genomic assembly required additional efforts (in *X. tropicalis*). For the remaining species, the periostin coding sequence was inferred using a combination of approaches (detailed in Methods). A detailed account of the origin and status of each species' periostin protein sequence is provided in Additional file 1, Table S1, and the actual sequences are provided in a comprehensive listing in Additional file 5. In the same table, we summarize existing evidence for periostin transcript variation and for other unusual features of periostin gene structure in the 12 species selected for detailed analysis.

Hoersch and Andrade-Navarro *BMC Evolutionary Biology* 2010, **10**:30
http://www.biomedcentral.com/1471-2148/10/30

We will return to a detailed analysis of the 13 amino acid repeat motif later in the manuscript and focus first on transcriptional and genomic evidence for periostin isoforms.

Alternative splicing events specific to the periostin C-terminal region are observed in all euteleostome lines with sufficient transcript evidence, including teleost fish

Alternative splicing affecting the region from exon 17 to exon 21 has been described for human [1] and murine periostin [2]. By evaluating available transcript evidence for the vertebrate species considered in this work, we universally found evidence for alternative splicing specific to this region for all species with non-trivial amounts of sequence data (see Additional file 1, Table S1 for details). This includes three fish species (zebrafish, stickleback, and medaka). Note that Additional file 1, Table S1 lists also transcript evidence for alternative periostin 5' and 3' ends observed in some species, including human.

Available periostin transcript evidence suggests that the set of exons subject to alternative splicing is also species-dependent (Additional file 1, Table S1). This observation is subject to considerable uncertainty due to limited transcript coverage, but human and murine periostin are compelling examples, having substantial amounts of transcript evidence. We found that in human, exons 17, 18, 19, and 21 - but never 20 - are alternatively spliced, but in mouse, alternative splicing does affect exon 20, as well as exons 17 and 21.

In summary, the aggregation of transcript data suggests that the occurrence of alternative splicing is a universal hallmark of this part of the periostin gene. This transcriptional variability could be important for a complete understanding of periostin function.

A 24th periostin exon is functional, expressed, and alternatively spliced in some tetrapod lines

Both human and murine periostin have been characterized as having a maximum of 23 exons ("canonical exons"), according to a substantial amount of periostin mRNA and EST sequences in these organisms.

Our analysis of the chicken periostin locus revealed the existence of an additional non-canonical exon situated between canonical exons 21 and 22, which we termed "21V22" to reflect this property. Exon 21V22 is present in 8 of the 15 mRNA/EST sequences covering this region. It is flanked by canonical splice sites and is 87 nucleotides long, allowing for maintenance of the reading frame without stop codon and thus translation into protein.

Sequence similarity searches with the protein sequence of this exon against genomic sequences produced clear matches in zebra finch (*Taeniopygia guttata*), lizard, opossum, and platypus, which are devoid of periostin transcript evidence. In all of these, the most basic hallmarks of a functional exon (maintenance of the reading frame without stop codon, canonical flanking splice sites) were also found, suggesting that this exon may also be functional in these species. However, equivalent searches failed to produce hits for this exon in frog (where we identified it separately, see below) and, interestingly, in placental mammals (human, mouse, dog, cow). While this echoes EST evidence for these species, we were surprised by the complete absence of even low quality hits in the genomic sequences, and performed a multiple alignment of periostin locus genomic sequence from 33 vertebrate species to get a clearer understanding of the "fate" of this exon across vertebrate taxa.

The genomic sequence alignment (Additional file 2, Figure S1) with chicken periostin as a reference shows a clear conservation peak for exon 21V22 not only for the aforementioned species for which this exon is known or highly likely to be expressed (chicken, finch, lizard, opossum, platypus; >75% conservation), but also for most other mammals tested (<75% conservation), including human.

We then analyzed the nucleotide sequences underlying the exon 21V22 conservation peaks (Additional file 3, Table S2A). We noted that for those species where the conservation peak was below 75%, hallmarks for a functional exon in general (maintenance of reading frame, no stop codons, flanking intronic GT-AG splice sites) were missing in many cases. In some others, these hallmarks are preserved, and the exon could be functional.

For example, in human, there could theoretically be exon 21V22 expression within full-length periostin transcripts and translation into protein product, as underscored by the GENSCAN [69] gene model NT_024524.504 (UCSC browser, Human Mar. 2006 Assembly), with a terminal exon identical to exon 21V22. The fact that it has never been experimentally observed leads us to speculate that the encoded protein sequence may either be expressed in unobserved conditions or be incompatible with human periostin function and remain unexpressed.

We conclude that exon 21V22 is a periostin exon not previously described. It is located between exons 21 and 22, and related in sequence to exons 17 - 21. Like these, it is subject to alternative splicing. Its expression is confined to a subgroup of vertebrates, including birds, where its expression is *de facto* observed in chicken, amphibians (see below), and putatively in lizards and non-placental mammals.

X. tropicalis expresses periostin exon 21V22 and a cluster of up to 8 additional genomically encoded exon 19 copies

When examined in the UCSC genome browser, the periostin locus of the claw frog *Xenopus tropicalis*

presents a notable feature: the number of (alternatively spliced) exons upstream of exons 21 and 21V22 by far exceeds the number of exons encountered in periostin from any other tetrapod species. Specifically, compared to mRNA sequence BC154911.1 as a reference, EST sequences indicate up to eight additional exons between exons 19 and 20. Analysis of these eight exon sequences revealed that they are extremely similar to *X. tropicalis* exon 19 (Figure 2), and we refer to these exons as exons 19A to exon19H. We are confident that these exons are not an artifact due to the preliminary nature of the *X. tropicalis* genome assembly, because they are present in both genomic and transcript sequence. Furthermore, we found a similar cluster of at least seven exon 19-like sequences among *Xenopus laevis* transcript sequences via similarity searches with *X. tropicalis* exons 19 and 19A - 19H as a query sequence (e.g. *X. laevis* mRNA GenBank:CB200763.2; data not shown).

An overview of all *X. tropicalis* ESTs covering the exon 19/20 region is given in Additional file 3, Table S3. Comparing the library origin between ESTs splicing directly from exon 19 to exon 20 (5 ESTs) to those that contain at least a subset of exons 19A - 19H (24 ESTs) revealed a strong bias for embryonic or metamorphic origin in the latter set. While far from definitive, this suggests the possibility of a developmental role of exons from the 19A - 19H cluster.

We furthermore analyzed the region without genomic match occurring in a majority of ESTs downstream of the exon 19 cluster (marked '*' in Figure 2), co-incident with a gap in the genomic assembly, and concluded that the unmatched region is periostin exon 21 and that the following exon (seen in 9 out of 33 transcript sequences covering this region) is actually exon 21V22 (see Methods).

The predicted basic secondary structure of the periostin C-terminus, consecutive beta sheets separated by turns, is conserved in periostin sequences across vertebrate species regardless of exon structure

We subjected periostin protein sequences exons 15 to 23 from 7 tetrapods (human, mouse, opossum, platypus, chicken, lizard, and frog) and 2 teleost fish species (zebrafish, locus 1 and stickleback, locus 1, see below) to a secondary structure prediction algorithm (See Methods for details). Prediction results are shown in Figure 3 in the context of a multiple sequence alignment of the nine sequences.

Generally, results show a universal multi-beta strand structure starting at the end of exon 15 and continuing into exon 22. Starting in exon 17, beta strands are usually separated by short coil stretches around proline residues occurring with a periodicity of 13 (or close to 13) in fish and at often somewhat larger intervals in tetrapods. (Note that for practical reasons, we are referring to the exon following exon 17 as exon "18=19", because for many of the teleost periostin sequences, it allows us to keep a numbering scheme on the basis of 23 periostin exons intact).

Evaluating the alignment and the secondary structure predictions in more detail, the following exceptions and refinements are notable:

-For tetrapod exons 17 to 21 (and 21V22, where applicable), each exon accommodates two beta-strands.
-Starting with exon 20 (middle, Xt: exon 21), the recurring proline residue is "lost" in tetrapod exons 20 and 21 and "reappears" universally in exon 22. It is also present in exon 21V22 (middle) found in only five of the seven species considered here.
-For the most part, the pattern of beta strands separated by coils continues through exons 20 and 21 despite the absence of a proline residue. However, interestingly, exon 20 (second half), and exon 21 (second half) show a predicted alpha helix instead of a beta sheet in some of the species.
-Due to the strongly conserved repeat structure in teleost fish, the beta strands are very regular here. Remarkably, they are also in phase with the tetrapod beta strands in the multiple sequence alignment.
-In the multiple sequence alignment, teleost exon 21 is aligned with tetrapod exon 21V22 and teleost exon 20 is aligned with tetrapod exon 21 (partial). This reflects both the length of teleost exon 18=19 accommodating, in the alignment, multiple tetrapod exons, and the uncertainty regarding a direct orthology between the short teleost exons 20 and 21 and tetrapod exons 20 and 21.

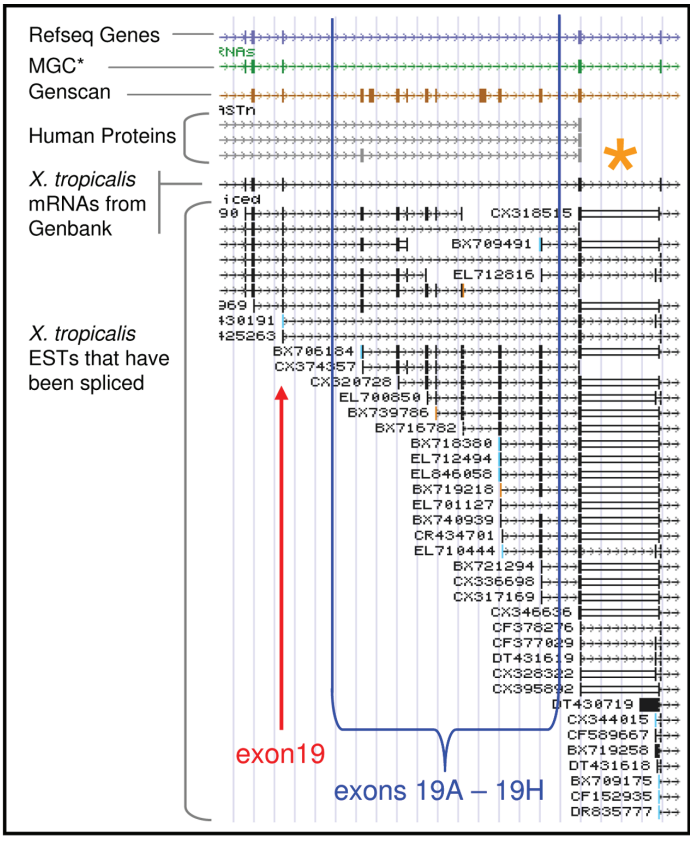
The predicted structure of two beta-strands per exon is interesting in light of the transcriptional variability these exons are subject to: Evidence for alternative splicing exists for each of the exons from 17 to 21V22 in at least one tetrapod species (Additional file 1, Table S1). The genomic duplications of exon 19 in *Xenopus tropicalis* are also alternatively spliced (Additional file 3, Table S3). Alternative in- or exclusion of any of these exons will then add or remove pairs of beta-strands to or from the protein.

Teleost fish genomes have two copies of periostin

Identifying the genomic periostin locus was straightforward in all tetrapod genomes, even in those lacking periostin transcript data (lizard, platypus, opossum) based on relevant multi- or cross-species tracks in the UCSC Genome Browser.

However, it is believed that the common ancestor of teleost fish has undergone whole-genome duplication (WGD) or at least a large-scale gene duplication event

Hoersch and Andrade-Navarro *BMC Evolutionary Biology* 2010, **10**:30
http://www.biomedcentral.com/1471-2148/10/30



B

exon19	DPSITRITKVIEGDPEFKLIKEGETRVTKVIQG
exon19A	EPTITKITRVIEGEPEFKLIKEGETRVTKVIQG
exon19B	EPTITKITRVIEGEPEFKLIKEGETRVTKVIQG
exon19C	EPTLTRITKVIGGESDLHLVKEGETRVTKVIQG
exon19D	EPTLTRITKVIGGEPDFHLVKEGETRVTKVIQG
exon19E	EP-ITRITKMIEGDQDFHLVREGETRVTKVIQG
exon19F	EPSITKITRVIEGEPDFKLIREGETRVTKVIQG
exon19G	EPTITKITRVIEGEPEFKLIKEGETRVTKVIQG
exon19H	EPIITKLTRVVEGEPDFRLIKEGETRVTKVIQG

Figure 2 The *Xenopus tropicalis* periostin C-terminal region in the UCSC genome browser. **(A)** The 'Spliced EST' track shows a cluster of eight additional genomic copies of exon 19 (referred to here as exons 19A - 19H) immediately downstream of exon 19. These exons, not represented in the RefSeq record, are subject to alternative splicing in various combinations. The asterisk (*) marks a region where multiple ESTs have subsequences unaligned to genomic sequence due to a gap in the assembly (not shown), which we resolved to be exon 21 (see main text). **(B)** Amino acid sequences of exons 19 and 19A - 19H. The very high degree of similarity among these is evident.

[70], complicating the situation. A systematic genome-wide sequence search (BLAT) with human periostin sequence against the five fish genomes resulted in typically three distinct matches. Analysis of these matches in terms of length and similarity indicated that the two stronger matches corresponded to periostin. This was confirmed by the reciprocal results of a genomic search with human TGFBI sequence. Here, the best match was always the third-ranked hit from the previous periostin genomic search (data not shown).

We conclude that teleost fish retained both copies of periostin after the duplication event. By contrast, the second copy of TGFBI was either universally lost (i.e. presumably before the teleost species radiation), or TGFBI was not part of the duplicated gene complement. Our designation of the two periostin loci in the fish species as "locus 1" and "locus 2" (see Additional file 1, Table S1 for genomic coordinates) was done on the basis of expression level as judged by the amount of transcript evidence mapped to either locus in the UCSC genome browser. Where possible (zebrafish and tentatively stickleback), we assigned the label "locus 1" to the locus found to be more highly expressed by that measure, coinciding in the case of zebrafish with the version of

Hoersch and Andrade-Navarro *BMC Evolutionary Biology* 2010, **10**:30
http://www.biomedcentral.com/1471-2148/10/30

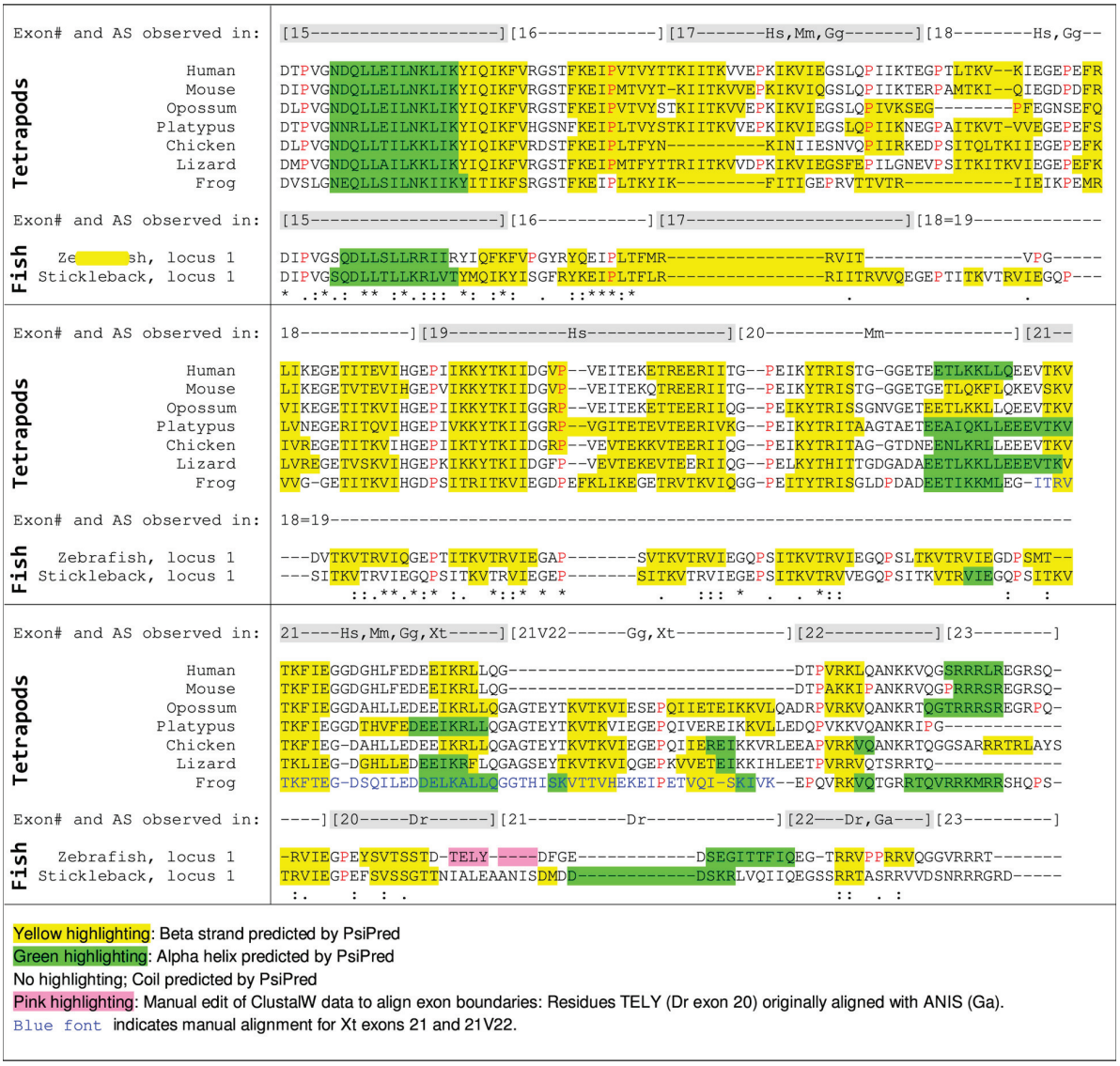


Figure 3 Multiple sequence alignment and secondary structure prediction results for the periostin C-terminal region. A ClustalW alignment for periostin's C-terminal region of seven tetrapods (Hs, Mm, Md, Oa, Gg, Ac, Xt) and two teleosts (Dr locus 1, Ga locus 1) is shown. The exon structure is indicated above the alignment, separately for tetrapods and teleosts, together with two-letter species codes, if alternative splicing is observed for a particular exon in that species. Residues are highlighted based on PsiPred secondary structure predictions performed separately for all nine sequences as indicated below the alignment. The general alignment of the secondary structure elements, mostly beta strands, across all species considered here, is obvious.

periostin described in the literature [21]. For the other fish species, we based the assignment on the grouping resulting from a multiple alignment-based dendrogram of teleost periostin protein sequences (see below).

We reconstructed as completely as possible periostin coding sequence from both loci for four of the five teleost fish species, relying mostly on sequence similarity searches with the known periostin sequence from zebrafish. In several cases, missing exons or other uncertainties remain due to gaps in genomic sequence and/or complete lack of any guiding transcript coverage. For one species (medaka), reconstruction from one periostin locus was not undertaken due to the fragmentary nature of the genomic sequence data.

A dendrogram based on an alignment of seven tetrapod periostin sequences (see above) together with the nine teleost periostin sequences shows, among the teleosts, a separation into two major branches by periostin locus 1 vs. 2 (Figure 4A) preceding the branching into the multiple teleost lineages, and thus echoing the genome duplication event. Within each branch (i.e., periostin locus), the dendrogram structure recapitulates the

Hoersch and Andrade-Navarro *BMC Evolutionary Biology* 2010, **10**:30
http://www.biomedcentral.com/1471-2148/10/30

phylogenetic relationship between the five teleost fish species (minus medaka for one locus), i.e., zebrafish as a member of the Ostariophysi lineage most distant from the other four species (Euteleostei), and the two puffers (Tetraodontidae) Takifugu and Tetraodon closest together.

Periostin sequence conservation is dramatically lower in its C-terminal region

Up to exon 16, periostin sequences are extremely well conserved between all euteleostomes, including the two periostin loci in teleosts. In fact, not counting exons not reconstructed due to missing genomic sequence and uncertainties around the boundaries of Ensembl-prediction based exons 15 and 16 for locus 2 in Tetraodon, a multiple sequence alignment for all seven tetrapod and nine teleost periostin sequences showed only 5 gapped positions from exon 2 through 16.

By contrast, for tetrapod periostin sequences alone, and not counting gaps corresponding to entire exons (see below), we counted 32 gapped columns in a multiple sequence alignment from exons 16 to 22. This increased speed at which the C-terminal part of

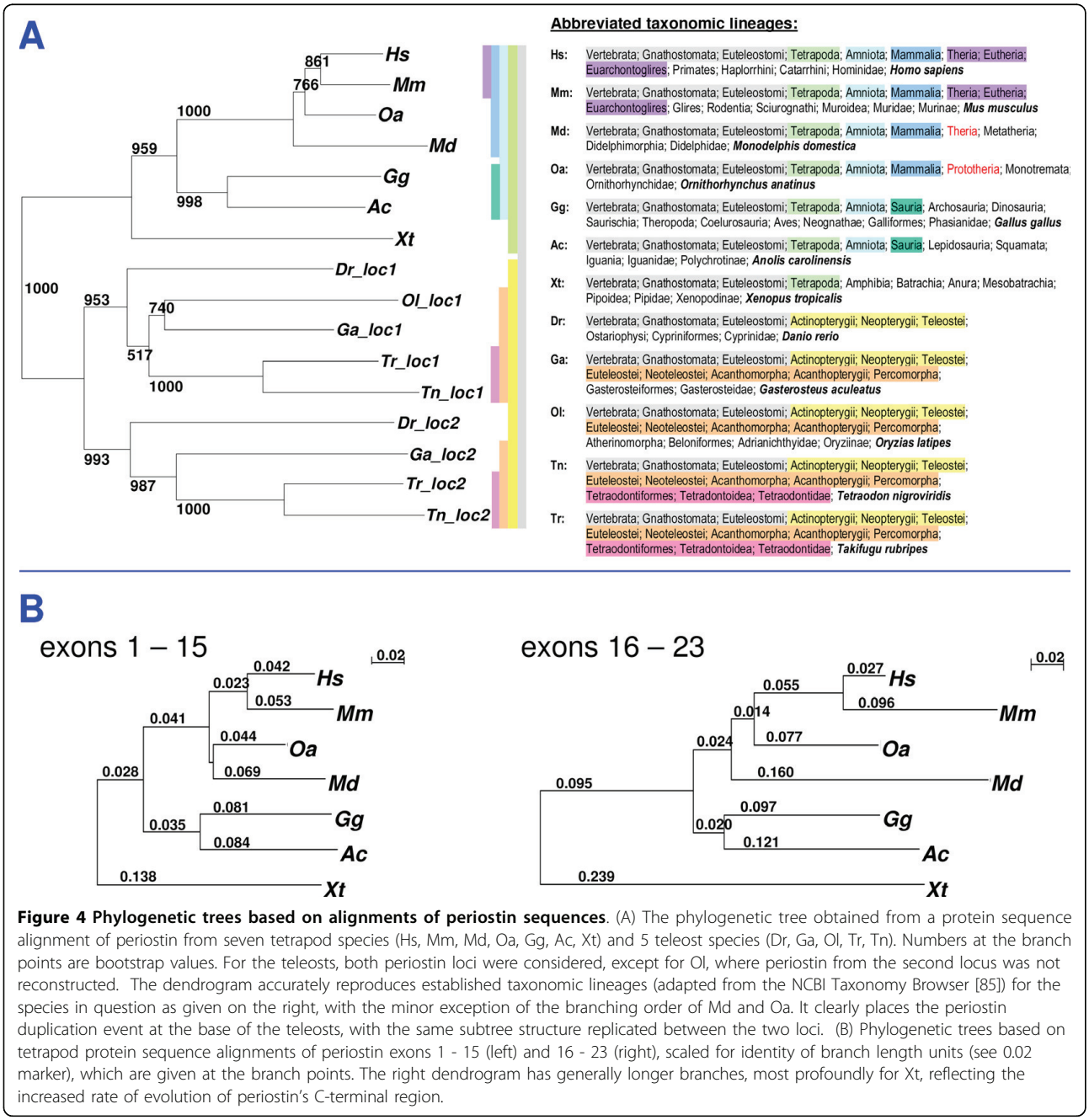


Figure 4 Phylogenetic trees based on alignments of periostin sequences. (A) The phylogenetic tree obtained from a protein sequence alignment of periostin from seven tetrapod species (Hs, Mm, Md, Oa, Gg, Ac, Xt) and 5 teleost species (Dr, Ga, Ol, Tr, Tn). Numbers at the branch points are bootstrap values. For the teleosts, both periostin loci were considered, except for Ol, where periostin from the second locus was not reconstructed. The dendrogram accurately reproduces established taxonomic lineages (adapted from the NCBI Taxonomy Browser [85]) for the species in question as given on the right, with the minor exception of the branching order of Md and Oa. It clearly places the periostin duplication event at the base of the teleosts, with the same subtree structure replicated between the two loci. (B) Phylogenetic trees based on tetrapod protein sequence alignments of periostin exons 1 - 15 (left) and 16 - 23 (right), scaled for identity of branch length units (see 0.02 marker), which are given at the branch points. The right dendrogram has generally longer branches, most profoundly for Xt, reflecting the increased rate of evolution of periostin's C-terminal region.

periostin evolves compared to the larger N-terminal portion comprising the EMI and fasciclin domains is also reflected in alignment-based dendrograms derived separately for these two sections of periostin sequence (Figure 4B). The overall tree topology is identical in both cases (with the exception of the relative positioning of opossum and platypus), but branch lengths are markedly longer for the C-terminal region-based alignment. Finally, further elements of genomic and transcriptional variation, discussed in detail below and exclusively found in the C-terminal periostin region, are adding to the picture of a part of the periostin gene that shows highly divergent and dynamic characteristics.

Starting with exon 17, we found in some cases dramatic variation with respect to exon number, exon length, and overall sequence conservation, some of which we described above for tetrapod species. The divergence for this region is, in some respects, even higher between teleost species.

In fish, we observed exon 17 to vary in length between 5 and 11 amino acids. Interestingly, short versions of exon 17 also occur among tetrapod periostin sequences, where the standard exon 17 length is 26 amino acids: birds (chicken and zebra finch) as well as frog (but not reptiles) have an exon 17 that is about 16 amino acids in length.

Exon 18=19 is especially remarkable in teleost fish: in periostin sequences from all species except medaka (*Oryzias latipes*), we found it to be longer than neighboring exons, sometimes remarkably so - for stickleback periostin from locus 2, it is 91 amino acids long (with EST coverage); for zebrafish locus 2, it appears to encode 187 amino acids (incomplete EST coverage). This exon is marked by a repeat structure that is obvious to unaided visual inspection and readily surfaced via high-quality secondary hits in genome-wide sequence searches with nucleotide or protein sequences containing the repeat. It was chiefly this exon (in zebrafish) which we had initially observed aligning to the human periostin query of exons 16 - 23 in a iterative sequence search, and it is this exon from which the repeat sequence for the profile-based homology search was taken.

Characterization of the periostin repeats

The periodicity of this repeat is generally 39 nucleotides/13 amino acids. Slight deviations from this length are observed in a fashion that keeps the reading frame intact. We used matrix dot plots to get a more detailed understanding of the repeat properties (see Methods for details). Interestingly, the repeat exhibits similarity to its reverse complement to varying degrees. The most notable example here is in stickleback periostin locus 1, where exon 18=19 reverse-complement similarity reaches 25/39 (64%) nucleotide identities with no gaps. This level of similarity is readily revealed in the matrix dot plots as lines orthogonal to the main diagonal. Other repeat instances show weaker levels of reverse-complement similarity that are not readily surfaced in the dot plots (e.g., 19/42 identities including a gap of 3 positions for zebrafish periostin exon 18=19 at locus 2 (Figure 5)).

Visualizing genomic periostin sequence extending beyond the boundaries of the repeat exon(s) in dot-matrix plots, we observed that the repeat pattern appears confined to the exons and does not extend into adjacent intronic sequence (Additional file 4, Figure S2).

The matrix dot plot in Figure 5 further illustrates our finding that the number of exons beyond exon "18=19" in teleosts is variable: Zebrafish locus 1 periostin has 4 more exons, but locus 2 periostin has as many as seven, the largest number we found. Many of these "extra" exons also show the repeat structure observed in exon 18=19 (Figure 5). On the other hand, periostin from the stickleback locus 2 has only 3 more exons after exon 18=19, one less than the canonical number.

On the protein level, the 39-nucleotide repeats encode a 13-amino acid sequence with the following consensus: ***PS***I***TK***V***TRV***I***EG***E. (Figure 6A; amino acids that in italics were found to be the most frequent at this position in every locus-specific repeat alignment considered, which were for both loci from zebrafish and stickleback, and for one locus from each of the two puffers, *T. nigroviridis* and *T. rubripes*). Its key characteristics as given here are: (i) a Pro at position 1, (ii) a stretch of hydrophobic amino acids (mostly Thr and Val) from positions 3 to 10, interspersed with the positively charged amino acids Lys and Arg in positions 5 and 8, and (iii) a final stretch of three amino acids that is dominated by negatively charged Glu and Asp (mostly at positions 11 and 13, with a Gly in between).

Tetrapod periostin sequences do not have a large exon with obvious repeats like most teleost fish. However, dot matrix plots readily reveal self similarities within the transcript region spanned by exons 17 - 22 (Additional file 4, Figure S3).

Similarities to the teleost repeat are likewise demonstrated: Figure 6B-K gives sequence logo representations [65] for the teleost repeat described above and for periostin exons 15 - 23 derived from the seven tetrapods considered here (Hs, Mm, Md, Oa, Gg, Ac, Xt). A color-coded version of a matrix dot plot is shown, generated by comparing 2 instances of the teleost repeat consensus (26 aa) against concatenated tetrapod consensus sequences from exons 15 - 23. The highest similarities to the 13 aa repeat are detected in exon 18, exon 19 (*X. tropicalis*, see below), exon 21, and exon 21V22 (see below), whereas we see somewhat weaker similarities in

Hoersch and Andrade-Navarro *BMC Evolutionary Biology* 2010, **10**:30
http://www.biomedcentral.com/1471-2148/10/30

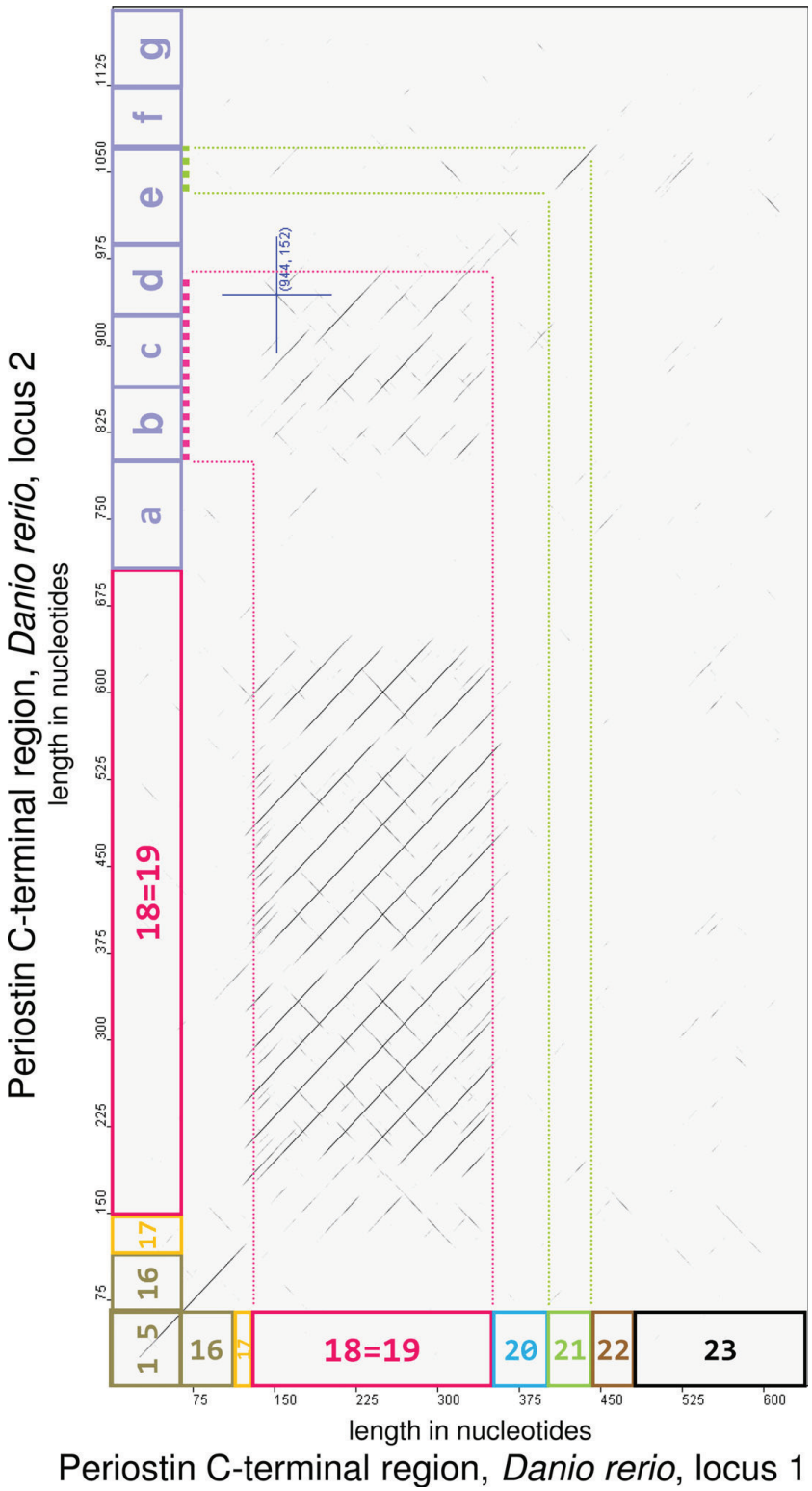


Figure 5 Matrix dot plot of periostin nucleotide sequences from the C-terminal region of zebrafish locus 1 (horizontal) versus locus 2 (vertical). Exon sizes and boundaries are indicated in either dimension, with locus 2 exons downstream of 18=19 marked here provisionally as a - g. The dot plot shows the 39 nucleotide repeat pattern within exon 18=19 and also reveals similarities to the repeat within locus 2 exons b, c, d as well as a correspondence (probable homology) of locus 1 exon 21 to locus 2 exon e (partial). Weaker reverse-complement similarities are also visible.

Hoersch and Andrade-Navarro *BMC Evolutionary Biology* 2010, **10**:30
http://www.biomedcentral.com/1471-2148/10/30

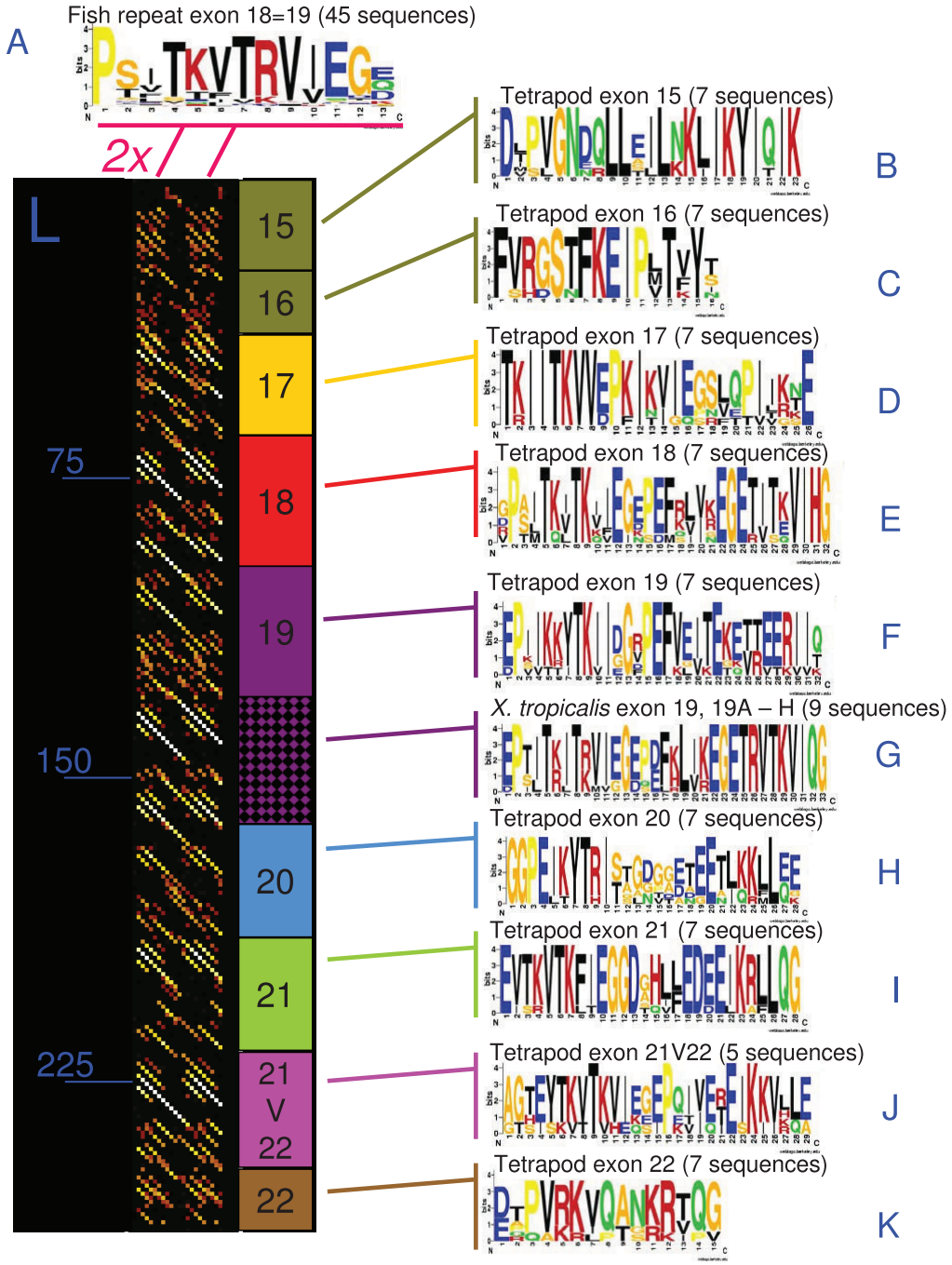


Figure 6 Sequence logo representations of the teleost fish periostin repeat consensus and of the consensus sequences from tetrapod exons of the periostin C-terminal region. **(A)** The periostin repeat sequence logo for teleosts was obtained from an alignment of the 13 aa repeat units of periostin exon 18=19 (see main text) from four teleost species (Dr, Ga, Tn, Tr). The starting amino acid (proline) is arbitrarily chosen (see also Methods). **(B - K)** The exon-specific sequence logo representations were obtained from alignments of the respective periostin exons from seven tetrapod species considered here (Hs, Mm, Md, Oa, Gg, Ac, Xt). Exceptions are (G), the Xt exon 19 cluster, where a sequence alignment of exons 19 and 19A - 19H (see main text) was the basis for the sequence logo, and (J), exon 21V22, which was based only on sequences from Md, Oa, Gg, Ac, and Xt. **(L)** A matrix dot plot showing similarities between the teleost repeat consensus (horizontal, as duplet for clarity) and the sequential tetrapod exon consensus sequences (vertical), with the exon sizes indicated on the side. The dot plot is color-coded (white-yellow-orange-red-black), with white indicating highest similarity and black indicating no similarity. Both this dot plot and the sequence logos indicate the strong similarities between the teleost repeats and the tetrapod C-terminal exons found in the alternatively spliced exons 17 - 21V22 (see main text).

Hoersch and Andrade-Navarro *BMC Evolutionary Biology* 2010, **10**:30
http://www.biomedcentral.com/1471-2148/10/30

exons 17, 19, and 20. Similarities in exons 15, 16, and 22 are markedly weaker than those found in the exons in-between.

The concept that tetrapod exons between 16 and 22 are homologous to the long repeat exon 18=19 in teleost fish is powerfully reinforced by the alignment given in Figure 3. This alignment shows not only the primary structure (sequence), but also the predicted secondary structure elements. Here, tetrapod exons 17 to 21V22 align with the teleost repeats such that the sequence alignment results also in an alignment of the secondary structure elements, i.e., the beta strands. Critically, tetrapod exon 21V22, which appears downstream of exon 21, shows an especially high similarity with the teleost repeat consensus: for example, the subsequence EYTKVTKVIEGEP from chicken exon 21V22 is 77% identical to the teleost repeat consensus ***S*I*TK*V*TRV*I*E-GE*P*** (rearranged here for a terminal P to align with the chicken sequence). Thus, exon 21V22 defines the range of tetrapod exons homologous to the long teleost repeat exon.

Discussion

The C-terminal region of periostin is likely a key to disambiguating periostin function

Despite ongoing discussions regarding the periostin alternatively spliced C-terminal region (e.g. [42,71]) dating back to its original description [1,2], this region has not been target of in-depth analyses to identify its biological function. Here, we presented evidence for its universally high degree of variability in a phylogenetic context, which we hope will assist future research on periostin. We laid out the multiplicity and, within the euteleostomes, phylogenetic universality of modalities observed (including alternative splicing as well as genomically encoded variable exon counts and lengths) to modify this part of periostin. Considering also that the functionally distinct periostin paralog TGFBI lacks an equivalent to periostin's C-terminal region, we propose that this region is of functional relevance, and that the sequence variation imparted by alternative splicing could be a modulator of periostin function.

In studies aimed at elucidating determinants of tumor-suppressive properties of periostin, the periostin C-terminal region was found to be sufficient to suppress anchorage-independent growth in T24 bladder cancer cells [72], to suppress cell invasiveness in SBT31A bladder cancer cells, and to abrogate the metastatic potential of highly metastatic B16F10 mouse melanoma cells in an *in vivo* assay [42]. While the tumor-suppressive properties documented for periostin in these studies remain to be reconciled with a large body of literature describing periostin as tumor promoting, these results demonstrate the functional significance of the periostin C-terminal region as a whole. This notion is also playing out in discussions in literature not explicitly concerned with decoding periostin function on the sequence or domain level. For example, the significance of a recent high-profile study on periostin as a potential therapeutic target after heart attack [73] has been questioned on the basis of this work being partly based on a version of periostin lacking the C-terminal region [13,25,74].

The situation is more complex when it comes to different periostin splicing isoforms characterized by individual presence or absence of cassette exons 17 - 21 (human and mouse). For the following discussion, we are using a notation of exon numbers preceded with a minus sign in superscript to characterize periostin isoforms by their absent exons relative to a full-length periostin, denoted $POSTN^{fl}$, with a full complement of exons, i.e. 23 for human and mouse.

Based on the common occurrence of gel pictures showing multiple periostin bands throughout the literature [e.g. 2,4,15,17,25,29,75,76], we can assume that studies on native periostin typically deal with a mixture of isoforms, although the exact number is usually unknown. In apparent agreement with the number of four isoforms originally reported [2], most gels show a cluster of 4 or sometimes 5 periostin bands. However, the similar length of the alternatively spliced exons gives rise to possible combinations that are very close in size, so that the actual number is potentially higher due to isoforms co-migrating in the same band. Indeed, studies have found higher isoform counts, for example 6 in human kidney cancer [76] and more than 8 in human breast tissue (Hoersch et al: Alternative splicing of periostin in human breast cancer, manuscript in preparation), and in both cases, gel pictures allowed distinction of only four or five variants.

Descriptions of experimental procedures manipulating periostin in some way, for example expressing in ectopically in a cell line, are not always explicit on the specific isoform used, and the attributes "full-length" or "wild type" are not always or unambiguously referring to the isoform with all exons present. In some instances, the relevant isoform can be traced via database accession numbers provided, but overall, this information is not consistently enough available to enable a meaningful isoform-specific meta-analysis.

More interestingly, biological observations regarding periostin have, in some cases, been explicitly tied to specific variants, as illustrated by the following three examples.

(i) In a study on periostin expression in bone [71], the designation "periostin-like factor" or "PLF" was introduced by the authors specifically for periostin isoform $POSTN^{-21}$, while the designation "periostin" refers to isoform $POSTN^{-17}$ in the same study. Using the same

designations, this group subsequently described differential spatiotemporal expression patterns of these two isoforms in the developing mouse embryo [77], although it has to be understood that the "isoform-specific antibodies" used in this work are technically only site-specific for exons 21 and 17, respectively, and could, in principle, report on multiple isoforms characterized by exon 17 or exon 21 presence.

(ii) Differential expression profiles over time were observed for different periostin isoforms after a myocardial infarct event in periostin-null mice, with isoform $POSTN^{-17-21}$ being expressed strongest initially and then becoming weaker after a sustained period of high expression as the expression level of other isoforms gradually increased [14].

(iii) In a continuation of earlier work, it was found that periostin isoforms $POSTN^{fl}$ and $POSTN^{-17-21}$ abrogate invasion in cell lines B16F10 and SBT991 and *in vivo* lung metastasis of B16F10 cells in mice, while isoform $POSTN^{-17-18-21}$ does not [51]. Along similar lines, but with opposite directionality, isoform-specific differences in the ability to promote migration and invasion were observed in breast cancer cells (Hoersch et al: Alternative splicing of periostin in human breast cancer, manuscript in preparation).

Apart from the often implicit issue of periostin isoform expression, several aspects of periostin are controversially described and discussed in the literature to date, prominently among them: (i) periostin subcellular localization, (ii) the cell types responsible for periostin expression and the physiological sites of periostin protein localization post secretion, and (iii), in the cancer literature, periostin function being tumor promoting or suppressing.

It will be interesting to see to what extent the variability of periostin isoform expression will contribute to resolving these issues, a possibility that is commonly raised in discussion sections in the periostin literature. Given the sequence and predicted structural similarities between periostin's cassette exons due to the underlying repeat structure laid out in this study, we detail below a functional interpretation that employs a dose-type rationale, i.e. a binding interaction as a function effectively tuned by exon count. This seems more compelling to us than a purely exon-specific mechanism, although the reality may well be a complex overlay of both mechanisms.

The periostin C-terminal region exhibits exceptional transcriptional and genomic variability

We demonstrated the homology of periostin's multi-exon C-terminal region in tetrapods and its C-terminal region in teleosts, where a 13 amino acid sequence is stacked into adjacent repeats. The majority of these is, in most of the teleost species studied, concentrated in one exon that is unusually large (dubbed "exon 18=19" in this work).

The absence of protein domain signatures in periostin's C-terminal region is thus no longer puzzling. We can assume that this region as observed today evolved from an ancestor repeat region, leading to its distribution over multiple exons, particularly in tetrapod species.

We demonstrated that the periostin C-terminal region is showing increased evolutionary plasticity relative to the remainder of the gene. This is evident from comparing branch lengths of phylogenetic trees based on separate sequence alignments of the C-terminal region and of the remainder of the gene (Figure 4B). Moreover, it is reinforced by the variability we found almost exclusively for periostin's C-terminal regions between different taxa that are not even reflected in these dendrograms: profoundly different exon structure (teleosts vs. tetrapods), marked divergence between the two periostin copies in teleosts, variable exon length (exon 17 across euteleostomes, exon 18=19 in teleosts), an additional exon not previously described (exon 21V22 in tetrapods except placental mammals), and additional genomic copies of exon 19 with a strong developmental implication (in frogs).

We also reported that alternative splicing is a universal hallmark of the periostin C-terminal region. While long established for human and murine periostin, we found evidence for alternative splicing events affecting multiple exons in periostin's C-terminal regions in all organisms studied here (including teleost fish) for which non-trivial amounts of transcript sequence data were publicly available. Exons 17, 18, 19, 20, 21, and 21V22 are all alternatively spliced cassette exons, although it appears that in any given species, at least one of them is constitutively expressed. Although uncertainties remain in some cases due to low sequence coverage, the predicted alternative splicing patterns were found to vary by species, sometimes subtly so (e.g. in human and mouse with regard to exon 20).

Taken together, these observations point to periostin's C-terminal region being exceptionally dynamic on different scales. On a transcriptional scale, alternative splicing could give rise to periostin variants in a tissue, development, or disease process dependent manner, the functional impact of which is only beginning to emerge. On a genomic, i.e., evolutionary scale, the increased plasticity of this region leads to periostin configurations that are specific to taxonomic lineages or groups.

The periostin C-terminal region and neofunctionalization

We understand this increased evolutionary plasticity as a manifestation of neofunctionalization [78] affecting euteleostome periostin where we observe variations in exon length, count, structure, and variable splicing patterns in the C-terminal region. We hypothesize that the repeats

Hoersch and Andrade-Navarro *BMC Evolutionary Biology* 2010, **10**:30
http://www.biomedcentral.com/1471-2148/10/30

were acquired as *de novo* sequence in an exonization event [79], either by ancestral periostin (after the periostin/TGFBI split) or by the common periostin/TGFBI ancestor gene, in which case the repeats were lost by TGFBI after the split.

Data available at this time is insufficient to distinguish between these two possibilities, chiefly because of the preliminary nature of relevant genome assemblies outside the euteleostomes (for details and Figures, see Additional file 6). This newly exonized repeat sequence was then relatively free to evolve, while the remainder of the gene remained much more conserved due to the structural and functional constraints imposed by the EMI and FAS1 domains.

Remarkably, our observations concerning the two periostin copies in teleost fish point to the same mechanism of neofunctionalization playing out here. Between the two periostin copies in teleosts, we are again seeing variations in exon length, count, and structure. In addition, to the extent that the EST record can serve as a guide, the expression level difference between the two periostin copies in teleosts might be species-specific (zebrafish and stickleback, Additional file 1, Table S1).

It is interesting to speculate that periostin's C-terminal region and its high evolutionary plasticity may explain why periostin, but not TGFBI, is duplicated in teleosts. The absence of a second TGFBI copy in teleosts might reflect the limited freedom of TGFBI to functionally differentiate from the "primary" copy due to the constraints imposed by the EMI and FAS1 domains, leading to loss of the redundant, if not harmful, second copy. By contrast, periostin's C-terminal region could be instrumental for the rapid evolvement of a functional role sufficiently distinct from the periostin paralog in question.

Additional file 6 summarizes further aspects of periostin's C-terminal region in particular and the periostin/TGFBI paralogy in general, placing periostin and TGFBI into a larger framework of chordate evolution.

Interpretations of the secondary structure prediction results for periostin's C-terminal region

Through secondary structure prediction on periostin's C-terminal region from seven tetrapod and two teleost species, we found consecutive beta strands as the predominant structural feature of this region, specifically, one beta strand per repeat unit. For tetrapods, this means that generally two beta strands are encoded per exon.

The universal splicing variations and the genomic variability observed across and between species could, in principle, point to a functional differentiation of the C-terminal region of periostin. This interpretation is especially attractive in the light of both literature reporting isoform-specific findings and our observation of the eight additional genomic copies of exon 19 in *X. tropicalis*, the expression of which might be tied to developmental stage, specifically (pre)metamorphic animals, according to EST data.

The extracellular matrix is rich in potential interaction and binding partners for periostin, a few of which have been elucidated in some detail. Most prominent among these are a number of integrins (e.g. [29-32]). Heparin binding has been shown, possibly mediated by a Cardin-Weintraub consensus sequence [37,80] encoded by exon 23.

More recently, periostin binding to collagen I was reported in the context of a phenotypic characterization of periostin-null mice [35], but the binding mechanism was not described. Similarly, periostin binding with fibronectin was observed, but not explained on a molecular or structural basis [38].

Since collagen and fibronectin binding have not been described for TGFBI, the periostin paralog lacking an extended and variable C-terminal region, we hypothesize that periostin might interact with these proteins in this region. Collagen I - fibronectin binding has recently been elucidated in complexes of two or four sequential fibronectin FN1 or FN2 domains, components of a fibronectin region called "gelatin-binding domain", with peptides from the collagen I α_1 or α_2 chain. With the two fibronectin domains in question being beta-sandwich structures, the collagen peptides assume, in the complex, a beta strand conformation with anti-parallel orientation to one of the FN domain beta strands (beta-zipper) [81].

This mode of binding is similar to earlier descriptions of the way bacterial cell wall proteins from *Staphylococcus aureus* and *Streptococcus pyogenes* attach to human fibronectin [82,83]: Here, bacterial repeats were found to form a tandem beta-zipper with two fibronectin FN1 domains. This finding and the repetitive arrangement of both the bacterial units and the FN1 domains then led the authors to propose a model of an extended beta-zipper, involving a larger number of FN1 domains and bacterial repeats. Therefore, since the periostin's C-terminal region has repeats reminiscent of these bacterial repeats, we hypothesize that they may assume a structure of multiple consecutive beta strands when binding to sequential beta strand elements of fibronectin domains (and possibly other ECM proteins) by way of an extended beta-zipper.

In such a model, the strength of this interaction would be influenced by the number of beta strands available for binding with - in the case of fibronectin - maximally 11 available sequential FN1/FN2 domains in the N-terminal region. The number of available beta strands from periostin's side is a function of exon count, which is in turn determined - and, in this model, calibrated or tuned - by alternative splicing as well as lineage-specific differences.

Conclusions

We believe that this study marks an important contribution to periostin research. Despite the considerable increase in published literature since periostin's discovery in an osteoblast cell line in 1993, especially over the past few years, some aspects have remained poorly understood. This paper provides evolutionary context for periostin's enigmatic C-terminal region and argues that it is likely key to a detailed understanding and disambiguation of periostin function, which is subject to some controversy in present literature. We put forward a number of findings and interpretations that will hopefully be taken on in further studies and experimentally in the future. We believe that our secondary structure prediction, the finding of additional genomic exon 19 copies in *X. tropicalis*, and the duplication of periostin in teleosts are particularly rewarding experimental targets.

Finally, we note that this work has benefited greatly from the growing collection of whole-genome sequences available, without which it would have been impossible in its present form. At the same time, it strongly supports the argument not only for continued new genome sequencing, especially of strategically selected representatives from key phyla, but importantly also for a sustained effort to finish, at least to a reasonable degree, existing draft assemblies. While it will not be practical in all cases to arrive at high-quality chromosome-by-chromosome assemblies that are subject to sustained curation efforts, highly fragmented genome sequences render detailed gene-level analyses impossible, or at least severely limit the power of their results. Some aspects of our work remained incomplete, mirroring the unfinished state of key genome assemblies, for example within the teleosts, but especially outside the euteleostome group.

Additional file 1: Table S1: Periostin in seven tetrapod and five teleost species. For periostin from seven tetrapod and five teleost species, this table summarizes data on chromosomal location, available transcript abundance and variation, and the periostin protein sequences reconstructed in the course of this work and listed in Additional file 5.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-30-S1.PDF]

Additional file 2: Supplementary Figure S1. Genomic sequence alignment of the periostin locus for 15 species and illustrating the variable conservation of exon21V22.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-30-S2.PDF]

Additional file 3: Supplementary Tables S2 and S3. Select alignments of genomic sequences comprising periostin exon 21V22 and exon 17 (Tables S2A, S2B) and overview of *Xenopus tropicalis* transcript sequence evidence covering the periostin C-terminal region (Table S3).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-30-S3.PDF]

Additional file 4: Supplementary Figures S2 and S3. Matrix dot plots of periostin C-terminal repeats in teleosts (Figure S2) and human and chicken (Figure S3).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-30-S4.PDF]

Additional file 5: FASTA-formatted sequence listing. Protein sequences of periostin and certain periostin homologs relevant for this work.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-30-S5.DOC]

Additional file 6: Additional evolutionary and phylogenetic considerations. This document summarizes additional evolutionary and phylogenetic considerations regarding periostin and its paralog, TGFBI, based on analyses of periostin/TGFBI homologs outside the euteleostomes. Contains Supplementary Figures S4 - S7.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-30-S6.PDF]

Acknowledgements
The authors thank Charles A. Whittaker and GuangJun Zhang (both at the Koch Institute for Integrative Cancer Research at MIT, Cambridge, USA) for helpful discussions and Barbara Bryant for critical reading of the manuscript. MAA acknowledges funding from the Helmholtz Alliance on Systems Biology (Helmholtz-Gemeinschaft Deutscher Forschungszentren) (including funding for open access charge).

Author details
[1]Bioinformatics and Computing Core, Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. [2]Bioinformatics Group, Max Delbrück Center for Molecular Medicine, Robert-Rössle-Strasse. 10, 13125 Berlin, Germany. [3]Computational Biology and Data Mining Group, Max Delbrück Center for Molecular Medicine, Robert-Rössle-Strasse. 10, 13125 Berlin, Germany.

Authors' contributions
SH conceived of and performed most of the analyses. MAA and SH discussed the research on an ongoing basis, and MAA helped with some analyses. SH wrote the manuscript with input from MAA, and both authors approved its final version.

Received: 21 August 2009
Accepted: 28 January 2010 Published: 28 January 2010

References

1. Takeshita S, Kikuno R, Tezuka K, Amann E: **Osteoblast-specific factor 2: cloning of a putative bone adhesion protein with homology with the insect protein fasciclin I.** *Biochem J* 1993, **294(Pt 1)**:271-278.
2. Horiuchi K, Amizuka N, Takeshita S, *et al*: **Identification and characterization of a novel protein, periostin, with restricted expression to periosteum and periodontal ligament and increased expression by transforming growth factor beta.** *J Bone Miner Res* 1999, **14**:1239-1249.
3. Oshima A, Tanabe H, Yan T, *et al*: **A novel mechanism for the regulation of osteoblast differentiation: transcription of periostin, a member of the fasciclin I family, is regulated by the bHLH transcription factor, twist.** *J Cell Biochem* 2002, **86**:792-804.
4. Kruzynska-Frejtag A, Wang J, Maeda M, *et al*: **Periostin is expressed within the developing teeth at the sites of epithelial-mesenchymal interaction.** *Dev Dyn* 2004, **229**:857-868.
5. Suzuki H, Amizuka N, Kii I, *et al*: **Immunohistochemical localization of periostin in tooth and its surrounding tissues in mouse mandibles during development.** *Anat Rec A Discov Mol Cell Evol Biol* 2004, **281**:1264-1275.

6. Rios H, Koushik SV, Wang H, *et al*: **periostin null mice exhibit dwarfism, incisor enamel defects, and an early-onset periodontal disease-like phenotype.** *Mol Cell Biol* 2005, **25**:11131-11144.
7. Hamilton DW: **Functional role of periostin in development and wound repair: implications for connective tissue disease.** *J Cell Commun Signal* 2008, **2**:9-17.
8. Kudo Y, Siriwardena BSMS, Hatano H, Ogawa I, Takata T: **Periostin: novel diagnostic and therapeutic target for cancer.** *Histol Histopathol* 2007, **22**:1167-1174.
9. Ruan K, Bao S, Ouyang G: **The multifaceted role of periostin in tumorigenesis.** *Cell Mol Life Sci* 2009, **66**:2219-2230.
10. Litvin J, Zhu S, Norris R, Markwald R: **Periostin family of proteins: therapeutic targets for heart disease.** *Anat Rec A Discov Mol Cell Evol Biol* 2005, **287**:1205-1212.
11. Dorn GW: **Periostin and myocardial repair, regeneration, and recovery.** *N Engl J Med* 2007, **357**:1552-1554.
12. Borg TK, Markwald R: **Periostin: more than just an adhesion molecule.** *Circ Res* 2007, **101**:230-231.
13. Conway SJ, Molkentin JD: **Periostin as a heterofunctional regulator of cardiac development and disease.** *Curr Genomics* 2008, **9**:548-555.
14. Shimazaki M, Nakamura K, Kii I, *et al*: **Periostin is essential for cardiac healing after acute myocardial infarction.** *J Exp Med* 2008, **205**:295-303.
15. Oka T, Xu J, Kaiser RA, *et al*: **Genetic manipulation of periostin expression reveals a role in cardiac hypertrophy and ventricular remodeling.** *Circ Res* 2007, **101**:313-321.
16. Litvin J, Blagg A, Mu A, *et al*: **Periostin and periostin-like factor in the human heart: possible therapeutic targets.** *Cardiovasc Pathol* 2006, **15**:24-32.
17. Yan W, Shao R: **Transduction of a mesenchyme-specific gene periostin into 293T cells induces cell invasive activity through epithelial-mesenchymal transformation.** *J Biol Chem* 2006, **281**:19700-19708.
18. Soltermann A, Tischler V, Arbogast S, *et al*: **Prognostic significance of epithelial-mesenchymal and mesenchymal-epithelial transition protein expression in non-small cell lung cancer.** *Clin Cancer Res* 2008, **14**:7430-7437.
19. Kruzynska-Frejtag A, Machnicki M, Rogers R, Markwald RR, Conway SJ: **Periostin (an osteoblast-specific factor) is expressed within the embryonic mouse heart during valve formation.** *Mech Dev* 2001, **103**:183-188.
20. Norris RA, Moreno-Rodriguez RA, Sugi Y, *et al*: **Periostin regulates atrioventricular valve maturation.** *Dev Biol* 2008, **316**:200-213.
21. Norris RA, Potts JD, Yost MJ, *et al*: **Periostin promotes a fibroblastic lineage pathway in atrioventricular valve progenitor cells.** *Dev Dyn* 2009, **238**:1052-1063.
22. Tkatchenko TV, Moreno-Rodriguez RA, Conway SJ, *et al*: **Lack of periostin leads to suppression of Notch1 signaling and calcific aortic valve disease.** *Physiol Genomics* 2009, **39**:160-168.
23. Zinn K, McAllister L, Goodman CS: **Sequence analysis and neuronal expression of fasciclin I in grasshopper and Drosophila.** *Cell* 1988, **53**:577-587.
24. Li G, Oparil S, Sanders JM, *et al*: **Phosphatidylinositol-3-kinase signaling mediates vascular smooth muscle cell expression of periostin in vivo and in vitro.** *Atherosclerosis* 2006, **188**:292-300.
25. Snider P, Hinton RB, Moreno-Rodriguez RA, *et al*: **Periostin is required for maturation and extracellular matrix stabilization of noncardiomyocyte lineages of the heart.** *Circ Res* 2008, **102**:752-760.
26. Wallace DP, Quante MT, Reif GA, *et al*: **Periostin induces proliferation of human autosomal dominant polycystic kidney cells through alphaV-integrin receptor.** *Am J Physiol Renal Physiol* 2008, **295**:F1463-1471.
27. Ji X, Chen D, Xu C, *et al*: **Patterns of gene expression associated with BMP-2-induced osteoblast and adipocyte differentiation of mesenchymal progenitor cell 3T3-F442A.** *J Bone Miner Metab* 2000, **18**:132-139.
28. Inai K, Norris RA, Hoffman S, Markwald RR, Sugi Y: **BMP-2 induces cell migration and periostin expression during atrioventricular valvulogenesis.** *Dev Biol* 2008, **315**:383-396.
29. Gillan L, Matei D, Fishman DA, *et al*: **Periostin secreted by epithelial ovarian carcinoma is a ligand for alpha(V)beta(3) and alpha(V)beta(5) integrins and promotes cell motility.** *Cancer Res* 2002, **62**:5358-5364.
30. Bao S, Ouyang G, Bai X, *et al*: **Periostin potently promotes metastatic growth of colon cancer by augmenting cell survival via the Akt/PKB pathway.** *Cancer Cell* 2004, **5**:329-339.
31. Baril P, Gangeswaran R, Mahon PC, *et al*: **Periostin promotes invasiveness and resistance of pancreatic cancer cells to hypoxia-induced cell death: role of the beta4 integrin and the PI3k pathway.** *Oncogene* 2007, **26**:2082-2094.
32. Butcher JT, Norris RA, Hoffman S, Mjaatvedt CH, Markwald RR: **Periostin promotes atrioventricular mesenchyme matrix invasion and remodeling mediated by integrin signaling through Rho/PI 3-kinase.** *Dev Biol* 2007, **302**:256-266.
33. Butcher JT, Norris RA, Hoffman S, Mjaatvedt CH, Markwald RR: **Periostin promotes atrioventricular mesenchyme matrix invasion and remodeling mediated by integrin signaling through Rho/PI 3-kinase.** *Dev Biol* 2007, **302**:256-266.
34. Doliana R, Bot S, Bonaldo P, Colombatti A: **EMI, a novel cysteine-rich domain of EMILINs and other extracellular proteins, interacts with the gC1q domains and participates in multimerization.** *FEBS Lett* 2000, **484**:164-168.
35. Norris RA, Damon B, Mironov V, *et al*: **Periostin regulates collagen fibrillogenesis and the biomechanical properties of connective tissues.** *J Cell Biochem* 2007, **101**:695-711.
36. Kim B, Olzmann JA, Choi S, *et al*: **Corneal dystrophy-associated H124H mutation disrupts TGFBI interaction with periostin and causes mislocalization to the lysosome.** *J Biol Chem* 2009, **284**:19580-19591.
37. Sugiura T, Takamatsu H, Kudo A, Amann E: **Expression and characterization of murine osteoblast-specific factor 2 (OSF-2) in a baculovirus expression system.** *Protein Expr Purif* 1995, **6**:305-311.
38. Takayama G, Arima K, Kanaji T, *et al*: **Periostin: a novel component of subepithelial fibrosis of bronchial asthma downstream of IL-4 and IL-13 signals.** *J Allergy Clin Immunol* 2006, **118**:98-104.
39. Kudo H, Amizuka N, Araki K, Inohaya K, Kudo A: **Zebrafish periostin is required for the adhesion of muscle fiber bundles to the myoseptum and for the differentiation of muscle fibers.** *Dev Biol* 2004, **267**:473-487.
40. Coutu DL, Wu JH, Monette A, *et al*: **Periostin, a member of a novel family of vitamin K-dependent proteins, is expressed by mesenchymal stromal cells.** *J Biol Chem* 2008, **283**:17991-18001.
41. Kim JE, Kim SJ, Lee BH, *et al*: **Identification of motifs for cell adhesion within the repeated domains of transforming growth factor-beta-induced gene, betaig-h3.** *J Biol Chem* 2000, **275**:30907-30915.
42. Kim CJ, Yoshioka N, Tambe Y, *et al*: **Periostin is down-regulated in high grade human bladder cancers and suppresses in vitro cell invasiveness and in vivo metastasis of cancer cells.** *Int J Cancer* 2005, **117**:51-58.
43. Dingwall C, Laskey RA: **Nuclear targeting sequences–a consensus?.** *Trends Biochem Sci* 1991, **16**:478-481.
44. Hynes RO: **Integrins: bidirectional, allosteric signaling machines.** *Cell* 2002, **110**:673-687.
45. Park SJ, Park S, Ahn H, Kim I, Lee B: **Conformational resemblance between the structures of integrin-activating pentapetides derived from betaig-h3 and RGD peptide analogues in a membrane environment.** *Peptides* 2004, **25**:199-205.
46. Klintworth GK: **The molecular genetics of the corneal dystrophies–current status.** *Front Biosci* 2003, **8**:d687-713.
47. Ma C, Rong Y, Radiloff DR, *et al*: **Extracellular matrix protein betaig-h3/TGFBI promotes metastasis of colon cancer by enhancing cell extravasation.** *Genes Dev* 2008, **22**:308-321.
48. Calaf GM, Echiburú-Chau C, Zhao YL, Hei TK: **BigH3 protein expression as a marker for breast cancer.** *Int J Mol Med* 2008, **21**:561-568.
49. Zhang Y, Wen G, Shao G, *et al*: **TGFBI deficiency predisposes mice to spontaneous tumor development.** *Cancer Res* 2009, **69**:37-44.
50. Norris RA, Kern CB, Wessels A, *et al*: **Detection of betaig-H3, a TGFbeta induced gene, during cardiac development and its complementary pattern with periostin.** *Anat Embryol* 2005, **210**:13-23.
51. Kim CJ, Isono T, Tambe Y, *et al*: **Role of alternative splicing of periostin in human bladder carcinogenesis.** *Int J Oncol* 2008, **32**:161-169.
52. Rani S, Barbe MF, Barr AE, Litvin J: **Periostin-like-factor and Periostin in an animal model of work-related musculoskeletal disorder.** *Bone* 2009, **44**:502-512.
53. Altschul SF, Madden TL, Schäffer AA, *et al*: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.

54. Larkin MA, Blackshields G, Brown NP, *et al*: **Clustal W and Clustal X version 2.0.** *Bioinformatics* 2007, **23**:2947-2948.
55. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**:755-763.
56. Kent WJ, Sugnet CW, Furey TS, *et al*: **The human genome browser at UCSC.** *Genome Res* 2002, **12**:996-1006.
57. Karolchik D, Kuhn RM, Baertsch R, *et al*: **The UCSC Genome Browser Database: 2008 update.** *Nucleic Acids Res* 2008, **36**:D773-779.
58. Kent WJ: **BLAT–the BLAST-like alignment tool.** *Genome Res* 2002, **12**:656-664.
59. Blanchette M, Kent WJ, Riemer C, *et al*: **Aligning multiple genomic sequences with the threaded blockset aligner.** *Genome Res* 2004, **14**:708-715.
60. Hubbard TJP, Aken BL, Ayling S, *et al*: **Ensembl 2009.** *Nucleic Acids Res* 2009, **37**:D690-697.
61. Alekseyenko AV, Kim N, Lee CJ: **Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes.** *RNA* 2007, **13**:661-670.
62. Perrière G, Gouy M: **WWW-query: an on-line retrieval system for biological sequence banks.** *Biochimie* 1996, **78**:364-369.
63. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I: **VISTA: computational tools for comparative genomics.** *Nucleic Acids Res* 2004, **32**: W273-279.
64. Brudno M, Malde S, Poliakov A, *et al*: **Glocal alignment: finding rearrangements during alignment.** *Bioinformatics* 2003, **19(Suppl 1)**:i54-62.
65. Crooks GE, Hon G, Chandonia J, Brenner SE: **WebLogo: a sequence logo generator.** *Genome Res* 2004, **14**:1188-1190.
66. Brodie R, Roper RL, Upton C: **JDotter: a Java interface to multiple dotplots generated by dotter.** *Bioinformatics* 2004, **20**:279-281.
67. Jones DT: **Protein secondary structure prediction based on position-specific scoring matrices.** *J Mol Biol* 1999, **292**:195-202.
68. Bryson K, McGuffin LJ, Marsden RL, *et al*: **Protein structure prediction servers at University College London.** *Nucleic Acids Res* 2005, **33**:W36-38.
69. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268**:78-94.
70. Taylor JS, Braasch I, Frickey T, Meyer A, Peer Van de Y: **Genome duplication, a trait shared by 22000 species of ray-finned fish.** *Genome Res* 2003, **13**:382-390.
71. Litvin J, Selim A, Montgomery MO, *et al*: **Expression and function of periostin-isoforms in bone.** *J Cell Biochem* 2004, **92**:1044-1061.
72. Yoshioka N, Fuji S, Shimakage M, *et al*: **Suppression of anchorage-independent growth of human cancer cell lines by the TRIF52/periostin/OSF-2 gene.** *Exp Cell Res* 2002, **279**:91-99.
73. Kühn B, del Monte F, Hajjar RJ, *et al*: **Periostin induces proliferation of differentiated cardiomyocytes and promotes cardiac repair.** *Nat Med* 2007, **13**:962-969.
74. Lorts A, Schwanekamp JA, Elrod JW, Sargent MA, Molkentin JD: **Genetic manipulation of periostin expression in the heart does not affect myocyte content, cell cycle activity, or cardiac repair.** *Circ Res* 2009, **104**: e1-7.
75. Shao R, Bao S, Bai X, *et al*: **Acquired expression of periostin by human breast cancers promotes tumor angiogenesis through up-regulation of vascular endothelial growth factor receptor 2 expression.** *Mol Cell Biol* 2004, **24**:3992-4003.
76. Castronovo V, Waltregny D, Kischel P, *et al*: **A chemical proteomics approach for the identification of accessible antigens expressed in human kidney cancer.** *Mol Cell Proteomics* 2006, **5**:2083-2091.
77. Zhu S, Barbe MF, Amin N, *et al*: **Immunolocalization of Periostin-like factor and Periostin during embryogenesis.** *J Histochem Cytochem* 2008, **56**:329-345.
78. Ohno S: **Evolution by Gene Duplication.** Berlin: Springer-Verlag 1970.
79. Zhang XH, Chasin LA: **Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons.** *Proc Natl Acad Sci USA* 2006, **103**:13427-13432.
80. Cardin AD, Weintraub HJ: **Molecular modeling of protein-glycosaminoglycan interactions.** *Arteriosclerosis* 1989, **9**:21-32.
81. Erat MC, Slatter DA, Lowe ED, *et al*: **Identification and structural analysis of type I collagen sites in complex with fibronectin fragments.** *Proc Natl Acad Sci USA* 2009, **106**:4195-4200.
82. Schwarz-Linek U, Werner JM, Pickford AR, *et al*: **Pathogenic bacteria attach to human fibronectin through a tandem beta-zipper.** *Nature* 2003, **423**:177-181.
83. Bingham RJ, Rudiño-Piñera E, Meenan NAG, *et al*: **Crystal structures of fibronectin-binding sites from Staphylococcus aureus FnBPA in complex with fibronectin domains.** *Proc Natl Acad Sci USA* 2008, **105**:12254-12258.
84. Hsu F, Pringle TH, Kuhn RM, *et al*: **The UCSC Proteome Browser.** *Nucleic Acids Res* 2005, **33**:D454-458.
85. Sayers EW, Barrett T, Benson DA, *et al*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2009, **37**:D5-15.

doi:10.1186/1471-2148-10-30
Cite this article as: Hoersch and Andrade-Navarro: Periostin shows increased evolutionary plasticity in its alternatively spliced region. *BMC Evolutionary Biology* 2010 **10**:30.

Table S1: Periostin in seven tetrapod and five teleost species

Species, locus (if > 1); UCSC assembly name/date	Chromosomal region	#Refseqs / mRNAs / ESTs	Observed non-canonical exons, alternative splicing (AS) [1] *Comments*	Periostin sequence based on	Missing exons [2]	Periostin protein sequence identifier [3, 4]
Human (Hs) (hg18/Mar2006)	chr13:37,034,001-37,082,000	4 / many / many	▪ Alt. 5'-most exon suggested ▪ **AS:** exons 17, 18, 19 (weakly supported), 21	NP_006466.1 [R]		>Hs_PNfl_ _NP_006466.1 [i]
Mouse (Mm) (mm9/July2007)	chr3:54,164,001-54,196,000	1 / many / many	▪ **AS:** exons 17, 20 (weakly supported), 21	NP_056599.1 [R], ESTs		>Mm_PNfl
Opossum (Md) (monDom4/Jan2006)	chr4:301,065,001-301,112,000	0 / none / none	▪ Exon 21V22 ▪ Shorter version of exon 18	Genomic sequence [H]		>Md_PNfl
Platypus (Oa) (ornAna1/Mar2007)	Ultra336:2,915,001-2,966,000	0 / none / none	▪ Exon 21V22	Genomic sequence [H]	23	>Oa_PNfl-ex23
Chicken (Gg) (galGal3/May2006)	chr1:176,287,001-176,325,000	1 / some / some	▪ Exon 21V22 ▪ **AS:** Exons 17, 18, 21, 21V22	NP_001025712.1 [R], ESTs		>Gg-PNfl
Lizard (Ac) (anoCar1/Feb2007)	scaffold_65:3,980,001-4,060,000	0 / none / none	▪ Exon 21V22	Genomic sequence [H]	23	>Ac_PNfl-ex23
Frog (Xt) (xenTro2/Aug2005)	scaffold_505:99,001-160,000	1 / some / many	▪ Exon 21V22 ▪ Cluster of 8 genomic exon 19 copies (exons 19A..H) ▪ **AS:** Exons 19A..H, 21, 21V22 ▪ *Appearance of an alt. 3'-most exon likely due to assembly artifact.* ▪ *Exon 21 obscured by gap in assembly, but evident in ESTs*	NP_001106376.1 [R], ESTs		>Xt_PNfl >Xt_PNfl+ex19A..H
Zebrafish (Dr), loc 1 (danRer5/July2007)	chr15:25,151,001-25,183,000	2 / very few / many	▪ Long repeat exon 18=19 ▪ Additional two 5'-most exons, observed in 5 ESTs ▪ **AS:** exons 20, 21, 22 ▪ *Apparent AS of exon 17 most likely a BLAT mapping artifact*	NP_981966.1 [R], ESTs		>Dr_PNfl.loc01
Zebrafish (Dr), loc 2 (danRer5/July2007)	chr10:19,780,001-19,840,000	0 / none / very few	▪ Long repeat exon 18=19 (*only partially covered by transcript evidence!*) ▪ 6 exons between 18=19 and 23, 4 of which show ~13aa repeat ▪ Alternative 5' end suggested	ESTs, genomic sequence [H]		>Dr_PNfl.loc02
Stickleback (Ga), loc 1 (gasAcu1/Feb2006)	chrI:13,650,001-13,661,000	0 / none / some	▪ Long repeat exon 18=19 ▪ **AS:** exon 22	ENSGACP00000015398 [E], ESTs		>Ga_PNfl.loc01
Stickleback (Ga), loc 2 (gasAcu1/Feb2006)	chrVII:20,256,001-20,268,000	0 / none / some	▪ Long repeat exon 18=19 ▪ Only 3 exons following exon 18=19	ENSGACP00000027192 [E], ESTs		>Ga_PNfl.loc02
Medaka (Ol), loc 1 (oryLat2/Oct2005)	chr13:7,474,001-7,487,000	0 / none / very few	▪ **AS:** exons 17, 18 (as long/short/absent), 19? ▪ No repeat exon (18=19) clearly identified. ▪ *Possibly only 20 exons total.*	ENSORLP00000004465 [E], ESTs, genomic sequence [H]	19	>Ol_PNprtl.loc01
Medaka (Ol), loc 2 (oryLat2/Oct2005)	scaffold4189:1-6,488	0 / none / very few	*Not assessable. POSTN is partly covered by short scaffold4189, which is flanked by assembly gaps.*	n/a		*Not reconstructed!*
***T. nigroviridis* (Tn), loc 1** (tetNig1/ Feb2004)	chr16:4887001-4897000	0 / very few / none	▪ Repeat exon 18=19 ▪ Only 3 exons following exon 18=19 (?)	CAG09019.1 [G]	1,2,7[p], 8	>Tn_PNprtl.loc01_ _CAG09019.1 [i]
***T. nigroviridis* (Tn), loc 2** (tetNig1/Feb2004)	chr7:2,694,001-2,709,000	0 / none / none	▪ Long repeat exon 18=19 (*not covered by transcript evidence!*)	GSTENT00018874001_prot [E], genomic sequence [B]	17	>Tn_PNprtl.loc02
***T. rubripes* (Tr), loc 1** (fr2/Oct2004)	chrUn:57,690,001-57,701,000	0 / none / none		NEWSINFRUP00000152208 [E]	1[p], 16ff	>Tr_PNprtl.loc01__ N.00000152208 [i]
***T. rubripes* (Tr), loc 2** (fr2/Oct2004)	chrUn:234,103,001-234,123,000	0 / none / none	▪ Long repeat exon 18=19 (*not covered by transcript evidence!*)	NEWSINFRUP00000144260 [E]	17, 20ff	>Tr_PNprtl.loc02__ N.00000144260 [i]

[1] Non-canonical exons and alternative splicing based on transcript data as available in UCSC browser. Non-canonical exons relative to periostin exon structure as established for human and mouse (23 exons). Listing of an exon under 'AS' implies alternative presence or absence unless otherwise noted. [2] Missing exons were not identified despite our efforts, were missing from public sequences (Tn loc 1, Tr), or were not incorporated into reconstructed sequence (Ol), but are thought to be present in reality. [3] All sequences provided in Additional file 5. [4] "fl" identifies full-length sequences (missing exon 23 acceptable), "prtl" identifies partial sequences.
[R] RefSeq database, [H] via homology searches, [E] Ensembl database, [G] GenPept database; [p] partial, [i] Identical to public sequence indicated in identifier

Periostin shows increased evolutionary plasticity in its alternatively spliced region
Sebastian Hoersch and Miguel A. Andrade-Navarro
Additional file 2

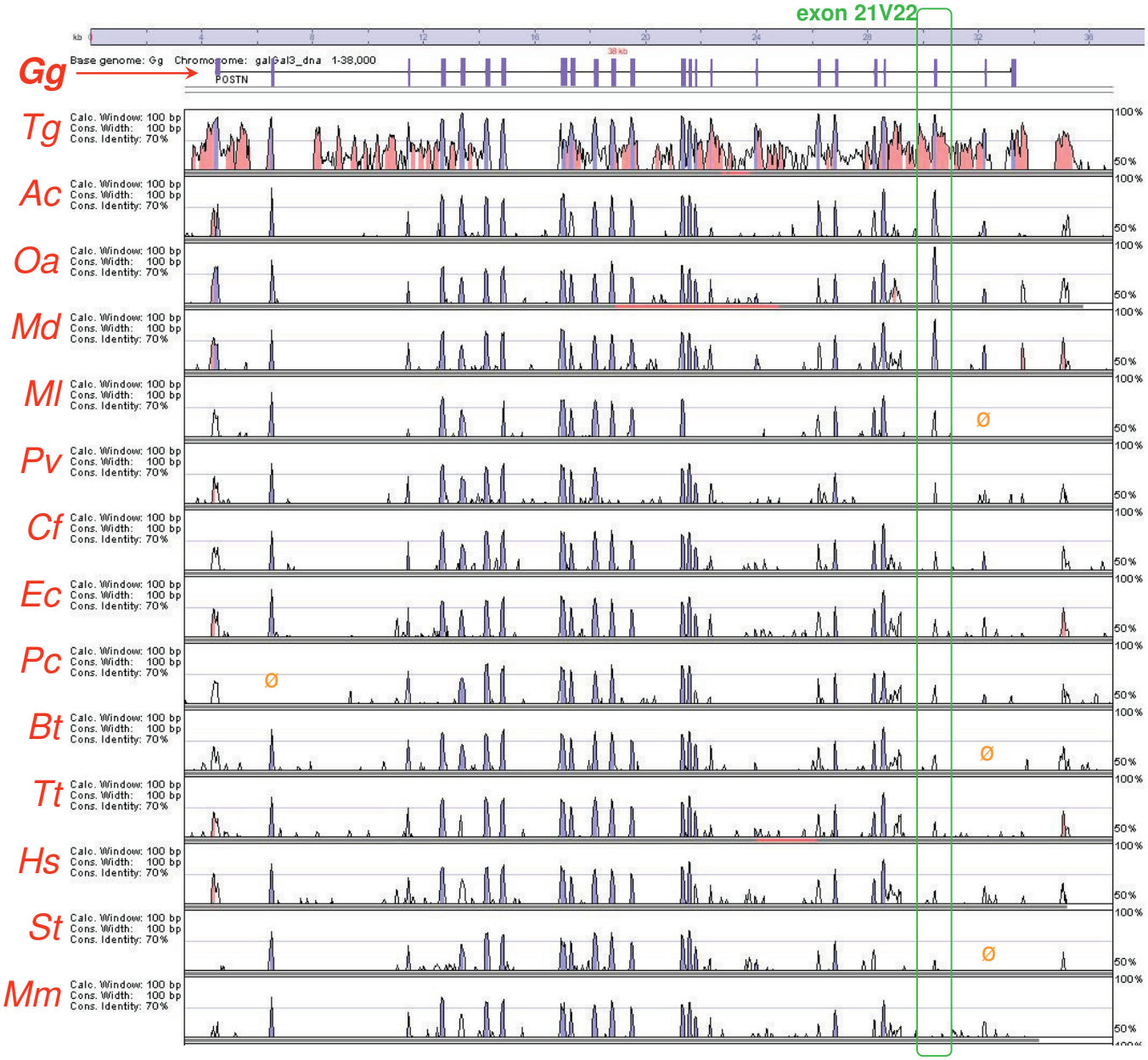


Figure S1: Genomic sequence alignment of the periostin locus of 15 species*) obtained with VISTA.
Alignments are relative to chicken periostin (G*g*, top). Conservation curves from 14 other species are displayed below, arranged manually by decreasing conservation of exon 21V22 (green frame). Conservation peaks of exons are colored blue, of non-coding sequence red. While conservation levels stay generally high (above the 70% mark indicated by the faint horizontal line) for most exons or, in some cases, are universally low due to exon shortness (for example, exon 17), the peak for exon 21V22 displays the widest range of conservation level of all exons, from very highly conserved (Tg, Ac, Oa, Md, coming close to the 100% line) to not registering in the VISTA output (Mm, not exceeding the 50% line), with other species on a continuum in-between.

*)***Gg**: chicken (*Gallus gallus*), ***Tg***: zebra finch (*Taeniopygia guttata*), ***Ac***: anole lizard (*Anolis carolinensis*), ***Oa***: platypus (*Ornithorhynchus anatinus*), ***Md***: opossum (*Monodelphis domestica*), ***Ml***: microbat (*Myotis lucifugus*), ***Pv***: megabat (*Pteropus vampyrus*), ***Cf***: dog (*Canis familiaris*), ***Ec***: horse (*Equus caballus*), ***Pc***: hyrax (*Procavia capensis*), ***Bt***: cow (*Bos taurus*), ***Tt***: dolphin (*Tursiops truncates*), ***Hs***: human (*Homo sapiens*), ***St***: squirrel (*Spermophilus tridecemlineatus*), ***Mm***: mouse (*Mus musculus*).

The **Ø** symbol indicates regions of partially missing genomic sequence.

Additional file 3: Tables S2 and S3

Table S2: Select genomic alignments of periostin exon 21V22 and exon 17 produced by VISTA

In the representation below, highly conserved regions were background highlighted by VISTA [63], blue for coding sequence, pink otherwise. Coding sequence for the respective exon was otherwise manually rendered in blue font. Intronic splice sites are highlighted in green, relevant stop codons in red, and the codon corresponding to the central chicken exon21V22 proline in yellow, if conserved, or magenta, if mutated. Translations of exon sequences (bold) are given below the alignment blocks, in red font where frameshifted relative to a functional exon (for exon21V22).

Table S2A: Exon 21V22

Top: chicken [>Gg galGal3_dna:30581-30931 (+)]
Bottom: zebra finch [>Tg taeGut1_dna|_RevComp_|:29612-29959 (+)]

```
000030581  AAGACCACAGCATGATTACTGTGACCTGCATACATCTACACTCCCAAAA-TGACATTCTT
>>>>>>>>>         | ||||||      ||   ||||| ||| |||  ||||||||| ||||||| ||
000029612  ------ATAGCATG------GTCGTCTGCA-ACACCTA--CTCCCAAAAATGACATTGTT

000030640  TCCACTGAAGATTTCT-------------ATGCTATCCCAGATTTAACTCATTTGTTTTT
>>>>>>>>>  ||| |||| |  | ||             ||||||| |||| ||||||||||||||||||
000029657  TCCTCTGAGGGCTCCTTGCATATGGAAAAATGCTATGCCAGCTTTAACTCATTTGTTTTT

000030687  AATGCTTATCATAGCTGGAACTGAGTACACCAAGGTTACTAAAGTAATTGAGGGAGAACC
>>>>>>>>>   ||||||| |||||| ||||||||||||||||||||||||||||||||||||||||||| ||
000029717  GATGCTTACCATAGCCGGAACTGAGTACACCAAGGTTACTAAAGTAATTGAGGGAGAGCC

000030747  ACAGATTATCGAGAGAGAAATCAAGAAAGTCCATTTGGAAGGTTAGTTTGTCTATAAATC
>>>>>>>>>  |||||||||||||||||||||||||||||| ||| |||||||||||||| || | ||||
000029777  ACAGATTATCGAGAGAGAAATCAAGAAAGTTCATCTGGAAGGTTAGTTTTTCAA-AAATT

000030807  CAATATCACTAAGGCCTCCTCTGTCTGAAGTAATCATTCTAGGTGCCAATTTATAGTATT
>>>>>>>>>   | |||||||||| | | || ||||  |||| ||||||| |||||     |||||  |||
000029836  AAGTATCACTAAGACTTGCTTTGTCCAAAGTGATCATTCCAGGTGAGGTTTTATCATATC

000030867  TATTCACAAAAGTTACTTAGGTCTGAAGTCAATATTTTAGCAAGAGTTTTATGTTTGATT
>>>>>>>>>  ||||||  ||| ||| ||||||||| || ||  | |||||||||||||||||  ||| ||
000029896  TATTCATTAAAATTA-TTAGGTCTGGAGCCAGCACTTTAGCAAGAGTTTTATAATTGTTT
```

```
hoersch@luria:~/bin/Tools/translateDNA.pl -ks -frames rna -minlength 5
>Tg_ex21V22 (len=87)
CCGGAACTGAGTACACCAAGGTTACTAAAGTAATTGAGGGAGAGCCACAGATTATCGAGAGAGAAATCAAGAAAGT
TCATCTGGAAG

>Tg_ex21V22|F3_1|[3-86]
GTEYTKVTKVIEGEPQIIEREIKKVHLE
||||||||||||||||||||||||||||
GTEYTKVTKVIEGEPQIIEREIKKVHLE (Gg)
```

Top: chicken [>Gg galGal3_dna:30581-30931 (+)]
Bottom: opossum [>Md monDom4_dna|_RevComp_|:11489-11808 (-)]

```
000030581  AAG--ACCACAGCA-------------TGATTACTG---TGACCTGCATACATCTACACT
>>>>>>>>>   |   |||| ||||              ||| || ||   |  |   |   ||    | |
000011808  GAATTACCAAAGCACATACACCTTTTTTGAGTAATGTTCTATCTCACTATCACACTCGCC

000030623  CCCAAAATGACATTCTTTCCACT-------GAAGATTTCTATGCTATCCCAGATTTAAC-
>>>>>>>>>   | ||||||       |||||         |||   |||| | | |    | ||| ||
000011748  TCTAAAATGCTCCCAATTCCATCATCCAAGGAAATGTTCTCTCCAAAATTATATTGAATT

000030675  ------------TCATTTGTTTTTAATGCTTATCATAGCTGGAACTGAGTACACCAAGGT
>>>>>>>>>             | | |||||| |||||||||| |||||||||||||||||||||||||
000011688  GTACAACTGTGATTACTTGTTTCTAATGCTTATCTTAGCTGGAACTGAGTACACCAAGGT

000030723  TACTAAAGTAATTGAGGGAGAACCACAGATTATCGAGAGAGAAATCAAGAAAGTCCATTT
>>>>>>>>>   |||||||||||||| | |||||||||||||||| |  ||||||||||||||||
000011628  CACTAAAGTAATTGAGAGTGAACCACAGATTATCGAAACGGAAATCAAGAAAGTCCTCCA

000030783  GGAAGGTTAGTTTGTCTATAAATCCAATATCACTAAGGCCTCCTCTGTCTGAAGTAATCA
>>>>>>>>>  || ||||||||||||||||   || || | | ||||||   |
000011568  GGCAGGTTAGTTTGTCTATTCCTCTAACAGCCCTAAGGTTCC------------------

000030843  TTCTAGGTGCCAATTTATAGTATTTATTCACAAAAGTTACTTAGGTCTGAAGTCAATATT
>>>>>>>>>                                        ||  |     ||| ||||
000011526  -------------------------------------TATCTTCACTTGATGTCA-----
```

```
hoersch@luria:~/bin/Tools/translateDNA.pl -ks -frames rna -minlength 5
>Md_ex21V22 (len=87)
CTGGAACTGAGTACACCAAGGTCACTAAAGTAATTGAGAGTGAACCACAGATTATCGAAACGGAAATCAAGAAAGT
CCTCCAGGCAG

>Md_ex21V22|F3_1|[3-86]
GTEYTKVTKVIESEPQIIETEIKKVLQA
|||||||||||| |||||| |||||
GTEYTKVTKVIEGEPQIIEREIKKVHLE (Gg)
```

Exon 21V22 *continued*

Top: chicken [>Gg galGal3_dna:30581-30931 (+)]
Bottom: *M. lucifugus* [>little.bat GeneScaffold_2401|_RevComp_|:245049-245249 (+)]

```
000030581  AAGACCACAGCATGATTACTGTGACCTGCATACATCTACACTCCCAAAATGACATTCTTT
>>>>>>>>>    |      |   || || | | |                    || | |||  |||||
000245049  CGGGTACTAATATTATGATTTT-------------------TCACTAAAATACATT----

000030641  CCACTGAAGATTTCT------------ATGCTATCCCAGATTTAACTCATTTGTTTTTAA
>>>>>>>>>         |||||  ||             ||| | || || | |   |    |   |||||| |
000245086  -------AGATTCCTTAACTAAAAAAAATGTTCTCACAAAATGCATAAA---GTTTTTCA

000030689  TGCTTATCATAGCTGGAACTGAGTACACCAAGGTTACTAAAGTAATTGAGGGAGAACCAC
>>>>>>>>>  | |||||| |||    |||||  |  |||||   |||||||||||| ||| ||||    |
000245136  TTCTTATCTTAGGCCAAACTGCATTTACCAAAAGTACTAAAGTAATTAAGGAAGAATATC

000030749  AGATTATCGAGAGAGAAATCAAGAAAGTCCATTTGGAAGGTTAGTTTGTCTATAAATCCA
>>>>>>>>>  | | ||| |||| || || ||| || |||    |||||||||
000245196  AAACTATGGAGAAAGGAACCAAAAAGGTCA---TGGAAGGTT------------------

000030809  ATATCACTAAGGCCTCCTCTGTCTGAAGTAATCATTCTAGGTGCCAATTTATAGTATTTA
>>>>>>>>>
000000000  ------------------------------------------------------------

000030869  TTCACAAAAGTTACTTAGGTCTGAAGTCAATATTTTAGCAAGAGTTTTATGTTTGATTTG
>>>>>>>>>  || |||||  | |||
000245235  TTGACAAATCTGACT-----------------------------------------------
```

```
hoersch@luria:~/bin/Tools/translateDNA.pl -ks -frames rna
>little.bat_ex21V22 (len=84)
GCCAAACTGCATTTACCAAAAGTACTAAAGTAATTAAGGAAGAATATCAAACTATGGAGAAAGGAACCAAAAAGGT
CATGGAAG

>little.bat_ex21V22|F3_1|[3-83]
QTAFTKSTKVIKEEYQTMEKGTKKV-ME
 |  || ||||  |  |  |   |||  |
GTEYTKVTKVIEGEPQIIEREIKKVHLE (Gg)
```

Top: chicken [>Gg galGal3_dna:30581-30931 (+)]
Botton: dog [>Cf canFam2_dna:38906-39134 (+)]

```
000030581  AAGACCACAGCATGATTACTGTGACCTGCATACATCTACACTCCCAAAATGACATTCTTT
>>>>>>>>>                                  ||||  ||
000038906  -------------------------------ACATTTA----------------------

000030641  CCACTGAAGATTTCTATGCTATCCCAGATTTAACTCATTTGTTTTTAATGCTTATCATAG
>>>>>>>>>                                                        ||  |||
000038913  ------------------------------------------------------TCTTAG

000030701  CTGGAACTGAGTACACCAAGGTTACTAAAGTAATTGAGGGAGAACCACAGATTATCGAGA
>>>>>>>>>   |  | ||   |  ||  |  |||||||||| || ||||||||   ||   ||| ||||
000038919  GTAAAGCTATATTTACTGAACCTACTAAAGTA-TTAAGGGAGAATATCAACCTATAGAGA

000030761  GAGAAATCAAGAA--AGTCCATTTGGAAGGTTAGTTTGTCTATAAATC---CAATATCAC
>>>>>>>>>  || ||  || ||  ||||  ||| || ||||      |  | |||||   ||||||  |
000038978  AAGGAACAAAAAAGAAGTCATTTTAGATGGTT------TTGACAAATCTGCCAATATAGC

000030816  TAAGGCCTCCTCTGTCTGAAGTAATCATTCTAGGTGCCAATTTATAGTATTTATTCACAA
>>>>>>>>>          ||| |||  | | ||         ||   ||  || |||   |   |
000039032  ATGTAAAGCCTGTGTAGGGAATA---------GGGATCAGCTTGTAGGTCTGGCTTT---

000030876  AAGTTACTTAGGTCTGAAGTCAA--TATTTTAGCAAG---AGTTTTATGTTTGATTTGGA
>>>>>>>>>     | | || ||    || ||||  || || || |||   | | || |  ||| |||
000039080  ---TGATTTGGGCTCTAATTCAAACTACTTAAGTAAGTTAATTCTTCTACTTGGTTTC--
```

```
hoersch@luria:~/bin/Tools/translateDNA.pl -ks -frames rna
>Cf_ex21V22 (len=88)
GTAAAGCTATATTTACTGAACCTACTAAAGTATTAAGGGAGAATATCAACCTATAGAGAAAGGAACAAAAAAGAAG
TCATTTTAGATG
>Cf_ex21V22|F3_1|[3-56]
KAIFTEPTKVLRENINL*
```

Exon 21V22 *continued*

Top: chicken [>Gg galGal3_dna:30581-30931 (+)]
Bottom: human [>Hs hg18_dna:43690-44217 (+)]

```
000030644  CT-------------GAAGATTTCTATGCTATCCCAG-----------------------
>>>>>>>>>  ||             | | ||||||     ||   ||
000043929  CTTTTGGTAATTTGAGTATATTTCTCAATCATTTAAGTCATGACTATTAACATGGCTAAC

000000000  ------------------------------------------------------------
>>>>>>>>>
000043989  TGGTACTTTGAGTACTGATGTTGTGATTTGTCACTGATACACTTCCAAGACCTTAACTGA

000030668  -----------ATTTAACTCATTTGTTTTTAATGCTTATCATAGCTGGAACTGAGTACAC
>>>>>>>>>             ||| ||  |||    ||||   ||||| | |||    || |  | |  |
000044049  AAAAATTTCTCATTAAATGCATAAAATTTTTGGGCTTACCTTAG----AATTATGCATTC

000030717  CAAGGTTACTAAAGTAATTGAGGGAGAACCACAGATTATCGAGAGAGAAATCA----AGA
>>>>>>>>>   |    ||| ||| |||| | ||||||  ||||| |||   || || ||  |    | |
000044105  TGAAACGACTCAAGCAATTAAAGGAGAATAACAGACTATGTCGAAAGGAACAACAACAAA

000030773  AAGTCCATTTGGAAGGTTAGTTTGTCTATAAATCCAATATCACTAAGGCCTCCTCTGTCT
>>>>>>>>>  |||||  ||| |||||||
000044165  AAGTCATTTTTGAAGGT-----------------------------------------

000030833  GAAGTAATCATTCTAGGTGCCAATTTATAGTATTTATTCACAAAAGTTACTTAGGTCTGA
>>>>>>>>>                                    ||| ||||  |
000044183  ------------------------------------TTCGCAAATCT-------------

000030893  AGTCAATATTTTAGCAAGAGTTTTATGTTTGATTTGGAA
>>>>>>>>>   | ||| ||      ||||| || |  ||||
000044194  -GCCAACAT------AAGAGGTTAAGATTTG--------
```

```
hoersch@luria:~/bin/Tools/translateDNA.pl -ks -frames rna -minlength 5
>Hs_ex21V22 (len=87)
AATTATGCATTCTGAAACGACTCAAGCAATTAAAGGAGAATAACAGACTATGTCGAAAGGAACAACAACAAAAAGT
CATTTTTGAAG
>Hs_ex21V22|F3_1|[3-86]
LCILKRLKQLKENNRLCRKEQQQKVIFE
```

Top: chicken [>Gg galGal3_dna:30581-30931 (+)]
Bottom: mouse [>Mm mm9_dna:27269-27702 (+)]

```
000030581  AAGACCACAGCATGATTACTGTGACCT-----GCATACATCTACACTCCCAAAATGACAT
>>>>>>>>>    |   |      |||   | ||| ||      ||  ||| | | ||| ||   | |||
000027269  CTGGAGA------GATGGATCTGAGCTTAAGAACACTCAT-TGCTCTCTCAGGGTAACAG

000030636  TCTTTCCACTGAAGATTTCTATGCTATCCCAGATTTAACTCATTTGTTTTTAATGCTTAT
>>>>>>>>>   || |          |||||  |||  | ||||     ||    ||  |||||  |   |
000027322  GCTCT---------GTTTCTCAGCTTCCACAGAGCAGCCTGGACTGCCTTTAACTCCAGT

000030696  CATAG-----CTGGAAC-----TGAGTACACCAAGGTTACTA-----AAGTAATTGAGGG
>>>>>>>>>     ||     |||| |      |  |  | |||  | |||||     || | ||  |
000027373  TCCAGGGGCTCTGGCATCCTTTTCTGGTCTCCACAGGTACTACATACAAATGATGCATAT

000030741  AGAACCACA------------------------GATTATCGAGAGAGAAATCAAGAAAG
>>>>>>>>>  | |  ||||                        || |      |||   |  | ||
000027433  AAACTCACACCCATGCACACACATACATACACATAATAAATAGAGGAGGCCTTTAAAAGA

000030776  TCCATTTGGAAGGTTAGTTTGT-CTATAAATCCAATATCACTAAGGCCTCCTCTGTCTGA
>>>>>>>>>  |  | |      |||| ||| | ||  |     || ||    |||||   ||   | |
000027493  TGTACTATAGTAACTAGTGTGTGCGATGTAGTTTCTAGCATCTGGGCCTGAGCTAGATAA

000030835  AGTAA-----TCATTCTAG---------------------------------GTGCCAA
>>>>>>>>>    ||      || ||||||                                  | | ||
000027553  GTTACCTTTCTCTTTCTAGCTCTCTAAATGTTTCTTCTCCCACATCCACTGGAGAGTCAC

000030856  TTTATAGTATTT-------------------ATTCACAAAAGTTACTTAGGTCTGAAGTC
>>>>>>>>>  ||  ||  | ||                   |  ||  ||||| |  | || ||     |
000027613  TTCCTAACACTTTCATGTTGAGATATCTGAGAGCCAGGAAAGTAAACTCGGGCT-----C
```

```
hoersch@HoerschBook:~/bin/Tools/translateDNA.pl -ks -frame rna
>Mm_exon21V22 (len=131 or 124?)
GGGCTCTGGCATCCTTTTCTGGTCTCCACAGGTACTACATACAAATGATGCATAT
AAACTCACACCCATGCACACACATACATACACATAATAAATAGAGGAGGCCTTTAAAAGA
TGTACTATAGTAACTA

>Mm_exon21V22|F3_1|[3-98]
ALASFSGLHRYYIQMMHINSHPCTHIHTHNK*
>Mm_exon21V22|F2_1|[2-91]
GSGILFWSPQVLHTNDAYKLTPMHTHTYT*
>Mm_exon21V22|F1_1|[1-48]
GLWHPFLVSTGTTYK*
```

Table S2B: Exon 17

```
Top: human [>Hs hg18_dna:36101-36400 (+)]
Bottom: mouse [>Mm mm9_dna:20236-20526 (+)]

000036101  ATCCTGTGCCTTAGTATCCTACAGAGACAAAAAAGAAAAGGAAAAAAAAAAAAAAACCTTT
>>>>>>>>>  | |   |   |||    |  | |||||| |||||| | |                  ||||
000020236  AGCTGATAAATTAACGCCTGATAGAGACCAAAAAGGACAA----------------CTTT

000036161  GTTCAGGTT---TTTATGTAAATATTTTTCATGTTCCACTTTCTATACTTTCAGCAACTA
>>>>>>>>>   | |  |||   | ||  ||||  |  |||||| ||||||||||||  ||||||||||||
000020280  ATCCGAGTTGTTTCTAGTTAAACTTCGTTCATGCTCCACTTTCTATGTTTTCAGCAACTA

000036218  AAATTATAACCAAAGTTGTGGAACCAAAAATTAAAGTGATTGAAGGCAGTCTTCAGCCTA
>>>>>>>>>  |||||||||||||||| |||||||||||||||||||| ||| ||||||||||||||||||
000020340  AAATTATAACCAAAGTCGTGGAACCAAAAATTAAAGTCATTCAAGGCAGTCTTCAGCCTA

000036278  TTATCAAAACTGAAGGTAAAAATCAATGTTATACATGGTTCCTT---TTATAGTAAAAAC
>>>>>>>>>  |||||||||| |||||||||   |||   | | ||||||||||   || | | ||||||
000020400  TTATCAAAACGGAAGGTAAAG-CCAACACTCTCCATGGTTCCTTACATTCTGGAAAAAAC

000036335  ATCTGTAAAAAT--ACTGATCTCTATGAATGTATCTACATTCTTTTAAAAACTAACATGT
>>>>>>>>>  || |     |||  ||| |||||| ||||| || ||||||||||||  ||||||||||
000020459  ATATCATGTAATCAACTCATCTCTCTGAATATACCTACATTCTTTTTGAAACTAACATGC

hoersch@luria:~/bin/Tools/translateDNA.pl -ks -frames rna -minlength 5
>Mm_ex17 (len=81)
CAACTAAAATTATAACCAAAGTCGTGGAACCAAAAATTAAAGTCATTCAAGGCAGTCTTCAGCCTATTATCAAAAC
GGAAG

>Mm_ex17|F3_1|[3-80] (len=81)
TKIITKVVEPKIKVIQGSLQPIIKTE
||||||||||||||| ||||||||||
TKIITKVVEPKIKVIEGSLQPIIKTE (Hs)
```

```
Top: human [>Hs hg18_dna:36101-36400 (+)]
Bottom: chicken [>Gg galGal3_dna:23299-23440 (+)]

000036101  ATCCTGTGCCTTAGTATCCTACAGAGACAAAAAAGAAAAGGAAAAAAAAAAAAAAACCTTT
>>>>>>>>>    || |    |||  |||                            |  ||  | ||  |
000023299  GACCAGATGATTACCATC-------------------------AGGAATTATAATTTCC

000036161  GTTCAGGTTTTTATGTAAATATTTTTCATGTTCCACTTTCTATACTTTCAGCAACTAAAA
>>>>>>>>>   || ||  |         ||| |  || ||  |||| ||||||||||||||||
000023333  ATTAAGAATG------GAATGTGCTTTATACTCCATTTTCTATACTTTCAG---------

000036221  TTATAACCAAAGTTGTGGAACCAAAAATTAAAGTGATTGAAGGCAGTCTTCAGCCTATTA
>>>>>>>>>                            |||||||  | ||||||  |||   | ||||||||  |
000023378  ------------------------AAATTAACATAATTGAAAGCAACGTACAGCCTATCA

000036281  TCAAAACTGAAGGTAAAAATCAATGTTATACATGGTTCCTTTTATAGTAAAAACATCTGT
>>>>>>>>>  ||| ||  |||||||||               |||  ||||
000023414  TCAGAAAGGAAGGTAAA--------------TGGGGCCTTC------------------

000036341  AAAAATACTGATCTCTATGAATGTATCTACATTCTTTTAAAAACTAACATGTTAAATGTT
>>>>>>>>>
000000000  ----------------------------------------------------------------

hoersch@luria:~/bin/Tools/translateDNA.pl -ks -frames rna -minlength 5
>Gg_ex17 (len=48)
AAATTAACATAATTGAAAGCAACGTACAGCCTATCATCAGAAAGGAAG


>Gg_ex17|F3_1|[3-47]
...........INIIESNVQPIIRKE
           |  ||   ||||  |
TKIITKVVEPKIKVIEGSLQPIIKTE (Hs)
```

Table S3: *X. tropicalis* transcript sequences covering the periostin C-terminal region.

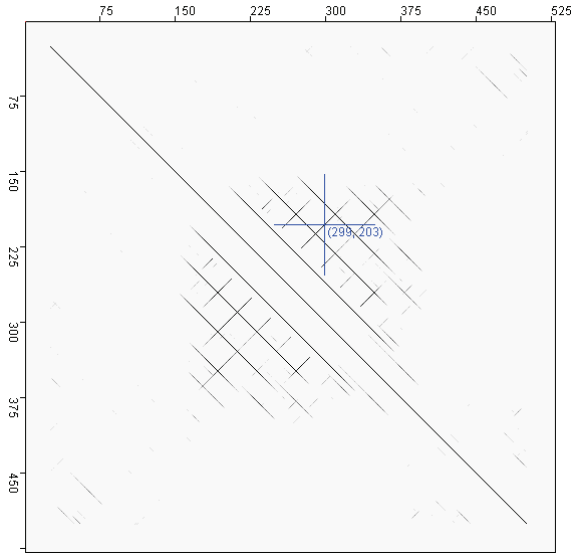
For all Genbank transcript sequences found in the UCSC genome browser (xenTro2/Aug2005), the table lists accession numbers, library information, and exon structure with respect to the cluster of exons 19A – 19H and to exons 21 and 21V22. With one exception (BC154911), ESTs from embryonic and metamorphic frogs show presence of at least a subset of exons 19A – H (yellow), while ESTs from adult frogs do not (blue). No such obvious correlation is apparent for exons 21 and 21V22, with exon 21V22 occurring in both developing and adult frogs. Exon 21 is observed in developing frogs only, but is not always present.

mRNA/EST accession	Library Info	Exon status re. the exon 19A-H cluster [1]	Exon status re. exons 21, 21V22 [1]
BC154911	NICHD_XGC_tropTail_m / tail, tga strain, metamorphic	19-20	20-22
DT424643	NIH_XGC_tropSki1 / Skin / Adult	19-20	n/a
CR412272	XGC-tailbud / tailbud (stage 28-30)	19-A-B-C^	n/a
CF345375	NICHD_XGC_Swb1N / whole body / 10 month old male	19-20	20-22
DR891672	NIH_XGC_tropTad5 / whole embryo / Tadpole (st. 36-41)	19-A-B-C-D^	n/a
BX739418	XGC-tadpole / tadpole (stage 35-40) /	19-A-B-D-E-F-20	n/a
CX380890	NIH_XGC_tropTad5 / whole embryo / Tadpole (st. 36-41)	19-A-B-C-D-E-F^	n/a
CX409969	NIH_XGC_tropTad5 / whole embryo / Tadpole (st. 36-41)	19-A-20	20-21-22
DT430191	NIH_XGC_tropSki1 / Skin / Adult	19-20	20-21V22-22
DT425263	NIH_XGC_tropSki1 / Skin / Adult	19-20	20-22
BX706184	XGC-tadpole / tadpole (stage 35-40)	^A-B-D-E-F-G-H-20	20-21-22
CX374357	NIH_XGC_tropTad5 / whole embryo / Tadpole (st. 36-41)	^A-B-C-D-E-F-G-H-20	na
CX320728	NIH_XGC_tropTad5 / whole embryo / Tadpole (st. 36-41)	^B-D-E-F-G-H-20	20-21-22
EL700850	NICHD_XGC_tropLimb_m / Limb / Metamorphic	^D-E-F-G-H-20	20-21-21V22-22
BX739786	XGC-tadpole / tadpole (stage 35-40)	^E-F-G-H-20	20-21-22
BX716782	XGC-tadpole / tadpole (stage 35-40)	^F-G-H-20	20-21-22
BX718380	XGC-tadpole / tadpole (stage 35-40)	^G-H-20	20-21-22
EL712494	NICHD_XGC_tropLimb_m / Limb / Metamorphic	^G-H-20	20-21-22
EL846058	NICHD_XGC_trop_25 / whole / Stage 25	^G-H-20	20-21-22
BX719218	XGC-tadpole / tadpole (stage 35-40)	^G-H-20	20-21-22
EL701127	NICHD_XGC_tropLimb_m / Limb / Metamorphic	^G-20	20-21-22
BX740939	XGC-tadpole / tadpole (stage 35-40)	^G-H-20	20-21-22
CR434701	XGC-tailbud / tailbud (stage 28-30)	^G-H-20	20-21-22
EL710444	NICHD_XGC_tropLimb_m / Limb / Metamorphic	^G-H-20	20-21V22-22
BX709491	XGC-tadpole / tadpole (stage 35-40)	^H-20	20-21-22
EL712816	NICHD_XGC_tropLimb_m / Limb / Metamorphic	^H-20	20-21V22-22
BX721294	XGC-tadpole / tadpole (stage 35-40)	^H-20	20-21-22
CX336698	NIH_XGC_tropTad5 / whole embryo / Tadpole (st. 36-41)	^H-20	20-21-22
CX317169	NIH_XGC_tropTad5 / whole embryo / Tadpole (st. 36-41)	^H-20	20-21-22
CX344015	NIH_XGC_tropTad5 / whole embryo / Tadpole (st. 36-41)	n/a	^21V22-22
CF589667	NICHD_XGC_Swb1N / whole body / 10 month old male	n/a	^21V22-22
CX318515	NIH_XGC_tropTad5 / whole embryo / Tadpole (st. 36-41)	n/a	20-21-22
CX346636	NIH_XGC_tropTad5 / whole embryo / Tadpole (st. 36-41)	n/a	20-21-22
CF378276	NICHD_XGC_Swb1N / whole body / 10 month old male	n/a	20-22
CF377029	NICHD_XGC_Swb1N / whole body / 10 month old male	n/a	20-21V22-22
DT431619	NIH_XGC_tropSki1 / Skin / Adult	n/a	20-21V22-22
CX328322	NIH_XGC_tropTad5 / whole embryo / Tadpole (st. 36-41)	n/a	20-21-21V22-22
CX395892	NIH_XGC_tropTad5 / whole embryo / Tadpole (st. 36-41)	n/a	20-21-22
DT431618	NIH_XGC_tropSki1 / Skin / Adult	n/a	^21V22-22

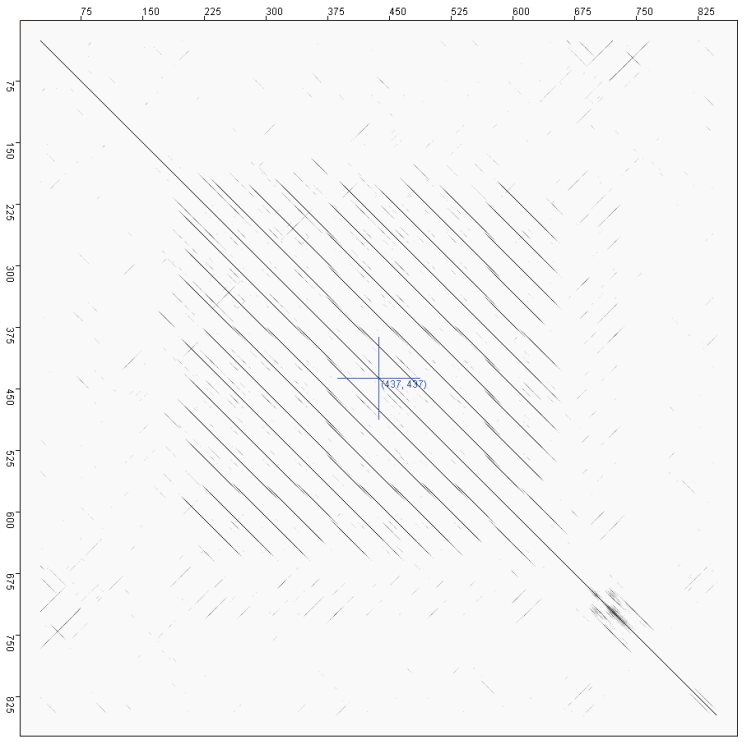
[1] A '^' symbol identifies exons partially covered.

Periostin shows increased evolutionary plasticity in its alternatively spliced region
Sebastian Hoersch and Miguel A. Andrade-Navarro
Additional file 4: Figures S2 and S3

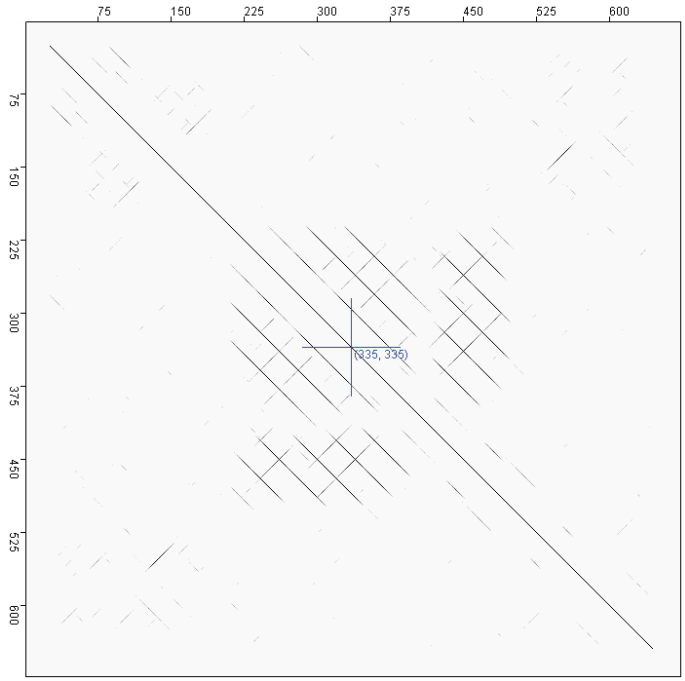
***D. rerio*, POSTN locus 1, exon 18=19 +/-150 nt's**



***D. rerio*, POSTN locus 2, exon 18=19 +/-150 nt's**



***T. nigroviridis*, POSTN locus 2, exon 18=19 +/-150 nt's**



***G. aculeatus* , POSTN locus 1, exon 18=19 +/-150 nt's**

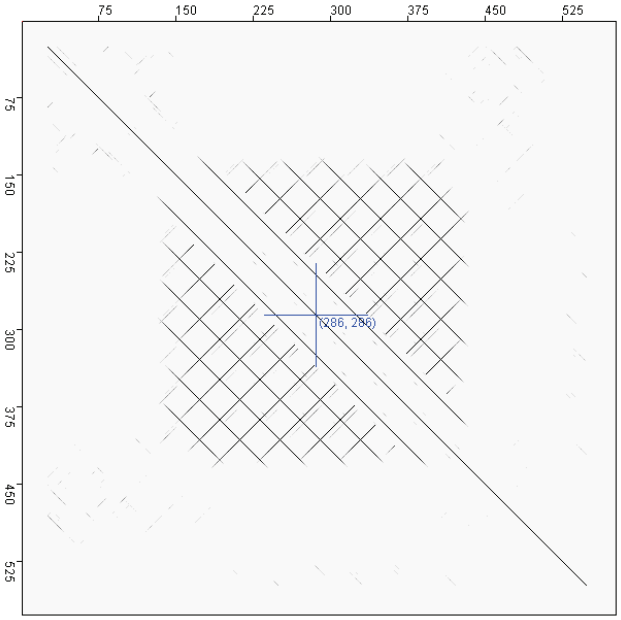
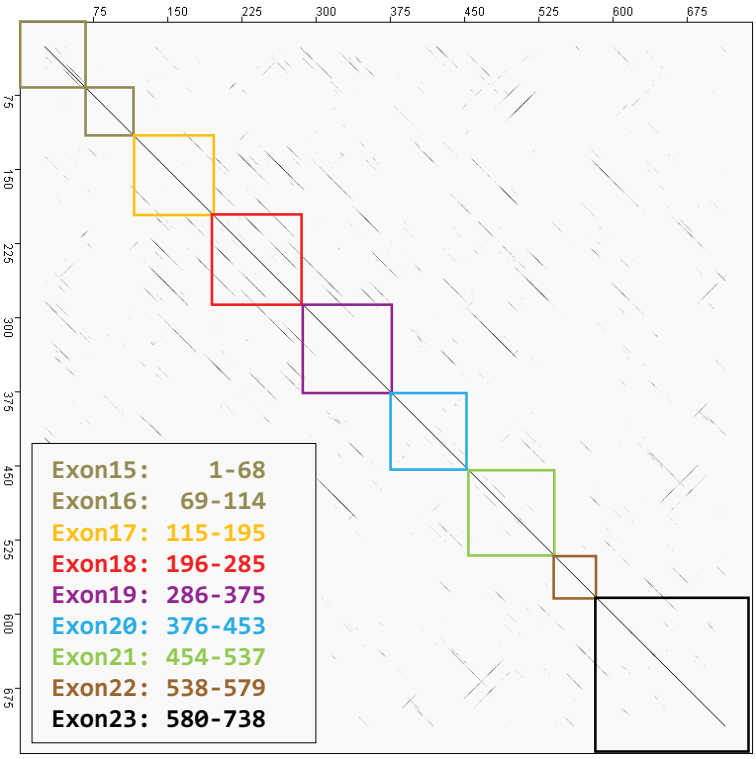


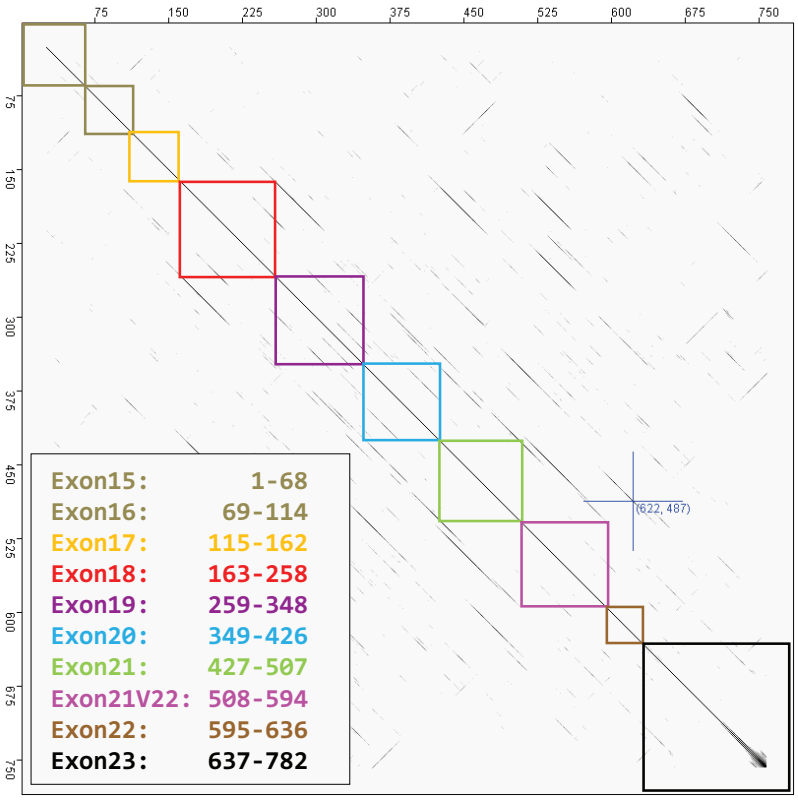
Figure S2: Matrix dot plots of teleost periostin nucleotide sequence comprising exon 18=19 against themselves.
In all four examples, exon 18=19 sequence is flanked by 150 nucleotides of intron sequence on either side. It is clearly visible that the repeat structure does generally not extent beyond the exon boundaries. The varying degree of similarity to the reverse complement repeat sequence is also obvious, being strongest in *G. aculeatus*, locus 1 and basically invisible in *D. rerio*, locus 2.

Figure S3: Matrix dot plots of human (top) and chicken (bottom) periostin nucleotide sequences, exons 15 – 23 against themselves.
Colored boxes delineate exons as indicated in the plots' legends. The repetitive structure and similarities within and among exons 17 and 21 (21V22 for chicken) are clearly visible.

Human periostin, exons 15 – 23



Chicken periostin, exons 15 – 23



Periostin shows increased evolutionary plasticity in its alternatively spliced region

Sebastian Hoersch and Miguel A. Andrade-Navarro

Additional file 6, text with Figures S4 – S7

Evolutionary and phylogenetic considerations regarding periostin and its paralog, TGFBI

According to our survey of the SMART domain database (http://smart.embl-heidelberg.de/, data not shown), periostin and TGFBI – with their sequence of secretion signal, EMI domain and tetrad of FAS1 domains – represent one of only two vertebrate contexts for FAS1 domains, the second being the presence of FAS1 domains in the scavenger receptors stabilin 1 and 2 (STAB1 and STAB2, also FEEL-1/2 and FELE-1/2), where as many as seven FAS1 domains occur interspersed with other domains.

Outside the vertebrate group, candidate genes encoding the domain architecture "signal sequence – EMI domain – 4x FAS1 domain" can additionally be found in cephalochordates (lancelet, *Branchiostoma floridae*; RefSeq:XP_002235318.1) and gastropods (California sea hare, *Aplysia californica*; Uniprot:Q8N0B2_APLCA). Thus it appears that this domain architecture predates the protostome / deuterostome split within the Coelomata.

This picture is muddled by a lack of obvious candidates for four-fold FAS1 domain proteins within deuterostome groups of varying proximity to the vertebrates: the comparatively well studied Tunicates and Echinoderms (represented by the genome assemblies for *Ciona intestinalis* and the sea urchin *Strongylocentrotus purpuratus*) do not appear to have them, even without EMI domain. We are assuming that this is due to a secondary loss in these groups, given that the four-fold FAS1 domain structure per se is phylogenetically very broadly encountered, although not ubiquitous. It was originally described in grasshopper (*Schistocerca americana*) and fruit fly (*Drosophila melanogaster*) [23] for the protein fasciclin (which provided the name for the FAS1 (fasciclin I) domain), and is generally common in arthropods. According to records in the SMART database, there are even a few known instances of four FAS1 domain proteins in bacteria (curiously, all in marine and/or psychrophilic species) and in fungi.

Given the framework outlined above, it was reasonable to hope that we might be able to phylogenetically pinpoint the periostin / TGFBI split and the appearance of the repeat structure for periostin, taking advantage of complete genome sequences from relevant species that have become available recently.

In this context, the elephant shark or ghost shark (*Callorhinchus milii*) was of interest as a cartilaginous fish and thus as representative of the Chondrichthyes, the non-euteleostome branch of the jawed vertebrates (Gnathostomata). We found evidence supporting the presence of both periostin and TGFBI in the elephant shark genome (see Figure S4).

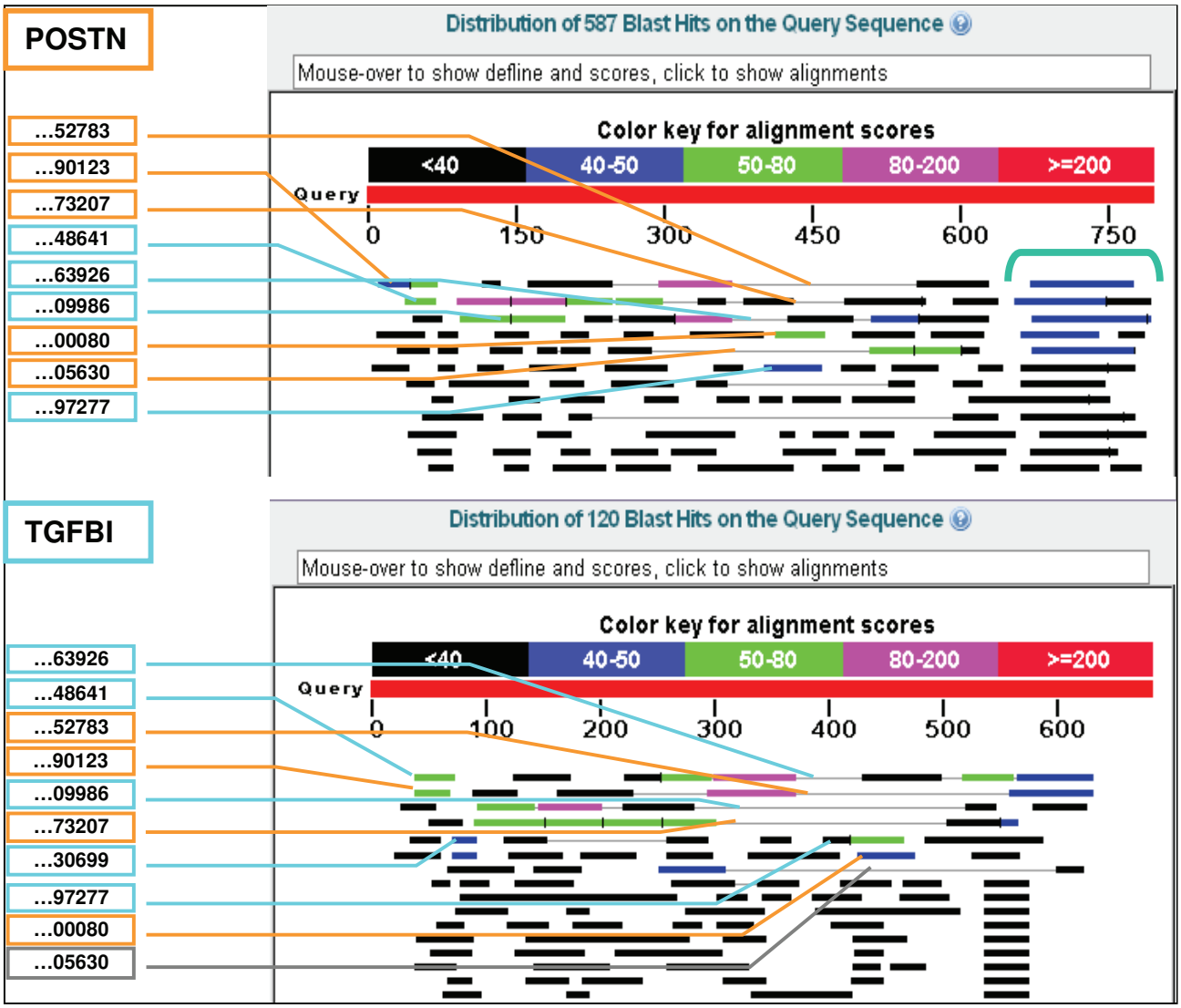


Figure S4: Graphical representation of the disambiguation process for separating POSTN and TGFBI in the elephant shark (*Callorhinchus milii*) genome (unfinished assembly).
TBLASTN searches were performed against *C. milii* "whole genome shotgun" sequences using the "genomic BLAST" server at http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi?organism=euk with *Danio rerio* (locus 1) periostin protein sequence (top panel) and TGFBI protein sequence (bottom panel) as queries. *C. milii* subject sequences (identified by the last five digits of their sequence accession number) were annotated with regard to being primary or secondary hits for the respective search. It emerged that by this metric alone, subject sequences, while commonly appearing in both searches, could be near-unambiguously assigned to either POSTN or TGFBI (reflected in the color coding, orange for POSTN, cyan for TGFBI). This, in turn, enabled the partial reconstruction of *C. milii* POSTN and TGFBI sequences. The much higher number of hits for the POSTN compared to the TGFBI query (587 vs. 120) is mostly due to the POSTN C-terminal region (blue-green bracket), which results in a disproportional number of hits.

This is borne out in an alignment-based phylogenetic tree with the partially reconstructed sequences (Figure S5), where the candidate periostin and TGFBI sequences from shark co-cluster with their respective euteleostome orthologs.

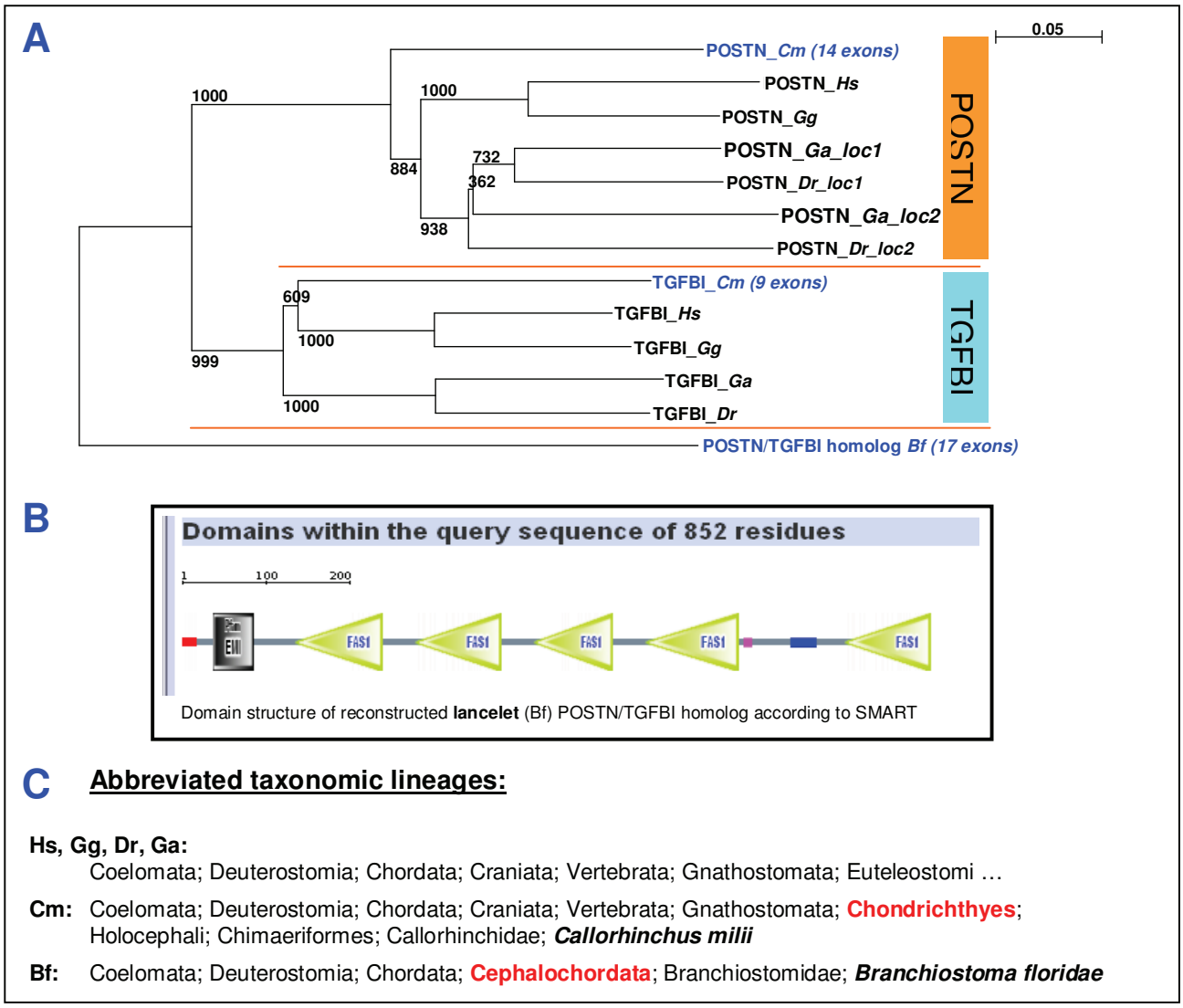


Figure S5: A POSTN/TGFBI homolog in lancelet (*Branchiostoma floridae*).
(A) Phylogenetic tree based on a ClustalW alignment of POSTN and TGFBI sequences from select euteleostome species, the reconstructed POSTN and TGFBI sequences from elephant shark (*C. milii*, see Fig.S4), and reconstructed POSTN/TGFBI homolog from the lancelet *B. floridae* (based on predicted lancelet protein Refseq:XP_002235318.1 and translated lancelet ESTs), which clearly falls outside the POSTN and TGFBI groups. Accession numbers for TGFBI sequences (except Cm): RefSeq:NP_000349.1 (Hs), RefSeq:NP_990367.1 (Gg), RefSeq:NP_878282.1 (Dr), Ensembl:ENSGACP00000021567 (Ga).
(B) Domain structure of the reconstructed lancelet POSTN/TGFBI homolog (as predicted by SMART, http://smart.embl-heidelberg.de/), which interestingly features a fifth C-terminal FAS1 domain. Note that the tree topology given in (A) is not attributable to this fifth domain, since the tree was calculated excluding gapped columns.
(C) Taxonomic lineages of the newly considered species *C. milii* and *B. floridae*.

However, the question if elephant shark periostin also has the repeat structure universally encountered in euteleostomes could not be resolved: While we identified multiple candidate repeat regions with a periodicity of 13 amino acids and similarity to a periostin repeat consensus based on teleost fish (Figure S6), the fragmentary nature of the elephant shark genome assembly made it impossible to establish or rule out an actual inter-exon connection between matches to repeats and those to other regions of the periostin gene.

A genome-based reconstruction of the sequence for the candidate periostin/TGFBI homolog from lancelet, a cephalochordate, and its alignment with periostin and TGFBI sequences showed it falling outside the periostin and TGFBI clusters in the resulting dendrogram (Figure S5).

A **Best repeat match in elephant shark:**

```
> AAVX01225691.1 <http://blast.fugu-sg.org/cgi-bin/scaff_shark.pl?scaffold=AAVX01225691.1>
          Length = 1839

 Score =  109 bits (240), Expect(3) = 3e-41
 Identities = 65/145 (44%), Positives = 78/145 (53%)
 Frame = -2

Query: 1    PSITKVTRVIEGEPSITKVTRVIEGEPSITKVTRVIEGEPSITKVTRVIEGEPSITKVTR 60
             PS    T++ E EPS    TR+ E EPS    TR+ E EPS     R+ E EPS    TR
Sbjct: 1742 PSRVAETKIREMEPSRVAETRIREMEPSRVAETRIREMEPSRVAEIRIREMEPSRVAETR 1563

Query: 61   VIEGEPSITKVTRVIEGEPSITKVTRVIEGEPSITKVTRVIEGEPSITKVTRVIEGEPSI 120
             + E EPS    T++ E EPS    TR+ E EPS    TR+ E EPS    TR+ E EPS
Sbjct: 1562 IWEMEPSRVAETKIREMEPSRVAETRIREMEPSRVAETRIREMEPSRVAETRIREMEPSG 1383

Query: 121  TKVTRVIEGEPSITKVTRVIEGEPS 145
                TR+ E EPS    TR+ E EPS
Sbjct: 1382 VAETRIREMEPSRVAETRIREMEPS 1308
```

B

Shark repeat consensus (11 sequences)

Teleost fish repeat exon 18=19 consensus (45 sequences)

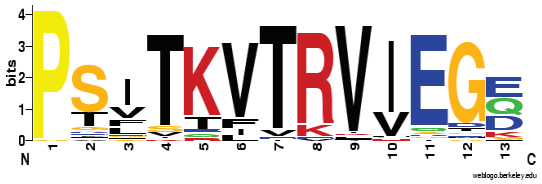
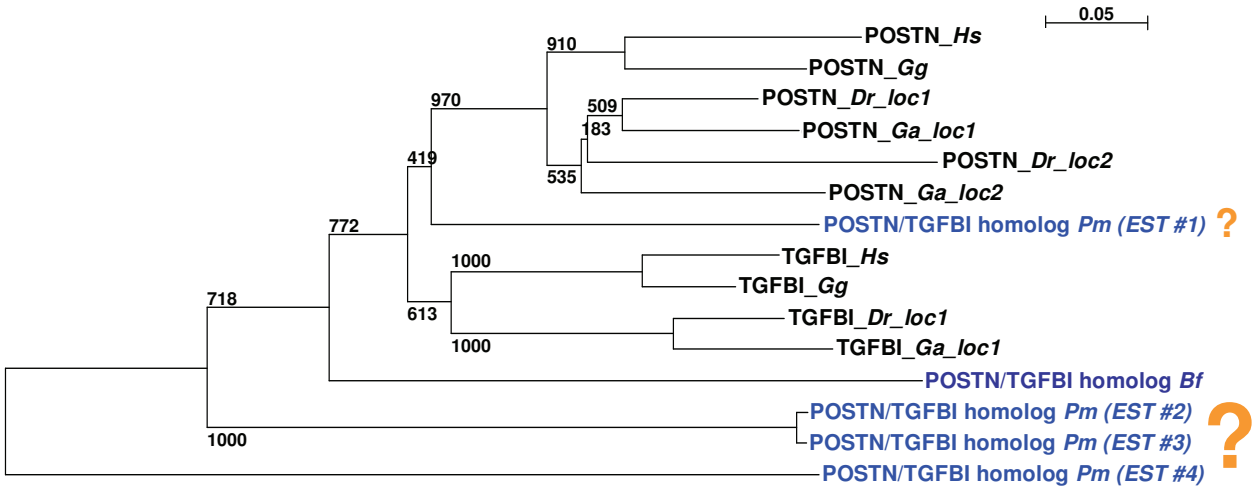


Figure S6: Best TBLASTN hit in the *C. milii* genome for a query of concatenated teleost fish repeat consensus sequences.
(A) A query of concatenated teleost fish repeat consensus sequences (based on 45 repeat units from different teleost fish species, see main text for details) was used as a query in a TBLASTN search against the elephant shark genome sequence (via http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi?organism=euk); the BLAST alignment for the best hit is shown .
(B) A comparison of the repeat consensus from this hit (top) to the teleost repeat consensus used for the TBLASTN search show (bottom) via sequence logo representation.

So far, these findings are consistent with the prevalent concept of whole genome duplication events within the chordates that provide the most straightforward explanation for the emergence of the periostin / TGFBI paralog pair: Apart from the most recent whole genome duplication (WGD) event at the base of teleost fish, two other WGD events, one at the base of the jawed vertebrates and another at the base of all vertebrates, are thought to have occurred [70].



Abbreviated taxonomic lineages:

Hs, Gg, Dr, Ga:
Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Euteleostomi …

Cm: Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; **Chondrichthyes**; Holocephali; Chimaeriformes; Callorhinchidae; ***Callorhinchus milii***

Pm: Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; **Hyperoartia**; Petromyzontiformes; Petromyzontidae; ***Petromyzon marinus***

Bf: Coelomata; Deuterostomia; Chordata; **Cephalochordata**; Branchiostomidae; ***Branchiostoma floridae***

Figure S7: Phylogenetic tree based on the alignment of select POSTN and TGFBI protein sequences and four translated EST sequences from putative lamprey (*Petromyzon marinus*) POSTN/TGFBI homologs.
EST #1 groups with POSTN, ESTs #2, #3, #4 form an outgroup relative to all other sequences, including the POSTN/TGFBI homolog from lancelet, which is contrary to the lamprey's phylogenetic position relative to euteleostomes and cephalochordates (see taxonomic lineages below). The interpretation of this result is difficult (see main text for details).
P. marinus EST accession numbers (and translated nucleotide ranges): EST #1: GenBank:DW022362 (2-886), EST #2: GenBank:DW021408 (2-841), EST #3: GenBank:FD721541 (1-855), EST #4: GenBank:FD706478 (3-830).

Crucially, searches in the genome sequence of the lamprey (*Petromyzon marinus*), a jawless vertebrate, and as such positioned after the first chordate WGD event affecting vertebrates, but not subject to the second one at the base of the jawed vertebrates, produced complex and ambiguous results (see Figure S7). Their interpretation was substantially hampered by the extremely preliminary status of the genome assembly.

Consequently, while successfully fitting periostin and TGFBI into the larger framework of chordate evolution, we are presently unable to precisely identify the WGD event responsible for the split between these two genes and to clarify the origin of periostin's C-terminal repeat region beyond the euteleostome group.

RNA interference and retinoblastoma-related genes are required for repression of endogenous siRNA targets in *Caenorhabditis elegans*

Alla Grishok[a,1], Sebastian Hoersch[a,b], and Phillip A. Sharp[a,2]

[a]Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02139; and [b]Bioinformatics Group, Max Delbrück Center for Molecular Medicine, 13125 Berlin, Germany

Contributed by Phillip A. Sharp, October 23, 2008 (sent for review August 15, 2008)

In *Caenorhabditis elegans*, a vast number of endogenous short RNAs corresponding to thousands of genes have been discovered recently. This finding suggests that these short interfering RNAs (siRNAs) may contribute to regulation of many developmental and other signaling pathways in addition to silencing viruses and transposons. Here, we present a microarray analysis of gene expression in RNA interference (RNAi)-related mutants *rde-4*, *zfp-1*, and *alg-1* and the retinoblastoma (Rb) mutant *lin-35*. We found that a component of Dicer complex RDE-4 and a chromatin-related zinc finger protein ZFP-1, not implicated in endogenous RNAi, regulate overlapping sets of genes. Notably, genes a) up-regulated in the *rde-4* and *zfp-1* mutants and b) up-regulated in the *lin-35*(Rb) mutant, but not the down-regulated genes are highly represented in the set of genes with corresponding endogenous siRNAs (endo-siRNAs). Our study suggests that endogenous siRNAs cooperate with chromatin factors, either *C. elegans* ortholog of acute lymphoblastic leukemia-1 (ALL-1)-fused gene from chromosome 10 (AF10), ZFP-1, or tumor suppressor Rb, to regulate overlapping sets of genes and predicts a large role for RNAi-based chromatin silencing in control of gene expression in *C. elegans*.

endo-siRNA | microarray | Rb | RNAi

Among the species with sequenced genomes the nematode *Caenorhabditis elegans* encodes the largest number of Argonaute proteins, which interact with short RNAs (1). Also, a large number of endogenous, short interfering RNAs (endo-siRNAs) have been cloned from *C. elegans* (2–5). They are distinct from microRNAs (miRNAs), are largely generated by RNA-dependent RNA polymerases (RdRP), and match thousands of genes. These observations suggest that multiple gene-regulatory networks involving Argonaute proteins and endo-siRNAs exist in the nematode.

We have characterized a system of RNAi-induced transcriptional gene silencing (RNAi-TGS) of a repetitive transgene expressed in the soma of *C. elegans* (6). Also, we found that RNAi pathway genes and *lin-35*(Rb) synergize in repressing the intestinal cell divisions and in repressing the cyclin E gene (*cye-1*) expression, likely through cooperative inhibition of *cye-1* transcription (7). Two chromatin-related genes, *zfp-1* and *gfl-1*, promote the RNAi process in *C. elegans*, either directly or indirectly, they also contribute to RNAi-TGS of a repetitive transgene (6, 8, 9). Interestingly, both genes were also found to antagonize the repressive function of LIN-35(Rb) (10, 11). Therefore, ZFP-1 and GFL-1 appear to regulate both RNAi and Rb target genes.

The *C. elegans* Rb protein LIN-35 represses inappropriate transcription of germline-specific genes (12) and growth factors (13) in differentiated somatic cells and functions redundantly with other transcriptional repressors (14). Also, *lin-35* mutants are more sensitive to exogenous RNAi than wild-type worms (11, 15). This might be partially because of the de-repression of germline-specific RNAi pathway genes in somatic cells.

Because RNAi genes were found to function in the same processes as *lin-35*, we conducted microarray experiments to find potential targets regulated by RNAi-TGS and *lin-35*. We used *rde-4* and *zfp-1* mutants affecting RNAi-TGS. RDE-4 is a dsRNA binding protein interacting with Dicer (16) whereas ZFP-1 is a nuclear protein that is likely to affect transcription directly. Our previous study indicated that miRNAs might have a role in promoting RNAi-TGS in *C. elegans* as well (6); therefore, we included miRNA pathway Argonaute mutant *alg-1* in our experiments.

Our analysis revealed *i*) that *zfp-1* and *rde-4* mutant animals have strikingly similar profiles of alterations in gene expression and *ii*) that there is an enrichment of genes with matching (antisense) endo-siRNAs (3–5) only among genes up-regulated, but not down-regulated, in *zfp-1* and *rde-4* mutants. These genes therefore might represent direct targets of chromatin-based silencing induced by endogenous RNAi pathways. Interestingly, endo-siRNAs matched not only genes negatively regulated by *rde-4* and *zfp-1*, but also those primarily inhibited by LIN-35(Rb).

We also report that *zfp-1*, unlike *rde-4*, opposes the repressive function of LIN-35 in controlling intestinal nuclear divisions and *cye-1* expression. Our results suggest that ZFP-1 may play both a positive and a negative role in regulating gene expression.

Results

Microarray Data Analysis. To find target genes regulated by RNAi and Rb, we performed a series of microarray experiments using RNA from L1-L2 larvae of the wild type and loss-of-function mutants *rde-4* (17), *zfp-1* (10), *alg-1* (7), and *lin-35* (18). We conducted pairwise comparisons of the levels of gene expression in each mutant compared with the wild type and selected statistically significant changes in gene expression by two-sample *t* test (*P* value <0.01), requiring in addition an expression difference of at least 1.5-fold between two group averages. Our microarray data are summarized in Dataset S1 and Dataset S2.

A majority of the genes changing expression in the *lin-35* mutant compared with the wild type (535 of 710) were up-regulated consistent with the repressive role of the LIN-35 protein (Table 1). Similar numbers of genes were either up-regulated or down-regulated in each of the RNAi-related mutants: 420 were "up" in *zfp-1* and 434 were "down" whereas 285 were "up" in *rde-4* and 219 were "down", and 170 were "up" in *alg-1* and 213 were "down." The numbers of genes similarly regulated in different mutants are listed in Table 1. Ten genes commonly up-regulated in all four mutants are described in Table S1.

Author contributions: A.G., S.H., and P.A.S. designed research; A.G. performed research; A.G. and S.H. analyzed data; and A.G., S.H., and P.A.S. wrote the paper.

The authors declare no conflict of interest.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE13258).

[1]Present address: Department of Biochemistry and Molecular Biophysics, College of Physicians and Surgeons, Columbia University, New York, NY 10032.

[2]To whom correspondence should be addressed at: Koch Institute for Integrative Cancer Research, MIT, 40 Ames Street, E17–529, Cambridge, MA 02139. E-mail: sharppa@mit.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0810589105/DCSupplemental.

Table 1. Numbers of genes changing expression compared with the wild type in indicated mutant backgrounds (top) and numbers of overlapping genes between indicated mutants (bottom)

Mutant backgrounds	# genes UP	Enrichment	*P* value	# genes DOWN	Enrichment	*P* value
lin-35	535			175		
zfp-1	420			434		
rde-4	285			219		
alg-1	170			213		
lin-35 and *zfp-1*	56	4.54	1.39×10^{-21}	65	15.59	2.15×10^{-60}
lin-35 and *rde-4*	40	4.78	2.07×10^{-16}	32	15.21	9.1×10^{-29}
lin-35 and *alg-1*	39	7.81	5.45×10^{-24}	40	19.55	1.58×10^{-40}
zfp-1 and *rde-4*	110	16.74	1.44×10^{-107}	138	26.45	2.78×10^{-174}
zfp-1 and *alg-1*	68	17.35	7.26×10^{-67}	131	25.81	7.40×10^{-163}
rde-4 and *alg-1*	41	15.41	3.77×10^{-37}	77	30.07	2.65×10^{-96}

Genes up-regulated in the mutants are listed in the "UP" column and down-regulated gene are listed in the "DOWN" column. Enrichment factors and *P* values for significance of overlaps are indicated. Descriptions of genes changing expression in indicated mutants are listed in Dataset S2. Overlaps between groups of genes misregulated in two or more mutants can be found by using the "data filter" option in the Excel file and selecting "1" in the columns representing conditions of interest.

***zfp-1* and *rde-4* Mutants Have Similar Gene Expression Profiles.** A comparison of gene sets misregulated in the studied mutants revealed a very significant overlap between genes regulated by *rde-4* and genes regulated by *zfp-1*, including genes both up-regulated and down-regulated in the mutants compared with wild type (Table 1). Fifty percent of genes regulated by *rde-4* (close to 250) are included in a group affected by *zfp-1*. This degree of overlap in transcriptome regulation has not been reported before for any pair of RNAi-related genes.

Next, we used the gene expression terrain map ("topomap") (19) as a platform for functional annotation of misregulated gene sets. In this work, based on the analysis of extensive microarray expression datasets, 17,658 C. *elegans* genes were divided into forty-five expression clusters ("mounts") of coregulated genes. Kim and colleagues also redundantly assigned membership in 56 functional categories to 5,615 functionally characterized *C. elegans* genes, resulting in 8,212 category assignments (19). We mapped our datasets of misregulated genes in various mutants to mounts and categories (Fig. 1). A heatmap representation with clustering dendrograms summarizing significant enrichment of genes from ours and other relevant studies in functional groups of genes (mounts and categories) defined by Kim and colleagues (19) is shown in Fig. 1 and, more completely, in Fig. S1. In this representation, related functional groups are clustered on the *y* axis and related datasets are clustered on the *x* axis. This allows functional annotation and comparison of multiple datasets. P-values for statistical significance and representation factors for gene enrichment in specific groups are listed in Dataset S3.

We chose topomap as a vehicle for functional analysis over possible alternatives, especially gene ontology (GO) annotation, for a number of reasons. Chief among them is the considerably greater coverage of *C. elegans* genes (77% for topomap compared with 46% for GO) that is—by nature of the "annotation process"—not restricted to known and characterized genes. Therefore, topomap-based functional assignment described in our study is not limited to well studied genes. Functional annotation of our expression data using GO platform (data not shown) was similar to that obtained with topomap, but we arrived at a more complete picture of gene expression by using topomap.

A comparison of the functional categories of genes changing expression in different mutants revealed a striking similarity between transcriptome profiles in *rde-4* and *zfp-1* mutants (Fig. 1 and Fig. S1). This similarity suggests that common biological processes are affected by both mutations. For example, certain germline-enriched and oocyte genes (mount #02) are overrepresented in groups of genes with higher expression levels in *zfp-1* and *rde-4* mutants and close to 20% of genes commonly up-regulated in both mutants belong to this category (Dataset S2). Indeed, functional annotation of the groups of genes commonly affected by each combination of two mutants (presented in Table 1) revealed the same categories of enrichment as those that were common between the two single mutant profiles (Fig. S2).

Therefore, two independent types of analyses: 1) a direct comparison of genes changing expression in two mutants (Table 1) and 2) functional annotation of misregulated genes (Fig. 1 and Figs. S1 and S2) strongly suggest that *zfp-1* and *rde-4* work in the same pathway (RNAi-TGS) and point to a very significant role of this pathway in biology of *C. elegans*.

The *rde-4* mRNA level was not changed in the *zfp-1* mutant and vice versa, indicating that a simple model of regulation of one gene by the product of another does not account for the correlation. We cannot exclude the possibility that protein levels of RDE-4 or ZFP-1 might change. However, these types of changes are not likely to be due to the direct regulation by RDE-4 or ZFP-1 because RDE-4 is known to interact with RNA and ZFP-1 is a chromatin factor.

Genes with higher expression in *zfp-1* and *rde-4* mutants were overrepresented among the functional groups 'protein expression,' 'germline-enriched,' 'biosynthesis,' 'mitochondrial,' and 'cell cycle,' whereas those genes that were down-regulated in the mutants frequently represented intestine-specific genes involved in metabolic processes (Fig. 1). Histone genes were also significantly enriched in the *rde-4* down-regulated gene set (Fig. 1). Importantly, ZFP-1 appears to have a larger role in gene expression regulation than RDE-4 (Fig. 1, Table 1, and Dataset S1). Consistent with these results, *zfp-1* mutants have some developmental phenotypes, such as slow growth and protruded vulva (10), whereas *rde-4* mutant worms are superficially normal.

A recent microarray study reported gene expression changes in the RNAi pathway mutants *rde-1*, *rde-4*, and *dcr-1* (20). We mapped the misregulated gene sets from this study to the functional groups of coregulated genes (Fig. 1) and found that genes down-regulated in the *rde-4* mutant were enriched in intestine-specific group contained significant number of histone genes and proteases. This signature corresponds to that of genes down-regulated in the *rde-4* mutant from our study (Fig. 1). However, genes found up-regulated in the *rde-4* mutant do not have a signature consistent with our findings (Fig. 1). One difference between the studies is that we used L1-L2 larva and the published report used adult worms (20). Because adult worms contain both differentiated somatic tissues and actively proliferating and specialized germline cells, whereas the L1-L2 larvae contain primarily somatic cells, the resulting "average" gene expression profile is likely to be different in adults and larvae. In addition, mutant backgrounds may have different effects on gene expression in somatic and germline tissues.

GENETICS

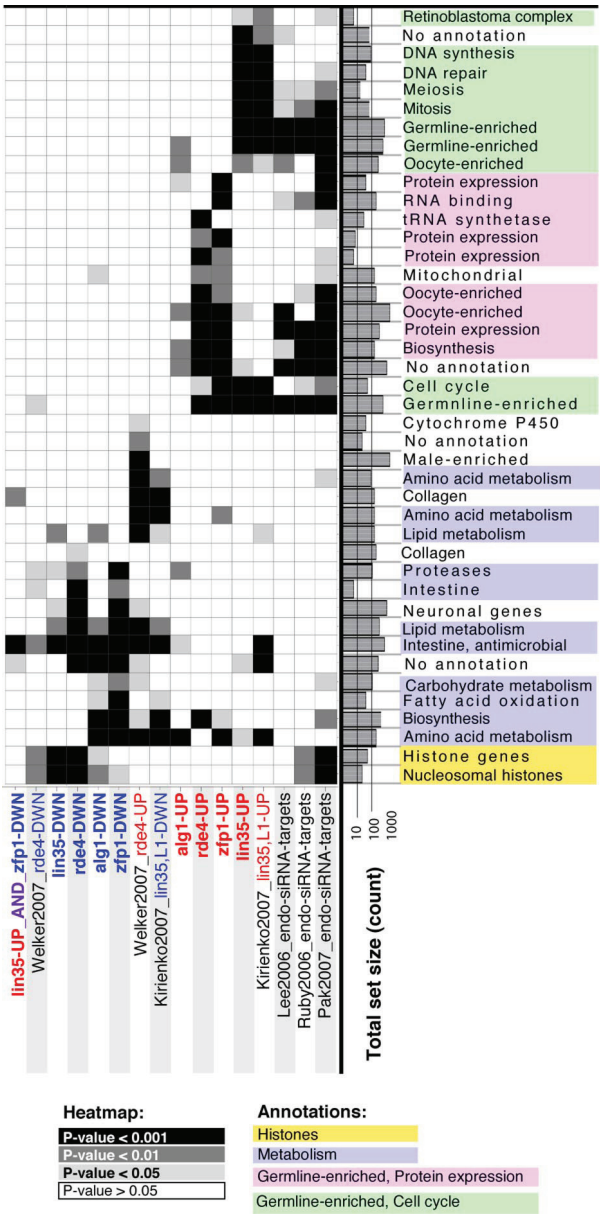


Fig. 1. Heatmap showing the enrichment of selected functional groups as defined by ref. 19 (*y* axis) in gene sets from various datasets (*x* axis). The gray shades in the heatmap indicate significance levels as indicated in the legend. The bars to the right of the heatmap indicate the total size of each functional group. For a comprehensive version of the figure including all functional groups with any significant enrichment and clustering dendrograms informing the order of groups and datasets, see Fig. S1. Enrichment factors and their *P* values are presented in Dataset S3.

Endogenous siRNA Preferentially Map to Genes Up-Regulated in *zfp-1*, *rde-4*, and *lin-35* Mutants. As mentioned, we found almost equal numbers of genes both positively and negatively regulated by RDE-4 and ZFP-1 (Table 1). Although it is conceivable that ZFP-1 might act directly as an activator or as a repressor on both types of target genes, RDE-4, required for the production of siRNAs (21), is much more likely to contribute directly only to gene silencing.

To gain insight about possible direct targets of RNAi-TGS, we mapped the endo-siRNA target genes according to three independent studies (3–5) to the coregulated groups defined by Kim and colleagues (Fig. 1). Interestingly, distribution of siRNA target genes was nonrandom and mostly consistent among the three studies (3–5) (Fig. 1). We found that siRNA-matching genes were very significantly enriched in the same functional groups as the genes up-regulated in *zfp-1* and *rde-4* mutants or *lin-35* mutants (Fig. 1), but not in the groups overrepresented in genes down-regulated in the same mutants, with the exception of histone genes (Fig. 1). This suggested that specific genes negatively regulated by *zfp-1* and *rde-4* or by *lin-35* might be more likely to have a matching endo-siRNA.

Indeed, direct comparison of endo-siRNA target gene sets and *zfp-1*, *rde-4*, *alg-1*, and *lin-35*-regulated genes revealed a statistically significant enrichment of genes with siRNAs only in the sets of genes up-regulated in the mutants but not in the down-regulated sets of genes (Table 2). Close to 50% of genes with increased expression in each of the mutant strains were reported to have a matching siRNA. These data strongly suggest that genes up-regulated in the mutants represent the direct targets repressed by RNAi and that the down-regulated genes might be affected by the mutations indirectly. A very large overlap between *rde-4* and *zfp-1*-regulated genes (Table 1) together with high significance of enrichment in siRNA targets of gene sets up-regulated in both mutants (Table 2) further strengthens the prediction of numerous target genes regulated by RNAi-TGS in *C. elegans*.

When functional annotation was done on the groups of genes representing overlaps between genes up-regulated in the mutants and siRNA target genes, the signatures of "UP in *zfp-1* and siRNA target" and "UP in *rde-4* and siRNA target" groups were found to be very similar and very close to "UP in *zfp-1* and *rde-4*" signature. On the contrary, the signature of the "UP in *lin-35* and siRNA target" group was distinct from "UP in *zfp-1* and *rde-4*" and very similar to that of "UP in *lin-35*", whereas "UP in *alg-1* and siRNA target" group had similarity to groups of genes regulated by *zfp-1* and *rde-4* and also regulated by *lin-35*. We conclude that siRNA targets overlapping with *lin-35*-regulated genes are distinct from the groups of siRNA targets regulated by *zfp-1* and *rde-4*. Although chromatin factor ZFP-1 may be directly involved in the endo-siRNA pathway as this gene was implicated in supporting RNAi, the overlap between *lin-35* and endo-siRNA target genes likely represents synergy between the two repressive pathways.

The main signature of the *lin-35* mutant is de-repression of germline-specific genes in somatic tissues of larvae (Fig. 1), which is consistent with previous findings (11, 22). There are three main groups of coregulated genes that represent germ line (mounts #02, #07, and #11, Fig. S1). Endogenous siRNAs are enriched in those same groups: of the total of 4,372 siRNA targets represented in the topomap dataset, 1,448 were found to belong to these germline mounts. Direct comparison of germline-specific siRNA target genes with the sets of genes changing expression in the mutants (Table S2) revealed 9× overrepresentation of genes up-regulated in *lin-35* larvae (154 compared with 17 expected by chance). This correlation may indicate that endo-siRNAs synergize with LIN-35 in repressing germline-specific fate in somatic tissues. Alternatively, although both LIN-35 targets and endo-siRNAs preferentially correspond to germline-enriched genes, LIN-35 and RNAi may regulate those genes independently in distinct tissues: soma and germ line, respectively.

When we compared nongermline siRNA target genes with gene sets changing expression in the mutants, genes up-regulated in *lin-35* were enriched modestly (2.5×) and enriched less than genes down-regulated in *lin-35* (3.3×) (Table S3). In contrast, genes up-regulated in *zfp-1* and *rde-4* were overrepresented among siRNA targets (4×), independently of their germline or nongermline classification (Tables S2 and S3). The corresponding sets of down-regulated genes were not overrepresented. This analysis further supports synergy between endo-siRNAs, *rde-4* and *zfp-1*, in gene expression regulation. Although *lin-35* and endo-siRNA targets do not appear to correlate outside of germline-enriched group of genes, the possibility of synergy between LIN-35 and endo-siRNAs in repressing germline fate in the soma still remains and needs to be studied further.

Table 2. endo-siRNA targets are overrepresented among genes up-regulated in RNAi and Rb mutants

	UP in *lin-35*	UP in *zfp-1*	UP in *rde-4*	UP in *alg-1*	DOWN in *lin-35*	DOWN in *zfp-1*	DOWN in *rde-4*	DOWN in *alg-1*
Totals in 18,459 ↓ →	428	333	232	142	143	371	191	185
Endo-siRNA targets 3,892	Overlap with endo-siRNA targets							
	223	149	107	61	37	74	41	48
Representation f-r	2.5	2.1	2.2	2	1.2	0.9	1	1.2
P value	1.8×10^{-46}	1.4×10^{-22}	9.6×10^{-18}	3.2×10^{-9}	0.09	0.73	0.48	0.064

Endo-siRNA target gene set, according to Pak and Fire (2) and gene sets determined to be ″UP″- or ″DOWN″-regulated in our expression data were mapped to 18,459 genes with TOPOMAP representation (with recalls ranging from 74% to 84%). For each group, total and overlap counts are listed as well as representation factors and *p* values for overlaps.

Notably, the cyclin E gene targeted by endo-siRNAs is expressed very highly during oogenesis and is categorized as 'germline-enriched.' Therefore, its repression by Rb and RNAi pathways in somatic tissues that we discovered genetically (7) may serve as an example of possible large-scale cooperation between endo-siRNAs and LIN-35 in repressing common targets.

Mutation in *zfp-1* Suppresses Extra Nuclei Division Phenotype and Enhanced Cyclin E Expression in *lin-35* Mutant Worms. We have found that the combination of the RNAi pathway mutants *rde-1*, *rde-4*, or the miRNAi pathway mutants *dcr-1* and *alg-1* with the *lin-35* mutation leads to a significant increase in postembryonic nuclear divisions in the intestine of the double mutant worms (Fig. 2 *A* and *B*) (7). Increases in cyclin E (*cye-1*) transcription under these conditions are at least partially responsible for this phenotype (7).

Because *rde-4* and *zfp-1* regulate many common genes, we tested whether ZFP-1 also cooperates with LIN-35 in repressing *cye-1*. Surprisingly, we found that combining the *zfp-1* mutation with *lin-35(lf)* did not lead to an increase in nuclear divisions. Instead, the *zfp-1* mutation suppressed extra nuclear divisions associated with the *lin-35; dcr-1*, *lin-35; alg-1* (Fig. 2*A*) and *lin-35; rde-1* double mutant combinations (Fig. 2*B*). This suppression by *zfp-1(lf)* of a phenotype associated with the lack of transcriptional repressors is comparable with its suppression of a multivulva phenotype (10, 11). In both cases it is likely that *zfp-1* function is required for an enhanced expression of the de-repressed target genes.

Because cyclin E is one of the target genes repressed by LIN-35, we tested whether enhanced expression of *cye-1* mRNA in *lin-35(lf)* worms requires ZFP-1. Indeed, we found that in the *lin-35; zfp-1* double mutant strain, the *cye-1* mRNA level was reduced as compared with that in *lin-35(lf)* (Fig. 2*C*). We did not observe a reduction in *cye-1* mRNA levels in the *zfp-1* mutant alone, indicating that its activity is not required for normal levels of expression of this gene.

Genes Repressed by *lin-35* and Activated by *zfp-1*. Our genetic studies of *cye-1* regulation and published reports (10, 11) indicate that *zfp-1* may act as an activator of LIN-35(Rb)-repressed genes (Fig. 2). However, the microarray results strongly suggest that ZFP-1 and RDE-4 have a direct repressive effect on a number of other targets, which are not regulated by LIN-35. We were interested in identifying an additional group of genes, those oppositely regulated by *lin-35* and *zfp-1*, and further selected for up-regulated expression in *lin-35(lf)* background and down-regulated expression in the *zfp-1* mutant with a change in expression intermediate between *zfp-1* and *lin-35* in *rde-4(lf)* and *alg-1(lf)* (see *SI Text*). Fifty-seven genes with expression profiles showing high similarity to this "custom expression profile" were identified (Dataset S1 and Fig. 1). Notably, three Argonaute genes were found in this group. This representation is statistically significant (enrichment factor 38×, *P* value 6.36×10^{-5}).

We used quantitative real-time PCR to analyze the expression levels of the candidate genes with the largest differences in expression between *zfp-1* and *lin-35* mutants (down-regulated in *zfp-1* and up-regulated in *lin-35*) or genes with smaller expression changes that we find interesting, such as Argonaute gene *csr-1* (1). The expression of these chosen genes was tested in mutants used for the array analysis and in *lin-35*; *rde-4* and *lin-35*; *zfp-1* double mutants that have limited viability (Fig. 3 *A–F*). A few genes showed suppression of their enhanced expression in *lin-35(lf)* background

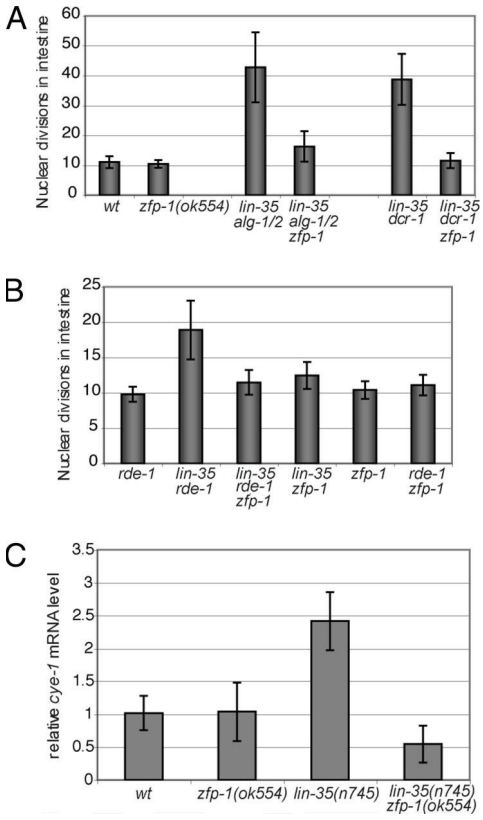


Fig. 2. Suppression of the increased nuclear division phenotype and increased *cye-1* mRNA expression in *lin-35* mutants by *zfp-1(ok554)*. (*A* and *B*) Quantification of postembryonic nuclear divisions in the intestine (number of nuclei in adult worms after subtraction of 20 nuclei present in L1) in different genetic backgrounds. Intestinal nuclei were counted in 15–30 worms and data for each genotype are presented as a mean number ± SD. The following mutants were used *lin-35(n745)unc-13(e1091)*, *zfp-1(ok554)*, *rde-1(ne300)*, *alg-1/2(RNAi)*, and *dcr-1(RNAi)*. Similar results to those shown in *A* were obtained with *lin-35(RNAi)* and *zfp-1(RNAi)*. (*C*) Real-time RT-PCR analysis of the expression levels of *cye-1* mRNA in different mutant backgrounds. Levels of *cye-1* mRNA were normalized to *ama-1* mRNA levels. Results of 2 independent experiments are shown as means and ranges of relative expression compared with wild type.

GENETICS

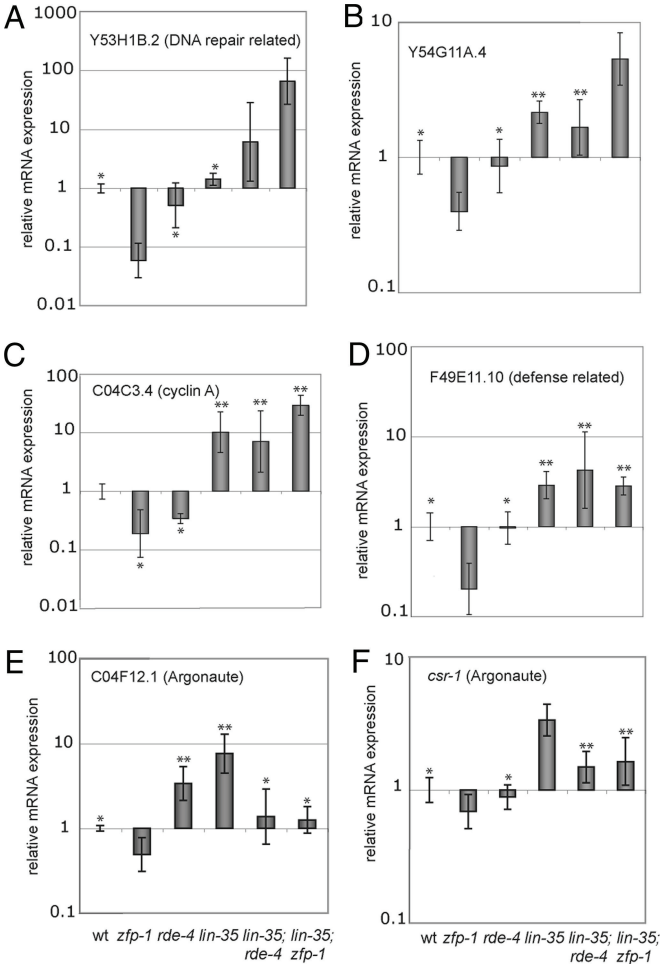


Fig. 3. ZFP-1 acts as a positive and negative regulator of genes repressed by LIN-35. Real-time RT-PCR analysis of the expression levels of indicated LIN-35 and ZFP-1 target mRNAs in different mutant backgrounds. (*A–F*) Examples of genes up-regulated in *lin-35* (−/−) and down-regulated in *zfp-1* (−/−). The order of mutants tested is presented the same in all images and is indicated in the bottom. The levels of tested mRNAs were normalized to *arx-2* mRNA levels. Results of 3 RT-PCR experiments are shown as means and ranges of relative expression compared with wild type. Groups of bars labeled * or ** in each image are not statistically different from each other, considering $P < 0.05$.

when the *zfp-1* mutation was added to the *lin-35* mutant (Fig. 3 *E* and *F*), whereas enhanced expression of other genes was not suppressed by *zfp-1(lf)* (Fig. 3 *A–D*). These results reveal a complex regulation of tested genes by LIN-35 and ZFP-1 and suggest that ZFP-1 may have a dual role (of an activator and repressor) in regulating expression of specific targets.

We also performed real time RT-PCR analysis of the expression of several germline-enriched genes with matching siRNAs and repressed by *lin-35*, but not affected by *zfp-1(ok554)*, similarly to *cye-1* (Fig. S3). The *lin-35*; *rde-4* and *lin-35*; *zfp-1* double mutants were also included in this analysis. Loss of *rde-4* or *zfp-1* does not appear to contribute very significantly to the dramatic de-repression of these targets in the *lin-35* mutant background.

Discussion

Our microarray study was motivated by the finding of cooperation between RNAi-TGS and Rb in cyclin E regulation (7). We aimed at identifying more targets of these repressive pathways.

The profile of genes up-regulated in the *lin-35* mutant larvae confirms its role in the repression of germline-specific fates in somatic cells (11). More than half of genes up-regulated in the *lin-35* mutant have matching endogenous siRNAs (311/535, enrichment factor 2.3×, *P* value 2×10^{-60}). Although we cannot exclude a possibility that these endo-siRNAs are produced in the germ line and also function in this tissue, it is equally possible that siRNAs generated in the germ line by RdRP are inherited and function along with LIN-35 to repress germline genes in the soma. Cyclin E is an example of a 'germline-enriched' gene repressed by LIN-35 and RNAi in the somatic tissues (7). This pattern of expression of Cyclin E is not unique to nematodes. Cyclin E expression has been shown to be continuous throughout the cell cycle in germline stem cells of *Drosophila* (23) and embryonic stem (ES) cells from mouse (24) and primates (25). High level of cyclin E was proposed to indicate "stemness" of the cells (23). In somatic cells in these organisms, constitutive Cyclin E expression is repressed with the onset of cell-cycle dependent regulation. Our results demonstrating repression of cyclin E in the soma along with other germline genes are consistent with this idea.

Interestingly, we found that many RNAi-related Argonaute genes (*ppw-1*, *sago-2*, C16C10.3, C04F12.1, and *csr-1*) are repressed by LIN-35. Argonaute proteins interact with siRNAs and are essential for the silencing process. *C. elegans* Argonaute genes *ppw-1*, *sago-2*, and C04F12.1 function redundantly in the RNAi process (1). The level of expression of these genes is elevated eight to ten fold in *lin-35(lf)* larvae. This finding may explain why the *lin-35* mutant is more susceptible to exogenous RNAi (11, 15).

We identified very significant enrichment of endo-siRNA target genes among genes up-regulated in *rde-4* (*P* value 9.6×10^{-18}) and *zfp-1* (*P* value 1.4×10^{-22}) mutants. Also, these mutants affected a large number of common genes. Previous studies aimed at identifying common misregulated targets among various endo-RNAi pathway mutants (3) have not detected large overlaps in misregulated genes or common functional signatures predicting biological pathways where regulation by endogenous RNAi may take place. Therefore, this is the first study demonstrating a connection between *zfp-1* function and endogenous RNAi processes and identifying specific genes that are 1) endo-siRNA targets, 2) up-regulated in *rde-4*(−/−), and 3) up-regulated in *zfp-1*(−/−) and belong to very specific functional groups, such as regulation of protein translation and germline function (Dataset S2).

We infer that genes commonly up-regulated in the *rde-4* and *zfp-1* mutants and containing matching siRNAs are the direct targets of nuclear RNAi. This prediction is based on the role of *rde-4* and *zfp-1* genes in our characterized system of transcriptional silencing of a transgene (6), the demonstrated requirement of RDE-4 for production of at least some endo-siRNAs (3, 26) and on the predicted nuclear function of the ZFP-1 protein. ZFP-1 is a homolog of mammalian protein AF10, which causes myeloid leukemia when fused to MLL (27). Both ZFP-1 and AF10 contain two N-terminal PHD zinc fingers and a C-terminal leucine zipper domain. Some PHD zinc fingers were recently recognized as histone-binding modules interacting with either methylated (28, 29) or unmethylated (30) lysine 4 of histone H3. The protein sequences of most terminal PHD fingers of ZFP-1 and AF10 align very well with histone-binding PHD fingers of other proteins, strongly suggesting that these proteins interact with chromatin via PHD domains. AF10 was shown to recruit histone H3 lysine 79 Dot1 methyltransferase via its leucine zipper domain (31) and to play a role in transcriptional elongation (32). It is possible that ZFP-1 binds histones with its N-terminal PHD domain and recruits different protein factors with its C-terminal domain. It could serve as an adaptor for both activators (Dot1) and repressors (RNAi factors) and regulate gene expression at the transcription elongation step.

The majority of the endo-siRNAs in *C. elegans* is antisense to mature mRNA sequences and is likely produced by RdRPs by using those mature RNAs as templates (2–5). A very recent discovery of an Argonaute protein NRDE-3 that binds endo-siRNAs and shuttles between the cytoplasm and the nucleus (26) further supports a

possibility that endo-siRNAs and ZFP-1 may work together in the nuclear RNAi pathway in *C. elegans*.

Materials and Methods

***C. elegans* Strains.** Worms were maintained on nematode growth medium plates seeded with OP50 bacteria. The strains used are listed in the *SI Text*. Adult or L4 worms were used for counting intestinal nuclei in strains containing *elt-2*::*gfp*/*LacZ* reporter. RNAi by feeding was performed as described (7). We used *lin-35(n745)* mutant linked to the weak *unc-13(e1091)* allele in our experiments to facilitate gene expression comparison between a *lin-35* single mutant and *lin-35*; *rde-4* and *lin-35*; *zfp-1* double mutants constructed in *unc-13(e1091)* background. Only one of eighteen *lin-35unc-13*-dependent genes that we tested by real-time RT-PCR, *sod-3*, had an increased expression in *unc-13(e1091)* background compared with wild type (data not shown). However, its expression was even higher in the *lin-35unc-13* strain. Because the functional categories of genes up-regulated in *lin-35* mutant were almost identical between our study and that of Kirienko and Fay (22) (Fig. 1), which used an unmarked *lin-35* mutant, we believe that the number of false positives in our study, due to *unc-13*, is very low.

***C. elegans* Collection for Microarray Experiments.** Nematodes were synchronized at L1 stage by hypochlorite treatment of gravid hermaphrodites and hatching their eggs overnight in liquid culture without food. Resulting populations were cultured on OP50 bacteria for 6–7 h and collected for RNA preparations.

RNA Preparation and Microarray Hybridization. Tri Reagent (MRC) was used for total RNA preparation from frozen worms resulting in 5–30 μg RNA per sample. The quality of RNA samples was confirmed by BioRad Bioanalyzer. Affymetrix GeneChip *C. elegans* Genome Arrays with a total of 22,625 probesets were hybridized with cDNA and scanned according to manufacturer's standard protocol. All conditions (WT and 4 mutants) were profiled in triplicate. Replicates were biological replicates (separately grown worm populations), with two exceptions: because of shortage of biological material, there were only two biological replicates available for the *lin-35(n745) unc-13(e1091)* I and *zfp-1(ok554)* III strains; for both, one biological sample was hybridized twice to set up a consistent triplicate structure across the dataset. Subsequent analysis of the replicate structure by using unsupervised hierarchical clustering showed that the agreements between technical replicates is in the same range as those between biological replicates, validating the approach taken. Raw data processing and normalization was performed by using the Bioconductor (33); http://www.bioconductor.org/) packages 'affy' and 'gcrma' to generate the dataset of GC-RMA expression measures (34) used for further analysis.

Data Analysis: Sets of Differentially Expressed Genes. Sets of probesets with up- or down-regulated expression in the mutants relative to WT were determined via *t* test (two-tailed, homoscedastic) with a *P* value cutoff of 0.01, requiring in addition an average expression difference of 1.5 or greater on the natural scale.

Complete data analysis description, which includes generation of idealized expression profile, gene assignment and mapping, topomap assignments and graphic generation, is presented in *SI Text*.

RT and quantitative real time PCR was performed as described in refs. 6 and 7.

ACKNOWLEDGMENTS. We thank Manlin Luo for performing cDNA labeling and microarray hybridizations and Charlie Whittaker for help with microarray data processing. We also thank Iva Greenwald, Oliver Hobert, Joel Neilson, and Anthony Leung for comments on the manuscript. The *elt-2*::*gfp*/*lacZ* strain was generated by Anne Hart (Massachusetts General Hospital). The *zfp-1(ok554)* strain was provided by the *C. elegans* Gene Knockout Project at Oklahoma Medical Research Foundation, which is part of the International *C. elegans* Gene Knockout Consortium. Some strains used in this study were obtained from the *Caenorhabditis* Genetics Center, which is funded by National Institutes of Health National Center for Research Resources. This work was supported by a Leukemia and Lymphoma Foundation Fellowship #3260–07 (to A.G.), United States Public Health Service Grant PO1-CA42063 from the National Cancer Institute (to P.A.S.), and partially by Cancer Center Support Grant P30-CA14051 from the National Cancer Institute.

1. Yigit E, *et al.* (2006) Analysis of the *C. elegans* Argonaute family reveals that distinct Argonautes act sequentially during RNAi. *Cell* 127:747–757.
2. Ambros V, *et al.* (2003) MicroRNAs and other tiny endogenous RNAs *in C. elegans*. *Curr Biol* 13:807–818.
3. Lee RC, Hammell CM, Ambros V (2006) Interacting endogenous and exogenous RNAi pathways in *Caenorhabditis elegans*. *RNA* 12:589–597.
4. Pak J, Fire A (2007) Distinct populations of primary and secondary effectors during RNAi in *C. elegans*. *Science* 315:241–244.
5. Ruby JG, *et al.* (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* 127:1193–1207.
6. Grishok A, Sinskey JL, Sharp PA (2005) Transcriptional silencing of a transgene by RNAi in the soma of *C. elegans*. *Genes Dev* 19:683–696.
7. Grishok A, Sharp PA (2005) Negative regulation of nuclear divisions in *Caenorhabditis elegans* by retinoblastoma and RNA interference-related genes. *Proc Natl Acad Sci USA* 102:17360–17365.
8. Dudley NR, Labbe JC, Goldstein B (2002) Using RNA interference to identify genes required for RNA interference. *Proc Natl Acad Sci USA* 99:4191–4196.
9. Kim JK, *et al.* (2005) Functional genomic analysis of RNA interference in *C. elegans*. *Science* 308:1164–1167.
10. Cui M, Kim EB, Han M (2006) Diverse chromatin remodeling genes antagonize the Rb-involved SynMuv pathways in *C. elegans*. *PLoS Genet* 2:e74.
11. Wang D, *et al.* (2005) Somatic misexpression of germline P granules and enhanced RNA interference in retinoblastoma pathway mutants. *Nature* 436:593–597.
12. Wang D, Ruvkun G (2004) Regulation of *Caenorhabditis elegans* RNA interference by the daf-2 insulin stress and longevity signaling pathway. *Cold Spring Harb Symp Quant Biol* 69:429–431.
13. Cui M, *et al.* (2006) SynMuv genes redundantly inhibit lin-3/EGF expression to prevent inappropriate vulval induction in *C. elegans*. *Dev Cell* 10:667–672.
14. Fay DS, Yochem J (2007) The SynMuv genes of *Caenorhabditis elegans* in vulval development and beyond. *Dev Biol* 306:1–9.
15. Lehner B, *et al.* (2006) Loss of LIN-35, the *Caenorhabditis elegans* ortholog of the tumor suppressor p105Rb, results in enhanced RNA interference. *Genome Biol* 7:R4.
16. Tabara H, Yigit E, Siomi H, Mello CC (2002) The dsRNA binding protein RDE-4 interacts with RDE-1, DCR-1, and a DExH-box helicase to direct RNAi in *C. elegans*. *Cell* 109:861–871.
17. Tabara H, *et al.* (1999) The rde-1 gene, RNA interference, and transposon silencing in *C. elegans*. *Cell* 99:123–132.
18. Lu X, Horvitz HR (1998) lin-35 and lin-53, two genes that antagonize a *C. elegans* Ras pathway, encode proteins similar to Rb and its binding protein RbAp48. *Cell* 95:981–991.
19. Kim SK, *et al.* (2001) A gene expression map for *Caenorhabditis elegans*. *Science* 293:2087–2092.
20. Welker NC, Habig JW, Bass BL (2007) Genes misregulated in *C. elegans* deficient in Dicer, RDE-4, or RDE-1 are enriched for innate immunity genes. *RNA* 13:1090–1102.
21. Parrish S, Fire A (2001) Distinct roles for RDE-1 and RDE-4 during RNA interference in *Caenorhabditis elegans*. *RNA* 7:1397–1402.
22. Kirienko NV, Fay DS (2007) Transcriptome profiling of the *C. elegans* Rb ortholog reveals diverse developmental roles. *Dev Biol* 305:674–684.
23. Hsu HJ, LaFever L, Drummond-Barbosa D (2008) Diet controls normal and tumorous germline stem cells via insulin-dependent and -independent mechanisms in *Drosophila*. *Dev Biol* 313:700–712.
24. Stead E, *et al.* (2002) Pluripotent cell division cycles are driven by ectopic Cdk2, cyclin A/E and E2F activities. *Oncogene* 21:8320–8333.
25. Fluckiger AC, *et al.* (2006) Cell cycle features of primate embryonic stem cells. *Stem Cells* 24:547–556.
26. Guang S, *et al.* (2008) An Argonaute transports siRNAs from the cytoplasm to the nucleus. *Science* 321:537–541.
27. Daser A, Rabbitts TH (2005) The versatile mixed lineage leukaemia gene MLL and its many associations in leukaemogenesis. *Semin Cancer Biol* 15:175–188.
28. Li H, *et al.* (2006) Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature* 442:91–95.
29. Pena PV, *et al.* (2006) Molecular mechanism of histone H3K4me3 recognition by plant homeodomain of ING2. *Nature* 442:100–103.
30. Lan F, *et al.* (2007) Recognition of unmethylated histone H3 lysine 4 links BHC80 to LSD1-mediated gene repression. *Nature* 448:718–722.
31. Okada Y, *et al.* (2005) hDOT1L links histone methylation to leukemogenesis. *Cell* 121:167–178.
32. Bitoun E, Oliver PL, Davies KE (2007) The mixed-lineage leukemia fusion partner AF4 stimulates RNA polymerase II transcriptional elongation and mediates coordinated chromatin remodeling. *Hum Mol Genet* 16:92–106.
33. Gentleman RC, *et al.* (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol* 5:R80.
34. Wu Z, *et al.* (2004) A model-based background adjustment for oligonucleotide expression arrays. *J Am Stat Assoc* 99:909–917.

Correction

GENETICS

Correction for "RNA interference and retinoblastoma-related genes are required for repression of endogenous siRNA targets in *Caenorhabditis elegans*" by Alla Grishok, Sebastian Hoersch, and Phillip A. Sharp, which appeared in issue 51, December 23, 2008, of *Proc Natl Acad Sci USA* (105:20386–20391; first published December 10, 2008; 10.1073/pnas.0810589105).

The authors note that an additional affiliation for Sebastian Hoersch was omitted from the article. Sebastian Hoersch's second affiliation is Bioinformatics Group, Max Delbrück Center for Molecular Medicine, 13125 Berlin, Germany. The corrected author and affiliation lines appear below. The online version has been corrected.

Alla Grishok[a], Sebastian Hoersch[a,b], and Phillip A. Sharp[a]

[a]Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02139; and [b]Bioinformatics Group, Max Delbrück Center for Molecular Medicine, 13125 Berlin, Germany

www.pnas.org/cgi/doi/10.1073/pnas.1110993108

Supporting Information

Grishok *et al.* 10.1073/pnas.0810589105

SI Text

The following strains were used: wild type Bristol strain N2, *lin-35(n745) unc-13(e1091)* I, *unc-13(e1091)* I, *rde-4(ne299)* III, *zfp-1(ok554)* III, *alg-1(gk214)* X, *elt-2*::*gfp*/*lacZ* X, *lin-35(n745) unc-13(e1091)* I; *rde-4(ne299)* III, *lin-35(n745) unc-13(e1091)* I; *zfp-1(ok554)* III, *lin-35(n745) unc-13(e1091)* I; *rde-1(ne300)* V; *elt-2*::*gfp*/*lacZ* X, *lin-35(n745) unc-13(e1091)* I; *zfp-1(ok554)* III; *rde-1(ne300)* V; elt-2::*gfp*/*lacZ* X, *zfp-1(ok554)* III; *elt-2*::*gfp*/*lacZ* X, *rde-1(ne300)* V; *elt-2*::*gfp*/*lacZ* X, *zfp-1(ok554)* III; *rde-1(ne300)* V; and *elt-2*::*gfp*/*lacZ* X.

Generation of "Custom" Expression Profile. Using the "Profile Search" functionality in the Spotfire DecisionSite data analysis software, an idealized expression profile was defined across all samples showing high expression in *lin-35(lf)* background, low expression in the *zfp-1* mutant, and intermediate expression in the RNAi pathway mutants. This profile was encoded numerically by using the values 1, −1, and 0, respectively. Probesets with expression profiles similar to this idealized profile were selected by requiring a Pearson's correlation $r > 0.6$ and in addition, to ensure preservation of the idealized profile's key features, P values <0.05 for differential expression between WT and *zfp-1* mutant and between WT and *lin-35* mutant.

Gene Assignments and Mapping. An Affymetrix-generated annotation file (version 15-Nov-2006), specifically data in the column "Ensembl," was used to map probeset identifiers to *Caenorhabditis elegans* genes. For probesets with multiple gene annotations (3.1% of total), all gene annotations were retained, as identifying a "correct" one is often impossible because of cross-hybridization issues. No attempt was made to universally assign genes to probesets without gene annotation (5.7% of total). Gene collections from other *C. elegans* studies used in comparative analyses were downloaded from the authors' websites or the publications' supplemental data sections, maintaining the original author's gene assignments.

Collections of siRNA target genes were, where possible, used as provided by the authors (1). Alternatively (2, 3), collections of siRNA sequences provided by the authors were, if necessary, preprocessed to eliminate linker sequences, and then subjected to a BLASTN (4) analysis against a database of *C. elegans* full-length transcripts obtained by using the UCSC Table Browser (5) [Jan 2007 (ce4) assembly, track "WormBase genes," resulting in 30,497 sequences]. Only matches with 100% identity/100% query sequence coverage were considered.

Topomap Assignments. Topomap raw data (6) were downloaded from the supplementary data website (http://www.sciencemag.org/feature/data/kim1061603/gl/gene_list.html) and assembled into a gene-centric annotation table comprising 18,459 gene entries, 17,658 of which had mount assignments, and 5,615 had (frequently more than one) category assignment. For functional studies, *C. elegans* gene sets were mapped to the Topomap representation, resulting in recovery rates between 67% and 80% for mounts. For categories, the recovery rates were typically lower and more diverse, between 37% and 72%. Representation factors r (as provided in Dataset S3) to measure the enrichment or impoverishment of gene sets of interest ("set 1") in mounts and biogroups ("set 2") were calculated according to $r = (n_{1,2})/[(n_1 \times n_2)/N]$, with $n_{1,2}$ = number of genes common to set 1 and 2, n_1 = number genes in set 1, n_2 = number of genes in set 2, n = total number of genes considered. Significance of enrichment and impoverishment was determined separately by using Fisher Exact test as implemented in the statistical computing package R (http://www.r-project.org/).

Graphics Generation. The heatmap (Fig. 1 and Figs. S1 and S2) were generated with Spotfire DecisionSite.

1. Lee RC, Hammell CM, Ambros V (2006) Interacting endogenous and exogenous RNAi pathways in *Caenorhabditis elegans*. *RNA* 12:589–597.
2. Pak J, Fire A (2007) Distinct populations of primary and secondary effectors during RNAi in *C. elegans*. *Science* 315:241–244.
3. Ruby JG, *et al.* (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* 127:1193–1207.
4. Altschul SF, *et al.* (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.
5. Karolchik D, *et al.* (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32:D493–496.
6. Kim SK, *et al.* (2001) A gene expression map for *Caenorhabditis elegans*. *Science* 293:2087–2092.

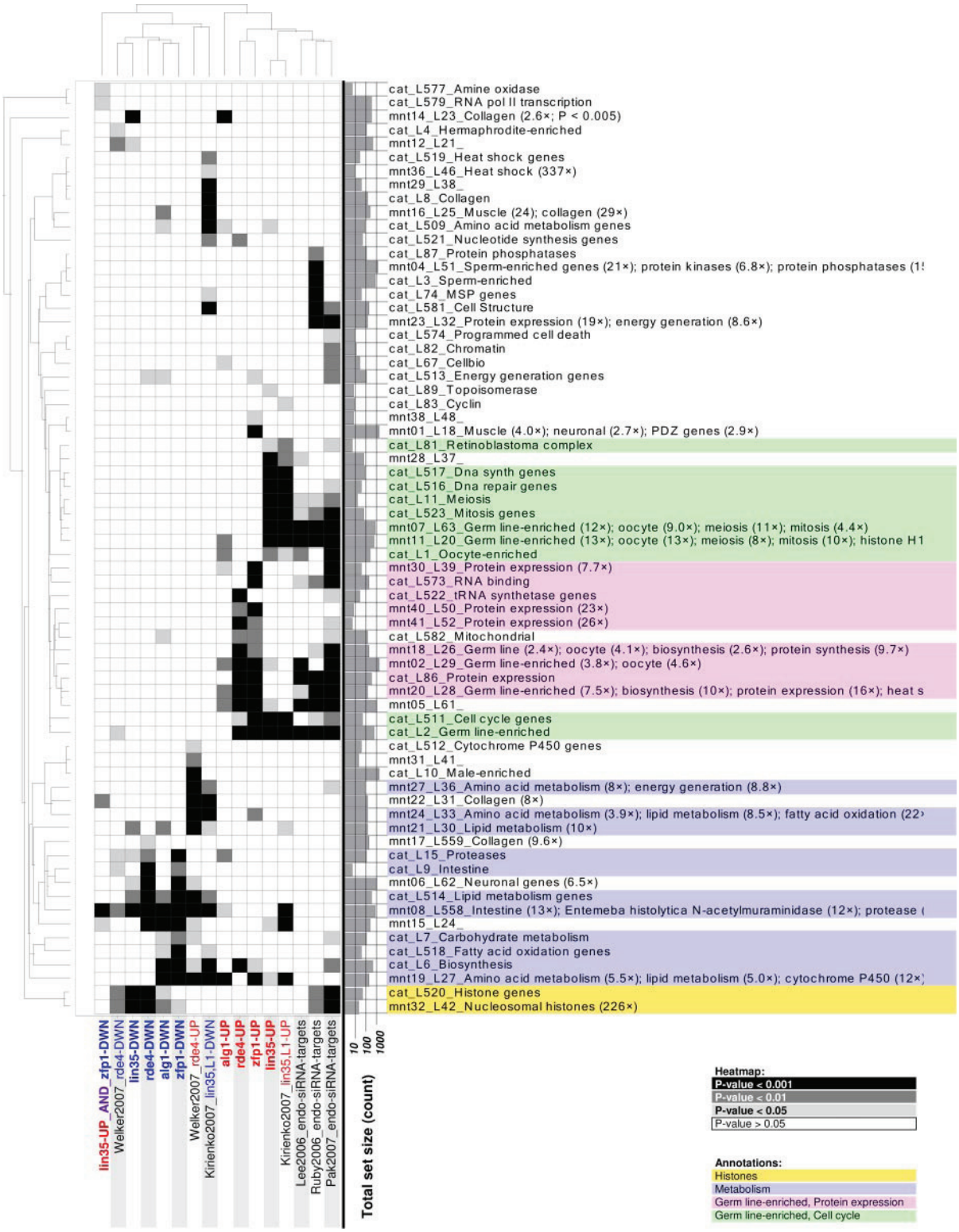


Fig. S1. Heatmap showing the enrichment of selected functional groups as defined by ref. 19 (*y* axis) in gene sets from various datasets (*x* axis). The gray shades in the heatmap indicate significance levels as indicated in the legend. The bars to the right of the heatmap indicate the total size of each functional group. *P* values for the significance of enrichment are presented in Dataset S3.

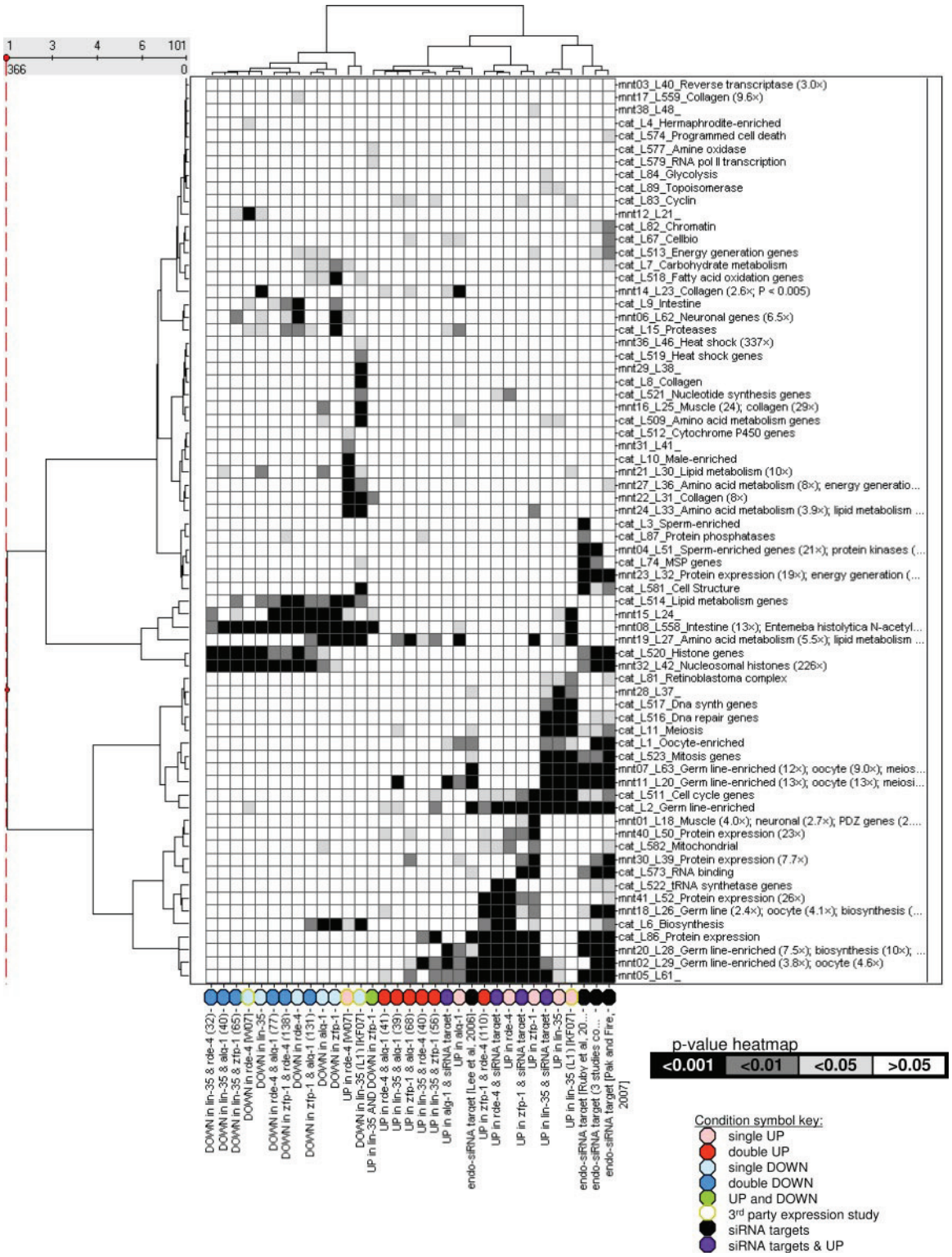


Fig. S2. Variation of a heatmap shown in Fig. S1 that includes additional conditions (gene sets) listed on *x* axis and defined by symbol key.

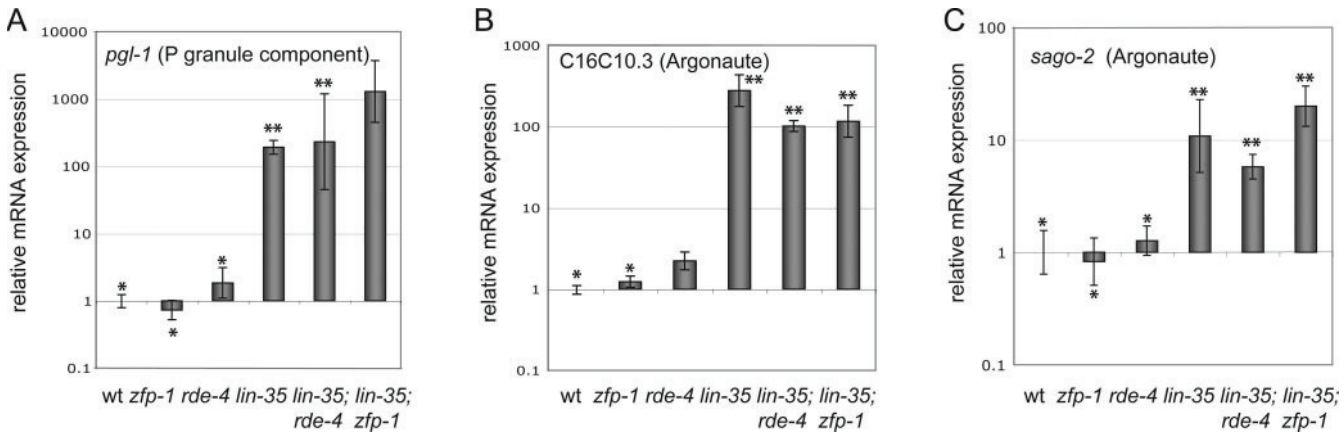


Fig. S3. Examples of genes strongly up-regulated in *lin-35* mutants. Results of three RT-PCR experiments are shown as means and ranges of relative expression compared with wild type.

Table S1. Genes commonly up-regulated in *rde-4(−/−), zfp-1(−/−), alg(−/−)* and *lin-35(−/−)* backgrounds

Cosmid ID	Gene name	Gene description	Chr	Human homolog*	Endo-siRNA target†
C05D10.1		Intestine enriched	III		
C11H1.3		Predicted E3 ubiquitin ligase	X	RNF157	
C14H10.3		Predicted pyridoxal-dependent decarboxylase	X	PDXDC1	
C52D10.2		Up-regulated in dcr-1(−/−) (21)	IV		
K08B4.1	*lag-1*	Transcription factor orthologous to members of the CSL (CBF1, Suppressor of Hairless, LAG-1) family	IV	RBPJ	Pak07
T07D4.4		DEAD(Asp-Glu-Ala-As) -box helicase	II	DDX19A	Lee06, Ruby06, Pak07
W03F9.1		C4-type Zn-finger protein	V	ZNF259	Lee06, Pak07
Y102E9.2		WD repeat protein WDR4	III	WDR4	Pak07
Y48G1C.7			I		Ruby06, Pak07
Y54G9A.4		Fe2+/Zn2 + regulated transporter	II	SLC39A3	

*Tentative human homologs of *C. elegans* genes were defined by a BLAST E-value cutoff of 10^{-10} (best hit).
†Monikers used flag genes being identified as endogenous siRNA target in three studies as follows: Lee06 = Lee *et al.* (1), Pak07 = Pak and Fire (2), and Ruby06 = Ruby *et al.* (3).

Table S2. Overlaps between germline genes (mounts #02, 07, and 11), endo-siRNA targets, and groups of genes up- and down-regulated in four mutants

Germline: 2,862 genes
Endo-siRNA target: 4,372 genes
In experimental group X: n genes

X	n	3-way overlap: observed	3-way overlap: expected*	Enrichment factor
lin-35 UP	413	154	16.6	9.3
lin-35 DOWN	137	10	5.5	1.8
zfp-1 UP	303	42	12.2	3.5
zfp-1 DOWN	348	16	14.0	1.1
rde-4 UP	208	33	8.3	4.0
rde-4 DOWN	177	12	7.1	1.7
alg-1 UP	134	25	5.4	4.6
alg-1 DOWN	175	13	7.0	1.9

*Based on a total of 17,658 genes with Topomap mount annotation

Table S3. Overlaps between nongermline genes in mounts #04, 05, 23, 27, 28, 30, 32, and 41 enriched in endo-siRNA targets (from Fig. S1), endo-siRNA target genes, and groups of genes up- and down-regulated in four mutants

Nongermline, siRNA enriched: 2,531 genes
Endo-siRNA target: 4,372 genes
In experimental group X: n genes

X	n	3-way overlap: observed	3-way overlap: expected*	Enrichment factor
lin-35 UP	413	36	14.66	2.46
lin-35 DOWN	137	16	4.86	3.29
zfp-1 UP	303	44	10.75	4.09
zfp-1 DOWN	348	8	12.35	0.65
rde-4 UP	208	28	7.38	3.79
rde-4 DOWN	177	6	6.28	0.96
alg-1 UP	134	12	4.76	2.52
alg-1 DOWN	175	10	6.21	1.61

*Based on a total of 17,658 genes with Topomap mount annotation

Other Supporting Information Files

Dataset S1 (XLS)
Dataset S2 (XLS)
Dataset S3 (XLS)

Highly aneuploid zebrafish malignant peripheral nerve sheath tumors have genetic alterations similar to human cancers

GuangJun Zhang[a], Sebastian Hoersch[a,b], Adam Amsterdam[a], Charles A. Whittaker[a], Jacqueline A. Lees[a], and Nancy Hopkins[a,1]

[a]Department of Biology and David H. Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02139; and [b]Bioinformatics Group, Max Delbrück Center for Molecular Medicine, 13125 Berlin, Germany

Contributed by Nancy Hopkins, August 17, 2010 (sent for review June 30, 2010)

Aneuploidy is a hallmark of human cancers, but most mouse cancer models lack the extensive aneuploidy seen in many human tumors. The zebrafish is becoming an increasingly popular model for studying cancer. Here we report that malignant peripheral nerve sheath tumors (MPNSTs) that arise in zebrafish as a result of mutations in either ribosomal protein (*rp*) genes or in *p53* are highly aneuploid. Karyotyping reveals that these tumors frequently harbor near-triploid numbers of chromosomes, and they vary in chromosome number from cell to cell within a single tumor. Using array comparative genomic hybridization, we found that, as in human cancers, certain fish chromosomes are preferentially overrepresented, whereas others are underrepresented in many MPNSTs. In addition, we obtained evidence for recurrent subchromosomal amplifications and deletions that may contain genes involved in cancer initiation or progression. These focal amplifications encompassed several genes whose amplification is observed in human tumors, including *met*, *cyclinD2*, *slc45a3*, and *cdk6*. One focal amplification included *fgf6a*. Increasing fgf signaling via a mutation that overexpresses *fgf8* accelerated the onset of MPNSTs in fish bearing a mutation in *p53*, suggesting that *fgf6a* itself may be a driver of MPNSTs. Our results suggest that the zebrafish is a useful model in which to study aneuploidy in human cancer and in which to identify candidate genes that may act as drivers in fish and potentially also in human tumors.

array comparative genomic hybridization | *fgf8* | *met* | Illumina | copy number alteration

Chromosomal instability is a hallmark of human cancer (1, 2). It results in aneuploidy (a nondiploid number of chromosomes) and subchromosomal abnormalities, including inversions, translocations, deletions, and amplifications (3–5). Aneuploidy is particularly common in solid tumors (6). For example, in one study, 85% of colorectal cancers were aneuploid and possessed an average of 60–90 chromosomes (7). Such tumors are frequently heterogeneous, with the number of chromosomes varying from cell to cell.

Most of the chromosomal changes seen in complex cancer genomes are likely to be nonspecific by-products of chromosomal instability. Others, however, are clearly drivers of the cancer phenotype, including certain whole-chromosome amplifications and certain subchromosomal translocations, amplifications, and deletions. A major goal of cancer research is to distinguish pathogenetically relevant alterations from passive changes (2).

Cytogenetic technologies such as chromosome banding, fluorescence in situ hybridization (FISH), and comparative genomic hybridization (CGH) have been adapted for the purpose of characterizing numerical and structural chromosome abnormalities in cancers (8, 9). With CGH, chromosomal abnormalities are measured as DNA copy-number alterations (CNA). Recent large-scale CGH studies have found that human tumors possess preferred whole-chromosome, chromosome-arm, and focal CNAs (10, 11).

The conservation of gene function has made it possible to use animal models to study human cancer. The most powerful model system has been the mouse. However, most mouse cancer models do not display the extensive aneuploidy seen in many types of human tumors. For this reason, researchers have engineered mouse cancer models with chromosomal instability. Such mice develop tumors that are highly aneuploid and also possess subchromosomal alterations. These alterations can then be exploited to identify genetic alterations that drive the cancer phenotype (12). Comparison of T-cell acute lymphoblastic leukemia (T-ALL) in this mouse model with human T-ALL using array CGH (aCGH) indicated important roles of FBXW7 and PTEN in this tumor type (12). In a manner similar to the use of the mouse model, comparison of CNAs in canine and human colorectal cancer has been used to extend the cross-species comparison strategy to identify human cancer genes (13).

The zebrafish is becoming a popular model organism for studying cancer, and a number of tumor models have been made by expression of oncogenes or the mutation of tumor suppressor genes (14). Zebrafish tumors have been shown to have similar gene expression signatures as human cancers (15). However, the nature of the zebrafish cancer genome, including numerical and structural changes, has been largely unexplored, although fluorescence-assisted cell sorting (FACS) analysis suggested that some tumors may be aneuploid (16–18) and low-resolution aCGH has indicated the presence of subchromosomal amplifications and deletions (19). Our laboratory previously reported that zebrafish heterozygous for mutations in any of 17 different *rp* genes develop malignant peripheral nerve sheath tumors (MPNSTs) (20, 21). This is an otherwise rare tumor type in our fish colony. Interestingly, fish homozygous for an inactivating mutation of *p53* also develop MPNSTs (16). Here we report that MPNSTs that arise in either *rp* or *p53* mutant zebrafish mimic human cancer in that they exhibit massive aneuploidy and heterogeneity within a single tumor. Furthermore, as in human cancers, custom-oligonucleotide aCGH and massively parallel synthetic sequencing reveal that despite their heterogeneity, fish MPNSTs display both preferred whole-chromosome copy-number alterations and significant focal copy-number alterations.

Results

Zebrafish MPNSTs Are Highly Aneuploid. To determine whether, like human MPNSTs, zebrafish MPNSTs that arise in either *rp* or *p53* mutant fish are aneuploid, we first investigated the DNA content

Author contributions: G.Z., A.A., J.A.L., and N.H. designed research; G.Z. and A.A. performed research; S.H. and C.A.W. contributed new reagents/analytic tools; G.Z., S.H., A.A., C.A.W., and N.H. analyzed data; and G.Z., S.H., A.A., and N.H. wrote the paper.

The authors declare no conflict of interest.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, http://www.ncbi.nlm.nih.gov/projects/geo (accession no. GSE23666).

[1]To whom correspondence should be addressed. E-mail: nhopkins@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1011548107/-/DCSupplemental.

of fish MPNSTs by FACS. Primary cells were isolated from large externally visible tumors and immediately stained with propidium iodide (PI) and then analyzed. Cells from most zebrafish MPNSTs contain at least one peak at a location expected for aneuploid cells, usually near 3N in DNA content and varying between 2N and 4N. In addition, there was usually a peak representing cells with 2N or near-2N DNA content in the same tumor, and the relative number of cells with 2N versus apparently aneuploid DNA content varied among tumors (Fig. 1 *A*, *D*, *G*, and *J*). We do not know whether the 2N peak represents contaminating normal cells or diploid tumor cells.

To exclude the possibility that the peak with higher DNA content was either an artifact of cell aggregation or 2N cells in the G2 or M phase of the cell cycle, we double-labeled tumor cells with PI and antibody to phosphorylated histone H3 (pH3), which stains cells in late G2 and M, and analyzed them by FACS. The majority of pH3-positive cells were present in a peak with twice the DNA content of the major ~3N peak, suggesting that the aneuploid ~3N cells are dividing. To label cells in S phase, BrdU was injected intraperitoneally into fish bearing tumors, and tumor cells from these fish were then costained with PI and analyzed by FACS. Consistent with the results of pH3 staining, most of the BrdU-positive cells had DNA content greater than that of cells in the presumptive aneuploid peak (Fig. S1). Taken together, these data suggest that the ~3N peak is not an artifact of cell aggregation but rather represents aneuploid cells in G1. Furthermore, the results of pH3 and BrdU labeling suggested that, even in samples with both 2N and ~3N peaks, most of the proliferating cells are derived from the ~3N population. Staining of tumor cells with antibody to γ-tubulin demonstrated that the aneuploid cells had centrosome abnormalities similar to multipolar spindles often found in human cancers (Fig. S1).

To further characterize the aneuploidy in fish MPNSTs, we prepared metaphase chromosome spreads from tumors of five fish heterozygous for mutations in *rp* genes. To prevent potential artifactual changes in chromosome number caused by cell culture, colchicine was injected intraperitoneally into tumor-bearing fish and both normal and tumor cells were harvested 4 h later and fixed. We counted the chromosome number in 19 cells from normal tissue and 100 cells from each of the five tumors. We consistently found 50 chromosomes (the 2N number) in metaphase spreads from normal tissue (Fig. 1 *B* and *C*). In contrast, the chromosome number in tumor cells averaged around 70 per cell in each of the five samples. Notably, the chromosome number varied widely from cell to cell within each tumor, from 48 to 124 chromosomes (Fig. 1 *E*, *F*, *H*, *I*, *K*, and *L*). Thus, we conclude that, like many human solid tumors, zebrafish MPNSTs are both highly aneuploid and heterogeneous in chromosome number.

Preferential Whole-Chromosome Alterations in Zebrafish MPNSTs. Aneuploidy is characteristic of most human solid tumors, and

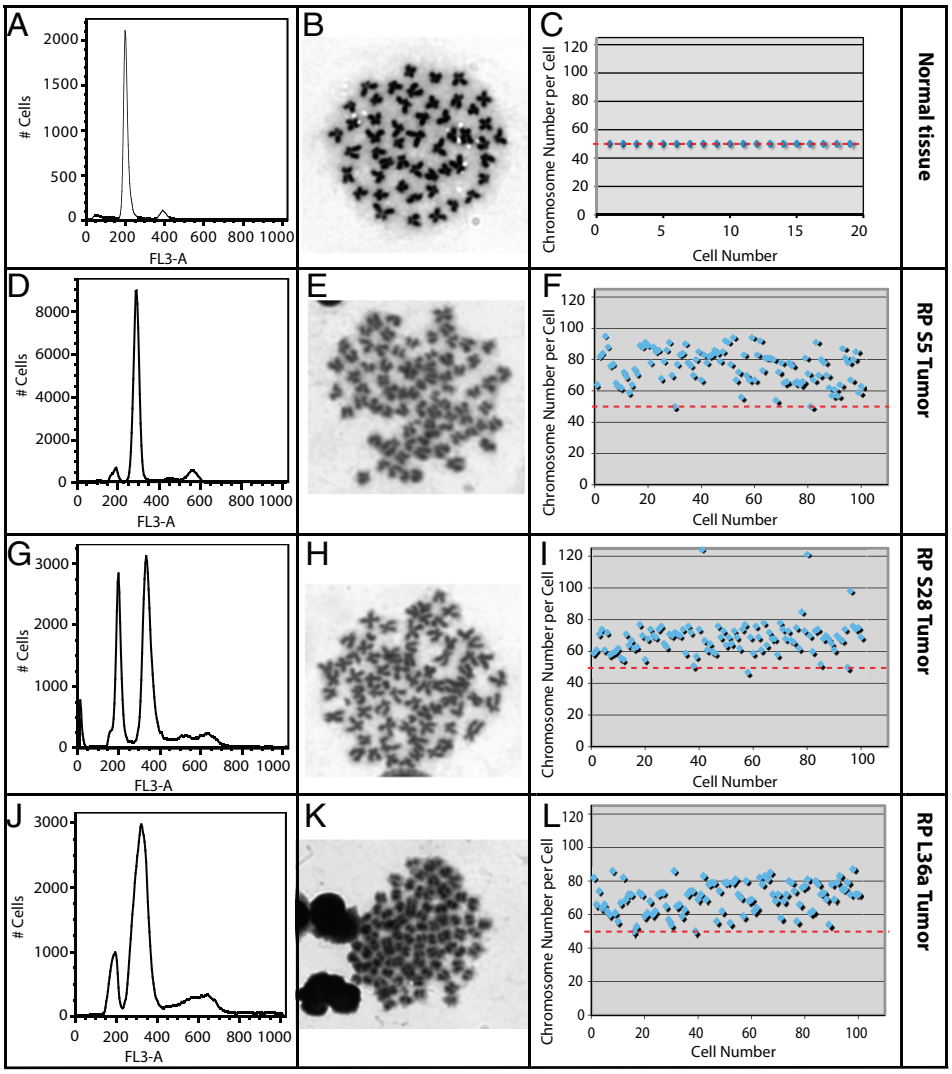


Fig. 1. Zebrafish MPNSTs are highly aneuploid and heterogeneous. FACS DNA content analysis (*A*, *D*, *G*, and *J*), metaphase chromosome spreads (*B*, *E*, *H*, and *K*), and number of chromosomes per cell in 19 normal cells (*C*) or 100 tumor cells (*F*, *I*, and *L*) from three MPNSTs arising in *rp* heterozygous fish or from normal tissue, as labeled on the right. Tumors possess an aneuploid peak and usually in addition a 2N peak that is located at around 200 on the *x* axis. Note that the number of chromosomes per cells varies dramatically within a single tumor.

preferential gain or loss of particular chromosomes has been observed for both specific tumor types and across many types of cancers. Based on this precedent, we sought to establish whether there are common chromosomal copy-number changes among zebrafish MPNSTs. To this end, we designed a custom DNA microarray for aCGH experiments. Our Agilent custom array, designed against the zebrafish Zv7/danRer5 genome assembly, comprised about 15,000 60-mer probes at an average separation of ≈100 kb. After the release of the Zv8 assembly, we remapped the positions of all of the probe sequences and eliminated any for analysis that were not assigned to an assembled chromosome, resulting in 13,646 usable probes. For the most part, the changes in the assembly did not affect the chromosomal coverage on our array, except that a large part of chromosome 4 unique to the Zv8 assembly turned out to be poorly represented.

We analyzed DNA from 36 tumors, 5 of them from fish homozygous for a mutation in *p53*, and 31 of them obtained from 13 different lines of *rp* heterozygotes. To avoid artifacts arising from polymorphisms in the fish genome, we used DNA from the tail of the same fish as the reference DNA for each tumor sample. Tumor samples and their respective reference DNAs were differentially labeled with Cy3 or Cy5, and hybridized to the same array. Data were normalized across the entire probe set and subjected to a circular binary segmentation algorithm. Thus, if a segment corresponding to an entire chromosome was found to be either above or below the baseline, we considered that to be a whole-chromosome gain or loss, respectively. Importantly, as these tumors are ~3N, the baseline is likely three copies. Thus, losses can represent two or fewer copies, whereas gains should represent four or more copies.

As in human tumors, we found that chromosome copy number changes in zebrafish MPNSTs were not random. The most dramatic and common change was a relative gain of chromosome 25 (Fig. 2). Other chromosomes that were frequently overrepresented were 10 and 11. The most commonly underrepresented chromosomes were 8 and 15. By contrast, chromosomes 3, 12, 13, 14, and 16, for example, showed relatively little deviation from baseline across the 36 samples or had similar-sized subsets of tumor samples showing over- and underrepresentation (Fig. 2). To confirm these findings, we sequenced several samples using the Illumina Genome Analyzer platform. Based on a total of at least 900,000 alignable reads for both normal and tumor tissues in three fish, we obtained nearly identical postsegmentation results for copy-number inter- and intrachromosomal variations as with aCGH (Fig. 3*A*). To confirm the aCGH results by yet another method, we used Southern blots to investigate an additional 14 tumors arising in *rp* heterozygous fish. For this analysis, we used probes against the commonly overrepresented chromosome 25, against the commonly underrepresented chromosome 15, and against chromosomes 13, 14, and

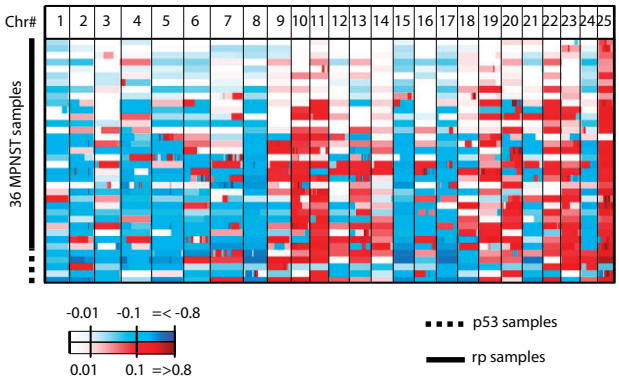


Fig. 2. Array CGH analysis of copy-number changes in zebrafish MPNSTs. Heatmap showing aCGH results indicating relative chromosomal gain (red) and loss (blue) in 36 zebrafish tumor/normal sample pairs across chromosomes 1–25. Thirty-one *rp* tumor samples are shown on the top and five *p53* tumor samples are shown on the bottom. The heatmap is colored according to the segment mean of the data as processed by a circular binary segmentation algorithm, with segment mean values between –0.01 and 0.01 displayed white, and segment mean values between 0.01 (–0.01) and 0.8 (–0.8) displayed in continuous shades of red (blue) on a quasilogarithmic scale (see color band at the bottom). Segment mean values above (below) 0.8 (–0.8) are rendered in the most saturated color value.

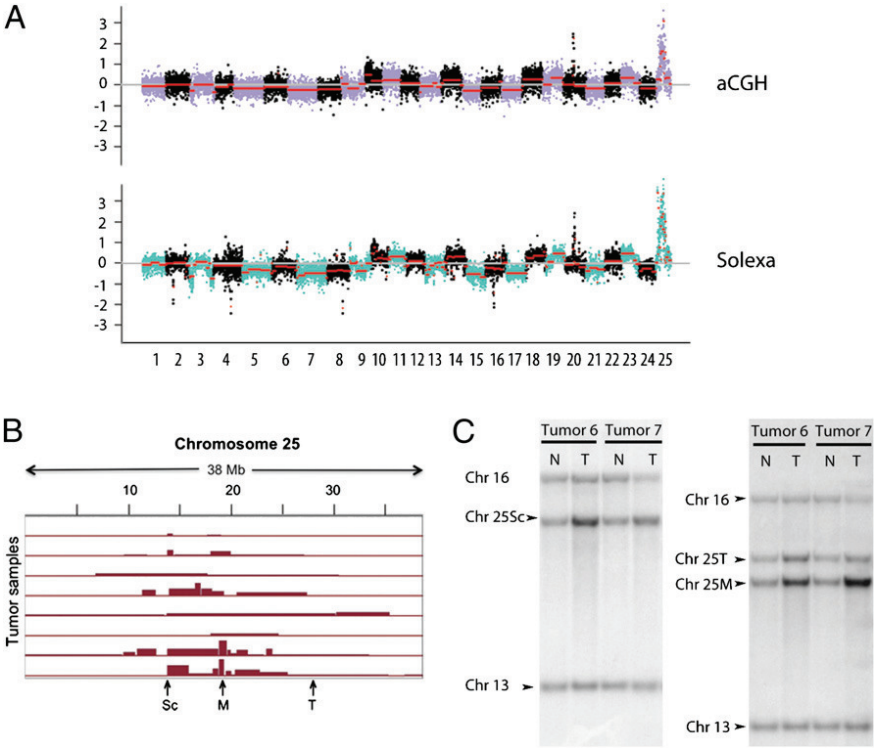


Fig. 3. Array CGH results are supported by Illumina sequencing and Southern blot. (*A*) Postnormalization data of one zebrafish tumor/normal sample pair analyzed by aCGH (*Upper*) and Illumina sequencing (*Lower*). The red line indicates the segment mean as obtained by a circular binary segmentation algorithm. Note that the results obtained by these two technologies with respect to both inter- and intrachromosomal changes are largely in agreement. (*B*) Array CGH analysis demonstrated that 8 of the 36 tumor samples showed subchromosomal amplifications on chromosome 25 beyond the chromosome-level gain. The positions of Southern blot probes to two areas most often overamplified (Sc and M) and an area not additionally amplified (T) are indicated by arrows. (*C*) Southern blots of the DNA isolated from tails (N) and tumors (T) from 2 of the 14 *rp* heterozygotes analyzed in this way. Three or four probes were used on each blot; the positions of bands to which probes from chromosomes 13 or 16 or various locations on chromosome 25 as indicated in *B* hybridized are marked. Note that in tumor 6, all of the chromosome 25 probes show more signal in the tumor sample than the tail sample; in tumor 7, whereas all chromosome 25 probes do so, the increase for the M probe (and to a lesser extent the Sc probe) is greater than that for the T probe. Quantitation for all of the 14 *rp* heterozygotes and various probes can be found in Figs. S2 and S3.

16, whose sequences show little tendency to be either over- or underrepresented. Quantitation of the hybridization signal from these probes confirmed that chromosome 15 was frequently underrepresented and chromosome 25 was frequently overrepresented in this independent sample of 14 tumors (Fig. S2); the amount of chromosomes 13 and 14 did not consistently vary (Fig. S3). Overall, these data suggest that certain chromosomes are preferentially gained or lost in highly aneuploid MPNSTs of zebrafish. We found the same trends in *p53* mutant tumors as in *rp* mutant tumors, although *p53* mutant tumors appear to additionally show common loss of chromosomes 3 and 8 and gain of chromosome 24 (Fig. 2). We would need to analyze more *p53* samples to be confident that these changes are consistently preferred in that genetic background.

Subchromosomal Amplifications on Chromosome 25. In addition to whole-chromosome changes, both aCGH and sequencing revealed subchromosomal areas of copy-number variation in fish MPNSTs (Figs. 2 and 3*A*). Striking areas of amplification occurred on the already overrepresented chromosome 25. To verify these changes by another method, we used Southern blot analysis of DNA from the 14 additional *rp* heterozygous tumors described above. As controls, we again used probes that hybridize to either chromosome 13, 14, or 16. We used three probes that hybridize to different regions of chromosome 25; two of the probes (Sc and M) hybridize to one or the other of the two regions found to be highly amplified on this chromosome by aCGH analysis, and one probe (T) hybridizes to an area that tracks with the rest of chromosome 25 in being overrepresented but not additionally amplified (see Fig. 3*B* for probe locations). In at least a third of the 14 tumors analyzed by Southern blot, sequences of chromosome 25 represented by the Sc and/or M probes were amplified relative to the T region on chromosome 25 (Fig. 3*C* and Fig. S4). As observed previously with aCGH on the initial panel of 36 tumors, sometimes both Sc and M regions were amplified, sometimes just one or the other was amplified.

Regions of Common Focal Amplification or Deletion in Fish MPNSTs. Identifying regions of focal amplification or deletion that are biologically relevant in cancers with complex genomes is not trivial, because of the large number of random alterations that occur in such tumors. Here we first identified regions that appeared to be either commonly gained or lost by visual inspection of heatmap displays of segmented aCGH data for each chromosome. To arrive at a final list, we then combined our findings with the results of a permutation-based test (22) to determine which regions met the criterion of statistical significance in terms of ruling out the null hypothesis that overlaps between segments in different samples are merely random. We used this two-pronged approach because, whereas statistical significance is obviously impossible to judge by eye, the algorithmic results themselves are subject to certain limitations and artifacts requiring human input. For example, a key parameter in such analysis is a fixed cutoff level that is required for segment means to be considered for candidate focal regions. Although we generally used a cutoff of ± 0.2 (segment mean), we found that for chromosome 25, a higher cutoff level (0.6) was needed in addition to adequately take into account the large dynamic range (up to a segment mean of 3.47) encountered here.

We identified two types of apparent focal changes. The first type was similar to what has been observed in human tumors: amplifications or deletions with varying boundaries from tumor to tumor, among which areas of overlap were evident. We found eight such regions on five different chromosomes among the 36 tumor samples analyzed. Together, they contain ≈220 genes (Dataset S1). Seven of these regions were gains, one was a loss. Two of the amplified areas identified lie on chromosome 25, one near the Sc probe and one near the M probe described above. The Sc region includes five to seven genes: *ccnd2a*, *tigara*, *fgf6a*, *slc45a3*, *c12orf4*, and possibly *dyrk4* and *nduf9* (these last two lie between the last probe in the segment and the first probe outside of the segment). The M region includes 14 genes, one of which is the proto-oncogene *met*, which has been previously identified as a gene frequently amplified in human MPNSTs and which is activated by mutation or overexpressed in perhaps 50% of human tumors (23, 24). The other genes in this region are *arntl2*, *loh12cr1*, *cep290*, *tmtc3*, *kitla*, *dusp6*, *poc1b*, *tsga14*, *cpa1*, *cpa5*, *tes*, *cav2*, and *cav1*.

In the second type of focal change, two to five fish had the same few (two to six) probes representing an amplification or deletion. We found eight cases of this type. Although we cannot rule out the possibility that these are genuine copy-number changes, it seems highly possible that these are cases where, even in the most recent assembly of the zebrafish genome (Zv8), the genomic positions of these probes are misassigned. Should these probes in reality be on chromosomes that are (for that tumor) up or down at the whole-chromosome level relative to the chromosome to which they have been misassigned, they would appear to represent a focal change. In initially working with the aCGH data with the probes mapped to the previous assembly (Zv7), we had observed multiple instances of this type of focal event, and nearly all of them disappeared after probes were reassigned based on the Zv8 assembly. Thus, we conclude that this type of "narrow" focal change cannot be called with any reliability under the present circumstances.

To determine whether genes in a focally amplified region may contribute to the initiation or progression of cancer requires a biological assay. As described next, we obtained preliminary evidence that *fgf6* might be a driver of zebrafish MPNSTs.

Preliminary Biological Validation of a Candidate Gene from a Focal Amplification: *fgf8* Overexpression Accelerates Tumorigenesis in a *p53* Mutant Background. The *fgf* gene family contains many members, but in mammals they can all bind to four FGF receptors, members of the receptor tyrosine kinase superfamily, suggesting that they might have similar ability to signal through common MAP-kinase signaling pathways. Although we do not have a mutant line of fish that overexpresses *fgf6a*, a mutant line designated Hag^{D1} overexpresses *fgf8a* as a result of a retroviral insertion (25). This line develops large neuroblastomas at a low frequency and rarely before 1 y of age (25). To determine whether *fgf8a* may be a driver gene for MPNSTs (as well as neuroblastomas), we introduced the Hag^{D1} mutation into fish heterozygous or homozygous for a mutation in *p53*. *fgf8a* overexpression accelerated the onset of MPNSTs in both *p53* heterozygous and homozygous backgrounds (Fig. 4). This is consistent with the possibility that, by analogy, *fgf6a* may also be a driver in MPNSTs, and it demonstrates the feasibility of testing candidate genes identified by genomic approaches.

Discussion

Our results suggest that the zebrafish may be a useful model in which to study aneuploidy in human cancer. As with many human cancers, we showed that zebrafish MPNSTs that arise in fish heterozygous for *rp* mutations or homozygous for a *p53* mutation are highly aneuploid, frequently possessing pseudotriploid genomes. Furthermore, as in human cancers, the number of chromosomes per cell is extremely heterogeneous within a single zebrafish MPNST. Finally, aCGH, massively parallel sequencing, and Southern blotting revealed that despite the heterogeneity seen in fish tumors, some copy-number changes dominate any given tumor, presumably because cells with those changes have been selected for better growth. These copy-number alterations include both whole-chromosome and subchromosomal regions.

The similarities in genomic changes between fish and human MPNSTs extend beyond single tumors to properties shared by multiple tumors. We observed that among the 36 zebrafish MPNSTs analyzed by aCGH, whole-chromosome copy number changes are far from random: Several chromosomes are frequently gained and seldom lost in many tumors, whereas others show frequent loss and

rare gain. Recent analysis of large numbers of human tumors by CGH and sequencing shows that many chromosomes or chromosome arms are preferentially gained or lost in particular tumor types, and some preferences are shared across many human cancers (10, 11). It is postulated that the preferential over- or underrepresentation of certain chromosomes or chromosome arms in both human and zebrafish tumors may reflect a growth advantage conferred by genes that lie on these chromosomes.

In addition to whole-chromosome changes, we also found subchromosomal segments of gain or loss in fish MPNSTs. Using statistical methods to identify common focal areas of copy-number variation in human tumors has been a powerful tool for distinguishing focal changes that contain driver genes from those that are merely a harmless consequence of genomic instability. We used similar methods here and identified eight statistically significant regions of gain or loss manifest in several tumors. An analysis of more tumors, as well as use of a higher-resolution platform (both the aCGH and sequencing approaches used here had windows of about 0.1 Mb) will likely identify additional focal areas as well as narrow their size. This analysis in zebrafish is currently hampered by the unfinished nature of the zebrafish genome assembly, as well as incomplete gene annotation, but both will surely improve. In addition, adapting sophisticated tools such as GISTIC (26) for use with zebrafish, especially in combination with an increased sample size, could improve the statistical robustness of our findings. Meanwhile, the focal changes we have found thus far, although certainly not complete, clearly contain genes whose amplification is observed in human tumors, such as *met*, *cyclinD2*, *slc45a3*, and *cyclin-dependent kinase 6*. *met* is of particular interest because it has been identified as a gene that is often overexpressed in human MPNSTs (23). Furthermore, an increase in copy number of an activated *met* oncogene is thought to underlie the common gain of an extra chromosome 7 in papillary renal carcinomas (27), and increased expression of *met* may in part explain the common overrepresentation of chromosome 7 in many human cancers.

The high degree of aneuploidy observed in fish MPNSTs could prove to be extremely beneficial in helping to identify important drivers in human cancer. Most murine cancer models do not show extensive aneuploidy. As a result, mouse cancer models have been engineered specifically to generate a greater degree of genomic instability (e.g., 12). Using such models, a comparison of syntenic regions between mouse and human chromosomes that show copy number variation in tumors in both organisms has been helpful in further identifying biologically relevant focal changes in human tumors (12, 28). A similar approach using shared copy-number alterations between fish and human cancers could be even more powerful. This is because mouse and human are relatively close evolutionarily, so focal changes shared by these two organisms tend to share nearly all of the same genes, making it difficult to separate driver genes from passengers that merely cosegregate due to proximity. In contrast, zebrafish and human are far more evolutionarily distant, and syntenic blocks between them tend to be much shorter (29). Thus, focal regions in each organism containing the same driver gene are unlikely to share many additional genes, making it much easier to narrow down the list of candidate driver genes within the shared focal region.

Even more difficult to analyze in human tumors has been the significance of preferred chromosome or chromosome arm-level changes. As with focal regions, because of the nature of the synteny between fish and human genomes, a comparison of the genes on chromosomes preferentially gained or lost in human and zebrafish tumors could severely limit the number of candidate genes to examine. Although some kind of biological assay would be needed, such as the effects upon growth and survival properties in tumor cell lines, it is far easier to study the effects of knockdown or overexpression of a few dozen genes on a chromosome than of several hundred. Furthermore, as the example here with overexpression of *fgf8a* in zebrafish with *p53* mutations shows, it is also possible to test the effects of copy-number alterations of specific genes on tumorigenesis in the fish itself, including in the context of different tumor-promoting mutations or transgenes.

The goal of much cancer research today is to identify the combination of driver genes that allow each and every human tumor to proliferate and evolve. The ability to type many highly aneuploid tumor genomes by next-generation sequencing and aCGH in zebrafish should help to parse tumors into groups with specific combinations of drivers generated by copy-number alterations and then to test combinations of inhibitors to treat the different subclasses of cancer.

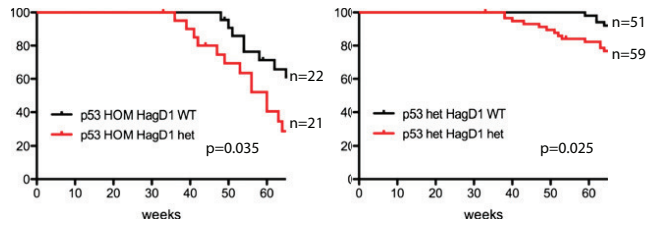


Fig. 4. Overexpression of *fgf8a* synergizes with mutation of *p53* in MPNST development. Kaplan–Meier curves indicate when externally visible tumors were first observed in fish with the indicated genotypes. All fish were progeny of the same four pairs born the same day. Tumor type was confirmed by histological analysis, and fish with neuroblastoma (1 in *p53* HOM, Hag^{D1} het, 1 in *p53* het, Hag^{D1} het) were excluded from the analysis. No MPNSTs were found in *p53* WT, Hag^{D1} heterozygous fish. *P* values comparing genotyped pairs are shown.

Materials and Methods

Zebrafish Lines. Tumor-prone lines of zebrafish used in these studies have been described previously. They include fish heterozygous or homozygous for a point mutation in *p53* (16), fish heterozygous for an insertional mutation in any one of 15 different ribosomal protein genes (21), and fish with the insertional allele Hag^{D1}, which overexpresses *fgf8a* (25). Stocks were maintained as described previously and genotypes were determined by PCR at 8–18 wk of age as described (30).

Genomic DNA Isolation and Southern Blot Analysis. Genomic DNA from tail tissue and tumors was prepared as in ref. 31. Approximately 5 μg of each sample was cut with HindIII and Southern blots were conducted as in ref. 32. Details of probe sequences and quantitation are provided in *SI Materials and Methods*.

Flow Cytometry. Tumors were dissected out after euthanization with 500 mg/L tricaine, and tail fins were clipped to obtain normal cells from the same fish. After euthanization, tumors were dissected out of *p53* homozygous or *rp* heterozygous fish between the ages of 8 and 24 mo. Single-cell suspensions were made by digestion at room temperature for 1 h in 2 U/mL dispase (Invitrogen) in PBS, and were strained through a 40-μm filter. Cell suspensions were then pelleted and washed one time with PBS before being fixed with 70% ethanol and stored at −20 °C. For BrdU labeling, 10 mM BrdU was injected into the peritoneal cavity of tumor-bearing fish, and the tumor was dissected out about 45 min later. BrdU, pH3, and propidium iodide (PI) double staining were performed as has been described for embryonic cells (33). FACS analysis was conducted by FACScan (Becton-Dickinson). DNA content, pH3, and BrdU data were analyzed by FlowJo (Tree Star).

Cytogenetics and Chromosome Counting. Colchicine (0.025%; 10 μL/g weight) was injected peritoneally into fish 4 h before dissection. After fish had been killed, tumors were dissected out and cell suspensions were prepared in the same way as for FACS. Cells were treated with 0.5% KCl and incubated at 35 °C for 30 min. Following hypotonic treatment, the cells were fixed in Carnoy's fixative (75% methanol, 25% acetic acid) and stored at −20 °C. Chromosome spreads were made as described (34). For chromosome counting, slides were stained with either DAPI or Giemsa. Pictures were taken at 1,000× magnification using a Zeiss Axioplan II upright microscope, and chromosome numbers were counted using ImageJ (National Institutes of Health).

Array CGH and Data Processing. Genomic DNA was isolated from tumors and from tail tissue of the same fish (used as reference) as in ref. 35. Samples were prepared and hybridized to the array containing 13,648 probes mapped to assembled chromosomes in Zv8/danRer6 according to the Agilent standard protocol with some modifications. The arrays were scanned with Agilent scanner G2505B, and Agilent Feature Extraction software v9.1.3.1 was used to obtain normalized data (column logRatio in output file) that were subsequently converted from $\log_{10}$ to $\log_2$ scale. Normalized $\log_2$-ratio data were submitted to the circular binary segmentation algorithm (36) as implemented in the BioConductor package DNAcopy (v1.16.0) and processed with default parameters. The segments and segment mean values obtained were the basis for subsequent analyses. For two sample pairs, an accidental dye swap was identified by analysis and comparison of single-channel data, and segment means for these samples were multiplied by −1 to correct this mistake. The STAC algorithm (22) in a Java implementation provided on the authors' Web site (http://cbil.upenn.edu/STAC; v1.2) was applied to assess whether the overlap of two or more subchromosomal segments was likely to be a chance event. Full datasets are publicly available at http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE23666;. Additional details on the design of the array, labeling and hybridization conditions, and STAC analysis are provided in *SI Materials and Methods*.

Massively Parallel Synthetic Sequencing. Samples for sequencing with the Illumina Genome Analyzer IIx system were prepared from the same batches of genomic DNA used for aCGH, according to published methods (37–39). Sequences were processed with either Bustard.py (OLB 1.6.0) or GERALD.pl (CASAVA 1.6.0), downloaded from the Illumina Web site (http://www.illumina.com/software/genome_analyzer_software.ilmn), with a produced read length of 41 nucleotides. Further details on the analysis of the sequencing data are provided in *SI Materials and Methods*. Sequencing data are available at http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE23666.

Note Added in Proof. In further support of the concern that the very small focal CNAs that were observed in several fish might have been an artifact of misassignment of the underlying probe sequences in the genome assembly, we found that probe sequences for 5 of the 8 cases map to a different chromosome in the pre-Ensembl release of the Zv9 assembly in July 2010.

ACKNOWLEDGMENTS. We thank Tim Angelini and Kate Anderson for maintenance of the zebrafish colony. We thank Eduardo Torres and Angelika Amon for assistance with aCGH experiments. We thank members of the Lees laboratory for helpful discussion, especially Christopher Sansam for help with the FACS analysis. We thank Kim Brown, Lynda Chin, Charles Lee, and members of their laboratories for helpful discussions. We thank Agilent Technologies for making available to us their zebrafish CGH/ChIP optimized probe database. This work was supported by a grant from Arthur C. Merrill to N.H., a National Institutes of Health–National Cancer Institute grant (CA106416) to J.A.L. and N.H., and a core grant (CA14051) to the Koch Institute for Integrative Cancer Research. J.A.L. is a Ludwig Scholar at the Massachusetts Institute of Technology.

1. Negrini S, Gorgoulis VG, Halazonetis TD (2010) Genomic instability—An evolving hallmark of cancer. *Nat Rev Mol Cell Biol* 11:220–228.
2. Lengauer C, Kinzler KW, Vogelstein B (1998) Genetic instabilities in human cancers. *Nature* 396:643–649.
3. Duesberg P, Rausch C, Rasnick D, Hehlmann R (1998) Genetic instability of cancer cells is proportional to their degree of aneuploidy. *Proc Natl Acad Sci USA* 95:13692–13697.
4. Rajagopalan H, Nowak MA, Vogelstein B, Lengauer C (2003) The significance of unstable chromosomes in colorectal cancer. *Nat Rev Cancer* 3:695–701.
5. Rajagopalan H, Lengauer C (2004) Aneuploidy and cancer. *Nature* 432:338–341.
6. Mitelman F, Johansson B, Mertens F (2010) Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. Available at http://cgap.nci.nih.gov/Chromosomes/Mitelman. Accessed August 30, 2010.
7. Pellman D (2001) Cancer. A CINtillating new job for the APC tumor suppressor. *Science* 291:2555–2556.
8. Speicher MR, Carter NP (2005) The new cytogenetics: Blurring the boundaries with molecular biology. *Nat Rev Genet* 6:782–792.
9. Pinkel D, Albertson DG (2005) Array comparative genomic hybridization and its applications in cancer. *Nat Genet* 37 (Suppl):S11–S17.
10. Beroukhim R, et al. (2010) The landscape of somatic copy-number alteration across human cancers. *Nature* 463:899–905.
11. Baudis M (2007) Genomic imbalances in 5918 malignant epithelial tumors: An explorative meta-analysis of chromosomal CGH data. *BMC Cancer* 7:226.
12. Maser RS, et al. (2007) Chromosomally unstable mouse tumours have genomic alterations similar to diverse human cancers. *Nature* 447:966–971.
13. Tang J, et al. (2010) Copy number abnormalities in sporadic canine colorectal cancers. *Genome Res* 20:341–350.
14. Feitsma H, Cuppen E (2008) Zebrafish as a cancer model. *Mol Cancer Res* 6:685–694.
15. Lam SH, et al. (2006) Conservation of gene expression signatures between zebrafish and human liver tumors and tumor progression. *Nat Biotechnol* 24:73–75.
16. Berghmans S, et al. (2005) tp53 mutant zebrafish develop malignant peripheral nerve sheath tumors. *Proc Natl Acad Sci USA* 102:407–412.
17. Langenau DM, et al. (2003) Myc-induced T cell leukemia in transgenic zebrafish. *Science* 299:887–890.
18. Langenau DM, et al. (2005) Cre/lox-regulated transgenic zebrafish model with conditional myc-induced T cell acute lymphoblastic leukemia. *Proc Natl Acad Sci USA* 102:6068–6073.
19. Freeman JL, et al. (2009) Construction and application of a zebrafish array comparative genomic hybridization platform. *Genes Chromosomes Cancer* 48:155–170.
20. Amsterdam A, et al. (2004) Many ribosomal protein genes are cancer genes in zebrafish. *PLoS Biol* 2:E139.
21. Lai K, et al. (2009) Many ribosomal protein mutations are associated with growth impairment and tumor predisposition in zebrafish. *Dev Dyn* 238:76–85.
22. Diskin SJ, et al. (2006) STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res* 16:1149–1158.
23. Mantripragada KK, et al. (2008) High-resolution DNA copy number profiling of malignant peripheral nerve sheath tumors using targeted microarray-based comparative genomic hybridization. *Clin Cancer Res* 14:1015–1024.
24. Knudsen BS, Vande Woude G (2008) Showering c-MET-dependent cancers with drugs. *Curr Opin Genet Dev* 18:87–96.
25. Amsterdam A, et al. (2009) Zebrafish Hagoromo mutants up-regulate fgf8 postembryonically and develop neuroblastoma. *Mol Cancer Res* 7:841–850.
26. Beroukhim R, et al. (2007) Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc Natl Acad Sci USA* 104: 20007–20012.
27. Zhuang Z, et al. (1998) Trisomy 7-harbouring non-random duplication of the mutant MET allele in hereditary papillary renal carcinomas. *Nat Genet* 20:66–69.
28. Alberici P, et al. (2007) Aneuploidy arises at early stages of Apc-driven intestinal tumorigenesis and pinpoints conserved chromosomal loci of allelic imbalance between mouse and human. *Am J Pathol* 170:377–387.
29. Catchen JM, Conery JS, Postlethwait JH (2009) Automated identification of conserved synteny after whole-genome duplication. *Genome Res* 19:1497–1505.
30. Golling G, et al. (2002) Insertional mutagenesis in zebrafish rapidly identifies genes essential for early vertebrate development. *Nat Genet* 31:135–140.
31. Amsterdam A, et al. (1999) A large-scale insertional mutagenesis screen in zebrafish. *Genes Dev* 13:2713–2724.
32. Gaiano N, Allende M, Amsterdam A, Kawakami K, Hopkins N (1996) Highly efficient germ-line transmission of proviral insertions in zebrafish. *Proc Natl Acad Sci USA* 93: 7777–7782.
33. Shepard JL, Stern HM, Pfaff KL, Amatruda JF (2004) Analysis of the cell cycle in zebrafish embryos. *Methods Cell Biol* 76:109–125.
34. Lee C, Smith A (2004) Molecular cytogenetic methodologies and a bacterial artificial chromosome (BAC) probe panel resource for genomic analyses in zebrafish. *Methods Cell Biol* 77:241–254.
35. Strauss WM (2001) Preparation of genomic DNA from mammalian tissue. *Curr Protoc Mol Biol* 2.2.1–2.2.3.
36. Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* 23:657–663.
37. Quail MA, et al. (2008) A large genome center's improvements to the Illumina sequencing system. *Nat Methods* 5:1005–1010.
38. Bentley DR, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59.
39. Quail MA, Swerdlow H, Turner DJ (2009) Improved protocols for the Illumina genome analyzer sequencing system. *Curr Protoc Hum Genet* 62:18.2.1–18.2.27.

Supporting Information

Zhang et al. 10.1073/pnas.1011548107

SI Materials and Methods

Immunohistochemistry and Immunofluorescence. Tissues were fixed in 4% paraformaldehyde at 4 °C for up to 2 d. The fish were dehydrated in alcohol, cleared in xylene, and infiltrated with paraffin. Cell suspensions were made before 4% paraformaldehyde fixation. Primary antibody against phosphorylated histone H3 (pH3) (Santa Cruz Biotechnology) was diluted 1:100 and the antibody against γ-tubulin (Sigma) was used at 1:100. Nuclei were counterstained with DAPI for visualization using a Zeiss microscope.

Array CGH and Data Processing. Array probes (15,104) against zebrafish chromosomes 1–25 were selected from a pool of more than 7 million zebrafish probes kindly provided by Agilent. Preprocessing was performed excluding probes containing AluI and RsaI restriction sites to ensure compatibility with an enzymatic comparative genomic hybridization (CGH) protocol for genomic DNA digestion for labeling purposes, applying a cutoff to Agilent-provided probe quality scores, and excluding probes exhibiting perfect secondary matches down to 21 nucleotides as assessed by a BLAT (1) search against zebrafish genomic sequence (Zv7/danRer5 assembly). Among the resulting set of 60,000 probes, the target number of about 15,000 probes for chromosomes 1–25 was reached by an iterative heuristic for selecting approximately equidistant probes. After the Zv8/danRer6 genome assembly was released, we remapped the probes by BLAT search against the improved assembly Zv8/danRer6. Probes were flagged as "unusable" if they mapped with high quality to more than one assembled chromosome (1–25) in Zv8/danRer6, resulting in 13,648 available probes. With the exception of a novel and highly repetitive portion on chromosome 4, probe coverage, although not as uniform as for the Zv7/danRer5 assembly underlying the original design, was satisfactory under the new assembly.

Genomic DNA was isolated from tumors and from tail tissue of the same fish (used as reference) as in ref. 2. Samples were prepared and hybridized to the arrays according to the Agilent standard protocol with some modifications. Briefly, after fragmentation of 5 μg of tumor or reference DNA, DNA samples were labeled with Cy3-dCTP or Cy5-dCTP by random priming. Unincorporated nucleotides were removed using a QIAquick PCR purification kit (QIAGEN). Before hybridization the probes were measured by NanoDrop 1000 (Thermo Scientific), and were denatured by incubation at 95 °C for 2 min followed by cooling to room temperature. Array hybridization was carried out at 65 °C with about 200 ng probes/array in Agilent HI-RPM hybridization buffer. After washing, the arrays were scanned with Agilent scanner G2505B, and Agilent Feature Extraction software v9.1.3.1 was used to obtain normalized data (column logRatio in output file) that was subsequently converted from $\log_{10}$ to $\log_2$ scale. Normalized $\log_2$-ratio data were submitted to the circular binary segmentation algorithm (3) as implemented in the BioConductor package DNAcopy (v1.16.0), and processed with default parameters. The segments and segment mean values obtained were the basis for subsequent analyses. For two sample pairs, an accidental dye swap was identified by analysis and comparison of single-channel data, and segment means for these samples were multiplied by −1 to correct this mistake. Full datasets are publicly available at http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE23666.

Assessment of the Significance of Focal Changes with STAC. The STAC algorithm (4) in a Java implementation provided on the authors' Web site (http://cbil.upenn.edu/STAC; v1.2) was applied to assess whether the overlap of two or more subchromosomal segments was likely to be a chance event, given the chromosome-specific incidence of observed segments (STAC "footprint *P* value"). Segments indicative of relative gain and relative loss were processed separately. Only segments smaller than 20 MB and with a mean value above (below) 0.2 (−0.2) were considered for analysis, with the exception of gains on chromosome 25, where a mean value cutoff of 0.6 was used to better account for the high density of segments with mean values above 0.2 and their unusually broad mean value range. STAC was run with a resolution of 50,000 nucleotide positions and 5,000 permutations.

Massively Parallel Synthetic Sequencing. Samples for sequencing with the Illumina Genome Analyzer IIx system were prepared from the same batches of genomic DNA used for array CGH (aCGH) according to published methods (5–7). All of the enzymes used in the preparation were from New England Biolabs, and the oligonucleotides were synthesized by Eurofins MWG Operon. Oligonucleotide sequences (ATTGGC; GATCTG; TCAAGT; CTGATC; AAGCTA; GTAGCC; TACAAG; CGTGAT; ACATCG; GCCTAA; TGGTCA; CACTGT) were added to the 3′ end of the Illumina adaptors used for the paired-end library preparation to serve as barcodes so that multiple samples could be sequenced in the same reaction. Sequencing was performed on an Illumina GAIIx Sequencer using Sequence Control software v2.6.26. Sequences were processed with either Bustard.py (OLB 1.6.0) or GERALD.pl (CASAVA 1.6.0), downloaded from the Illumina Web site (http://www.illumina.com/software/genome_analyzer_software.ilmn), with a produced read length of 41 nucleotides.

FASTQ sequence files were split into sample sets according to a barcode strategy and then barcodes were trimmed from the sequences using the applications fastx_barcode_splitter.pl and fastx_trimmer from the FASTX-Toolkit-0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit) package. Illumina sequence qualities were converted to Sanger sequence qualities using the MAQ application fq_all2std.pl (http://maq.sourceforge.net/qual.shtml) (8). Sequences were then aligned to chromosomes in the zebrafish Zv8/danRer6 assembly using bwa0.5.7 (9). Unassembled "scaffold" and "NA" fragments were not included in the alignment target. Alignments with quality scores of 10 or greater were extracted from the resulting Sequence Alignment/Map files and counts were obtained for the number of reads aligning to consecutive, adjacent 100-kb physical windows along the genome. Per-window counts per sample were rescaled to a total of 1 million reads, and $\log_2$ ratios were calculated for tumor/normal pairs for all 100-kb windows (or set to 0, if the read count for one or both samples was 0). The $\log_2$ ratios were then submitted to the circular binary segmentation algorithm in the same manner as described for the aCGH data (see above). Sequencing data are available at http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE23666.

Southern Blot Analysis. Probes for Southern blots were ≈1 kb in length; fragments were amplified by PCR and subcloned into pBluescriptIIKS+ (Stratagene) or SP72 (Promega) before labeling. Sequences of the primers used to amplify the probe fragments are as follows:

Chr 13: AATGGCATCATCGTTCAGAC and TCTAGCCCTGCCACTGAATA
Chr 14: GGGAGAGAGGCTGACAAGTT and ACTTCTCTTGTGCGCATGTT
Chr 15: CCTGCATATGCCTACCACAT and CGTTTCCATGTGTGTCACTG

Chr 16: GGGCAAGATCACCACTCATA and ATCACCAACAGCCACATCTC
Chr 25Sc: GGGACAACAGGAACTCATCA and GCCCATTTCCTACAGTTTCC
Chr 25M: CTTCTGCAATGCAAGGAAAC and AGCTTCACGAACCAAGACAA
Chr 25T: TCAGCAAACTGGCTTGATTT and GATTGTGTGATGGCAGGAGT.

Hybridization signals were quantitated on a phosphoimager using Imagequant software (Molecular Dynamics). For each probe, a ratio was determined between the signal in the tumor and the signal in the tail. The average of this ratio for the "neutral" probes (Chr 13, Chr 14, and Chr 16 probes) was then used to normalize the ratio for all of the probes for each tumor/tail pair.

1. Kent WJ (2002) BLAT—The BLAST-like alignment tool. *Genome Res* 12:656–664.
2. Strauss WM (2001) Preparation of genomic DNA from mammalian tissue. *Curr Protoc Mol Biol* 2.2.1–2.2.3.
3. Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* 23:657–663.
4. Diskin SJ, et al. (2006) STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res* 16:1149–1158.
5. Quail MA, et al. (2008) A large genome center's improvements to the Illumina sequencing system. *Nat Methods* 5:1005–1010.
6. Bentley DR, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59.
7. Quail MA, Swerdlow H, Turner DJ (2009) Improved protocols for the Illumina genome analyzer sequencing system. *Curr Protoc Hum Genet* 62:18.2.1–18.2.27.
8. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18:1851–1858.
9. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.

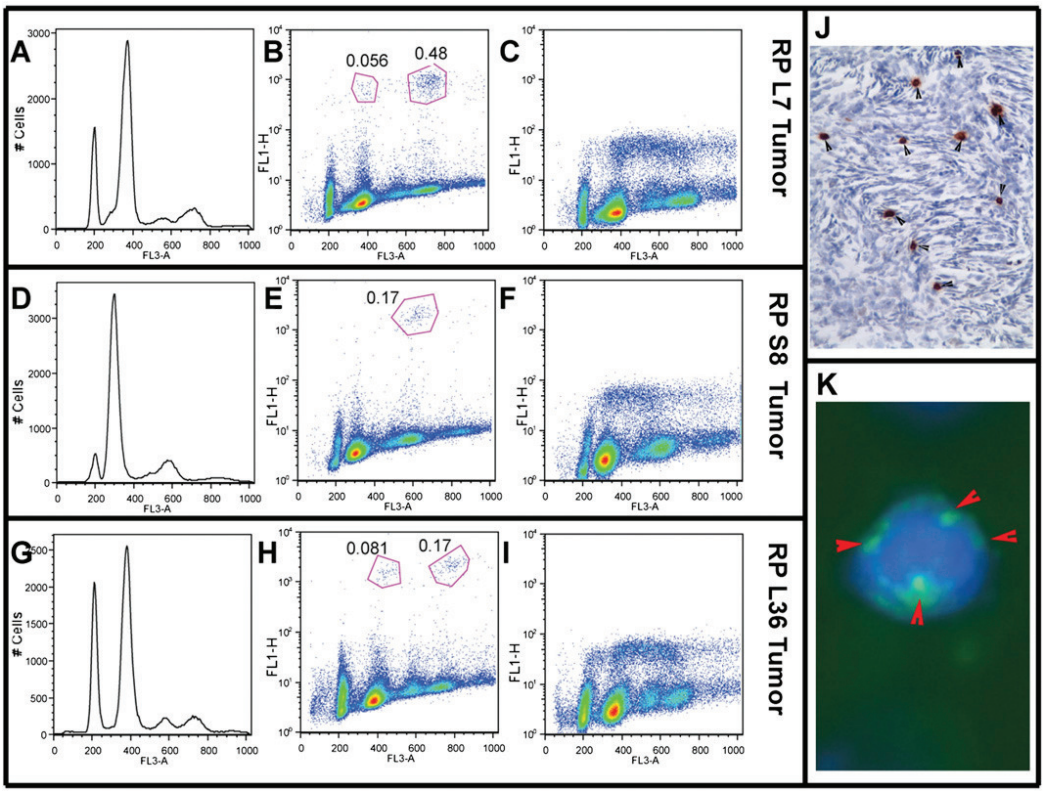


Fig. S1. Cell-cycle properties of zebrafish malignant peripheral nerve sheath tumor (MPNST) cells. Results are shown for cells from three different MPNSTs from different *rp* heterozygotes: *rpL7* (*A*–*C*), *rpS8* (*D*–*F*), and *rpL36* (*G*–*I*). (*A*, *D*, and *G*) DNA content distribution by FACS analysis of propidium iodide (PI)-labeled cells. On the *x* axis, 200 represents the normal 2N position, which was calibrated by running a mixture of normal and tumor cells. (*B*, *E*, and *H*) FACS analysis of cells double-stained with PI (*x* axis) and antibody to pH3 (*y* axis). The mitotic cells are pH3-positive, and are circled in red. The percentages of the cell populations are indicated beside the circles. (*C*, *F*, and *I*) S-phase labeling with a 30-min BrdU pulse. The positions of major horseshoe shapes indicate that the majority of proliferating cells are aneuploid. (*J*) Histological sections from the *rpL35* tumor were stained with antibodies to pH3. Mitotically active tumor cells were stained brown. (*K*) Tumor cells from *rpL35* were double-stained with DAPI and anti-γ-tubulin. The multiple centrosomes are indicated by red arrowheads.

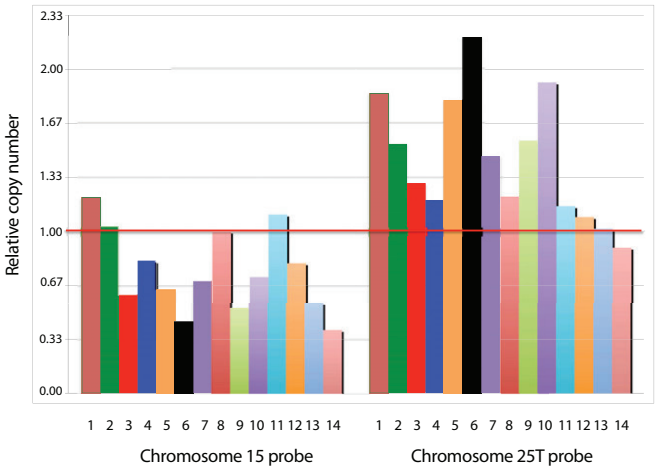


Fig. S2. Southern blot analysis confirms that chromosome 15 is often underrepresented and chromosome 25 is often overrepresented in zebrafish MPNSTs. Southern blots were performed with DNA from tail and tumor tissue from 14 *rp* heterozygotes. Signals for probes to chromosomes 13, 14, 15, 16, and 25 (T; Fig. 3*B*) were quantified, and a tumor/tail ratio was determined for each probe. The values for the chromosome 13, 14, and 16 probes were used to normalize the values for the chromosome 15 and 25 probes.

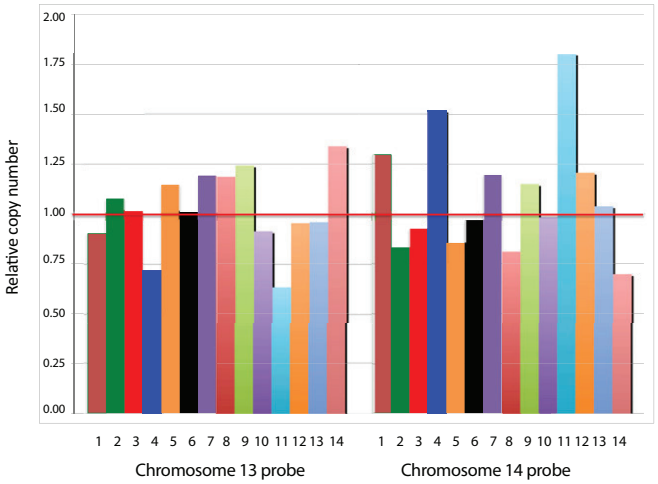


Fig. S3. Southern blot analysis confirms that chromosomes 13 and 14 do not vary in copy number consistently across zebrafish MPNSTs. Southern blots were performed with DNA from tail and tumor tissue from 14 *rp* heterozygotes. Quantitation of the normalized tumor/tail ratios for chromosome 13 and 14 probes was performed from the same Southern blots used in Fig. S2.

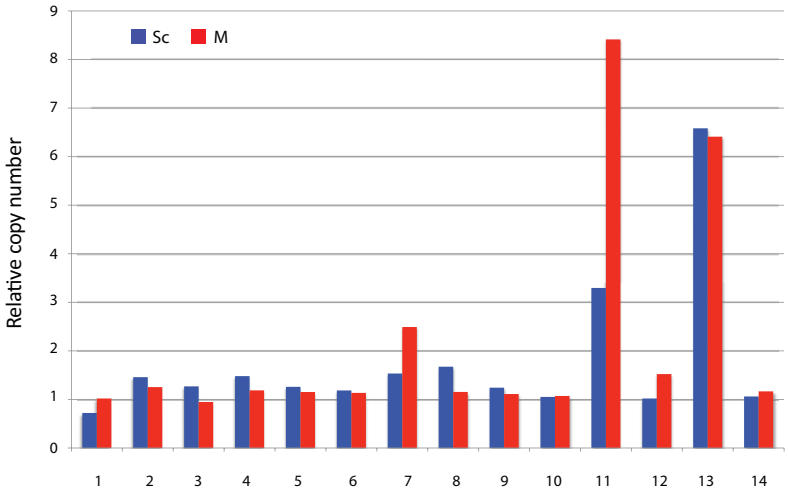


Fig. S4. Southern blot analysis confirms subchromosomal amplifications on chromosome 25 in zebrafish MPNSTs. The same Southern blots described in S2 were additionally probed with sequences from different positions on chromosome 25 (Fig. 3*B*) and analyzed as in Fig. S2. For each tumor, the ratio of the value for probe Sc or probe M to probe T is shown. Seven of 14 tumors show greater amplification for the Sc probe and/or the M probe than the T probe.

Table S1. Coordinates of common focal changes and genes contained therein

Cutoff	0.2	0.2	0.2	0.2	0.2	0.2	0.6	0.6
Change	GAIN	LOSS	GAIN	GAIN	GAIN	GAIN	GAIN	GAIN
CHR	chr9	chr9	chr13	chr13	chr17	chr19	chr25	chr25
segment coordinates (outside/ inside-inside/ outside)*	0/ 90,897 - 942,637/ 1,022,533	26,184,442/ 26,280,956 - 26,784,577/ 26,837,792	36,381,421/ 36,485,738 - 36,947,490/ 37,028,635	38,398,270/ 38,963,490 - 41,969,088/ 42,111,394	28,622,850/ 28,706,826 - 35,513,416/ 35,610,681	38,060,565/ 38,172,964 - 45,022,722/ 45,121,465	13,991,469/ 14,053,597 - 14,445,547/ 14,580,496	18,244,527/ 18,326,711 - 19,316,686/ 19,567,146
genes in segment**	*actr5*	aplp1	*pcnx*	bai3	*eml1*	*bai2*	ccnd2a	arntl2
	bzw1	kirrel3	*med6*	lmbrd1	pkdcc	tgm1	tigara	loh12cr1
	igsf3		*ttc9*	col9a1	disp2	rabggta	fgf6a	cep290
	cldn10		*map3k9*	fam135a	rpusd2	stxbp3	slc45a3	tmtc3
	map3K12		slc8a3	frmd1	pcmtl	sfpq	c12 orf4	kitla
	pcbp2		cox16	tarbp1	ccdc32	zmym4	*dyrk4*	dusp6
	kif5a		actn1	slc35f3	prlhr	illr4	*ndufa9*	poc1b
	stac3		dcaf5	sfrp5	bahd1	c7orf64		tsga14
	map4k4		pacs2	loxl4	ivd	efcab1		cpa1
	crfb5		sfrs5a	pyroxd2	itpka	grn1		cpa5
	lims1		synj2bp	nkx2.3	ltk	grn2		tes
	ranbp2		gmfb	got1	rpap1	cdk6		cav2
	slc5a7		cnih	slc25a28	adprhl2	ccdc132		cav1
	fam136a		cdkn3	hpse2	cinp	pdk4		met
	pcyox1		sos2	hps1	vsx2	asb4		
	hint2		*l2hgdh*	ankrd2	lin52	ppp1r9a		
	ppp2cb		*atp5s*	morn4	pomt2	sgce		
	wrn		*cdkl1*	nkx1.2la	c14orf126	casd1		
	nrg1		*map4k5*	prkg1a	nubpl	col1a2		
	fut10			*pcdh15*	arhgap5	bet1		
					fam175b	gngt1		
					metll10	tfpi2		
					rab40b	calcr		
					cox11	slc25a13		
					stxbp4	shfm1		
					hlf	eif2c2		
					ctsl.1	gatad2b		
					vgll2b	tinagl1		
					ythdf2	trio		
					adam12	kalrn		
					dhx32	nfyc		
					klf11b	jtb		
					grhl1	anp32e		
					ywhaqb	plekho1		
					fam176a	vps45		
					ctsba	psmb4		
					kcnk2	sytl1		
					cenpf	fkbp9		
					smyd2a	crtap		
					prox1	fbxl2		
					snx9	clasp2		
					pomca	ubp1		
					erf3b	pum1		
					dnajc27	nkain1		
					adcy3	snrnp40		
					xdh	zcchc17		
					acbd3	fabp3		
					ccdc28a	dclk3		
					snap23	cyhr1		
					rin3	pdcd6ip		
					tip39	eef1a1		
					tmem121	sdc3		
					dnal1	matn1		
					fut8	sesn2		
					gphn	yrk		
					mpp5a	wasf2		
					eif2s1	cap1		
					id2a	ppt1		
					kidins220	rpl11		
					mboat2	tceb3		
					asapa2	trps1		
					itgb1bp1	csmd3		
					iah1	rad21		
					adam17a	mterfd1		
					rnf144a	uqcrb		
					rsad2	gatad1		
					sox11a	*ankib1*		
					colec11	*lypa2*		
					htrib	*c1orf128*		
					impg1	*cap2*		
					myo6b			
					mapre3			
					dpysl5			
					dnmt3a			
					dtnb			
					asxl2			
					kif3c			
					slc30a1			
					hadha			
					tmem62			
					zadh1			
					trerf1			
					crip1			
					crip2			
					mta1			
					tmem229b			
					zfp36l2			

* Coordinate boundaries are expressed as a range between the first or last probe showing the gain/loss to the neighboring probe that does not, separated by a slash.

** Genes in the uncertain boundary region - in between the first or last probe showing the gain/loss to the neighboring probe that does not - are shown in blue and italics.
Gene identities were manually curated from the UCSC genome browser, including all zebrafish RefSeq genes as well as all putative genes based upon mRNA evidence and homology to a human gene. In cases where the zebrafish did not have a refseq gene, or the gene had a zgc number for a name, the name of the human homolog was used.

Note: This table was reformatted for reprint.
The original spreadsheet version is available at:
http://www.pnas.org/content/suppl/2010/09/06/1011548107.DCSupplemental/sd01.xlsx

Lebenslauf

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

Publikationsliste

- Cecere G, Hoersch S, Jensen MB, Dixit S, Grishok A. ZFP-1(AF10)/DOT-1 complex and endogenous RNAi oppose H2B ubiquitination to negatively modulate transcription in C. elegans. (Manuscript in preparation)
- Zhang G, Hoersch S, Amsterdam A, Whittaker CA, Beert E, Catchen JM, Farrington S, Postlethwait JH, Legius E, Hopkins N, Lees JA. Comparative oncogenomic analysis of copy number alterations in human and zebrafish malignant peripheral nerve sheath tumors. (Manuscript submitted)
- Porter CJ, Palidwor GA, Sandie R, Price F, Krzyzanowski PM, Muro EM, Hoersch S, Smith M, Campbell PA, Perez-Iratxeta C, Rudnicki MA, Andrade-Navarro MA. Paired SAGE-microarray expression data sets reveal antisense transcripts differentially expressed in embryonic stem cell differentiation. In: Computational Biology of Embryonic Stem Cells. Bentham Scientific Publishers. (In press)
- Naba A, Clauser KR, Hoersch S, Liu H, Carr SA, Hynes RO. Proteomic analysis of extracellular matrices of normal tissues and tumors. Mol Cell Proteomics 2011 Dec.
- Ma L, Tan Z, Teng Y, Hoersch S, Horvitz HR. In vivo regulation of intron retention and exon skipping by U2AF large subunit and SF1/BBP in the nematode Caenorhabditis elegans. RNA 2011 Dec;17(12):2201-2211.
- Mansisidor AR, Cecere G, Hoersch S, Jensen MB, Kawli T, Kennedy LM, Chavez V, Tan M-W, Lieb JD, Grishok A. A Conserved PHD Finger Protein and Endogenous RNAi Modulate Insulin Signaling in C. elegans. PLoS Genetics. PLoS Genetics 2011 Sep;7:e1002299.
- Winslow MM, Dayton TL, Verhaak RGW, Kim-Kiselak C, Snyder EL, Feldser DM, Hubbard DD, DuPage MJ, Whittaker CA, Hoersch S, Yoon S, Crowley D, Bronson RT, Chiang DY, Meyerson M, Jacks T. Suppression of lung adenocarcinoma progression by Nkx2-1. Nature 2011 May;473(7345):101-104.
- Schlegel R, Chen Y, Zhao X, Monahan JE, Kamatkar S, Gannavarapu M, Glatt K, Hoersch S. **United States Patent: 7846737** - Genes, compositions, kits, and methods for identification, assessment, prevention and therapy of cervical cancer. 2010.
- Zhang G, Hoersch S, Amsterdam A, Whittaker CA, Lees JA, Hopkins N. Highly aneuploid zebrafish malignant peripheral nerve sheath tumors have genetic alterations similar to human cancers. Proc Natl Acad Sci U S A 2010 Sep;107(39):16940-16945.
- Endege WO, Ford D, Gannavarapu M, Glatt K, Hoersch S, Kamatkar S, Monahan JE, Schlegel R, Xu YY, Zhao X. **United States Patent: 7799518** - Nucleic acid molecules and proteins for the identification, assessment, prevention, and therapy of ovarian cancer. 2010.
- Lillie J, Gannavarapu M, Glatt K, Hoersch S, Kamatkar S, Mertens, Hattersley M, Monahan JE, Myer V, Wang Y, Xu Y, Zhao X, Meyers RE, Bast, Jr. RC, Hortobagyi GN, Pusztai L. **United States Patent: 7705120** - Compositions, kits, and methods for identification, assessment, prevention, and therapy of breast cancer. 2010.
- Hoersch S, Andrade-Navarro MA. Periostin shows increased evolutionary plasticity in its alternatively spliced region. BMC Evol. Biol 2010;10:30.
- Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, Schinzel AC, Sandy P, Meylan E, Scholl C, Fröhling S, Chan EM, Sos ML, Michel K, Mermel C, Silver SJ, Weir BA,

Reiling JH, Sheng Q, Gupta PB, Wadlow RC, Le H, Hoersch S, Wittner BS, Ramaswamy S, Livingston DM, Sabatini DM, Meyerson M, Thomas RK, Lander ES, Mesirov JP, Root DE, Gilliland DG, Jacks T, Hahn WC. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 2009 Nov;462(7269):108-112.

- Monahan JE, Hoersch S, Anderson DL, Endege WO, Ford D, Glatt K, Gorbatcheva BO, Kamatkar S, Xu YY, Gannavarapu M, Zhao X, Schlegel R, Mertens M, Bast, Jr. RC, Hortobagyi GN, Pusztai L. **United States Patent: 7601505** - Compositions, kits, and methods for identification, assessment, prevention, and therapy of breast cancer. 2009.
- Vasudevan KM, Barbie DA, Davies MA, Rabinovsky R, McNear CJ, Kim JJ, Hennessy BT, Tseng H, Pochanard P, Kim SY, Dunn IF, Schinzel AC, Sandy P, Hoersch S, Sheng Q, Gupta PB, Boehm JS, Reiling JH, Silver S, Lu Y, Stemke-Hale K, Dutta B, Joy C, Sahin AA, Gonzalez-Angulo AM, Lluch A, Rameh LE, Jacks T, Root DE, Lander ES, Mills GB, Hahn WC, Sellers WR, Garraway LA. AKT-independent signaling downstream of oncogenic PIK3CA mutations in human cancer. Cancer Cell 2009 Jul;16(1):21-32.
- Grishok A, Hoersch S, Sharp PA. RNA interference and retinoblastoma-related genes are required for repression of endogenous siRNA targets in Caenorhabditis elegans. Proc. Natl. Acad. Sci. U.S.A 2008 Dec;105(51):20386-20391.
- Schlegel R, Chen Y, Zhao X, Monahan JE, Kamatkar S, Gannavarapu M, Glatt K, Hoersh S. **United States Patent: 7125663** - Genes, compositions, kits and methods for identification, assessment, prevention, and therapy of cervical cancer. 2006.
- De Smet F, Pochet NLMM, De Moor BLR, Van Gorp T, Timmerman D, Vergote IB, Hartmann LC, Damokosh AI, Hoersch S. Independent test set performance in the prediction of early relapse in ovarian cancer with gene expression profiles. Clin. Cancer Res 2005 Nov;11(21):7958-7959; author reply 7959.
- Stec J, Wang J, Coombes K, Ayers M, Hoersch S, Gold DL, Ross JS, Hess KR, Tirrell S, Linette G, Hortobagyi GN, Fraser Symmans W, Pusztai L. Comparison of the predictive accuracy of DNA array-based multigene classifiers across cDNA arrays and Affymetrix GeneChips. J Mol Diagn 2005 Aug;7(3):357-367.
- Hartmann LC, Lu KH, Linette GP, Cliby WA, Kalli KR, Gershenson D, Bast RC, Stec J, Iartchouk N, Smith DI, Ross JS, Hoersch S, Shridhar V, Lillie J, Kaufmann SH, Clark EA, Damokosh AI. Gene expression profiles predict early relapse in ovarian cancer after platinum-paclitaxel chemotherapy. Clin. Cancer Res 2005 Mar;11(6):2149-2155.
- Hoersch S, Leroy C, Brown NP, Andrade MA, Sander C. The GeneQuiz web server: protein functional analysis through the Web. Trends Biochem. Sci 2000 Jan;25(1):33-35.
- Andrade MA, Brown NP, Leroy C, Hoersch S, de Daruvar A, Reich C, Franchini A, Tamames J, Valencia A, Ouzounis C, Sander C. Automated genome sequence analysis and annotation. Bioinformatics 1999 May;15(5):391-412.
- Horn F, Weare J, Beukers MW, Hörsch S, Bairoch A, Chen W, Edvardsen O, Campagne F, Vriend G. GPCRDB: an information system for G protein-coupled receptors. Nucleic Acids Res 1998 Jan;26(1):275-279.
- Diplomarbeit: "Modellierung eines Triazin-spezifischen Antikörpers und der Wechselwirkung mit seinem Hapten". Universität Stuttgart, Oktober 1996
- Studienarbeit: "Entwicklung und Programmierung eines Genetischen Algorithmus' im Rahmen der Auslegung eines Fuzzy-Reglers zur Führung einer biologischen Abwasserbehandlungsanlage". Universität Stuttgart, 1994.

Erklärung

„Ich, Sebastian Hörsch, erkläre, dass ich die vorgelegte Dissertation mit dem Thema: **Modellorganismen und Genomik in der Krebsforschung** selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, ohne die (unzulässige) Hilfe Dritter verfasst und auch in Teilen keine Kopien anderer Arbeiten dargestellt habe."

_____

Sebastian Hörsch (Datum)

Danksagungen / Acknowledgements

Diese Dissertation hat eine langjährige Geschichte, und ich bin dementsprechend vielen Menschen zu Dank verpflichtet, angefangen mit meinem Doktorvater, Professor Jens G. Reich, der mich und mein Ziel, einschließlich all der Richtungsänderungen, über die Jahre unterstützte. Ich danke Barbara Bryant für ihre Bereitschaft, hier in den Vereinigten Staaten meine Beraterin zu sein und für die vielen wertvollen Diskussionen. Miguel A. Andrade-Navarros Enthusiasmus bei gemeinsamen Forschungsprojekten war in ähnlicher Weise unersetzlich.

Den Professoren Nancy Hopkins, Alla Grishok und Phillip A. Sharp danke ich für die Gelegenheit, an ihren Forschungsprojekten teilzuhaben. Ihre ausführlichen und bedachten Kommentare zur Zusammenfassung meiner Dissertation beziehungsweise anderweitige Unterstützung waren ausgesprochen hilfreich.

Meine Jahre am M.I.T., am Koch Institut für Integrative Krebsforschung, waren für diese Dissertation von grundlegender Bedeutung. Dafür gebührt insbesondere Professor Jacqueline A. Lees und meinem Kollegen Charles A. Whittaker Dank.

Edelgard Wolf danke ich für all ihre praktische und administrative Hilfe in der Schlussphase der Dissertation.

Es ist leider so, dass ich in etlichen Fällen vielverspechende Forschungsansätze nicht – oder zumindest noch nicht – erfolgreich zu Ende führen konnte. Meine Dankbarkeit gilt all jenen, die an solchen Projekten beteiligt waren, in ganz besonderem Maße aber Peter Storz, John E. Monahan und Stephen Tirrell.

Für ihre Unterstützung über all die Jahre danke ich schließlich meiner Familie – meinen Eltern Ingrid Sophie und Walter und meiner Tochter Aurelia, vor allem aber meiner Frau Maggie. Ohne ihre ungebrochene Unterstützung wäre dies in der Tat nicht möglich gewesen.

Over the considerable time that this dissertation has been in the making, I've become indebted to many people, starting with my advisor, Professor Jens G. Reich, who has supported me and this endeavor over the years and through multiple changes of focus and direction. I thank Barbara Bryant for acting as my local advisor here in the U.S. and for the many invaluable discussions. Similarly, Miguel A. Andrade-Navarro's enthusiastic contributions in joint research and writing have been irreplacable.

I am grateful to Professors Nancy Hopkins, Alla Grishok , and Phillip A. Sharp for the opportunities to be part of their research efforts. Their detailed and thoughtful comments on the thesis summary and, respectively, help during a critical moment have been invaluable.

My years at the M.I.T. Koch Institute for Integrative Cancer Research have been empowering and enabling in critical fashion, and for this I would like to especially thank Professor Jacqueline A. Lees and my colleague Charles A. Whittaker.

I thank Edelgard Wolf for all her practical and administrative help during the final phase of the dissertation.

I am sad to say that, in several instances, once promising research avenues have ultimately not – or at least not yet – come to fruition. My gratitude goes to all that have been involved in such endeavors, in particular to Peter Storz, John E. Monahan, and Stephen Tirrell.

For bearing with me through all these years, I finally thank my family – my parents Ingrid Sophie and Walter, and my daughter Aurelia, but foremost my wife Maggie. Without her unwavering support, this truly would not have been possible.

Sebastian Hörsch

Cambridge, August 2011