

Evolutionäre und funktionelle Transkriptomanalyse der Protisten

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Christian Wöhle
aus Bochum

Düsseldorf, Mai 2014

aus dem Institut für Molekulare Evolution
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. William Frank Martin
Korreferent: apl. Prof. Dr. Ing. Gerhard Steger

Tag der mündlichen Prüfung: 26.06.2014

Für meine Mutter

Im Laufe dieser Arbeit wurden, mit Zustimmung des Betreuers, folgende Beiträge veröffentlicht oder zur Veröffentlichung eingereicht:

Publikationen aus dieser Arbeit

Woehle C, Dagan T, Martin WF & Gould SB (2011). Red and problematic green phylogenetic signals among thousands of nuclear genes from the photosynthetic and apicomplexa-related *Chromera velia*. *Genome Biol Evol*, 3:1220–1230.

Kusdian G, Woehle C, Martin WF & Gould SB (2013). The actin-based machinery of *Trichomonas vaginalis* mediates flagellate-amoeboid transition and migration across host tissue. *Cell Microbiol*, 15:1707–1721.

Gould SB, Woehle C, Kusdian G, Landan G, Tachezy J, Zimorski V & Martin WF (2013). Deep sequencing of *Trichomonas vaginalis* during the early infection of vaginal epithelial cells and amoeboid transition. *Int J Parasitol*, 43:707–719.

Woehle C, Kusdian G, Radine C, Graur D, Landan G & Gould SB (2013). The excavate parasite *Trichomonas vaginalis* expresses thousands of pseudogenes and long non-coding RNAs independently from neighboring genes. *BMC Genomics*, Eingereicht.

Weitere Publikationen

Müller M, Mentel M, van Hellemond JJ, Henze K, Woehle C, Gould SB, Yu R, van der Giezen M, Tielens AGM & Martin WF (2012). Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiol Mol Biol Rev*, 76:444–495.

Martin WF, Roettger M, Kloesges T, Thiergart T, Woehle C, Gould S & Dagan T (2012). Modern endosymbiotic theory: Getting lateral gene transfer into the equation. *J Endocytobiosis Cell Res*, 23:1–5.

de Vries J, Habicht J, Woehle C, Huang C, Christa G, Wägele H, Nickelsen J, Martin WF & Gould SB (2013). Is *ftsH* the key to plastid longevity in sacoglossan slugs? *Genome Biol Evol*, 5:2540–2548.

Maier U, Zauner S, Woehle C, Bolte K, Hempel F, Allen JF & Martin WF (2013). Massively convergent evolution for ribosomal protein gene content in plastid and mitochondrial genomes. *Genome Biol Evol*, 5:2318–2329.

Christa G, Zimorski V, Woehle C, Tielens AGM, Wägele H, Martin WF & Gould SB (2014). Plastid-bearing sea slugs fix CO₂ in the light but do not require photosynthesis to survive. *Proc Biol Sci*, 281:20132493.

Präsentationen

Woehle C, Dagan T, Martin WF & Gould SB (2011). Chromalveolates from the perspective of an apicomplexan alga. *SMBE Annual Meeting*. Kyoto, Japan. Posterpräsentation.

Woehle C, Landan G, Gould SB & Martin WF (2012). Problematic distribution of plastid gene signals among “Chromalveolates” without plastids. *SMBE Annual Meeting*. Irland, Vereinigtes Königreich. Posterpräsentation.

Woehle C, Kusdian G, Landan G, van Aerle R, van der Giezen M, Martin WF & Gould SB (2013). The RNA world of *Trichomonas vaginalis*. *ICOP Meeting XIV*. Vancouver, Kanada. Posterpräsentation.

Woehle C, Kusdian G, Landan G, Martin WF & Gould SB (2013). The RNA world of *Trichomonas vaginalis*. *Institut für Allgemeine Mikrobiologie, Christian-Albrechts-Universität*. Kiel, Deutschland. Vortrag.

Woehle C, Maier UG, Zauner S, Bolte K, Hempel F, Allen JF & Martin WF (2014). Convergent evolution for ribosomal protein retention in organellar genomes. *SMBE Satellite Meeting*. Kiel, Deutschland. Posterpräsentation.

INHALTSVERZEICHNIS

1	Zusammenfassung	1
2	Summary	3
3	Hintergrund	5
3.1	Phylogenie der Protisten	5
3.1.1	Klassifizierung und Evolution der Alveolaten	5
3.1.2	<i>Trichomonas vaginalis</i> und andere Excavaten	8
3.2	Sequenziermethoden der nächsten Generation	9
3.2.1	Methoden zur Sequenzierung im Vergleich	10
3.2.2	Auswertung von Sequenzierungen	13
3.3	Sequenzierung von Transkriptomen	14
3.3.1	Sequenzierung kurzer RNAs	15
3.3.2	3'-Sequenzierung	16
3.3.3	RNA-Seq	17
3.4	Transkriptome der Eukaryoten	18
3.4.1	RNA-Typen in der Translation	19
3.4.2	Kurze nicht-kodierende RNA	20
3.4.3	Lange nicht-kodierende RNA	21
3.4.4	Pseudogene	23
3.5	Zielsetzung	24
4	Publikationen	25
4.1	Red and problematic green phylogenetic signals among thousands of nuclear genes from the photosynthetic and apicomplexa-related <i>Chromera velia</i>	25
4.2	The actin-based machinery of <i>Trichomonas vaginalis</i> mediates flagellate-amoeboid transition and migration across host tissue ..	37
4.3	Deep sequencing of <i>Trichomonas vaginalis</i> during the early infection of vaginal epithelial cells and amoeboid transition	53

4.4	The excavate parasite <i>Trichomonas vaginalis</i> expresses thousands of pseudogenes and long non-coding RNAs independently from neighboring genes	67
5	Zusammenfassung der Ergebnisse	95
	Literatur	99

1 ZUSAMMENFASSUNG

Jahrzehnte lang war die Sanger-Sequenzierung die Grundlage der Ermittlung von Nukleotidabfolgen, beispielsweise bei Genomen. Erst in den letzten Jahren kamen neue Methoden hinzu, die in einem einzigen Durchlauf eine hundertfach erhöhte Datenmenge generieren konnten. Infolge dessen nahmen Sequenzierungen an Genomen deutlich zu. Ihre grundlegenden Techniken wurden auf Transkriptome übertragen, welche nicht nur die Entschlüsselung der exprimierten Gene, sondern auch die Charakterisierung von Expressionsmustern ermöglichen. Diese Verfahren lassen sich sowohl auf Modellorganismen als auch auf weniger erforschte Arten anwenden. Dadurch eignen sie sich ideal für Analysen von Protisten, die ein generell wenig untersuchtes und diverses Paraphylum darstellen.

Eine Sequenzierung des Transkriptoms der erst kürzlich beschriebenen Alge *Chromera velia*, ohne verfügbares Genom, gab Einblicke in die Struktur und phylogenetische Einordnung der nukleär kodierten Gene. Dabei konnte die Gruppierung dieser Alge mit den parasitär lebenden Apicomplexa bestätigt werden. Zusätzlich ergab sich eine nahe Verwandtschaft zu *Perkinsus marinus*, einem Pathogen der Auster. Die Ursprünge der Gene plastidärer Herkunft konnten im Verhältnis 1:1 auf Grün- und Rotalgen zurückgeführt werden. Da allerdings die Datendichte der Rotalgen im Vergleich zu Grünalgen (und Landpflanzen) unterrepräsentiert war und multiple Ursprünge der plastidären Gene zu komplexen Szenarien führen, bleibt deren Deutung problematisch.

Die Wandlung vom flagellierten zum amöboiden Zustand des menschlichen Parasiten *Trichomonas vaginalis* mit Fokus auf Proteine des Zytoskeletts wurde im Rahmen einer weiteren Studie betrachtet. Dabei konnte gezeigt werden, dass das Fimbrin-Protein TvFIM1 während der Morphogenese an der Bündelung der Aktinfilamente beteiligt ist. Bei Kontakt mit Vaginalepithelzellen ordnen sich diese, vorher ungeordnet verteilten, Proteine zusammen in der Zellperipherie an. Diese und andere Proteine des Zytoskeletts zeigen eine weite Verbreitung unter den Gruppen der Eukaryoten, wodurch ein gemeinsamer Ursprung und eine funktionelle Konservierung anzunehmen ist.

Eine Untersuchung der transkriptionellen Veränderungen an *T. vaginalis*, unter Nutzung der verfügbaren Genomsequenz, gab neue Einblicke in die Infektionsbio-

logie dieses Parasiten. Eine substantielle Transkription konnte für mehr als 24.000 Gene dokumentiert werden, dabei wurden für einige Gene Expressionsveränderungen um mehr als das Hundertfache bei Kontakt mit Wirtszellen und unter Einfluss von Sauerstoff ermittelt. Ein Vergleich der Expressionsmuster ergab, dass die Reaktion auf Sauerstoff über jene auf Wirtszellen dominiert. Dies verdeutlicht die Notwendigkeit für *T. vaginalis* auf rapide Veränderungen der Umgebung reagieren zu können, wie sie auch Bestandteil seiner Infektionsbiologie sind. Darüber hinaus zeigten Gene von höherer Sequenzähnlichkeit Tendenzen einer vergleichbaren Expression.

Analysen zur allgemeinen Expression von Genomen in verschiedenen Eukaryoten fanden weit mehr Bereiche exprimiert, als bekannte Modelle zur Genvorhersage vermuten lassen. In *T. vaginalis* konnten tausende Transkripte beschrieben werden, die intergenischen Bereichen zugeordnet werden konnten. Die Hälfte von ihnen wies Homologien zu Protein-kodierenden Genen auf und war daher möglicherweise auf exprimierte Pseudogene zurückzuführen. Weitere Analysen an diesen Transkripten fanden – im Gegensatz zu solchen in Hefe – eine Expression unabhängig von genomisch nah gelegenen Genen, die durch eigene Promotorsequenzen induziert wird. Die große Menge exprimierter Pseudogene und umfangreiche Genfamilien deuten darauf hin, dass sich das zugrunde liegende Genom in einem stetigen Wandel befindet.

Alle diese Auswertungen konnten moderne Sequenziermethoden nutzen, um durch verschiedenen Analyseverfahren neue Erkenntnisse zu sammeln. Mit oder ohne Referenzgenom zeigen sich diese Technologien dazu geeignet auch umfassende Studien an weniger erforschten Organismen durchzuführen. Dabei sind besonders Sequenzierungen am Transkriptom dazu in der Lage, Informationen über die bloße Basenabfolge der Gene hinaus zu ermitteln. Diese Techniken sind innerhalb kürzester Zeit zu zuverlässigen, zeitsparenden Methoden geworden und ermöglichen nun ganz neue Sichtweisen in der Forschung.

2 SUMMARY

For decades, the sequencing of DNA was based on the Sanger-sequencing method. In the recent years new methods have emerged that bring along the ability to generate a hundred times the amount of data in a single run. Next generation sequencing methods (NGS) have dramatically accelerated genome sequencing in general. NGS methods have also been adapted to analyse transcriptomes of cells, which not only reveal the gene content expressed, but additionally enable the characterisation of gene expression patterns. These methods can be applied to model organisms, as well as less-explored species, which make them a perfect tool to analyse a diverse range of protists, which represent a generally little studied and diverse paraphylum.

Transcriptome sequencing of the recently described alga *Chromera velia*, for which no genome sequence is available, gave an insight into the structure and phylogenetic heritage of nuclear encoded genes. The grouping of this alga with apicomplexan parasites was established and additionally a close relationship of this cryptic species with *Perkinsus marinus*, an oyster pathogen, was found. The evolutionary relationship of nuclear genes of complex plastid origin could be traced back to green and red algae in a 1:1 ratio. However, since available sequence data for red algae is underrepresented in comparison to that of green algae (and land plants), and multiple origins of plastid genes underpin complex scenarios, their interpretation remains problematic.

The morphogenesis of the human parasite *Trichomonas vaginalis* from a flagellated to an amoeboid form, focusing on cytoskeleton proteins, was examined in a further study. It was demonstrated that the fimbrin protein *TvFIM1* is involved in the bundling of actin filaments during morphogenesis. Contact with vaginal epithelial cells induces a clustering of actin and actin-bundles at the cell's periphery, that are otherwise randomly dispersed across the cytosol. These and other cytoskeletal proteins have a wide distribution among eukaryotic supergroups, which indicates for an ancient origin and functional conservation.

A study on transcriptional changes of *T. vaginalis*, whose genome sequence is available, revealed new insights into the infection biology of this parasite. The transcription of more than 24,000 genes was documented, with some of them

showing expression changes of more than a hundred fold during attachment to host cells and under the influence of oxygen. Results were compared and it was found that the response to oxygen was severe, and required by the parasite to survive rapid environmental changes that are part of its infection biology. In addition, genes of higher sequence similarity showed trends of a more comparable expression.

Expression analyses of various eukaryotic genomes have revealed that far more regions are expressed than gene model predictions would suggest. In *T. vaginalis*, thousands of transcripts were identified that mapped to intergenic regions. About half of them had homology to protein-coding genes, suggesting they represent expressed pseudogenes. Characterising these transcripts in more detail revealed they are – in contrast to those found in yeast – expressed independent from neighbouring genes and through their own promoters. The large amount of expressed pseudogenes, and extensive gene families, suggest that the genome is in a steady state of changing.

All presented studies were able to provide new findings through various analyses, using modern sequencing methods. Independent from having a sequenced genome to work with, these technologies possess the capabilities for the sophisticated analysis of less-studied organisms. In particular transcriptome sequencing is able to provide information that reaches far beyond the nucleotide sequence of the genes themselves. These tools have quickly become a reliable and quick method that have changed the way of doing research.

3 HINTERGRUND

3.1 Phylogenie der Protisten

Der Begriff Protist hat keine allgemein anerkannte Definition. Verbreitete Beschreibungen erklären, dass Protisten ein Paraphylum darstellen, bestehend aus Vertretern mehrerer eukaryotischer Gruppen (O'Malley *et al.*, 2012). Zur genaueren Klassifizierung werden sie entweder a) von Tieren, (Land-)Pflanzen und (echten) Pilzen separiert oder b) anhand fehlender Mehrzelligkeit unterschieden. Beide Definitionen sind problematisch und können unterschiedlich ausgelegt werden. Heute wird der Begriff Protist zwar noch benutzt um die vielen Gruppen der eukaryotischen Einzeller zusammenzufassen, allerdings werden sie in aktuellen Studien zusammen mit den höheren Eukaryoten klassifiziert und anhand ihrer tatsächlichen phylogenetischen Verwandtschaft gruppiert (Adl *et al.* 2005, 2012; siehe Abbildung 3.1 auf der nächsten Seite). Nach Adl *et al.* (2012) können Eukaryoten in die drei Übergruppen Amorphea, Excavata und Diaphoretickes klassifiziert werden. Das Taxon Amorphea lässt sich wiederum aufspalten in Amoebozoa und Opisthokonta, wozu unter anderem auch Menschen, Tiere und Pilze zählen. Des Weiteren besteht das Phylum Diaphoretickes, zusammen mit wenigen anderen Taxa, aus SAR und Archaeplastida. Zu Letzterem zählen auch Landpflanzen. Die phylogenetische Einordnung der verschiedenen einzelligen Eukaryoten ist teilweise noch umstritten und regelmäßig werden neue Arten beschrieben (Glücksman *et al.*, 2011; Moore *et al.*, 2008; Yabuki *et al.*, 2013). Verdeutlicht wird dies durch starke Variationen der Klassifizierung während der letzten Jahre sowie eukaryotischer Taxa die nicht klar zugeordnet werden können (Adl *et al.*, 2005, 2012). Diese Arbeit beschäftigt sich mit Vertretern der Gruppen Excavata und SAR auf die im Folgenden näher eingegangen wird.

3.1.1 Klassifizierung und Evolution der Alveolaten

Die Alveolata bilden zusammen mit den Rhizaria und den Stramenopiles das Taxon SAR (Benannt nach den Anfangsbuchstaben der drei Taxa). Es charakterisiert sich durch Vorhandensein sogenannter Alveolen, einmembraniger sack-

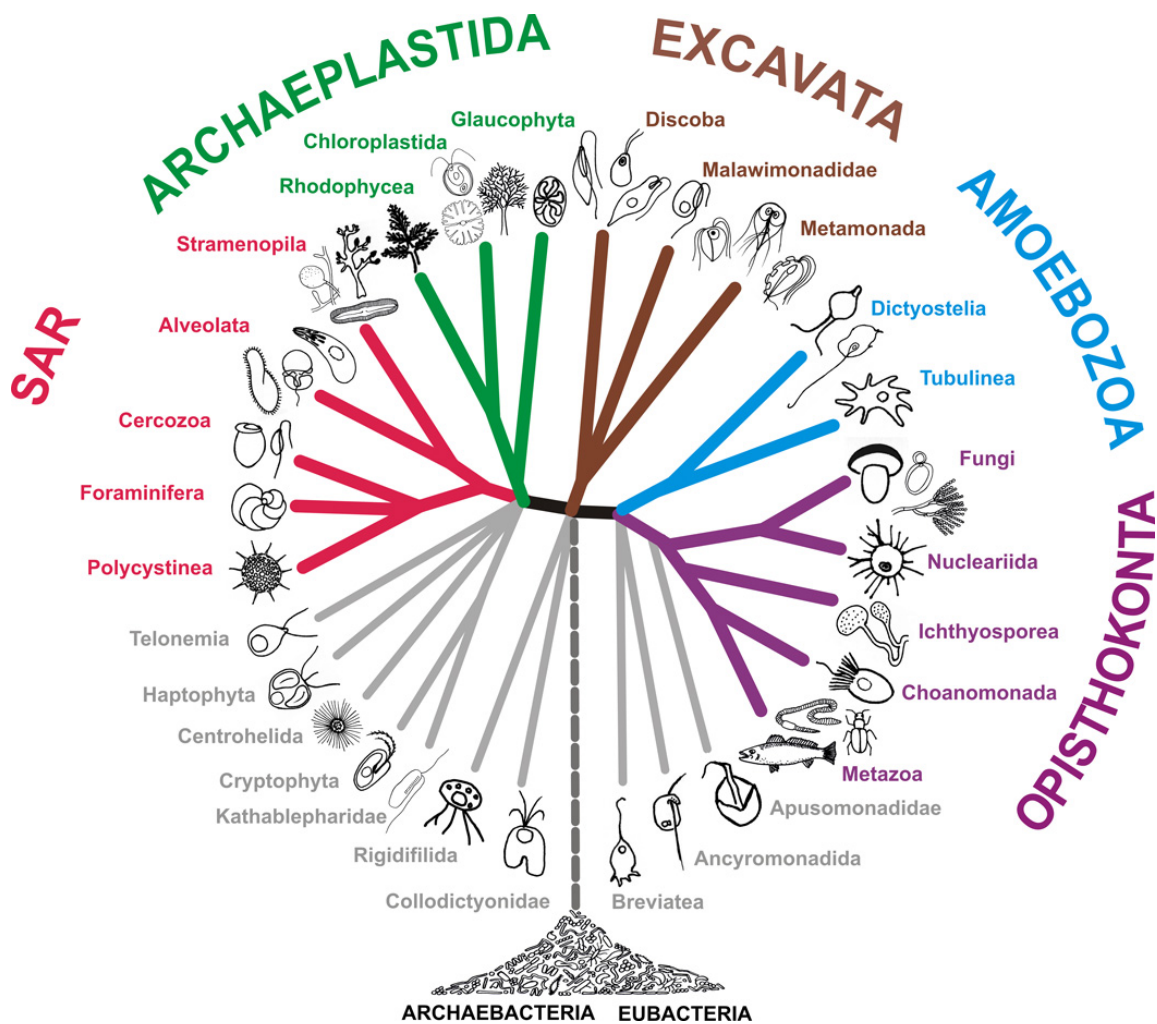


Abbildung 3.1: Klassifizierung eukaryotischer Phyla. Übernommen aus Adl *et al.* (2012).

artiger Kompartimente direkt unter der Plasmamembran. Trotz dieser gemeinsamen Charakteristika war diese Klassifizierung lange Zeit umstritten, da die Alveolaten aus teilweise sehr heterogenen Organismen zusammengesetzt sind (Gould *et al.*, 2008a). Wissenschaftliches Interesse liegt hier vorwiegend beim Phylum Apicomplexa, welches neben dem Malaria-Erreger noch andere bedeutende Pathogene des Menschen und der Nutztiere beinhaltet. Des Weiteren gehören den Alveolata die Dinoflagellata an, deren Vertreter im Gegensatz zu ihren eben genannten Verwandten teilweise phototroph sind und einen wichtigen Teil des marinen Phytoplanktons darstellen, den Ciliophora, die weitestgehend aus Mikropredatoren aufgebaut sind und den Protalveolata, welche die verbliebenden Taxa vereinigen (Adl *et al.*, 2012). Die meisten Alveolaten beherbergen Plastiden,

wenn auch teilweise nicht mehr photosynthetisch aktiv, wie zum Beispiel die Apicoplasten der Apicomplexa (McFadden, 2011). Nur in Ciliaten (Ciliophora) konnte eine plastidäre Endosymbiose nicht eindeutig nachgewiesen werden (Eisen *et al.*, 2006; Reyes-Prieto *et al.*, 2008). Die Plastiden der Alveolaten gehen, im Gegensatz zu denen der Pflanzen, aus einer oder mehreren sekundären Endosymbiosen von Rotalgen (Rhodophyceae) hervor (Archibald, 2009; Gould *et al.*, 2008b). Die Ausnahme bilden hier viele Dinoflagellaten bei denen Endosymbiosen höherer Ordnung, einschließlich Grünalgen (Chloroplastida) und andere photosynthetische Protisten, vorgekommen sind (Schnepf & Elbrächter, 1999).

Eine besondere Stellung in der Erforschung der Ursprünge der Plastiden in den Alveolaten nimmt eine kürzlich entdeckte Untergruppe der Protalveolata ein: Chromerida. Bei den zwei bekannten Vertretern dieses Taxons, *Chromera velia* und *Vitrella brassicaformis*, handelt es sich um marine Algen und wahrscheinlich die nächsten photoautotrophen Verwandten der Apicomplexa (Janouškovec *et al.*, 2010; Moore *et al.*, 2008; Oborník *et al.*, 2012). Dabei sind ihre Plastiden nicht reduziert und weisen weder die mehrfachen unabhängigen Endosymbiosen noch die problematische genomische Organisation (Howe *et al.*, 2008) der Dinoflagellaten auf. Andere Alveolaten sind wenig erforscht oder weisen, wenn überhaupt vorhanden, nur reduzierte Plastiden auf. So stellen die Chromerida unter den Alveolaten die beste Grundlage für phylogenetische Studien an diesen Organellen und deren Ursprüngen dar. Unter anderem konnte so auch bestätigt werden, dass die Plastiden der Apicomplexa auf Rotalgen zurückgehen (Janouškovec *et al.*, 2010). Allerdings ist der Ablauf der Endosymbiose noch umstritten. Nach der Chromalveolaten-Hypothese von Cavalier-Smith (1999) gehen die Plastiden von Alveolata, Stramenopiles und weniger anderer Taxa auf ein gemeinsames Ereignis zurück, bei dem ein Eukaryot eine Rotalge aufgenommen hat. Diese Hypothese ist in den vergangenen Jahren immer mehr in Kritik geraten (Baurain *et al.*, 2010; Parfrey *et al.*, 2010; Stiller *et al.*, 2009), weshalb das Taxon Chromalveolata auch aus der neuesten Version der eukaryotischen Phylogenie entfernt wurde (Adl *et al.*, 2012). Einige Arbeitsgruppen vermuten in den Chromalveolaten Überreste „grüner“ Plastiden, die den heutigen Plastiden „roten“ Ursprungs vorangingen und durch diese ersetzt wurden. Dabei stützen sie ihre Vermutung auf phylogenetische Analysen plastidärer Gene, bei denen teilweise eine direkte phylogenetische Nachbarschaft zu „grünen“ Pflanzen gefunden wurde (Moustafa *et al.*, 2009). Allerdings war bis vor kurzem nur das reduzierte nukleäre Genom einer einzigen Rotalge bekannt (Matsuzaki *et al.*, 2004) und es konnte gezeigt werden, dass

die Unterrepräsentation dieses Taxons die phylogenetischen Analysen beeinflusst (Burki *et al.*, 2012; Dagan & Martin, 2009; Woehle *et al.*, 2011).

3.1.2 *Trichomonas vaginalis* und andere Excavaten

Das Taxon Excavata charakterisierte sich ursprünglich durch Morphologie, wie die Gestalt des Zellmundes und das Aussehens des damit verbundenen Zytoskeletts, dabei gibt es durchaus Fälle sekundärer Reduzierung (Cavalier-Smith, 2002; Simpson, 2003). Später konnte diese Gruppierungen auch anhand phylogenetischer Analysen unterstützt werden (Hampl *et al.*, 2005, 2009; Moreira *et al.*, 2007). Zu den bekanntesten und am besten untersuchten Vertretern der Excavaten gehören Parasiten, wie zum Beispiel die Erreger der Schlaf- und Chagaskrankheit aus der Gattung *Trypanosoma* (Berriman *et al.*, 2005; El-Sayed *et al.*, 2005). Erwähnenswert wäre auch die Alge *Euglena*, die einen gut untersuchten Modellorganismus darstellt und zur Gruppe photoautotropher Excavaten gehört (Adl *et al.*, 2012; Hallick *et al.*, 1993). Ein paar Excavaten schienen Mitochondrien zu fehlen, während ihr Genom meist eine Reduzierung aufwies. Unter anderem führte dies zu der Hypothese, dass diese Taxa die ursprünglichsten der Eukaryoten darstellen könnten. Damit hätten sich diese amitochondriellen Organismen von den restlichen Eukaryoten getrennt noch vor der Endosymbiose eines Prokaryoten, welcher sich zum heutigen Mitochondrium entwickelte (Cavalier-Smith, 1987; Hashimoto *et al.*, 1994; Sogin, 1989, 1991; Sogin *et al.*, 1989; Yamamoto *et al.*, 1997). Allerdings stellte sich später heraus, dass wahrscheinlich alle bekannten Eukaryoten ursprünglich Mitochondrien besaßen und ihre Reduzierung sekundär geschah (Embley & Martin, 2006; McInerney *et al.*, 2014; Roger *et al.*, 1998; Shiflett & Johnson, 2010; Simpson *et al.*, 2006). Neue Einsichten ergab die erste Genomsequenzierung eines freilebenden Vetreters der Excavaten: *Naegleria gruberi* (Fritz-Laylin *et al.*, 2010). Im Gegensatz zur vorherigen Annahme, nach denen die ältesten gemeinsamen Vorfahren der Eukaryoten reduzierten und amitochondriellen Protisten ähneln sollten, kam die Hypothese auf, dass gerade durch seine Komplexität *N. gruberi* dem eukaryotischen Ursprung nahe kommen dürfte (Fritz-Laylin *et al.*, 2010; Koonin, 2010a,b; Wolf & Koonin, 2013).

Ein Vertreter der Excavata, welcher früher als dem gemeinsamen Vorfahr aller Eukaryoten ähnlich angesehen wurde, ist *Trichomonas vaginalis*. Bei diesem Organismus handelt es sich um den Erreger der Trichomoniasis, der wahrscheinlich verbreitetsten Geschlechtskrankheit weltweit (Carlton *et al.*, 2007; Petrin *et al.*, 1998; Schwebke & Burgess, 2004). Der Parasit nistet sich extrazellulär im Urogenital-

trakt von Männern und Frauen ein. Sichtbare Symptome gehen von kompletter Symptomfreiheit bis zu schweren Entzündungen und Irritationen. Des Weiteren wird ein höheres Risiko für Komplikationen während der Schwangerschaft (Cotch *et al.*, 1997) sowie für eine Infizierung mit dem humanen Immundefizienz-Virus (HIV; Sorvillo *et al.* 1998, 2001) mit dieser Krankheit in Verbindung gebracht. *T. vaginalis* unterscheidet sich von den meisten Protisten durch das Fehlen klassischer Mitochondrien. Stattdessen befinden sich hier die stark abgewandelten Hydrogenosomen, die allerdings ihren Ursprung mit den Mitochondrien teilen (Embley *et al.*, 2003). Mit zirka 60.000 annotierten Genen im Genom weist *T. vaginalis* die größte Menge kodierter Gene in einem sequenzierten Genom auf (Carlton *et al.*, 2007). Eine spätere Studie konnte diese Zahl, durch Zusammenfassen von Fragmenten identischer Gensequenzen, auf immerhin noch ~46.000 verringern (Smith & Johnson, 2011). Damit wäre die Menge der kodierten Gene immer noch zu den umfangreichsten zu zählen, besonders für einen Parasiten, da bei solchen meist eher reduzierte Genome beobachtet werden (Moran, 2002; Wolf & Koonin, 2013). Es ist allerdings auch noch nicht geklärt, wie viele der vorhergesagten Gene tatsächlich funktionell sind. Gould *et al.* (2013) konnten die Transkription von maximal der Hälfte der annotierten Gene nachweisen. In der Datenbank TrichDB (Aurrecochea *et al.*, 2009) lassen sich die Proteinsequenzen von *T. vaginalis* in ~10.000 orthologe Genfamilien einteilen (Chen *et al.*, 2006), dabei können diese Gruppen aus mehreren tausend Genen bestehen, die sich in der Nukleotidsequenz teilweise kaum unterscheiden. Ursache oder Funktion des großen Kodierungspotentials von *T. vaginalis* ist unklar. In Carlton *et al.* (2007) wurde hypothetisiert, dass diese massive Expansion des Genoms mit der Anpassung von Verdauungs- auf Urogenitaltrakt einherging und der mit der Expansion verbundene Größenzuwachs Vorteile bei der Phagozytose gebracht haben könnte. In einer anderen Arbeit wurde die massive Genomgröße teilweise auf große instabile Genfamilien zurückgeführt, die sich ständig wandeln (Cui *et al.*, 2010).

3.2 Sequenziermethoden der nächsten Generation

Die Sequenzierung von Nukleinsäuren ist heute ein grundlegender Bestandteil der Biologie. Erste Methoden dazu wurden 1977 veröffentlicht. Dabei handelte es sich um die Didesoxy-Methode nach Sanger (Sanger *et al.*, 1977) und die chemische Degradation an spezifischen Basen nach Maxam und Gilbert (Maxam & Gilbert, 1977). Die Sanger-Methode setzte sich durch und wird bis heute genutzt. Bei

dieser Technik werden bei der Amplifikation einer Nukleotidsequenz, in vier unterschiedlichen Ansätzen, an zufälligen Positionen verschiedene Didesoxynukleosidtriphosphate (ddNTPs) eingebaut, die die weitere Synthese zum Stoppen bringen. Es entstehen Fragmente unterschiedlicher Länge, die sich mittels Gelelektrophorese auftrennen lassen. Anhand der Lauflänge kann dann die Reihenfolge der Nukleotide nachvollzogen werden. Später wurde diese Technik kommerzialisiert, automatisiert und immer weiter verfeinert, wodurch diese Methodik für ein breites Spektrum der Forschung zugänglich gemacht wurde (Ansorge *et al.*, 1986, 1987; Pareek *et al.*, 2011)

Eine besondere Herausforderung für die Sanger-Technik stellte die Sequenzierung des menschlichen Genoms dar. Mehr als zehn Jahre vergingen und fast drei Milliarden Dollar wurden verbraucht bis das Humangenomprojekt (HGP) vollendet war (McGinn & Gut, 2013; Pareek *et al.*, 2011). Schließlich wurde als erste Sequenziermethode der nächsten Generation (engl. *next generation sequencing*; NGS) 2005 die Pyrosequenzierung von 454 lifescience¹ eingeführt (Margulies *et al.*, 2005; Pareek *et al.*, 2011). Bei dieser Methode wird direkt während der Synthese eines Gegenstranges der Einbau eines neuen Nukleotids über ein Lichtsignal detektiert (Ronaghi *et al.*, 1996). Zusätzlich sorgten massiv automatisierte und parallelisierte Arbeitsabläufe schon damals für eine hundertfach erhöhte Datenmenge im Vergleich zur Sanger-Methode (Margulies *et al.*, 2005). In den folgenden Jahren kamen zahlreiche weitere Sequenziermethoden hinzu und die Datenmenge erhöhte sich weiter, während die Kosten zurückgingen. Mit heutigen verfügbaren Mitteln lässt sich ein menschliches Genom binnen zehn Tagen für weniger als \$10,000 sequenzieren (McGinn & Gut, 2013). Bei andauerndem technischen Fortschritt wird damit gerechnet, dass dieser Betrag in den nächsten Jahren noch unter \$1000 fällt.

3.2.1 Methoden zur Sequenzierung im Vergleich

Heute gibt es zahlreiche verschiedene Sequenzierverfahren zu unterscheiden. McGinn & Gut (2013) klassifizieren vier Generationen der Nukleotidsequenzierung. Die klassische Methodik nach Sanger sowie Maxam & Gilbert stellen hier die erste Generation dar. Die zweite Generation zeichnet sich aus durch den parallelen Einsatz vieler Kopien der Ausgangssequenzen gefolgt von einer enzymatisch katalysierten Replikation. Dabei erfolgt der Einbau komplementärer

¹<http://www.454.com>

Nukleotide schrittweise in sich wiederholenden Zyklen. In jedem Durchlauf läßt sich der Einbau verschiedener Nukleotide anhand spezifischer Signale detektieren. Schließlich wird eine Sequenz durch die Verbindung der Signale der einzelnen Durchläufe ermittelt. Die meisten der heute genutzten Sequenziermethoden werden zu dieser Generation gezählt. Dazu gehört das bereits erwähnte 454, Illumina² (Bentley *et al.*, 2008), SOLiD³ (McKernan *et al.*, 2009) und, als eher neuere Technologie, IonTorrent³ (Rothberg *et al.*, 2011). Zur Kategorie der dritten Generation gehören Sequenziermethoden, die individuelle Moleküle ohne Einsatz enzymatischer Reaktionen entschlüsseln können. Technologien mit Nanoporen sind bisher das einzig erwähnenswerte passende System, welches sich allerdings noch in der Entwicklung befindet. Bei dieser Technik werden winzige Veränderungen in den elektrischen Strömen an einer Membran beim Durchqueren von DNA-Molekülen durch Poren detektiert (Branton *et al.*, 2008; Kasianowicz *et al.*, 1996). Als eine Methode die einzelne Ausgangssequenzen benutzt, aber noch auf enzymatische Reaktionen während der Sequenzierung angewiesen ist, kann PacBio⁴ (Eid *et al.*, 2009) zwischen der zweiten und dritten Generation eingeordnet werden. Die vierte Generation bezeichnet eine Sequenzierung direkt im lebenden Organismus. Diese Methode ist kaum erprobt und bisher wurde lediglich das Prinzip getestet (Larsson *et al.*, 2010). Allerdings wird dadurch verdeutlicht, welche Formen die Sequenzierung von Nukleotiden in Zukunft noch annehmen kann.

Aktuelle Sequenziermethoden befinden sich im ständigen Wandel. Jedes Jahr werden verbesserte und neue Varianten vorgestellt und die Technik noch weiter verfeinert. Das macht es schwierig die Leistungsfähigkeit der verschiedenen Sequenzierer zu vergleichen. Allerdings haben verschiedenen Methoden ihre Stärken und Schwächen, wodurch sich ihre Anwendungsschwerpunkte unterscheiden. Der Sanger-Sequenzierer 3730xl beispielsweise ist fähig Nukleotidfolgen von hoher Länge (900 Basenpaare) und Genauigkeit zu analysieren, allerdings ist die Sequenz- ausbeute pro Durchlauf im Kilobasenbereich gering, verglichen zu NGS-Systemem (Liu *et al.*, 2012). Er findet heute noch Anwendung bei der Sequenzierung von Einzelsequenzen bei denen es sich nicht lohnt eine umfangreiche Sequenzierung durchzuführen und bei der eine hohe Genauigkeit benötigt wird. Zum Vergleich ist die Ausgabe des 454 GS FLX+ mit 700 Millionen Nukleotiden bei ähnlich geringer Fehlerquote und einer Länge von bis zu 1 Kilobase um das hundert-

²<http://www.illumina.com>

³<http://www.lifetechnologies.com>

⁴<http://www.pacificbiosciences.com>

fache größer (Dark, 2013; Loman *et al.*, 2012). Diese Sequenziermaschine eignet sich ausgezeichnet zur Erstsequenzierung von Genomen oder Transkriptomen (*de novo*).

Zur quantitativen Analyse der Sequenzen, wie sie bei der Charakterisierung von Transkriptomen angewandt werden, gibt es besser geeignete Methoden. Hier werden eher größere Datenmengen benötigt um möglichst viele Transkripte ausreichend abdecken zu können, dagegen sind geringe Fehlerraten und lange Lese-längen weniger ausschlaggebend, wenn sie ausreichen um die Ausgabe eindeutig auf einer Referenzsequenz zuordnen zu können. Der Illumina HiSeq 2500 zeigt etwas geringere Genauigkeit und deutlich kürzere Sequenzierlängen von hundert Nukleotiden, dagegen produziert er bei jedem Durchlauf mehrere hundert Giga-basen an Daten. Der SOLiD 5500xl W erzeugt pro Durchlauf etwas weniger Daten bei noch kürzeren Sequenzen, aber besticht durch eine sehr hohe Genauigkeit. Dadurch ist er besonders gut geeignet zur Detektierung von Einzelnukleotid-Polymorphismen (engl. *single nucleotide polymorphism*; SNP). Die relativ neue IonTorrent-Technologie befindet sich mit Ion 318 chip v2 bei einer Datenprodukti-on von mehreren Gibabasen und Sequenzlänge von bis 400 Basenpaaren zwischen den zuletzt genannten Systemen. Da PacBio auf einen Amplifikationsschritt ver-zichtet, fällt eine mögliche Fehlerquelle weg. Allerdings wird bei der aktuellen Genauigkeit des PacBio RS II mit schätzungsweise 86% eine so hohe Sequen-zierabdeckung benötigt, dass daraus kein Nutzen mehr gezogen werden kann. Dafür ist die Sequenzierlänge hier mit mehreren Kilobasen deutlich höher als bei den bisherigen Methoden. Als erster Sequenzierer der dritten Generation, könnte die Nanoporen-Technologie die Sequenzierung noch einmal revolutionieren. Der geplante Sequenzierer GridOn⁵ soll hohe Datenmengen generieren können, ver-gleichbar mit Illumina, bei Sequenzlängen, die mit den heutigen Sequenzierern nicht zu vergleichen sind. Lediglich eine hohe Fehlerrate von 4% war 2012 noch in den Griff zu bekommen (Loman *et al.*, 2012).

Dass sich manche Systeme für bestimmte Anwendungen besser eignen, be-deutet nicht, dass sie nicht auch durch andere Techniken sinnvoll umgesetzt werden können (Loman *et al.*, 2012). So werden auch Illumina-Systeme durch ihre hohen Datenmengen, kombiniert mit der Möglichkeit der Sequenzierun-gen der DNA-Fragmente von beiden Enden her (engl. *paired-end*), dazu genutzt, *De-novo*-Sequenzierungen durchzuführen (Grabherr *et al.*, 2011; Simpson *et al.*, 2009). Zusätzlich können Kombinationen verschiedener Systeme genutzt wer-

⁵<https://www.nanoporetech.com>

den, um Nachteile auszugleichen und Vorteile mehrere Technologien zu vereinen (Diguistini *et al.*, 2009; Reinhardt *et al.*, 2009; Wang *et al.*, 2012). Beispielsweise konnte so die hohe Fehlerrate bei PacBio-Systemen mit anderen NGS-Anwendungen ausgeglichen werden (Bashir *et al.*, 2012). Allerdings entstehen durch weitere Sequenzierungen auch zusätzliche Kosten.

3.2.2 Auswertung von Sequenzierungen

Eine besondere Herausforderung stellt die Aufbereitung und Auswertung der Rohdaten dar, die anhand von NGS-Systemen produziert wurden. Heutige Sequenzierungen können hunderte von Gigabasen an Daten produzieren (siehe Abschnitt 3.2.1). Allein das Speichern und Archivieren dieser Datenmengen stellt hohe Ansprüche an die informationstechnische Infrastruktur (Loman *et al.*, 2012; Zhu *et al.*, 2013). Des Weiteren unterscheiden sich die Ausgaben der verschiedenen Sequenziertechniken und bei Genomen, denen unterschiedliche Sequenziermethoden zugrunde liegen, kann es problematisch sein sie miteinander zu vergleichen oder zu kombinieren. Rohdaten müssen oft noch aufbereitet werden. Die Ausgabe der Sequenzierer enthält neben Nukleotidfolgen auch passende Qualitätswerte für jede Base (Cock *et al.*, 2010; Patel & Jain, 2011). Dadurch lassen sich problematische Sequenzen und Nukleotide für spätere Analysen entfernen. Bei den geringen Sequenzierlängen der verschiedenen NGS-Systeme können weder Transkripte noch Genome aus der Entschlüsselung kompletter Sequenz hervorgehen, daher werden diese meist in zufällige Fragmente zerlegt (engl. *shotgun sequencing*). Die resultierenden Nukleotidabschnitte können sich an verschiedenen Positionen überlappen und darüber verbunden werden. Bei dieser sogenannten Assemblierung werden die Sequenzen verglichen und zu längeren fortlaufenden Nukleotidabfolgen (engl. *contiguous*; Contig) vereinigt (Dark, 2013; Loman *et al.*, 2012; Mardis, 2008).

Grundsätzlich können zwei Methoden der Auswertung unterschieden werden, je nachdem ob eine Referenzsequenz zur Verfügung steht oder nicht. Eine Referenz kann als Rückgrat dienen, um die Ergebnisse der Sequenzierung einordnen zu können. Dabei werden die resultierende Abschnitte an der Referenz zugeordnet (engl. *mapping*) und können weiter assembliert (Trapnell *et al.*, 2010) oder direkt anhand der Genomannotation ausgewertet werden (Gould *et al.*, 2013; Marioni *et al.*, 2008). Auf diesem Weg lassen sich Unterschiede zum Genom leicht detektieren oder die Expression von Genen charakterisieren, aber auch bisher unbekannte Transkripte können identifiziert werden (Trapnell *et al.*, 2010; Xu *et al.*,

2013). Diese Vorgehensweise ist sinnvoll bei Modellorganismen mit vorliegenden Genomsequenzen oder nah verwandten Arten. Da bei vielen Organismen keine Referenz verfügbar ist, müssen hier Informationen zur Anordnung der kurzen Fragmente aus der Sequenzierung selbst bezogen werden. Bei einer solchen *De-novo*-Assemblierung werden Nukleotidfragmente lediglich untereinander aligniert um Contigs zu bilden. Das Ergebnis ist im Optimalfall ein komplettes Genom oder zahlreiche vollständige Transkripte, je nachdem was sequenziert wurde.

Da die Nukleotide allein noch keine Aussage über funktionelle Elemente der Genome oder Transkriptome geben können, folgt meist noch ein Annotations-schritt. Dieser kann über einfache Homologiesuchen, wie BLAST (engl. *basic local alignment search tool*; Altschul *et al.* 1997) oder speziellen Programmen zur Vorhersage von Genen, wie GLIMMER (Delcher *et al.*, 1999) bei prokaryotischen Genomen, geschehen. Bei Genomsequenzierungen werden darüber hinaus Ergebnisse von Transkriptomsequenzierungen einbezogen (Curtis *et al.*, 2012; Denoeud *et al.*, 2008; Paterson *et al.*, 2012; Price *et al.*, 2012). Schließlich werden die Daten veröffentlicht und damit der wissenschaftlichen Gemeinschaft zur Verfügung gestellt (Kodama *et al.*, 2012; Pagani *et al.*, 2012). Dies können die Rohdaten oder resultierende Assemblierungen sein, die für spätere Analysen als Referenz genutzt werden können.

3.3 Sequenzierung von Transkriptomen

Die vorgestellten Sequenziermethoden lassen sich auf Genome und Transkriptome anwenden. Transkriptome (siehe dazu Abschnitt 3.4) repräsentieren den exprimierten Teil eines Genoms. Ähnlich zu Letzterem können sie genutzt werden um Nukleotidabfolgen von Genen zu untersuchen. Darüber hinaus geben sie quantitative Rückschlüsse auf deren Expression. So ermöglichen Sequenzierungen des Transkriptoms, Expressionsmuster in verschiedenen Geweben und unterschiedlichen physiologischen Bedingungen zu charakterisieren (Costa *et al.*, 2010; Wang *et al.*, 2009). Eine frühe Methode der Transkriptomanalyse war die Sequenzierung von ESTs (engl. *expressed sequence tags*), dabei wurden RNAs vom 5'- oder 3'-Ende her sequenziert (Nagaraj *et al.*, 2007). Die klassische Sequenzierung nach Sanger kam dabei zum Einsatz und führte zu Fragmenten von einigen hundert Nukleotiden Länge. Das Ergebnis waren qualitative Angaben über vorhandene Transkripte und deren Sequenzen. Aufgrund der geringen Datendichte und hohen

Kosten waren Aussagen zur Expression problematisch (Adams, 1996; Huang *et al.*, 2012; Wang *et al.*, 2009). Methoden zur Quantifizierung von Transkriptomen waren die quantitative Echtzeit-PCR (engl. *quantitative real-time polymerase chain reaction*; qPCR), Hybridisierungs-basierende Anwendungen wie *Microarrays* (Mardis, 2008; Mutz *et al.*, 2013; Wang *et al.*, 2009) und die Sequenzierung von enzymatisch abgespaltenen kurzen Fragmenten wie bei SAGE oder CAGE (engl. *serial/cap analysis of gene expression*; Harbers & Carninci 2005; Kodzius *et al.* 2006; Shiraki *et al.* 2003; Velculescu *et al.* 1995). Ersteres ist zu teuer und aufwendig um es für große Datenmengen anwenden zu können, wobei sich Hybridisierungs-basierende Methoden nur auf bekannte Gene beschränken und unter Kreuzhybridisierungen leiden, die mit starkem Hintergrundrauschen einhergehen (Okoniewski & Miller, 2006; Royce *et al.*, 2007; Wang *et al.*, 2009). SAGE und CAGE leiden nicht unter dieser Problematik und konnten auch schon auf neue Sequenziermethoden adaptiert werden, allerdings reichen die kurzen Fragmente von anfangs 14 und später 26 Nukleotiden (Matsumura *et al.*, 2005, 2003; Saha *et al.*, 2002) oft nicht für eine eindeutige Zuordnung aus (Wang *et al.*, 2009). Die Anwendung von Sequenzierern der nächsten Generation bot klare Vorteile gegenüber den bisherigen Methoden und bereitete den Weg für eine umfangreiche Anwendung (Costa *et al.*, 2010). Im Folgenden werden neue Techniken beschrieben, die mit der Entwicklung von NGS-Systemen einhergingen.

3.3.1 Sequenzierung kurzer RNAs

Zu Beginn war es mit den Sequenzierlängen der neuen Sequenzierer kaum möglich komplette Transkripte zu gewinnen (Margulies *et al.*, 2005). Besonders die Illumina-Sequenzierung zeichnete sich schon bei ihrer Markteinführung durch hohen Datendurchsatz aus, während ihre Sequenzlängen (35 Nukleotide) sehr kurz ausfielen (Bentley *et al.*, 2008). Diese waren hervorragend geeignet zur Sequenzierung kurzer nicht-kodierender RNAs (engl. *small RNA*; sRNA), einem Überbegriff funktionaler RNAs, die in ihrer Größe meist bei 20-30 Nukleotiden liegen (Lu *et al.* 2005; McCormick *et al.* 2011; siehe Abschnitt 3.4.2). Durch ihre passende Länge entfällt die Notwendigkeit einer Assemblierung und die Sequenzen können direkt an passende Referenzgenome aligniert werden. Darüber lassen sich die Mengen der unterschiedlichen sRNAs ermitteln. Diese Methodik wurde genutzt um zahlreiche neue Kandidaten funktioneller sRNA in den verschiedensten Organismen zu beschreiben und ihre Expression näher zu untersuchen (Chen *et al.*, 2009; Diermann *et al.*, 2010; Kozomara & Griffiths-Jones, 2014; Ruby *et al.*, 2006).

Eine Separierung der kurzen RNAs vom Rest des Transkriptom erfolgt meist lediglich durch eine Auftrennung nach Länge anhand einer Gelelektrophorese. Dadurch stellt es eine besondere Herausforderung dar, sie von Degradationsprodukten zu unterscheiden und genauer zu bestimmen (McCormick *et al.*, 2011). Zusätzlich spiegelt sich oft der hohe Anteil ribosomaler und Transfer-RNAs (rRNA beziehungsweise tRNA) im Transkriptom auch in den kurzen RNAs wider (Chen *et al.*, 2009; Costa *et al.*, 2010; Diermann *et al.*, 2010; Karpinets *et al.*, 2006; McCormick *et al.*, 2011), auch wenn gezeigt werden konnte, dass auch scheinbare Degradationsprodukte von tRNAs funktionell sein können (Garcia-Silva *et al.*, 2010; Lee *et al.*, 2009; Phizicky & Hopper, 2010).

3.3.2 3'-Sequenzierung

Eine Methode ähnlich dem SAGE, stellt eine Sequenzierung der Transkripte direkt vom 3'-Ende her dar. Sequenziererergebnisse können direkt Genen eines Genoms zugeordnet werden und benötigen lediglich eine Normalisierung zwischen den Proben (zum Vergleich siehe Abschnitt 3.3.3). Anhand des Poly-A-Schwanzes der Eukaryoten lassen sich Protein-kodierende Transkripte unkompliziert isolieren (Derti *et al.*, 2012; Wang *et al.*, 2009). Im Gegensatz zum SAGE kann hier weitgehend auf enzymatisches Schneiden verzichtet werden. Die Länge der Fragmente ist abhängig vom Sequenzierer (siehe Abschnitt 3.2.1). Demnach ergaben Illumina-Sequenzierungen schon bei Markteinführung längere Fragmente als jene durch SAGE und waren somit schon damals besser zuzuordnen (Bentley *et al.*, 2008). Mit heutigen Sequenzierlängen von 100 und mehr Nukleotiden (Dark, 2013) lassen sich die Sequenzen nicht nur besser einordnen, darüber hinaus ist es sogar möglich die 3'-Enden der Gene weiter zu charakterisieren (Jan *et al.*, 2011). Es wird geschätzt, dass mehr als die Hälfte der Gene im menschlichem Genom und mindestens ein Drittel der Mausgene unterschiedlich polyadenylierte mRNAs produzieren (Tian *et al.*, 2005). Verschiedene Protokolle zur Durchführung dieser Technik wurden weiterentwickelt und fanden Anwendung in verschiedenen Analysen (Derti *et al.*, 2012; Fox-Walsh *et al.*, 2011; Gould *et al.*, 2013; Hoque *et al.*, 2013; Shepard *et al.*, 2011; Wilkening *et al.*, 2013; Yoon & Brem, 2010). Da diese Methodik direkt am Poly-A-Schwanz ansetzt, können spezifisch mRNAs gewonnen werden, ohne von der großen Menge anderer RNA-Typen beeinflusst zu werden (Costa *et al.*, 2010; He *et al.*, 2010; Zhao *et al.*, 1999). Allerdings schränkt dies auch die Anwendbarkeit in Bezug auf unterschiedliche RNA-Typen und Prokaryoten stark ein, da ihnen diese 3'-Adenosine meist fehlen. Zudem zeigen aktuelle Sequen-

zierer noch Leselängen, die wenig geeignet sind um komplette Gene abzudecken. Daher ist das Vorhandensein einer Genomsequenz essentiell und Aussagen über den Aufbau der kompletten Transkripte sind nur eingeschränkt möglich.

3.3.3 RNA-Seq

Bei RNA-Seq (RNA-Sequenzierung) oder der sogenannten Gesamt-Transkriptom-*Shotgun*-Sequenzierung (engl. *whole transcriptome shotgun sequencing*; WTSS) handelt es sich um eine Methode die Menge aber auch die Sequenz von Transkripten zu ermitteln (Costa *et al.*, 2010; Morin *et al.*, 2008; Mortazavi *et al.*, 2008; Wang *et al.*, 2009). Diese Methodik nutzt den hohen Datendurchfluß moderner Sequenziermethoden optimal und basiert auf der Anwendung von Verfahren, die ursprünglich zur Sequenzierung von Genomen entwickelt wurden (Costa *et al.*, 2010; Marguerat & Bähler, 2010; Wang *et al.*, 2009). Dabei werden RNA-Sequenzen, vor oder nach dem Umschreiben in cDNA, in kleinere, zufällige Fragmente degradiert. Diese werden sequenziert und anschließend assembliert. Im Optimalfall ergeben sich komplette Transkripte zusammen mit der Menge zugehöriger Rohdaten zur Ermittlung der Expression (Martin & Wang, 2011). Unter Verwendung verschiedener Techniken der Sequenzierung (Cloonan *et al.*, 2008; Jenjaroenpun *et al.*, 2013; Mortazavi *et al.*, 2008; Sugarbaker *et al.*, 2008) kam RNA-Seq schon bei den verschiedensten Organismen zum Einsatz. Dazu zählen Mensch (Jenjaroenpun *et al.*, 2013; Marioni *et al.*, 2008; Morin *et al.*, 2008; Sugarbaker *et al.*, 2008), Maus (Cloonan *et al.*, 2008; Mortazavi *et al.*, 2008), Hefen (Nagalakshmi *et al.*, 2008; Wilhelm *et al.*, 2008), Pflanzen (Denoeud *et al.*, 2008; Lister *et al.*, 2008) und Prokaryoten (Passalacqua *et al.*, 2009; Perkins *et al.*, 2009). Auch Organismen ohne sequenzierte Genome konnten bereits auf diese Weise charakterisiert werden (Lowe *et al.*, 2011; Pallavicini *et al.*, 2013; Vera *et al.*, 2008).

Die Expression der einzelnen Elemente kann von den zugehörigen sequenzierten Fragmenten abgeleitet werden. Die einfachste Variante wäre es, die Zahl der Einzelfragmente einfach zu summieren. Aus methodischen Gründen wäre es allerdings sinnvoll die Zahlen weiter zu normalisieren (Costa *et al.*, 2010). So können aus längeren Transkripten mehr Fragmente hervorgehen als bei Kürzeren, wodurch die Expression automatisch höher erscheint (Mortazavi *et al.*, 2008). Des Weiteren ist es sinnvoll die Daten an den Gesamtumfang der Sequenzierung anzupassen um verschiedene Proben vergleichen zu können. Eine häufig genutzte Einheit, die auf diesen Annahmen beruht, ist Fragmente pro Kilobasen Länge pro Millionen zugeordneter Fragmente (engl. *reads/fragments per kilobase per million*;

RPKM beziehungsweise FPKM; Mortazavi *et al.* 2008; Trapnell *et al.* 2010). Eine Alternative stellt die Expression nach Transkripten pro Millionen (TPM) dar, welche beim Vergleich mehrerer Experimente genauere Ergebnisse ergeben soll (Li & Dewey, 2011; Li *et al.*, 2010; Wagner *et al.*, 2012).

RNA-Seq geht heute in seinem Nutzen über die bloße Bestimmung der Expression von Transkripten hinaus. Durch ihren relativ niedrigen Preis und hohen Datendurchsatz hat diese Technik zahlreiche Anwendungsmöglichkeiten gefunden (Costa *et al.*, 2010; Marguerat & Bähler, 2010; Mutz *et al.*, 2013; Wang *et al.*, 2009). RNA-Seq hat zur Charakterisierung zahlreicher unbekannter Gene geführt und dabei das Verständnis zu nicht-kodierenden RNAs (engl. *non-coding RNAs*; ncRNAs) erweitert (He *et al.*, 2008; Jacquier, 2009; Kutter *et al.*, 2012; Nagalakshmi *et al.*, 2008). Des Weiteren konnten auch bekannten Genen neue Spleißvarianten zugeordnet werden (Morin *et al.*, 2008; Mortazavi *et al.*, 2008; Sultan *et al.*, 2008). SNP-Analysen lassen sich einfach durchführen (Cloonan *et al.*, 2008; Morin *et al.*, 2008; Quinn *et al.*, 2013) und selbst bei relativ unbekanntem Organismen stößt diese Technik nicht an ihre Grenzen (Lowe *et al.*, 2011; Pallavicini *et al.*, 2013; Vera *et al.*, 2008). Sogar bei Sequenzierungen von Genomen kommt mittlerweile RNA-Seq zum Einsatz um Gene zu annotieren (Curtis *et al.*, 2012; Denoeud *et al.*, 2008; Paterson *et al.*, 2012; Price *et al.*, 2012). Mit der stetig steigenden Menge verfügbarer Sequenzierdaten werden noch umfangreichere Analysen ermöglicht. Diese werden nur noch durch die rechenintensive bioinformatische Aufbereitung der Daten beschränkt.

3.4 Transkriptome der Eukaryoten

Als Transkriptom wird die komplette Zusammensetzung der RNA einer Zelle zu einem bestimmten Zeitpunkt bezeichnet (Wang *et al.*, 2009). Es ist unter anderem abhängig vom Entwicklungsstand und von physiologischen Einflüssen. Aufgrund der Extrapolation von Wissen über Prokaryoten wurde angenommen, dass auch Transkriptome höherer Eukaryoten vorwiegend aus ribosomaler RNA (~80%), Transfer-RNA (~15%) und mRNA (engl. *messenger RNA*; 2-4%) bestehen (Lindberg & Lundeberg, 2010). Anderen RNA-Arten wurde lediglich eine untergeordnete Rolle zugesprochen. Erst in jüngster Zeit, mit der Entwicklung hochauflösender Analysemethoden, konnten zahlreiche neue, nicht Protein-kodierende Transkripte (engl. *non-coding RNA*, ncRNA) beschrieben werden.

Überraschende Ergebnisse ergaben die ersten umfassenden Transkriptomsequenzierungen. Die Hälfte der exprimierten Genomregionen wurden außerhalb

Protein-kodierender Bereiche gefunden (Bertone *et al.*, 2004; Okazaki *et al.*, 2002). Studien, die sich mit 1% des menschlichen Genoms befassten, konnten dort die Transkription von fast jeder Base (93%) nachweisen (ENCODE Project Consortium, 2007). Diese Menge wurde später für das gesamte Genom auf immerhin noch gut zwei Drittel verringert (Djebali *et al.*, 2012). Im GENCODE-Projekt (Harrow *et al.*, 2012) waren Anfang 2014 für den Menschen 57.820 Gene aufgeführt, von denen nur 35% Proteine kodierten. Bei der Maus waren es noch 58%. In den wenigen Transkriptomanalysen an Protisten konnte eine ähnlich umfangreiche Transkription des Genoms wie bei Säugetieren beschrieben werden (Nagalakshmi *et al.*, 2008; Wilhelm *et al.*, 2008; Xiong *et al.*, 2012). Dennoch führt die Tatsache, dass der Anteil nicht-kodierender DNA (ncDNA) in Einzellern generell geringer ausfällt, zu der Annahme, dass sich hier auch weniger ncRNA-Gene befinden (Taft *et al.*, 2007). Im Folgenden werden die grundlegenden Bestandteile des Transkriptoms näher beschrieben.

3.4.1 RNA-Typen in der Translation

Das zentrale Dogma der Molekularbiologie, formuliert durch Francis Crick, besagt, dass der Informationsfluss stets von Nukleinsäuren zu Proteinen geht (Sharp, 2009). Gleich mehrere RNA-Typen sind an diesem Prozess beteiligt. Die Boten-RNA (mRNA) fungiert als direkter Vermittler der Information vom Gen zum Protein. Die mRNA besteht vorwiegend aus mindestens einem Protein-kodierenden Bereich, welcher mit einem Startkodon „AUG“ beginnt und mit dem nächsten Stoppkodon abschließt (Kozak, 1983). Des Weiteren weisen sie bei Eukaryoten meistens am 5'-Ende ein modifiziertes Guanin-Nukleotid auf und zeigen am 3'-Ende einen Poly-A-Schwanz, der aus einer Folge von Adenin-Nukleotiden besteht. Beide Charakteristika haben in erster Linie die Funktion die RNA-Stabilität zu verbessern (Furuichi *et al.*, 1977; Gonatopoulos-Pournatzis & Cowling, 2014; Huez *et al.*, 1981; Zeevi *et al.*, 1982). Die Übersetzung der mRNA in ein Protein findet an den Ribosomen statt. Diese setzen sich aus ribosomalen Proteinen und einem anderen RNA-Typ, der ribosomalen RNA (rRNA), zusammen. Von Letzterer können grundsätzlich vier Typen eukaryotischen (18S, 25/28S, 5.8S, 5S) und drei Typen prokaryotischen (16S, 23S, 5S) Ursprungs unterschieden werden (Brosius *et al.*, 1981; Fromont-Racine *et al.*, 2003; Venema & Tollervy, 1999). Sie machen mit ~70% bis über ~90% die größte Masse der RNA im Transkriptom aus (Costa *et al.*, 2010; He *et al.*, 2010; Karpinets *et al.*, 2006).

Ähnlich wie rRNAs sind auch Transfer-RNAs (tRNA) essentielle Bestandteile der Translation. Sie sind die direkte Verbindung der Basentriplets (Kodons) der mRNA mit den passenden Aminosäuren (Rich & RajBhandary, 1976). Letztere sind kovalent an das 3'-Ende der tRNA gebunden und werden während der Translation an das wachsende Polypeptid angehängt. In der tRNA sind auch ungewöhnliche Basen, wie Inosin, Pseudouridin oder Lysidin vertreten (Marck & Grosjean, 2002; Murphy & Ramakrishnan, 2004). Das dem Guanin ähnelnde Inosin ist beispielsweise mitverantwortlich dafür, dass mehrere Kodons zu der gleichen Aminosäure führen können. Im Gegensatz zu den Typen der folgenden Abschnitte sind die hier beschriebenen RNAs schon länger bekannt und ihre Synthese und Wirkungsweise wurde umfangreich erforscht.

3.4.2 Kurze nicht-kodierende RNA

Zahlreiche, nicht-kodierende RNAs in verschiedenen Größen sind Bestandteil des Transkriptoms. Darunter sind auch viele kürzere RNA-Fragmente (engl. *small RNA*; sRNA), die trotz ihrer geringen Größe wichtige Aufgaben bei der Regulation von Genen und Genomen erfüllen (Carthew & Sontheimer, 2009; Chapman & Carrington, 2007; Jacquier, 2009; Mattick & Makunin, 2005). Dazu können auch ncRNAs von wenigen hundert Nukleotiden Länge, wie die snRNAs und snoRNAs gezählt werden, die unter anderem wichtige Aufgaben beim Aufbau der Ribosomen und im Spleißosom erfüllen (Holley & Topkara, 2011; Kiss, 2002; Maniatis & Reed, 1987; Mattick & Makunin, 2005). Allerdings werden als sRNAs meist wesentlich kürzere RNAs mit einer Länge im Bereich von 20 bis 30 Nukleotiden angesprochen (Carthew & Sontheimer, 2009; Chapman & Carrington, 2007; Jacquier, 2009). Diese können wiederum in die drei Hauptgruppen miRNA (engl. *micro RNA*), siRNA (engl. *small interfering RNA*) und piRNA (engl. *piwi-interacting RNA*) aufgeteilt werden. Während bei miRNAs eine kurze doppelsträngige Vorläufer-RNA aus dem komplementären Bereich einer Haarnadelstruktur (engl. *hairpin-structure*) prozessiert wird, geht ein vergleichbarer Vorläufer der siRNAs meist aus endogener oder exogener doppelsträngiger RNA (dsRNA) hervor. Im Gegensatz dazu entstehen piRNAs aus einzelsträngigen Transkripten. Der funktionelle Schritt ist in allen drei Fällen charakterisiert durch die einzelsträngige sRNA assoziiert mit Proteinen der Argonauten-Familie im RISC-Komplex (engl. *RNA-induced silencing complex*; Hutvagner & Simard 2008; Meister 2013). Dabei können unterschiedliche sRNAs verschiedenen Vertretern dieser Proteingruppe zugeordnet sein (Qi *et al.*, 2006; Siomi *et al.*, 2011; Yigit *et al.*, 2006). Die

sRNA führt den Komplex über komplementäre Basenpaarungen an eine Ziel-RNA (Carthew & Sontheimer, 2009; Chapman & Carrington, 2007; Meister, 2013). Anschließend inaktiviert RISC das Ziel über Degradation oder blockiert seine weitere Prozessierung (engl. *gene silencing*). Anfangs wurde angenommen, dass siRNAs primär in der Abwehr von Fremd-RNA fungieren, während miRNAs an der posttranskriptionellen Regulation der eigenen Gene beteiligt sind (Carthew & Sontheimer, 2009; Golden *et al.*, 2008). Später wurde jedoch klar, dass auch siRNAs an der zelleigenen Regulation mitwirken (Lippman & Martienssen, 2004; Vazquez *et al.*, 2004). Dagegen ist die Funktion der piRNAs mit der Inaktivierung von Transposons in der Keimbahn bei Tieren deutlich spezifischer definiert (Siomi *et al.*, 2011). Die hier beschriebenen kurzen RNAs konnten bisher nur in Eukaryoten nachgewiesen werden und Erkenntnisse beziehen sich meist auf Metazoa und höheren Pflanzen (Carthew & Sontheimer, 2009; Hutvagner & Simard, 2008). Die Verbreitung der Argonauten-Familie in Eukaryoten läßt auf einen gemeinsamen Ursprung schließen (Cerutti & Casas-Mollano, 2006), allerdings gibt es Unterschiede bei miRNAs von Tieren und Pflanzen bezüglich Wirkungsweise und Synthese (Axtell *et al.*, 2011; Jones-Rhoades *et al.*, 2006). Des Weiteren wurde den meisten miRNAs in Protisten eine unzureichende Klassifikation zugeschrieben (Tarver *et al.*, 2012).

3.4.3 Lange nicht-kodierende RNA

Lange nicht-kodierende RNAs (engl. *long non-coding RNA*, lncRNA) bilden zusammen mit kurzen RNAs die große Menge wenig erforschter ncRNAs, die aus der unerwartet umfassenden Transkription von eukaryotischen Genomen hervorgehen (Kapranov *et al.* 2007; Kung *et al.* 2013; Mercer *et al.* 2009; siehe Abschnitt 3.4). Um sRNAs von lncRNAs zu unterscheiden wird meist eine Mindestlänge von 200 Nukleotiden angewandt (Kapranov *et al.*, 2007). Im Gegensatz zu Ersteren sind Wirkungsweisen für lncRNAs lediglich an wenigen Beispielen untersucht (Amaral *et al.*, 2010; Niazi & Valadkhan, 2012). Es ist umstritten, ob die meisten lncRNAs überhaupt eine Funktion erfüllen und nicht nur ein Hintergrundrauschen der Transkription repräsentieren, welches durch die Ungenauigkeit der RNA-Polymerase II (Struhl, 2007) oder der Expression umliegender Gene verursacht wird (Ebisuya *et al.*, 2008; Wang *et al.*, 2011). Eine geringe Sequenzähnlichkeit zwischen verschiedenen Spezies und das Fehlen bekannter struktureller Charakteristika erschweren eine funktionelle Einordnung (Kung *et al.*, 2013; Mercer *et al.*, 2009).

Ansätze dazu beziehen sich auf eine geringe aber doch vorhandene evolutionäre Konservierung (Guttman *et al.*, 2009; Ponjavic *et al.*, 2007), differentielle Expression unter verschiedenen Faktoren (Amaral & Mattick, 2008; Dinger *et al.*, 2008) und der Möglichkeit als reverses Komplement (engl. *antisense*) eine mRNA von der gegensträngigen DNA zu regulieren (Faghihi & Wahlestedt, 2009; He *et al.*, 2008). Die wenigen näher untersuchten Beispiele zeigen, dass zumindest ein Teil der lncRNAs funktionell ist. Die *Xist* (engl. *X inactive-specific transcript*) RNA ist beispielsweise an der Inaktivierung des X-Chromosoms in Säugetieren beteiligt (Kung *et al.*, 2013; Ohhata *et al.*, 2008). *HOTAIR* (engl. *HOX antisense intergenic RNA*) RNA kann die Transkription des Homeobox-Proteins HOXD über eine Interaktion mit dem Protein-Komplex PRC2 reprimieren (Rinn *et al.*, 2007) und *MALAT1* (engl. *Metastasis associated in lung adenocarcinoma transcript*) ist assoziiert mit der Lokalisierung von Spleißfaktoren (Ji *et al.*, 2003; Tripathi *et al.*, 2010). Weitere Funktionen von lncRNAs konnten in der Organisation von Chromosomen, Telomeren und der Zellstruktur nachgewiesen werden (Amaral & Mattick, 2008; Hu *et al.*, 2012; Mercer *et al.*, 2009). Dabei sind gerade einmal 200 dieser RNAs näher untersucht, wovon sich die Hälfte allein auf den Menschen bezieht (Amaral *et al.*, 2010; Niazi & Valadkhan, 2012). Ausserhalb der Gruppe der Metazoa gibt es wenige Studien, die sich mit diesem RNA-Typ befassen (Au *et al.*, 2011).

Auch funktionslose Transkripte könnten eine biologische Bedeutung haben (Koonin & Wolf, 2010; Kung *et al.*, 2013; Lynch, 2007). Sie könnten als Basis funktioneller Innovation für die Zelle fungieren und so vielleicht später eine Funktion erhalten. Carvunis *et al.* (2012) konnten in Hefe die Translation vieler kurzer, kaum konservierter Peptide nachweisen und stellten die Hypothese auf, dass es verschiedene Intermediate zwischen genomischer DNA, zufällig transkribierten Sequenzen und funktionellen Peptiden geben könne. Da eindeutige Sequenzmerkmale fehlen, werden lncRNAs anhand ihres Bezugs zu bekannten Genen klassifiziert (Kung *et al.*, 2013). Demnach befinden sich intergenische lncRNAs (engl. *large/long intergenic/intervening noncoding RNA*; lincRNA) außerhalb kodierender Bereiche (Cabili *et al.*, 2011; Guttman *et al.*, 2009; Ulitsky *et al.*, 2011), *Antisense*-Transkripte liegen auf dem Gegenstrang anderer Gene (Faghihi & Wahlestedt, 2009; He *et al.*, 2008), intronische lncRNAs sind in Introns kodiert (Louro *et al.*, 2009; Rearick *et al.*, 2011) und exprimierte Pseudogene ähneln Protein-kodierenden Genen (Pink *et al.*, 2011; Poliseno, 2012).

3.4.4 Pseudogene

Als Pseudogene werden Gene bezeichnet, welche ihre Fähigkeit als Vorlage für ein funktionelles Produkt verloren haben (Pink *et al.*, 2011; Poliseno, 2012; Zheng *et al.*, 2007). Meist bezieht sich dieser Begriff auf die Fähigkeit Proteine zu kodieren. Dabei verhindert der Verlust der Promotorsequenz, Einfügen interner Stoppkodons, Veränderungen im Leserahmen und an den Spleißstellen die Transkription und Translation in ein funktionelles Polypeptid. Die Menge der Pseudogene im Menschen erreicht mit zehn- bis zwanzigtausend schon fast die der Proteinkodierenden Sequenzen (Pei *et al.*, 2012; Pink *et al.*, 2011; Zhang & Gerstein, 2004). In unizellulären Organismen wurden deutlich weniger Pseudogene beschrieben (Berriman *et al.*, 2005; Lawrence *et al.*, 2001; Zhang & Gerstein, 2004). Vor allem in den kompakten Genomen der Prokaryoten scheint für sie kein Platz zu bestehen (Kuo & Ochman, 2010; Lawrence *et al.*, 2001). In einigen Fällen zeigen Pseudogene Anzeichen von Expression. Für den Menschen gibt es verschiedene Schätzungen, die 2% bis 20% aller Pseudogene Transkription zusprechen (Yano *et al.*, 2004; Zheng *et al.*, 2007, 2005). Die Expression liegt dabei meist deutlich unter der des funktionellen Gegenstücks, kann in Beispielen dieser aber auch gleichkommen oder sie sogar übertreffen (Poliseno *et al.*, 2010; Sorge *et al.*, 1990; Zheng & Gerstein, 2007). Auch eine differentielle Expression in unterschiedlichen Geweben und unter verschiedenen physiologischen Bedingungen wurde beobachtet (Poliseno *et al.*, 2010; Zeller *et al.*, 2009; Zheng *et al.*, 2007). Allgemein wird angenommen, dass Pseudogene nur funktionslose Überreste von Genen repräsentieren und daher neutral evolvieren (Li *et al.*, 1981), allerdings zeigen manche Pseudogene neben der differentiellen Expression noch andere Anzeichen biologischer Funktion. So konnte gezeigt werden, dass gegensträngige Transkripte von Pseudogenen zu kurzen, regulatorischen RNAs prozessiert werden können (Tam *et al.*, 2008; Watanabe *et al.*, 2008) oder sich komplementär an ihre funktionellen Gegenstücke anlagern, um ihre Expression zu beeinflussen (Hawkins & Morris, 2010; Korneev *et al.*, 1999).

3.5 Zielsetzung

Sequenziermethoden der neuesten Generation haben Analysen auf Nukleotidebene revolutioniert. Es wurde möglich, ganze Genome binnen kurzer Zeit und unter geringem Kostenaufwand zu entschlüsseln. Eine andere Anwendungsmöglichkeit ist die Sequenzierung von Transkriptomen (siehe Abschnitt 3.4). So lassen sich nicht nur Basenfolgen bestimmen, darüber hinaus ist es auch möglich Expressionswerte abzuleiten. Ihre universelle Anwendbarkeit macht sie auch für weniger erforschte Organismen interessant, bei denen beispielsweise Genomsequenzen fehlen (Lowe *et al.*, 2011; Pallavicini *et al.*, 2013; Vera *et al.*, 2008). Ein Vorteil der besonders bei der Arbeit mit Protisten hilft, denn bis auf wenige Modellorganismen und einige Pathogene ist die Menge sequenzierter Genome hier noch gering.

Dabei zeigt gerade diese Gruppe die größte Diversität unter den Eukaryoten. Dies wird schon dadurch verdeutlicht, dass in der aktuellen Taxonomie Tiere, Pflanzen und Pilze nur zwei der fünf Super-Gruppen abdecken (Adl *et al.*, 2012). Protisten können uns wichtige Einblicke in den Ursprung der Eukaryoten geben (Fritz-Laylin *et al.*, 2010; Koonin, 2010a,b). Darüber hinaus befinden sich unter ihnen bedeutende Mikroorganismen. Diverse Gruppen von Pathogenen für Mensch, Tier und Pflanze sind hier vertreten (Dame *et al.*, 1996; El-Sayed *et al.*, 2005; Tyler *et al.*, 2006). Des Weiteren machen sie einen großen Teil des Phytoplanktons aus, der die Lebensgrundlage im Meer darstellt, während wenige uns sogar als Nahrungsmittel dienen (Blouin *et al.*, 2011; Hackett *et al.*, 2004).

NGS-Systeme wurden schon mehrfach bei Analysen an einzelligen Eukaryoten eingesetzt (Kim *et al.*, 2014; Teixeira *et al.*, 2012). Im Rahmen dieser Arbeit wird die Möglichkeit der Anwendung von Transkriptomsequenzierungen erprobt, um neue Einsichten über die weniger untersuchten Gruppen der Protisten zu gewinnen. Dabei wird eine *De-novo*-Sequenzierung der erst kürzlich beschriebenen Alge *Chromera velia* ausgewertet sowie referenz-basierende Methoden genutzt um das Transkriptom des Parasiten *Trichomonas vaginalis* unter unterschiedlichen Einflüssen zu charakterisieren. Eine darauf aufbauende Analyse beschäftigt sich mit der Suche nach bisher unbeschriebenen Transkripten. Ferner werden Transkriptsequenzen auch als Datenbank genutzt um auf das Vorhandensein bestimmter Gene in Organismen ohne sequenzierte Genome zu prüfen. Die beschriebenen Analysen sollen zeigen inwieweit die neuen Technologien geeignet sind um verschieden Aspekte an diesen Mikroorganismen zu erforschen.

4 PUBLIKATIONEN

4.1 Red and problematic green phylogenetic signals among thousands of nuclear genes from the photosynthetic and apicomplexa-related *Chromera velia*

Christian Woehle, Tal Dagan, William F. Martin & Sven B. Gould

Der vorliegende Artikel wurde 2011 in der Fachzeitschrift *Genome Biology and Evolution* (Impact factor 4,8) veröffentlicht.

Ergänzendes Material steht zur Verfügung durch die Webseiten des Verlegers⁶.

Beitrag von Christian Wöhle, Erstautor:

Versuchsplanung	30 %
Datenanalyse	80 %
Verfassen des Manuskripts	10 %

⁶<http://gbe.oxfordjournals.org/content/3/1220/suppl/DC1>

Red and Problematic Green Phylogenetic Signals among Thousands of Nuclear Genes from the Photosynthetic and Apicomplexa-Related *Chromera velia*

Christian Woehle, Tal Dagan, William F. Martin, and Sven B. Gould*

Molecular Evolution (Botanik III), Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany

*Corresponding author: E-mail: sbgould@gmail.com.

Accepted: 22 September 2011

Abstract

The photosynthetic and basal apicomplexan *Chromera velia* was recently described, expanding the membership of this otherwise nonphotosynthetic group of parasite protists. Apicomplexans are alveolates with secondary plastids of red algal origin, but the evolutionary history of their nuclear genes is still actively discussed. Using deep sequencing of expressed genes, we investigated the phylogenetic affinities of a stringent filtered set of 3,151 expressed sequence tag-contigs by generating clusters with eukaryotic homologs and constructing phylogenetic trees and networks. The phylogenetic positioning of this alveolate alga was determined and sets of phyla-specific proteins extracted. Phylogenetic trees provided conflicting signals, with 444 trees grouping *C. velia* with the apicomplexans but 354 trees grouping *C. velia* with the alveolate oyster pathogen *Perkinsus marinus*, the latter signal being reinforced from the analysis of shared genes and overall sequence similarity. Among the 513 *C. velia* nuclear genes that reflect a photosynthetic ancestry and for which nuclear homologs were available both from red and green lineages, 263 indicated a red photosynthetic ancestry, whereas 250 indicated a green photosynthetic ancestry. The same 1:1 signal ratio was found among the putative 255 nuclear-encoded plastid proteins identified. This finding of red and green signals for the alveolate mirrors the result observed in the heterokont lineage and supports a common but not necessarily single origin for the plastid in heterokonts and alveolates. The inference of green endosymbiosis preceding red plastid acquisition in these lineages leads to worryingly complicated evolutionary scenarios, prompting the search for other explanations for the green phylogenetic signal and the amount of hosts involved.

Key words: Chromera, Apicomplexa, Alveolata, chromalveolata, apicoplast, protist evolution.

Introduction

The Apicomplexa are a group of parasite protists that, with the exception of intestinal parasites from the genus *Cryptosporidium*, house a relict plastid known as the apicoplast (reviewed in McFadden 2010). The organelle does not perform photosynthesis but is nevertheless essential for ultimate parasite survival and propagation. This can probably be attributed to the number of biochemical pathways the apicoplast contains, which include parts of the fatty acid and isopentenyl diphosphatase synthesis (Waller et al. 1998; Jomaa et al. 1999), the assembly of iron–sulfur complexes (Seeber 2002), and segments of heme biosynthesis which is most likely carried out in conjunction with the mitochondria (Ralph et al. 2004). The recent discovery of *Chromera velia* has added the first nonparasitic autotroph with a photosynthetically active plastid to the base of the

apicomplexan phylum (Moore et al. 2008). At least one more photosynthetic basal apicomplexan has since been described, and collectively, they are currently designated as “chromerids” (Janouskovec et al. 2010; Obornik et al. 2011). Chromerid algae are suspected to be a missing link, connecting the parasitic Apicomplexa with their evolutionary past and algal relatives (Moore et al. 2008).

Apicomplexa belong to the alveolates, a group that includes the dinoflagellates and the ciliates, as well as other less intensely studied lineages such as the Perkinsidae (Gould, Waller, et al. 2008, Zhang et al. 2011). The infrakingdom Alveolata is characterized by the presence of cortical alveolae, a one-membrane bound compartment lying below the plasma membrane and together with longitudinal microtubules and an electron-dense layer of mainly unknown composition (referred to as epiplasm or subpellicular

The Author(s) 2011. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

network) forms a multilayered cell pellicle (Cavalier-Smith 1991; Gould, Tham, et al. 2008). The evolutionary history of photosynthesis among the alveolates involves the acquisition of red algal plastids via secondary endosymbiosis (Stoebe and Maier 2002; Gould, Waller, et al. 2008; Archibald 2009; Janouskovec et al. 2010; Keeling 2010). Most apicomplexans studied to date have highly reduced plastid genomes, whereas *Cryptosporidium* has lost the organelle entirely and only comparatively few genes have been retained in this case, which betray the previous existence of a photosynthetic organelle (Huang et al. 2004). No ciliate has yet been identified that possesses a plastid or a relict thereof, and to our knowledge, only one report claims to have identified phylogenetic evidence for the past presence of such an organelle based on 16 ciliate nuclear genes (Reyes-Prieto et al. 2008). Eisen et al. (2006) noticed a similar weak green signal in their earlier genome analysis manuscript of *Tetrahymena thermophila*, but in contrast concluded this signal is not above “background noise” that are expected at random in the analysis of 10,000 of genes. In conclusion, there is currently no credible evidence that ciliates ever were secondarily photosynthetic. Finally, Perkinisidae, like Apicomplexa, have nonphotosynthetic plastids that in addition seem to lack DNA and, hence, must have all of their required protein content encoded by nuclear genes (Matsuzaki et al. 2009).

The evolutionary origin of alveolates and their plastid(s) is furthermore coupled to the waning dispute over the “chromalveolate” hypothesis, which proposed that a single secondary endosymbiotic event gave rise to all protist lineages harboring a secondary plastid of red algal origin (Cavalier-Smith 1999). Some analyses supported the chromalveolate concept (Bachvaroff et al. 2005; Harper et al. 2005; Patron et al. 2007), but more current data indicates that it is wrong with respect to the prediction of a single secondary symbiosis (Sanchez-Puerta and Delwiche 2008; Baurain et al. 2010; Felsner et al. 2011). The monophyletic origin was challenged earlier on by others, who also proposed an alternative evolutionary model (Bodl 2005; Bodl et al. 2009). In conclusion, a clarification of the specific evolutionary relationships between these complex phyla remains to be provided (reviewed in Gould, Waller, et al. 2008; Sanchez-Puerta and Delwiche 2008; Archibald 2009; Keeling 2010).

A phylogenetic analysis of the two chromerid plastid genomes leaves no doubt that they are of red algal origin, highly reduced compared with red algal plastids in general and larger than those of dinoflagellate plastids and apicomplexan apicoplasts (Janouskovec et al. 2010). But plastid genomes reflect neither the phylogenetic position nor the evolutionary history of the host. A recent analysis on the two sequenced genomes of the diatoms (stramenopiles) *Thalassiosira* and *Phaeodactylum* has added a new twist to the chromalveolate conundrum (Moustafa et al. 2009). They found that approximately 16% of the proteins

potentially encoded by the nuclear genome of stramenopiles were found to reflect a green algal origin, that is, they indicated a closer phylogenetic affinity to the green lineage of primary plastids than to the red. Not unreasonably, they interpreted that observation as evidence for a green photosynthetic ancestry of the diatom host prior to the acquisition of the red algal plastid, as predicted in theory earlier on (Häuber et al. 1994, Becker et al. 2008, Frommolt et al. 2008), but with several caveats, in particular concerning lineage sampling (Dagan and Martin 2009). With the goal of better understanding the phylogenomic position and photosynthetic history of protists with red secondary plastids, we have employed deep sequencing to investigate the phylogeny of *C. velia* expressed nuclear genes.

Materials and Methods

Cell Culture, mRNA Processing, and Library Assembly

Cells were grown at 25 °C with a 16 h light and 8 h dark cycle in Tropic Marin PRO-REEF (Tropic Marin, Germany) supplemented with f/2 AlgaBoost (AusAqua, Australia). Cells of 800 ml culture (about 5×10^5 cells/ml) from three different time points (every 8 h starting 1 h before the light turned on) were harvested by centrifugation at $3,000 \times g$ for 20 min. RNA of those three samples was isolated separately with TRIzol (Invitrogen, Germany) following the manufactures protocol with the following modification: the cell pellet was grinded in the presence of liquid nitrogen for 5–10 min before TRIzol was added. After RNA quantification, the samples were pooled so that an equal amount of each was present and sent on dry ice for further processing to GATC-Biotech (Germany). At GATC, the RNA was amplified using their standard protocol for “True-Full-Length cDNA” and then additionally normalized before sequencing 2 million reads on a Titanium GS FLX (Roche). Trimming of adapter sequences, primary clustering, and assembly of the reads was performed by GATC-Biotech. Sequencing resulted in 2502269 reads with an average length of 239 bases, which were assembled into 29,856 contigs. Additionally, we included 2,854 *C. velia* expressed sequence tags (ESTs) from GenBank (Benson et al. 2009). Multiple copy proteins were unified and EST-contigs shorter than 100 nt removed. Furthermore, such EST-contigs with BlastN hits to the plastid genome of *C. velia* (e value cutoff 10^{-10} , downloaded from RefSeq, Pruitt et al. 2007) or the Rfam database (Gardner et al. 2009) were deleted in order to remove remnants of chloroplast-encoded transcripts and non-coding RNA families. All sequences have been deposited under JO786643–JO814452.

Database Preparation

The protein database sequences were obtained from either EuPathDB (Aurrecochea et al. 2007) RefSeq or in the case of *Cyanidioschyzon merolae* (Matsuzaki et al. 2004),

Ectocarpus siliculosus (Cock et al. 2010), and *Emiliania huxleyi* (<http://genome.jgi-psf.org/Emihu1/Emihu1.download ftp.html>) from their corresponding genome project homepages. From the downloaded files, we removed C-terminal stop codons and replaced selenocysteins by Xs. In cases where no adequate number of protein sequences was available, EST-contigs were used instead or in addition. For this purpose, we created an EST-contig database by downloading ESTs for all lineages with >1,000 entries from GenBank, with exception of the *Galdieria* ESTs, which were downloaded from the *Galdieria sulphuraria* genome project homepage (Weber et al. 2004). For further information and a list of organisms, see [supplementary information \(Supplementary Material online\)](#). The EST-contigs were translated into proteins by the method described below and merged with the protein database.

Chromera EST-contigs were translated in a protein sequence similarly to the method described in Min et al. (2005). The EST sequences were blasted (BlastX; Altschul et al. 1997), using *e* value threshold $\leq 1 \times 10^{-5}$ to the protein database and SwissProt database (Boeckmann et al. 2003). For sequences with blast hits, we translated the EST-contigs using the reading frame of the best blast hit (BBH). Sequences lacking a blast hit were predicted de novo by searching for the open reading frame (ORF) yielding the longest polypeptide (using both sense and antisense). In ORFs lacking an N-terminal methionine, the first codon in the EST-contig was translated into the first amino acid. When a C-terminal STOP codon was missing, the last codon in the EST-contig was translated into the last amino acid. Translated EST-contigs of *C. velia* were clustered into cognates of nearly identical EST-contigs by CDHIT (Weizhong and Godzik 2006) with a 95% amino acid sequence identity as a threshold, using the slow mode ($-g$ 1). For the remaining EST-contigs, a search for reciprocal BBH (rBBH; Tatusov et al. 1997) with an *e* value cutoff of $<1 \times 10^{-10}$ was performed against the protein/EST data set of each species/genus. In case of multiple BBH having identical *e* values, all hits were retained. In this case, the rBBH approach was used to reduce redundant hits within the ESTs of the same gene. Pairwise alignments of *Chromera* EST-contigs and their rBBH were reconstructed with Needleman and Wunsch alignment algorithm (Needleman and Wunsch 1970) using Needle (EMBOSS; Rice et al. 2000). Pairs with a global amino acid identity $\geq 25\%$ (excluding external gapped positions) were retained for further analysis. In case of multiple equally similar hits per one *Chromera* EST-contig or per one protein within the *Chromera* EST-contigs, the rBBH with the highest global similarity was used. Clusters of homologous proteins were constructed for *Chromera* EST-contigs and their homologs in all species data sets. An exclusion of 359 clusters comprising only EST-contigs yielded 3,151 clusters in total.

Phylogenetic Trees and Splits Networks

To reconstruct phylogenetic trees, all “nonchromalveolate” sequences except for one outgroup (the one showing the higher sequence similarity to the *Chromera* EST-contigs) were excluded from the clusters. Clusters having <4 remaining members were omitted. A total of 3,151 clusters of homologous proteins were aligned by MAFFT (Katoh and Toh 2008) using the default parameters. Multiple alignment quality was assessed using Guidance (Penn et al. 2010). Gapped alignment positions were removed and 86 short alignments (<10 positions) were excluded from further analysis. Phylogenetic trees were reconstructed from 2,258 multiple sequence alignments with PhyML (Guindon and Gascuel 2003) using the best fit model as inferred by ProtTest 3 (Darriba et al. 2011) using the Akaike information criterion (Akaike 1974) measure. For the reconstruction of a splits network, all splits within the phylogenetic trees were extracted using a Perl script and converted into a binary pattern that included 37 digits. If the split contained taxon *i* then digit *x*; in the corresponding pattern was set to “1,” otherwise it was “0.” Taxa that were missing in a tree were indicated by a “?” The resulting patterns were summarized in a splits network using SplitsTree (Huson and Bryant 2006).

To find *Chromera* sequences of green or red origin, only 1,174 clusters including proteins from Rhodophyta and Chloroplastida were used. All nonrhodophyta and nonchloroplastida sequences were removed from the clusters, except for those of *Chromera*. As an outgroup for each tree, the BBH to *C. velia* was used, which did not belong to Rhodophyta, Chloroplastida, a translated EST-contig or any organisms with a red algae as secondary endosymbiont. Phylogenetic trees were reconstructed from the resulting alignments (having ≥ 50 positions) using the same methodology described above, yielding 813 trees with an outgroup in total. The nearest neighbor to *Chromera* within each tree was determined by searching for the smallest clade that included *C. velia* and either only rhodophyta (red signal) or chloroplastida (green signal) and did not include the outgroup. For the determination of the position of *C. velia* in the trees as sister group or inside the red or green clades, we rooted the trees by the outgroups and searched for the second nearest neighbors using Newick Utilities package (Junier and Zdobnov 2010). Extraction of the longest branches to assess long-branch attraction was performed by the same package. Additional two split networks were reconstructed from trees sorted into red or green nearest neighbor using a composite outgroup regardless of the outgroup identity in each single tree.

Absence/Presence of Homologs in Other Species

In addition to the rBBH approach, homologs to *Chromera* EST-contigs within each species were identified by Blasting the clustered *Chromera* EST-contigs against the species data

set. BBHs with an e value $\leq 1 \times 10^{-10}$ were aligned with their *Chromera* homolog using Needle (EMBOSS; Rice et al. 2000). Global pairwise alignments resulting in $\geq 25\%$ amino acid identity after removal of external gapped positions were classified as a present homolog. The global amino acid identities presented in figure 2 were extracted from the pairwise alignments. The clusters that are shown along the y axis are sorted as follows: 1) all clusters specific for the apicomplexan phylum, 2) clusters of all members, 3) clusters that, except for *C. velia*, do have members just outside of apicomplexa. Within the three categories, the clusters were sorted by ascending number of present homologs within the Apicomplexa and descending number of present homologs within the non-Apicomplexa.

Prediction of Plastidal and Secretory Proteins

For the prediction of a signal peptide, only EST-contigs that were translated into a protein that started with a methionine were used. SignalP V3.0 (Emanuelsson et al. 2007) was used to find sequences with potential plastidal signal peptides. *Chromera* sequences having homologs (see "Database Preparation") that were annotated as plastid targeted were classified as plastidal proteins as well. All 657 detected sequences were then manually inspected, and an analysis including BlastP, SignalP, and TargetP (Emanuelsson et al. 2007) was used to determine the cleavage sites and distinguish plastidal from other secretory proteins. A sequence logo of the targeting signal was created using Weblogo (Crooks et al. 2004) from positions -20 to $+20$ in respect to the predicted cleavage site.

Annotation of Sequences

KEGG annotations were determined by using KAAS (Moriya et al. 2007) using translated *Chromera* sequences as query against the KEGG maps of 27 eukaryotes including (for the complete species name, see <http://www.genome.ad.jp/tools/kaas/>): hsa, dme, cel, ath, osa, olu, cme, sce, ddi, ehi, pfa, pyo, pkn, tan, tpv, bbo, cpv, cho, tgo, tet, ptm, tbr, tcr, lma, tva, pti, and tps. Protein functional categories were summarized as follows: KOs were mapped to the corresponding annotations obtained from KEGG FTP Server (<http://www.genome.jp/kegg/download/>). The main categories "Cellular Processes" and "Environmental Information Processing" were merged into "Cellular Processing and Signaling." Proteins in the "Unclassified, poorly characterized" category were classified as "Unclassified." All other "Unclassified" categories were added to subcategory "Other" of the corresponding main classification. Genes potentially associated with photosynthetic were identified by searching for the KEGG categories "Photosynthesis" and "Photosynthetic."

Results and Discussion

To obtain a broad sample of expressed genes, we isolated the RNA from exponentially growing cells every 8 h from

three different time points, covering light and dark cycle. The culture contained mainly nonflagellated immotile cells, although motile cells were also observed. The RNA was enriched for full-length transcripts and normalized before library sequencing. After assembly and filtering, 32,020 contigs with a balanced GC content of 50.76% and an average length of 827 bases were used to predict the protein sequences by BlastX, using a database containing the swissprot database and 34 selected genomes (for details, see Materials and Methods). To reduce redundancy, the predicted proteins were clustered by 95% identity, and homologous clusters formed by reciprocal blast to the protein database that additionally included predicted proteins from ESTs of lineages from which no genomic data were available, such as dinoflagellates. As a result, we obtained 3,151 clusters encoding on average 239 amino acids, which were then used for all subsequent analyses shown and discussed below. The predicted ORFs use all regular codons to encode the 20 standard amino acids, and no significant preference for certain codons was observed (supplementary table 1, Supplementary Material online).

Using Conserved Targeting to Identify Plastid Proteins

In order to screen for nuclear-encoded plastid proteins, we analyzed whether the targeting signal of these proteins—having to cross four membranes to reach the stroma—is as conserved as in many other organisms harboring a plastid of red algal origin (Patron and Waller 2007; Gould, Waller, et al. 2008). The plastid targeting signals of these organisms are well conserved, the translocon components involved are potential drug targets in Apicomplexa, and they have, hence, been a central topic of research. Furthermore, do they provide a molecular nontree-based evidence for the common ancestry—though not necessarily single origin—of the secondary plastids in the group (Gould, Sommer, Hadfi, et al. 2006; Patron and Waller 2007; Sommer et al. 2007; Lim et al. 2009; Spork et al. 2009; McFadden 2010).

First, we collected all contigs retrieving homologs with keywords such as plastid, chloroplast, or apicoplast within their annotation and analyzed their 5' end for an encoded signal peptide and its predicted cleavage site. From more than a 100 initial sequences, it became apparent that *Chromera* encodes a bipartite targeting signal (BTS) with a conserved cleavage motif (Ala-Phe) between signal and transit peptide. This position is crucial for correct targeting across the second innermost membrane of the plastids and present in cryptophytes, heterokontophytes, haptophytes, many dinoflagellates, and to a certain degree in the apicomplexan *Toxoplasma* (Gould, Sommer, Kroth, et al. 2006; Gruber et al. 2007; Patron and Waller 2007). It varies only a little, allowing to a lesser degree other bulky aromatic amino acids such as leucine, tyrosine, or tryptophane at the $+1$ position (Gruber et al. 2007; Patron and Waller 2007). The features of the subsequent transit peptide vary significantly more, even among apicomplexa

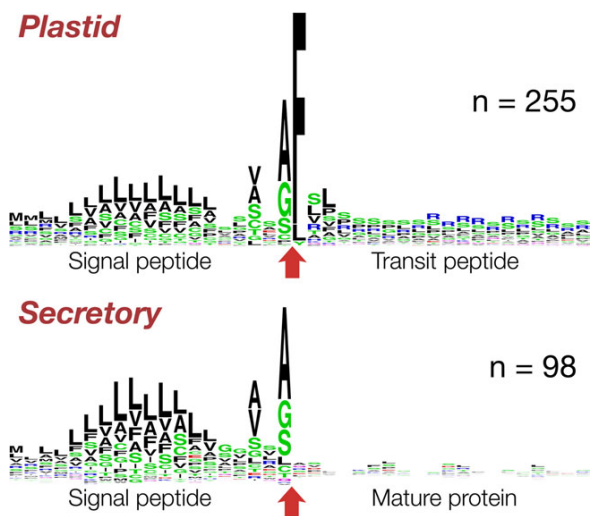


FIG. 1.—Sequence logo of the BTS of nuclear-encoded plastid proteins. The logo was curated based on 255 sequences, which encode an N-terminal signal peptide followed by a transit peptide. The $-20/+20$ positions relative to the cleavage site (red arrow) between the two parts of the BTS are shown. Secretory and plastid proteins both encode an almost identical signal peptide but only in the latter case a transit peptide follows. The N-terminal part of the transit peptide is enriched in serine residues and the C-terminal end with positively charged arginine residues.

themselves, but generally the level of phosphorylatable amino acids (serine/threonine) and positively charged amino acids (lysine and arginine) are elevated (Patron and Waller 2007).

In total, we collected 255 nuclear-encoded plastid proteins from our data set with a full-length 5' end encoding a BTS, from which we generated a sequence logo (fig. 1; supplementary table 2, Supplementary Material online). In 88.6% of them, the +1 position—that is, the first amino acid of the transit peptide—was a phenylalanine, in 7% a leucine, and in 2% a tyrosine. Compared with other transit peptides targeting to secondary red plastids, the current sample from *C. velia* represents a mix of features individually found in the transit peptides of other lineages with red plastids. They feature both enriched level of serine residues at the beginning and a stretch of positively charged amino acids that follows (fig. 1). For the latter, arginine instead of lysine residues are used when compared with *Plasmodium*. The latter can most likely be attributed to the high AT content of the *Plasmodium* genome. We found only one secretory nonplastid protein (a cathepsin homolog) with an Ala-Phe cleavage site. In this case, though, no transit peptide was predicted to succeed the signal peptide. Only a minor amount, less than 2% of likely plastid proteins such as a thylakoid lumen protein or a uroporphyrinogen III synthase (con11984 and con06800, respectively), encode amino acids other than F, L, or Y at the +1 position of the transit peptide. Apart from a wrong targeting signal prediction, there is, furthermore, the possi-

bility that some of those proteins are translocated across only the first 2 of the 4 membranes into the periplastidal compartment, present also in *C. velia* (Moore et al. 2008). These proteins harbor a BTS but different amino acids at the +1 transit peptide position (Gould, Sommer, Kroth, et al. 2006; Gruber et al. 2007). Of the latter, two ubiquitin-conjugating enzymes (con13687 and con23963) are of special interest as such enzymes are involved in protein translocation across the secondary plastids of red algal origin (McFadden 2010; Felsner et al. 2011; Moog et al. 2011).

The transit peptides of nuclear-encoded apicomplast proteins are generally characterized by a simple set of parameters, of which an overall positive charge is important (Tonkin et al. 2008). The chromerid BTS is an extraordinary example with chimeric characters, individually conserved in the different phyla and genera housing a red algal endosymbiont. The nature of the apicomplexan transit peptide holds the key to ultimately understanding how the proteins are selected from other secretory proteins in order to be transported to the apicomplast. *Chromera velia*, with its wealth of new sequences and a conserved targeting motif, offers a chance to commence a new search for the components involved once the entire nuclear genome becomes available.

Phyla Affinity and Phylogenetic Positioning

Evolution of protists with secondary plastids has generated a smorgasbord of organisms whose genomes show phylum-specific expansion of certain protein families and reduction of others, in Apicomplexa often reflecting the specialization of parasite–host interactions (Martens et al. 2008). *Chromera velia* is a nonparasitic phototrophic and basal apicomplexan and allows to investigate the question of what degree photosynthesis loss has in fact shaped apicomplexan parasites and their genomes compared with their photosynthetic relative.

Using 25% amino acid sequence identity as a cutoff, we found 151 *C. velia* EST-contigs that are unique to apicomplexa, and on the opposite almost 42% of our filtered EST-contigs retrieved homologs only outside the Apicomplexa (fig. 1 and supplementary information, Supplementary Material online). Twenty sequences are exclusively shared with *Perkinsus marinus* and 11 with ciliates. Thirty-five *C. velia* EST-contigs are exclusively shared with *P. marinus* and Apicomplexa and 13 with *P. marinus* and dinoflagellates, 80 with Apicomplexa, *P. marinus*, and dinoflagellates. In sum, 367 sequences of *Chromera* are exclusively shared with at least one other alveolate and five sequences were found unique to all alveolates. Expanding onto other phyla with secondary plastids of red origin (Haptophyta, Stramenopiles, and Cryptophyta), we find 143 EST-contigs exclusively shared with these. One hundred and ninety-nine EST-contigs of *Chromera* find homologs only outside of the alveolate, haptophyte, heterokont, and cryptophyte phyla.

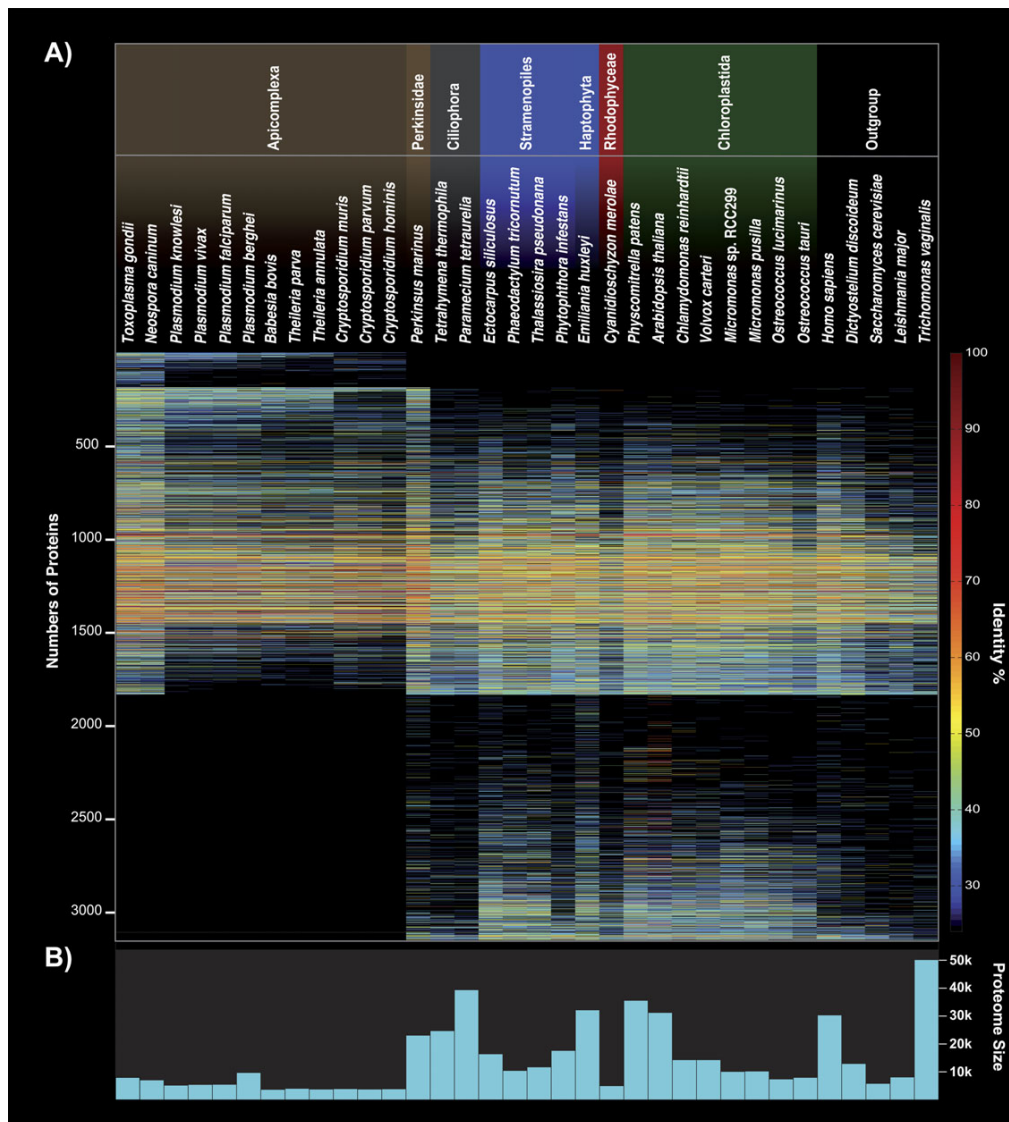


FIG. 2.—Presence/absence pattern and identity of the nuclear-encoded *Chromera velia* ESTs compared with 34 organisms. (A) The 3,151 sequences are sorted by their specificity and frequency to other Apicomplexa sequences. One hundred and fifty-one sequences have homologs only in Apicomplexa, whereas 1,316 sequences had homologs only in organisms other than Apicomplexa. Note that outside the Apicomplexa, *C. velia* shares the highest amount of overall identity with *Perkinsus marinus*. In (B), the potential amount of proteins encoded within the genomes used in the analysis.

As expected, the vast majority of the phyla-specific hits are proteins of unknown function (supplementary table 3, Supplementary Material online). Hence, an interpretation of what protein families might have expanded early within the apicomplexan phylum based on our EST-contig data would be unreasonable. Nevertheless, the amount of *Chromera*-encoded proteins that identify homologs only in organisms other than Apicomplexa, with 1,316 of 3,151, is huge. Martens et al. (2008) noticed a massive loss of genes encoding proteins involved especially in amino acid, carbohydrate, and lipid metabolisms and attributed

this to the parasitic lifestyle of apicomplexa. Indeed, approximately one-third of our 1,316 EST-contigs with a KEGG annotation retrieve KEGG annotations belonging to the three metabolic categories mentioned above and only 44 of them were classified in categories associated with photosynthesis. This confirms that losing photosynthesis (not the plastid) and giving up a host-independent lifestyle has had massive impact on the parasitic apicomplexan coding capacity.

The overall identity of the nuclear-encoded EST-contigs was compared with 34 organisms and summarized in a quantifying presence/absence pattern (fig. 2). The highest

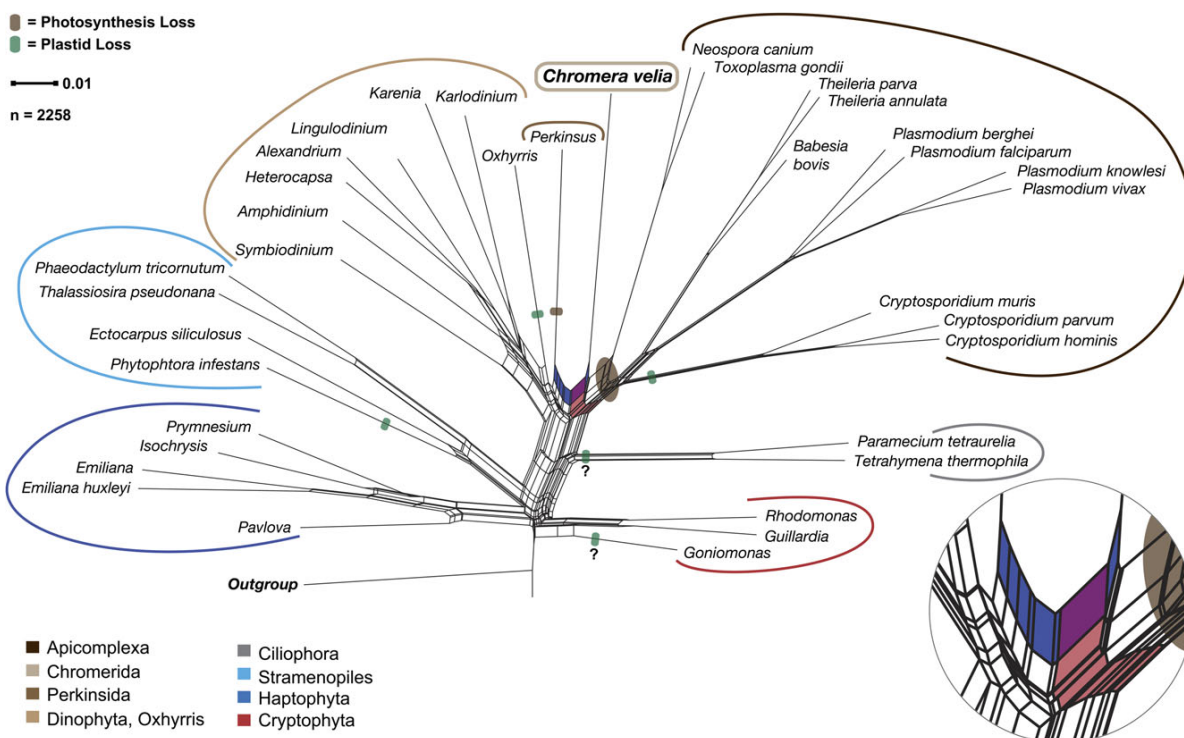


FIG. 3.—Splits network of distances derived from a matrix representation of all splits from the 2,258 homolog cluster trees generated. The net places the apicomplexan *Chromera velia* between nonphotosynthetic organisms. Bottom right shows an enlargement of the two splits that on the one side unites *C. velia*'s nuclear gene phylogeny with the Apicomplexa (light red)—whereby *C. velia* shows a basal position—and on the other highlights the signal linking it with the nonphotosynthetic *Perkinsus marinus* (blue split). Not only is this seen in the phylogeny above but also clearly in the gene distribution pattern in figure 1. The question marks indicate the two cases (Ciliates and Goniomonas), where it is disputed whether they lost or never had a plastid.

overall identity is shared with nonphotosynthetic Apicomplexa; by far with *Toxoplasma gondii* and another sarcocystidae, *Neospora canium*. The loss of the plastid in the genus *Cryptosporidium* has not affected the overall amount of sequence identity with *Chromera* as much as expected, when compared with the apicoplast bearing Aconoidasida, such as *Plasmodium*, *Babesia*, or *Theileria*. Generally, the amount of sequence identity of expressed *Chromera* genes can neither be directly linked to photosynthesis ability nor genome size (fig. 2). Notably and unexpectedly, the highest fraction of overall identity outside of the apicomplexan phylum is shared with *P. marinus*—a nonphotosynthetic but plastid bearing oyster pathogen.

A split network of distances derived from a matrix representation of all splits from the 2,258 homolog cluster trees with at least four members supports the monophyletic origin of alveolates and positions *C. velia* as the most basal apicomplexan (red split in fig. 3), but there is a conflicting split that links *C. velia* with *P. marinus*. Hence, the phylogenomic analysis is consistent with the presence/absence and sequence similarity of genes shown in figure 2. This, furthermore, raises the question as to whether chromerids should not only be pigeonholed as a basal apicomplexan but might

need to be treated as a separate lineage, sitting between the Perkinsidae and Apicomplexa, as already suggested by Moore et al. (2008). This is supported by our results, but should only be answered with confidence once more chromerid sequences, such as those of CCMP3155, become available. Our results, furthermore, suggest the possibility that a eukaryote–eukaryote endosymbiosis involving a red alga occurred after the ciliate phyla branched off independently, which included maybe a red alga phylogenetically linked but not identical to the one engulfed by the heterokont ancestor.

Green and Red Phylogenetic Signals among Nuclear-Encoded Proteins

The chromalveolate hypothesis posits that all members of this superphylum are united by the monophyletic origin of their secondary plastid from a red alga (Cavalier-Smith 1999). The plastid genome of the two chromerids supports a monophyletic rise of the currently present plastid in alveolates and heterokonts (Janouskovec et al. 2010) but that analysis did not include nuclear-encoded genes. Genome-sequencing projects of the two diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* and the oomycete

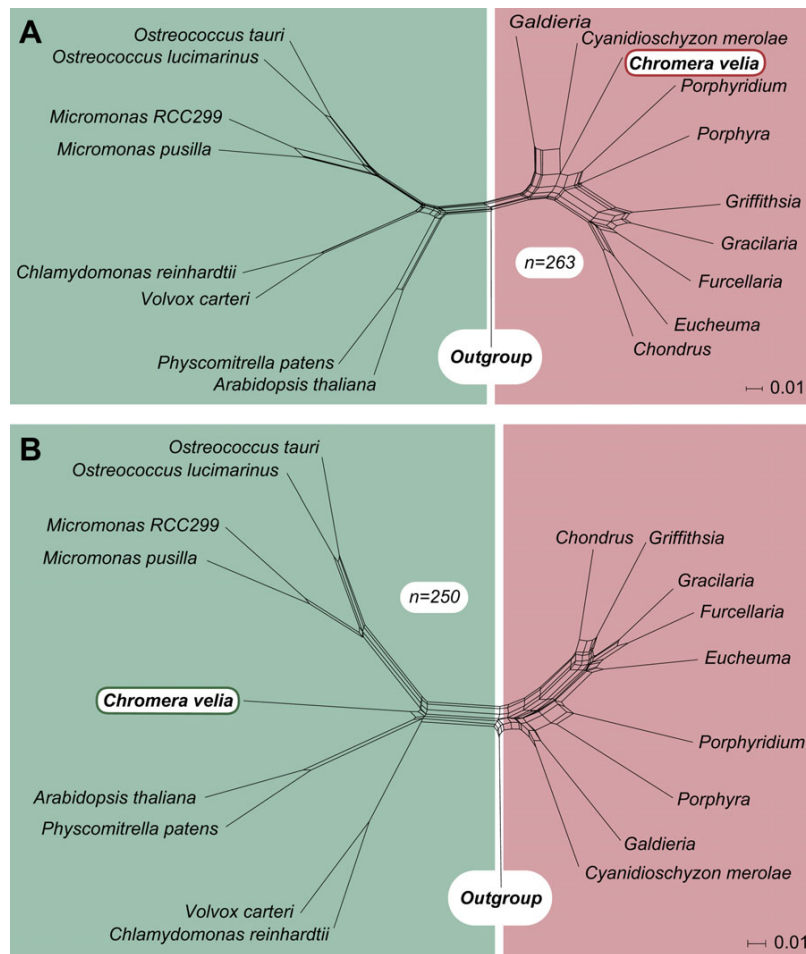


Fig. 4.—Comparison of the red and green signal of nuclear-encoded *Chromera* genes. Five hundred and thirteen phylogenetic trees contained genes of green and red origin and also an outgroup. Those part almost 50–50 into trees, in which the nearest neighbor of the *Chromera velia* homolog is either of red or green origin. In two splits networks combining all the red and green trees separately, the apicomplexan alga is unambiguously positioned among either the rhodophytes (A) or the chloroplastida (B).

Phytophthora had noticed a green signal in their phylogenetic analyses (Armbrust et al. 2004; Tyler et al. 2006; Bowler et al. 2008), but it was not until a genome-wide search for a green signal in diatom genomes, that the idea of a more complex evolutionary scenario specifically involving green endosymbionts was formulated explicitly (Moustafa et al. 2009). The authors claim 16% of the diatoms nuclear genome could be of green origin and suggest that more than 1,700 genes were introduced into the diatom genomes by a green algal endosymbiont preceding the red one.

From our set of *Chromera* clusters, 831 sequences had homologs only to the chlorophytes (chloroplastida/viridiplantae), whereas 176 retrieve a phylogenetic association with only the rhodophytes. As this result could strongly be influenced by the difference in gene sample size available for chlorophytes and rhodophytes, we compared only those

EST-contigs, for which homologs were present in both. From those sequence clusters, we generated 1,053 individual alignments (minimum of 50 amino acid positions) and maximum likelihood trees that contain a red as well as a green homologue, whereby 813 of them—comprising in total 93,745 aligned amino acid positions—furthermore, contained an outgroup. From those with an outgroup, 263 nearest neighbors were of the red, 250 of the green lineage. Furthermore, 55 of the 263 red and 86 of 250 green signals were positioned inside the red or green clades, respectively. A ratio of 1:1 was also found for the nuclear-encoded plastid proteins, where 16 proteins have a red and 16 proteins have a green affiliation. Based on the nearest neighbor trees, we generated two splits networks that reflect the position of *Chromera* within either the red or the green group, which themselves are clearly separated (fig. 4).

Thus, we can confirm the presence of green and red phylogenetic signals in chromalveolate genomes, as found by Moustafa et al. (2009), but the interpretation of that observation becomes very complex. The single origin of a secondary red algal plastid in the common ancestor of haptophytes, heterokontophytes, and cryptophytes and alveolates is rejected by the most recent molecular data (Stiller et al. 2009; Baurain et al. 2010; Felsner et al. 2011) and because the ciliates lack both green and red signals in their nuclear genes, a single origin of the green signal in the common ancestor of diatoms (Moustafa et al. 2009) and *Chromera* (this paper) can be excluded. Thus, if we interpret the green signal as evidence for a symbiosis and gene transfer, then two independent origins of the green signal must be postulated. In the simplest scenarios, this could entail 1) independent secondary symbioses of green algal symbionts in the ancestors of the *Chromera* and diatom lineages followed by replacement of the green plastid with additional independent red secondary symbioses (four secondary symbioses total) or 2) origin of the green signal via secondary symbiosis in a common ancestor of the red plastid donor for the diatom and *Chromera* lineages, in which case these would be tertiary plastids, counter to conventional wisdom (and three symbioses at the minimum are required, two of which entail closely related endosymbionts).

In general, that seems to be quite a bit of symbiosis and gene transfer in parallel, so it is prudent to question the premise that the green signal does in fact represent evidence for a biological event rather than being a manifestation of sampling, random, or other bias in the data. Because the red signal can be readily attributed to the origin of the red plastid, it is the green signal that is suspect, as it is the only reason to entertain the possibility of a large number of inferred symbioses that are otherwise not supported by any independent data. We looked to see if there was a tendency for the green alignments to be shorter, less reliable, or more poorly conserved, such that these factors might generate spurious phylogenetic signal. No such tendency was detected. We looked to see if amino acid content of the green versus red genes was significantly different and again no such tendency was detected (supplementary table 4, Supplementary Material online). We looked to see whether a strong skew existed with respect to functional categories, but we observed none (supplementary table 5, Supplementary Material online). To test for a possible long-branch attraction caused by using only one outgroup sequence, we checked if the tree root is located between the two longest branches in the tree. Long branch attraction was observed in only 10 red and 14 green phylogenies. Furthermore, we tested for differences regarding organism distribution, which were used as an outgroup, and found no significant differences.

Could the green signal both in diatoms and in *Chromera* simply be a random phylogenetic error? This is a possibility. How so? If we go back to Moustafa et al. (2009), what they

reported was a collection of green phylogenetic signals corresponding to diatom nuclear genes that branch with chlorophyte, streptophyte, and prasinophyte homologues. At face value, their data indicated three independent green secondary endosymbiotic events (at least), but the simplest and most reasonable interpretation—and the one that they favored—was that it was in fact only one green event with an endosymbiont (donor) of probably prasinophyte-like phylogenetic identity, whereby the streptophyte and chlorophyte signals represent, by inference, random phylogenetic error. But only one branch removed from the green lineage resides the red lineage. In other words, in the interpretation of Moustafa et al. (2009) regarding diatoms, one green endosymbiosis gave rise to three different green signals, two of which are the result of phylogenetic error (and very implicitly, the later red endosymbiosis for which we have evidence in the form of the plastid gave rise to no error at all). In our current interpretation of the diatom and the *Chromera* data, one red endosymbiosis each gave rise to the red signal in those lineages, but each red signal also contained error, namely all three green signals that Moustafa et al. (2009) observe (not just the two that they assume to be in error). Accordingly, in *Chromera*, the green signal is best interpreted as a phylogenetic error, in toto. Indeed, Moustafa et al. (2009) found about 1,700 green and about 400 red genes (a ratio of 4:1) in diatoms, and in our analysis, with slightly improved sampling, we see a ratio of about 1:1 (250:263). When we performed the same analysis with just the red algal genome of *C. merolae*, as Moustafa et al. (2009) did, the green signal increased and the red signal decreased by about 7% (56% and 44% vs. 49% and 51%). So, the green signal is attributable to sampling. A report by Stiller et al. (2009), which focused on red signals within the organisms having potentially lost their red algal endosymbiont, describes a similar correlation and they conclude: “to move away from a posteriori data interpretations and toward direct tests of explicit predictions from standing and future evolutionary hypothesis.” Hence, we expect that with improved sampling—especially more than one red algal genome available—and with more refined phylogenetic methods, the green signal in both the diatoms and *Chromera* should continue to decline. Whether the green signal is then reduced to nothing more but “background noise” remains to be seen.

In general, the more genes that are investigated to explain the origin of complex plastids, the more conflict is observed in the data (reviewed, e.g., in Gould, Waller, et al. 2008; Sanchez-Puerta and Delwiche 2008; Archibald 2009; Keeling 2010). The more organisms and genomic data are studied, the more apparent it becomes that a monophyletic scenario summarized in the chromalveolate hypothesis—although maybe attractive—must be rejected. The origin of organisms with secondary red plastids might entail similar but nonidentical hosts (that of heterokonts, haptophytes,

and cryptophytes) and similar but nonidentical endosymbionts (that of heterokonts and alveolates). Untangling these branches, keeping random phylogenetic errors in mind, remains a substantial challenge.

Supplementary Material

supplementary information and supplementary tables 1–5 are available at *Genome Biology and Evolution* online (<http://www.gbe.oxfordjournals.org/>).

Acknowledgments

We thank Jan Slapeta for discussing RNA isolation methods. This work is funded by Fit for Excellence (grant number 38700018). S.B.G. was furthermore funded by the Strategischer Forschungsfonds (grant number 3702008), both of the HH-University (HHU) Dusseldorf. Computational support and infrastructure was provided by the Zentrum für Informations- und Medientechnologie of the HHU Dusseldorf.

Literature Cited

- Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans Automat Contr.* 19:716–723.
- Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.
- Archibald JM. 2009. The puzzle of plastid evolution. *Curr Biol.* 19:81–88.
- Armbrust EV, et al. 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* 306:79–86.
- Aurrecochea C, et al. 2007. ApiDB: integrated resources for the apicomplexan bioinformatics resource center. *Nucleic Acids Res.* 35:427–430.
- Bachvaroff TR, Puerta MVS, Delwiche CF. 2005. Chlorophyll c-containing plastid relationships based on analyses of a multigene data set with all four chromalveolate lineages. *Mol Evol Biol.* 22:1772–1782.
- Baurain D, et al. 2010. Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. *Mol Evol Biol.* 27:1698–1709.
- Becker B, Hoef-Emden K, Melkonian M. 2008. Chlamydial genes shed light on the evolution of photoautotrophic eukaryotes. *BMC Evol Biol.* 8:203.
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2009. GenBank. *Nucleic Acids Res.* 37:26–31.
- Bodl A. 2005. Do plastid-related characters support the chromalveolate hypothesis? *J Phycol.* 41:712–719.
- Bodl A, Stiller JW, Mackiewicz P. 2009. Chromalveolate plastids: direct descent or multiple endosymbioses? *Trends Ecol Evol.* 24:119–121.
- Boeckmann B, et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31:365–370.
- Bowler C, et al. 2008. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* 456:239–244.
- Cavalier-Smith T. 1991. Cell diversification in heterotrophic flagellates. In: Patterson D, Larsen J, editors. *The biology of free-living heterotrophic flagellates*. Oxford: Clarendon Press. p. 113–131.
- Cavalier-Smith T. 1999. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J Eukaryot Microbiol.* 46:347–366.
- Cock J, et al. 2010. The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature* 465:617–621.
- Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res.* 14:1188–1190.
- Dagan T, Martin W. 2009. Seeing green and red in diatom genomes. *Science* 324:1651–1652.
- Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27:1164–1175.
- Eisen JA, et al. 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol.* 4:e286.
- Emanuelsson O, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc.* 2:953–971.
- Felsner G, et al. 2011. ERAD components in organisms with complex red plastids suggest recruitment of a preexisting protein transport pathway for the periplastid membrane. *Genome Biol Evol.* 3:140–150.
- Frommolt R, et al. 2008. Ancient recruitment by chromists of green algal genes encoding enzymes for carotenoid biosynthesis. *Mol Biol Evol.* 25:2653–2667.
- Gardner PP, et al. 2009. Rfam: updates to the RNA families database. *Nucleic Acids Res.* 37:136–140.
- Gould SB, Sommer MS, Hadfi K, et al. 2006. Protein targeting into the complex plastid of cryptophytes. *J Mol Evol.* 62:674–681.
- Gould SB, Sommer MS, Kroth PG, et al. 2006. Nucleus-to-nucleus gene transfer and protein retargeting into a remnant cytoplasm of cryptophytes and diatoms. *Mol Biol Evol.* 23:2413–2422.
- Gould SB, Tham WH, Cowman AF, McFadden GI, Waller RF. 2008. Alveolins, a new family of cortical proteins that define the protist infrakingdom Alveolata. *Mol Biol Evol.* 25:1219–1230.
- Gould SB, Waller RF, McFadden GI. 2008. Plastid evolution. *Annu Rev Plant Biol.* 59:491–517.
- Gruber A, et al. 2007. Protein targeting into complex diatom plastids: functional characterisation of a specific targeting motif. *Plant Mol Biol.* 64:519–530.
- Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.
- Harper JT, Waanders E, Keeling PJ. 2005. On the monophyly of chromalveolates using a six-protein phylogeny of eukaryotes. *Int J Syst Evol Microbiol.* 55:487–496.
- Häuber MM, Müller SB, Speth V, Maier UG. 1994. How to evolve a complex plastid?—A hypothesis. *Bot Acta.* 107:383–386.
- Huang J, et al. 2004. Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum*. *Genome Biol.* 5:R88.
- Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol.* 23:254–267.
- Janouskovec J, Horak A, Obornik M, Lukes J, Keeling PJ. 2010. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proc Natl Acad Sci U S A.* 107:10949–10954.
- Jomaa H, et al. 1999. Inhibitors of the non-mevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science* 285:1573–1576.
- Junier T, Zdobnov EM. 2010. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* 26:1669–1670.
- Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform.* 9:286–298.

- Keeling PJ. 2010. The endosymbiotic origin, diversification and fate of plastids. *Philos Trans R Soc Lond B Biol Sci.* 365:729–748.
- Lim L, Kalanon M, McFadden GI. 2009. New proteins in the apicoplast membranes: time to rethink apicoplast protein targeting. *Trends Parasitol.* 25:197–200.
- Martens C, Vandepoele K, Van de Peer Y. 2008. Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species. *Proc Natl Acad Sci U S A.* 105:3427–3432.
- Matsuzaki M, et al. 2004. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature* 428:653–657.
- Matsuzaki M, et al. 2009. A DNA-lacking plastid in the oyster pathogen *Perkinsus marinus*. *Phycologia* 48:82–83.
- McFadden GI. 2010. The apicoplast. *Protoplasma.* doi: 10.1007/s00709-010-0250-5.
- Min XJ, Butler G, Storms R, Tsang A. 2005. OrfPredictor: predicting protein-coding regions in EST-derived sequences. *Nucleic Acids Res.* 33:677–680.
- Moog D, Stork S, Zauner S, Maier UG. 2011. In silico and in vivo investigations of proteins of a minimized eukaryotic cytoplasm. *Genome Biol Evol.* 3:375–382.
- Moore RB, et al. 2008. A photosynthetic alveolate closely related to apicomplexan parasites. *Nature* 451:959–963.
- Moriya Y, Itoh M, Okuda S, Yoshizawa A, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 35:182–185.
- Moustafa A, et al. 2009. Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* 324:1724–1726.
- Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol.* 48:443–453.
- Obornik M, et al. 2011. Morphology and ultrastructure of multiple life cycle stages of the photosynthetic relative of Apicomplexa, *Chromera velia*. *Protist* 162:115–130.
- Patron NJ, Inagaki Y, Keeling PJ. 2007. Multiple gene phylogenies support the monophyly of cryptomonad and haptophyte host lineages. *Curr Biol.* 17:887–891.
- Patron NJ, Waller RF. 2007. Transit peptide diversity and divergence: a global analysis of plastid targeting signals. *Bioessays* 29:1048–1058.
- Penn O, Privman E, Landan G, Graur D, Pupko T. 2010. An alignment confidence score capturing robustness to guide-tree uncertainty. *Mol Biol Evol.* 27:1759–1767.
- Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35:61–65.
- Ralph SA, et al. 2004. Tropical infectious diseases: metabolic maps and functions of the *Plasmodium falciparum* apicoplast. *Nat Rev Microbiol.* 2:203–216.
- Reyes-Prieto A, Moustafa A, Bhattacharya D. 2008. Multiple genes of apparent algal origin suggest ciliates may once have been photosynthetic. *Curr Biol.* 18:956–962.
- Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. *Trends Genet.* 16:276–277.
- Sanchez-Puerta MV, Delwiche CF. 2008. A hypothesis for plastid evolution in chromalveolates. *J Phycol.* 44:1097–1107.
- Seeber F. 2002. Biogenesis of iron-sulphur clusters in amitochondriate and apicomplexan protists. *Int J Parasitol.* 32:1207–1217.
- Sommer MS, et al. 2007. Der1-mediated preprotein import into the periplastid compartment of chromalveolates? *Mol Biol Evol.* 24:918–928.
- Spork S, et al. 2009. An unusual ERAD-like complex is targeted to the apicoplast of *Plasmodium falciparum*. *Eukaryot Cell.* 8:1134–1145.
- Stiller JW, Huang J, Ding Q, Tian J, Goodwillie C. 2009. Are algal genes in nonphotosynthetic protists evidence of historical plastid endosymbioses? *BMC Genomics.* 10:484.
- Stoebe B, Maier UG. 2002. One, two, three: nature's tool box for building plastids. *Protoplasma* 219:123–130.
- Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. *Science* 278:631–637.
- Tonkin CJ, et al. 2008. Evolution of malaria parasite plastid targeting sequences. *Proc Natl Acad Sci U S A.* 105:4781–4785.
- Tyler BM, et al. 2006. Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* 313:1261–1266.
- Waller RF, et al. 1998. Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum*. *Proc Natl Acad Sci U S A.* 95:12352–12357.
- Weber AP, et al. 2004. EST-analysis of the thermo-acidophilic red microalga *Galdieria sulphuraria* reveals potential for lipid A biosynthesis and unveils the pathway of carbon export from rhodoplasts. *Plant Mol Biol.* 55:17–32.
- Weizhong L, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659.
- Zhang H, Campbell DA, Sturm NR, Dungan CF, Lin S. 2011. Spliced leader RNAs, mitochondrial gene frameshifts and multi-protein phylogeny expand support for the genus *Perkinsus* as a unique group of alveolates. *PLoS One.* 6:e19933.

Associate editor: John Archibald

4.2 The actin-based machinery of *Trichomonas vaginalis* mediates flagellate-amoeboid transition and migration across host tissue

Gary Kusdian, Christian Woehle, William F. Martin & Sven B. Gould

Der vorliegende Artikel wurde 2013 in der Fachzeitschrift *Cellular Microbiology* (Impact Factor 4,8) veröffentlicht.

Ergänzendes Material steht zur Verfügung durch die Webseiten des Verlegers⁷.

Beitrag von Christian Wöhle, Zweitautor:

Versuchsplanung	10 %
Datenanalyse	30 %
Verfassen des Manuskripts	10 %

⁷<http://onlinelibrary.wiley.com/doi/10.1111/cmi.12144/supinfo>

The actin-based machinery of *Trichomonas vaginalis* mediates flagellate-amoeboid transition and migration across host tissue

Gary Kusdian, Christian Woehle, William F. Martin and Sven B. Gould*

Institute for Molecular Evolution,
Heinrich-Heine-University Düsseldorf, 40225 Düsseldorf,
Germany.

Summary

Trichomonas vaginalis is the most widespread non-viral pathogen of the human urogenital tract, infecting ~3% of the world's population annually. At the onset of infection the protist changes morphology within minutes: the flagellated free-swimming cell converts into the amoeboid-adherent stage. The molecular machinery of this process is not well studied, but is thought to involve actin reorganization. We have characterized amoeboid transition, focusing in particular on TvFim1, the only expressed protein of the fimbrin family in *Trichomonas*. Addition of TvFim1 to actin polymerization assays increases the speed of actin filament assembly and results in bundling of F-actin in a parallel and anti-parallel manner. Upon contact with vaginal epithelial cells, the otherwise diffuse localization of actin and TvFim1 changes dramatically. In the amoeboid TvFim1 associates with fibrous actin bundles and concentrates at protrusive structures opposing the trailing ends of the gliding amoeboid form and rapidly redistributes together with actin to form distinct clusters. Live cell imaging demonstrates that *Trichomonas* amoeboid stages do not just adhere to host tissue, rather they actively migrate across human epithelial cells. They do so in a concerted manner, with an average speed of 20 $\mu\text{m min}^{-1}$ and often using their flagella and apical tip as the leading edge.

Introduction

Trichomonas vaginalis is the most widespread non-viral and sexually transmitted human parasite worldwide (Benchimol, 2004). Its genome encodes approximately

60 000 proteins, at least twice as many as its mammalian host (Carlton *et al.*, 2007). Of the many hundred million *T. vaginalis* infections that occur annually, the majority is asymptomatic, while a minority results in trichomoniasis with urogenital tract swelling and inflammatory discharge. However, asymptomatic infections decrease fertility, increase the risk of acquiring HIV and elevate the risk of prostate and cervical cancer (Petrin *et al.*, 1998; Stark *et al.*, 2009; Ryan *et al.*, 2011). Trichomoniasis is commonly treated with nitroimidazole derivatives, but resistant strains are on the rise (Kulda, 1999; Upcroft and Upcroft, 2001; Benchimol, 2008).

As an essential component of *T. vaginalis* urogenital tract colonization, minutes after contact with human urogenital tract cells, *T. vaginalis* undergoes a radical change in cell morphology (Lal *et al.*, 2006). The free-swimming flagellate cell flattens and spreads out over the host cell tissue to become an amoeba. This ability to shift from a flagellate-motile to an adherent-amoeboid cell stage is rare among protists and, outside the trichomonads, only known among a few species such as *Naegleria gruberi* and *Physarum polycephalum* (Fritz-Laylin *et al.*, 2010a; Ryan *et al.*, 2011). With only a few minutes to complete, this amoeboid transition is the fastest described and far more rapid than the *N. gruberi* transformation, which can require more than an hour (Fulton, 1993; Fritz-Laylin *et al.*, 2010b).

Little is known about the molecular machinery behind amoeboid transition, how it is orchestrated or the components involved. Clues come however from the circumstance that the amoeboid transition connects the two types of cellular locomotion known among eukaryotes: tubulin-based flagellar swimming and actin-based amoeboid gliding. Actin and tubulin are the core components of these locomotion types and of the eukaryotic cytoskeleton in general. Although the proteins associated with these two types of locomotion are ubiquitous among eukaryotic genomes, few species exhibit both types during their life cycle (Fig. 1). In amoeba and apicomplexan parasites, the actin-based cytoskeleton is the key component of gliding locomotion (Fukui, 2002; Baum *et al.*, 2008). The apicomplexan machinery is known as the glideosome and becomes active when the parasites need to overcome biological barriers or invade new host cells (Santos *et al.*, 2009). Previous reports localized the actin-binding

Received 27 December, 2012; revised 16 March, 2013; accepted 20 March, 2013. *For correspondence. E-mail gould@hhu.de; Tel. (+49) 211 811 3983; Fax (+49) 211 811 3554.

© 2013 John Wiley & Sons Ltd

includes genes for capZ, cofilin, formin and the Arp2/3 complex, in addition to at least 29 genes (five orthologous groups) encoding proteins of the actin family itself. Overall, all actins share a 40% global sequence identity and expression evidence exists for almost all of them at TrichDB (Aurrecochea *et al.*, 2009). The major actin-bundling protein of filopodia, fascin (Jansen *et al.*, 2011), appears absent from the *T. vaginalis* genome (Fig. 1). The parasite encodes two putative proteins of the fimbrin family; *TvFim1* (TVAG_351310) and *TvFim2* (TVAG_116370). Both contain two tandem actin-binding domains (ABDs), but only *TvFim1* contains the N-terminal Ca⁺⁺-binding EF-hand, typical for proteins of the fimbrin/plastin family (Korenbaum and Rivero, 2002). Compared with *TvFim1* and the human homologue *HsFim1* (L-Plastin), *TvFim2* appears to have experienced a 5' truncation, which extends into the N-terminal region of the first CH domain (Fig. S1B). Overall, the two *Trichomonas* fimbrins share a sequence identity of 74%, which increases to 82% when considering only the region containing the two ABDs.

At TrichDB we identified 145 expressed sequence tags (ESTs) for *TvFim1*, but none for *TvFim2*, despite the ESTs originating from mRNA obtained under different conditions such as normal and low-glucose culture conditions or fibronectin-mediated cytoadherence. To confirm we performed reverse-transcriptase PCR on isolated RNA from exponentially growing cells of *T. vaginalis*, both flagellate and amoeboid stage, and were only able to demonstrate expression of *TvFim1* (Fig. 2A). To determine whether the only expressed fimbrin gene is upregulated upon contact with host tissue we performed quantitative real-time PCR on RNA isolated from the free-swimming stage and at 5, 20 and 60 min after fibronectin induced morphogenesis. In two individual experiments, each using biological and technical triplicates, we observed no significant upregulation of *TvFim1* over the first hour compared with the slightly upregulated 40S ribosomal protein S5 (Fig. 2B). Hence, the parasite expresses only its full-length fimbrin copy and this gene does not appear to experience a significant upregulation upon contact with host tissue.

Overall, *TvFim1* shares a global sequence identity with the human homologue *HsFim1* of 43%. An alignment of *TvFim1* and its homologues from a range of eukaryotes shows that a high amount of conservation is found among the actin binding CH domains themselves (Fig. S1A). The amount of conservation of the individual amino acids thought to interact with actin (Klein *et al.*, 2004; Galkin *et al.*, 2008) is comparable to that of other fimbrins. Reconstructing the interaction of the actin-binding domains of *TvFim1* with actin using the known crystal structure of the *Schizosaccharomyces pombe* fimbrin protein (PDB accession 1RT8) and the F-actin-fimbrin

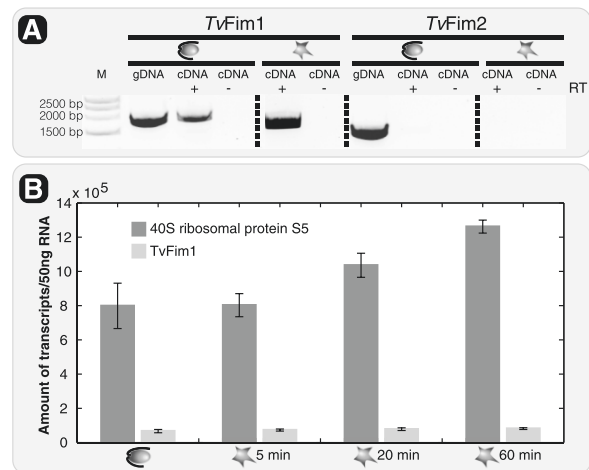


Fig. 2. A. Reverse transcriptase (RT) PCR on complementary DNA (cDNA) generated from motile and adherent cells (indicated by symbols as in Fig. 1) in the presence (+ RT) or, as control, absence of the reverse transcriptase enzyme (- RT), demonstrates the expression of *TvFim1* only. Control PCR was performed on genomic DNA (gDNA). B. Quantitative real-time PCR shows *TvFim1* not to be significantly upregulated during infection compared with a copy of the ribosomal protein S5 (TVAG_158720).

ABD2 complex (PDB accession 3BYH), confirms the alignment sequence identity. The superposition of the two fimbrins of *S. pombe* and *T. vaginalis* reveals no obvious difference and interacting residues within the CH3 and CH4 domain, such as E448, E523 or F602 (positions based on the alignment of Fig. S1A) are well conserved (Fig. 3).

TvFim1 increases the actin polymerization rate and bundles F-actin

Previous work on human and yeast fimbrins showed this protein family to bundle actin filaments (Namba *et al.*, 1992; Prassler *et al.*, 1997; Skau *et al.*, 2011). To verify *TvFim1*'s function as a fimbrin protein, and analyse the potential influence of *TvFim1* on actin polymerization, we heterologously overexpressed a full-length C-terminally HIS-tagged version of *TvFim1* in *Escherichia coli* and purified the protein using fast protein liquid chromatography (Fig. S2). All attempts to purify *T. vaginalis* actin failed to deliver sufficient quantities for *in vitro* studies, and we hence used actin purified from rabbit muscle and the amoeboid protist *Acanthamoeba* instead. We measured assembly kinetics of pyrene-labelled actin by fluorescence spectrophotometry and performed total internal reflection fluorescence microscopy (TIRF).

The pyrene assay showed that only *Acanthamoeba* actin polymerization significantly increased in the presence of *TvFim1* and saturation of polymerization was

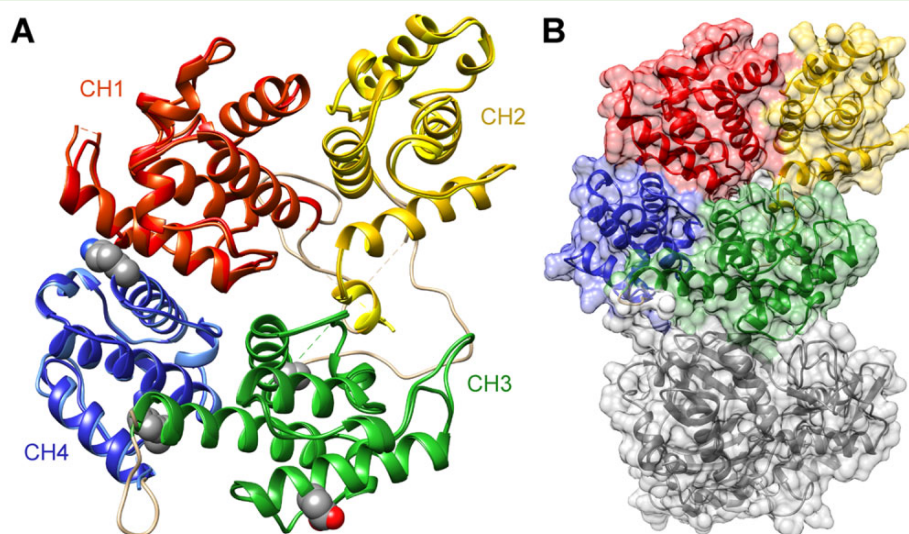


Fig. 3. Tertiary structure prediction of *TvFim1*.

A. Ribbon representation of the predicted tertiary conformation of *TvFim1* (darker colours) superimposed onto the known crystal structure of the *S. pombe* fimbrin (lighter colours). Four exemplary amino acids that have been implicated to interact with F-actin are highlighted (grey, sphere representation) in the yeast sequence.

B. Semitransparent surface representation of *TvFim1* (same colour code as in A) showing the interaction of the two CH domains 3 and 4 with an actin of *T. vaginalis* (grey; TVAG_337240). For details please refer to the text.

reached several minutes earlier (Fig. 4A). *TvFim1* seems to stimulate the rate of *Acanthamoeba* actin assembly, but with no apparent polymerization effect on rabbit actin. However, TIRF imaging performed with rabbit actin demonstrated an intense parallel and anti-parallel bundling of actin filaments through *TvFim1* (Fig. 4B). Dense bundles of many individual filaments were observed that continued to elongate at both ends of the bundles (Fig. 4C). We did not notice a favoured orientation of the aligned filaments, which is consistent with the bundling behaviour observed for yeast fimbrin (Skau *et al.*, 2011).

Fimbrin and actin dynamics upon contact with host tissue

We generated a polyclonal peptide antibody directed against an epitope derived from the first ABD. The antibody detects a single band migrating below the 70 kDa marker lane, fitting well with the predicted mass of *TvFim1* of approximately 68 kDa (Fig. S3A and B). Additionally we generated *Trichomonas* clones expressing haemagglutinin (HA) and green fluorescent protein (GFP)-tagged *TvFim1* to further validate antibody specificity, avoid potential cross-talk of our fimbrin antibodies with the host cells and perform live imaging. In protein extract from cells expressing C-terminally HA-tagged *TvFim1*, a band migrating slightly slower compared with the endogenous copy – due to the HA-tag – was identified plus two faster

migrating bands indicating putative proteolytic processing of the fusion protein (Fig. S3B).

In the free-swimming stage, *TvFim1* localized to the cell periphery in a gradient-like manner and did not distinctly colocalize with actin (Fig. 5A–C). Identical localization was observed for the endogenously expressed protein, as well as for the HA-tagged copy, indicating that the C-terminal HA-tag did not influence the fusion-proteins localization (Fig. S3D). In adherent-amoeboid cells, which were induced through the exposure to a monolayer of vaginal epithelial cells (VECs), the observed localization pattern of *TvFim1* changed dramatically. *TvFim1* now begins to colocalize more specifically with actin, also in the still pyriform stage (Fig. 5E–G), but is predominantly observed to form clusters subtending the plasma membrane (Fig. 5M–O) and to associate with filamentous arrays (Figs 5I–K and S4E–G). We further observed filamentous structures (Figs 5I–K and S4E–G) similar to ‘fibrous arrays’ previously observed during the analyses of a coronin homologue in *Trichomonas* (Bricheux *et al.*, 2000). Live imaging shows that the peripheral clusters at the migration front actually correspond to waves of fimbrin (Fig. 6). Specifically, at the contact sites clusters were observed and further around a contractile ring-like structure of a daughter cell budding from a multinuclear cell.

The time-lapse videos allowed to document different motion patterns of *T. vaginalis*, most importantly amoeboid migration across host tissue, only minutes upon

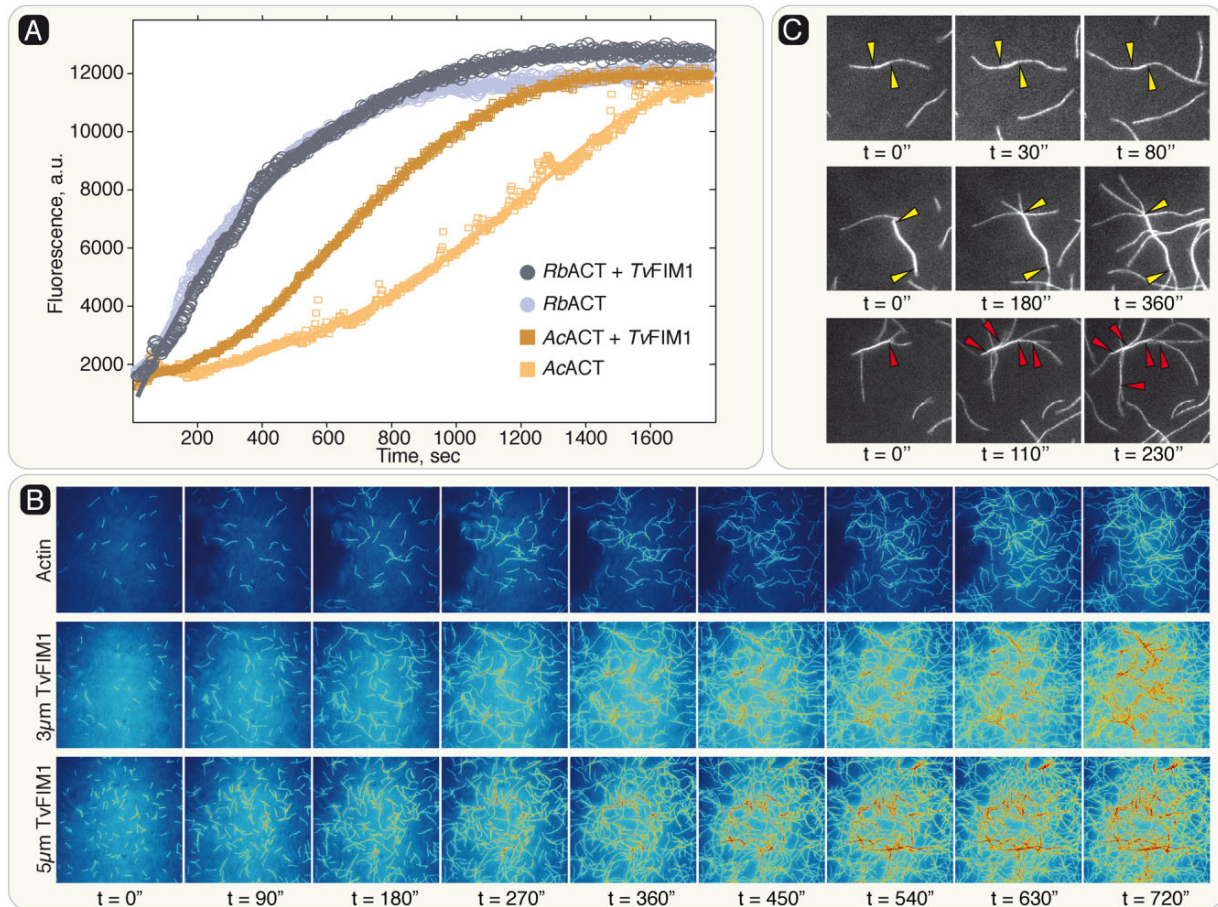


Fig. 4. Effect of TvFim1 on *in vitro* actin polymerization.

A. Polymerization assays of pyrene-labelled *Acanthamoeba* (AcACT) and rabbit actin (RbACT) in the presence and absence of 4 μ M TvFim1. Whereas no noticeable difference was observed for the polymerization rate of rabbit actin, the polymerization rate of the protozoan *Acanthamoeba* actin was increased.

B. Rabbit muscle actin polymerization was monitored in the absence (Actin) and presence (3 and 5 μ M) of purified TvFim1 using total internal reflection fluorescence microscopy and revealed a higher polymerization rate, as well as a strong parallel and anti-parallel bundling of actin filaments (see also Movies S2 and S3).

C. Details of F-actin bundled by TvFim1. Parallel and anti-parallel bundling is revealed by bundled filaments continuing to polymerize on both ends, marked by yellow arrow heads and at many branching points, marked by red arrow heads.

exposure to VECs (Movies S7–S10). *T. vaginalis* actively roams across the VEC-monolayer with an average speed of 20.2 μ m min⁻¹ (Table S3) and appears to use its flagella and apical tip as the guiding end (Fig. S5), which also appears to downright penetrate areas of adjacent and adherent VECs. One hour post infection we observed some amoeboid cells to massively increase their cell mass and adherent surface (Fig. S5); a process we refer to as juggernauting. The cells display multiple flagella pockets and nuclei (Figs S5 and S4K), which is consistent with previous observations (Yusof and Kumar, 2011). We observed individual *T. vaginalis* cells rapidly budding off from multinuclear cells, similar to what was described for *Tritrichomonas foetus* (Pereira-Neves and Benchimol,

2009). Overall, the cells appear highly agile while they migrate and scavenge from host tissue.

To determine whether the association of fimbrin and actin is specific for phenotypic plasticity induced through VECs, we investigated what happens during phagocytosis of *Saccharomyces*. Exponentially growing parasites were exposed to yeast cells in a 1:50 parasite/yeast ratio and incubated for 60 min in the CO₂ incubator – to be comparable to the other experiment – before fixation. We observed *T. vaginalis* to incorporate entire and multiple yeast cells at the same time, as previously documented (Pereira-Neves and Benchimol, 2007). As for the amoeboid form of the parasite, fimbrin now clearly colocalizes with actin, forming distinct rings around the large

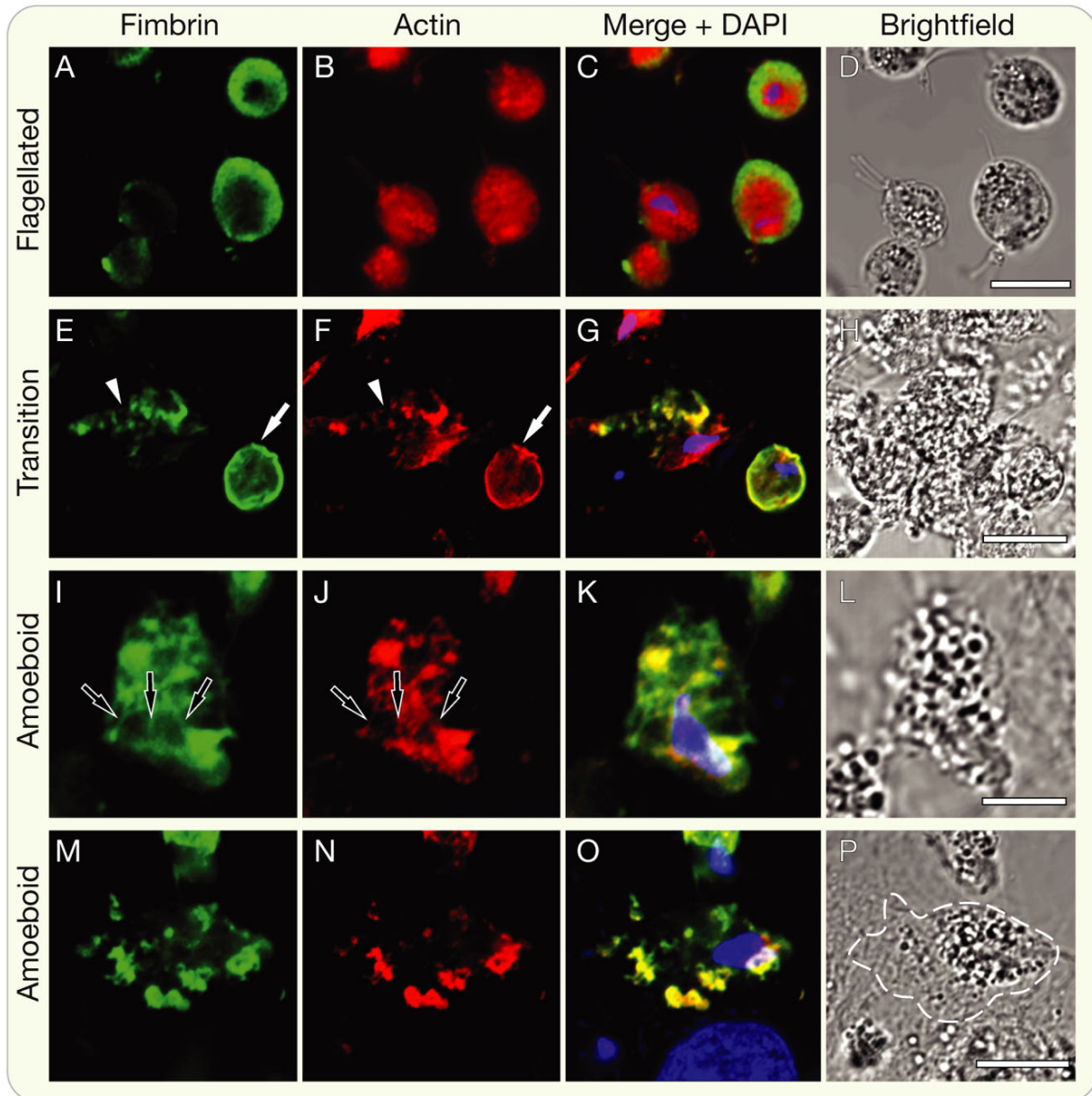


Fig. 5. Immunolocalization comparison of *TvFim1* in free-swimming and adherent-amoeboid cells. In flagellated motile cells fimbrin localizes to the periphery of the cells in a gradient-like manner with no obvious actin colocalization (A–D). In cells exposed to vaginal epithelial cells (E–P) fimbrin colocalizes with actin, associates with structures reminiscent of actin cables (arrows in I and J) and shows peripheral clustering together with actin (M–O). In (E–G) a still pyriform cell can be seen, in which actin and fimbrin already localize in a sharp ring at the cells periphery (arrow) and next to a fully adherent cell with a very different labelling pattern (arrow head). The dashed line in (P) outlines the parasite based on the fluorescent channels of fimbrin and actin. Scale: 10 μm .

phagocytic vesicle, but at the same time highlights a range of other subcellular structures, including some punctuate staining and again fibrous patterns (Fig. 7).

Discussion

Through the characterization of the fimbrin protein family and live cell imaging of *T. vaginalis* we provide detailed

evidence that the parasite's amoeboid morphogenesis during infection is accompanied by a rapid reorganization of the actin cytoskeleton. The flagellate to amoeboid transformation occurs several times faster in *T. vaginalis* than in *N. gruberi* (Arroyo *et al.*, 1993; Fritz-Laylin *et al.*, 2010a). We suggest that this might be attributable to the prey-induced activation of 'sleeping' components, such as

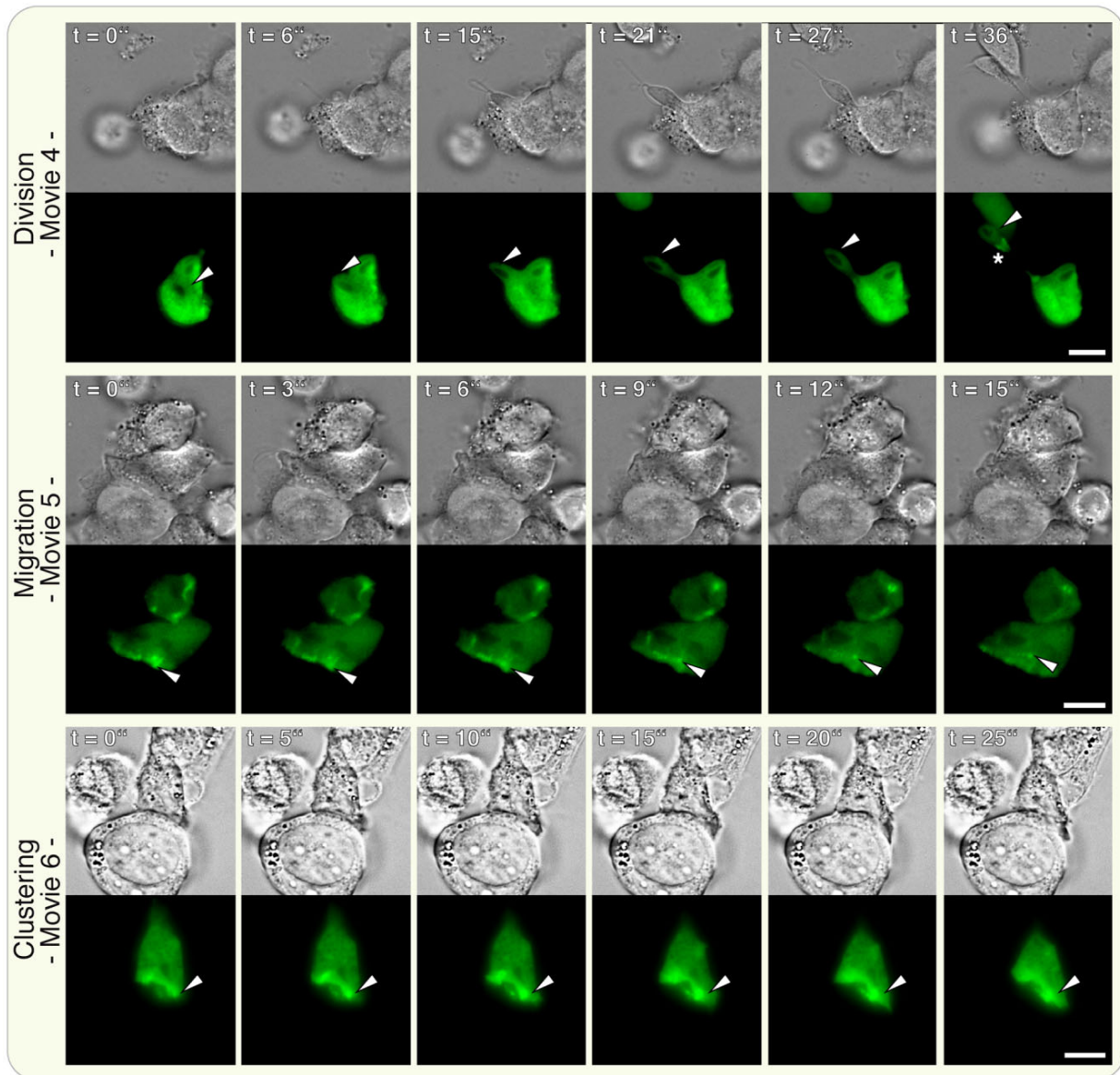


Fig. 6. Live imaging of *TvFim1::GFP* on host tissue. In the top panel a daughter cell can be seen to bud from the mother cell. *TvFim1::GFP* concentrates around a contractile ring-like structure (indicated by an asterisk) towards the basal end of the daughter cell during separation. During amoeboid migration across host tissue individual waves of *TvFim1::GFP* were clearly visible (wave indicated by arrow heads in the central panel). Clustering of *TvFim1::GFP* was predominantly found to occur around the initial host contact sites (indicated by arrow heads in the bottom panel). Images were taken 15 min (for clustering) and 70 min (for division and migration) after inoculation of host tissue with *T. vaginalis* respectively. Scale: 10 μm .

the *T. vaginalis* fimbrin protein *TvFim1*. Accordingly, the abundance and gradient-like localization of the actin-binding protein below the plasma membrane in the flagellate form (Fig. 5A–C) appears to be a prearrangement that keeps the free-swimming parasite primed for actin cytoskeleton mediated phenotypic plasticity upon contact with host tissue. With the molecular machinery required already in place, only the final signal at the end of the

activation cascade (such as calcium depletion or an altered phosphorylation state of the protein; see below) is required to trigger morphogenesis. This would explain why no significant upregulation of *TvFim1* during infection is detected, when compared with a copy of the 40S ribosomal protein S5. By contrast, upregulation of alpha-actinin using semi-quantitative PCR has been reported to occur during infection (Noël *et al.*, 2010), whereby expression

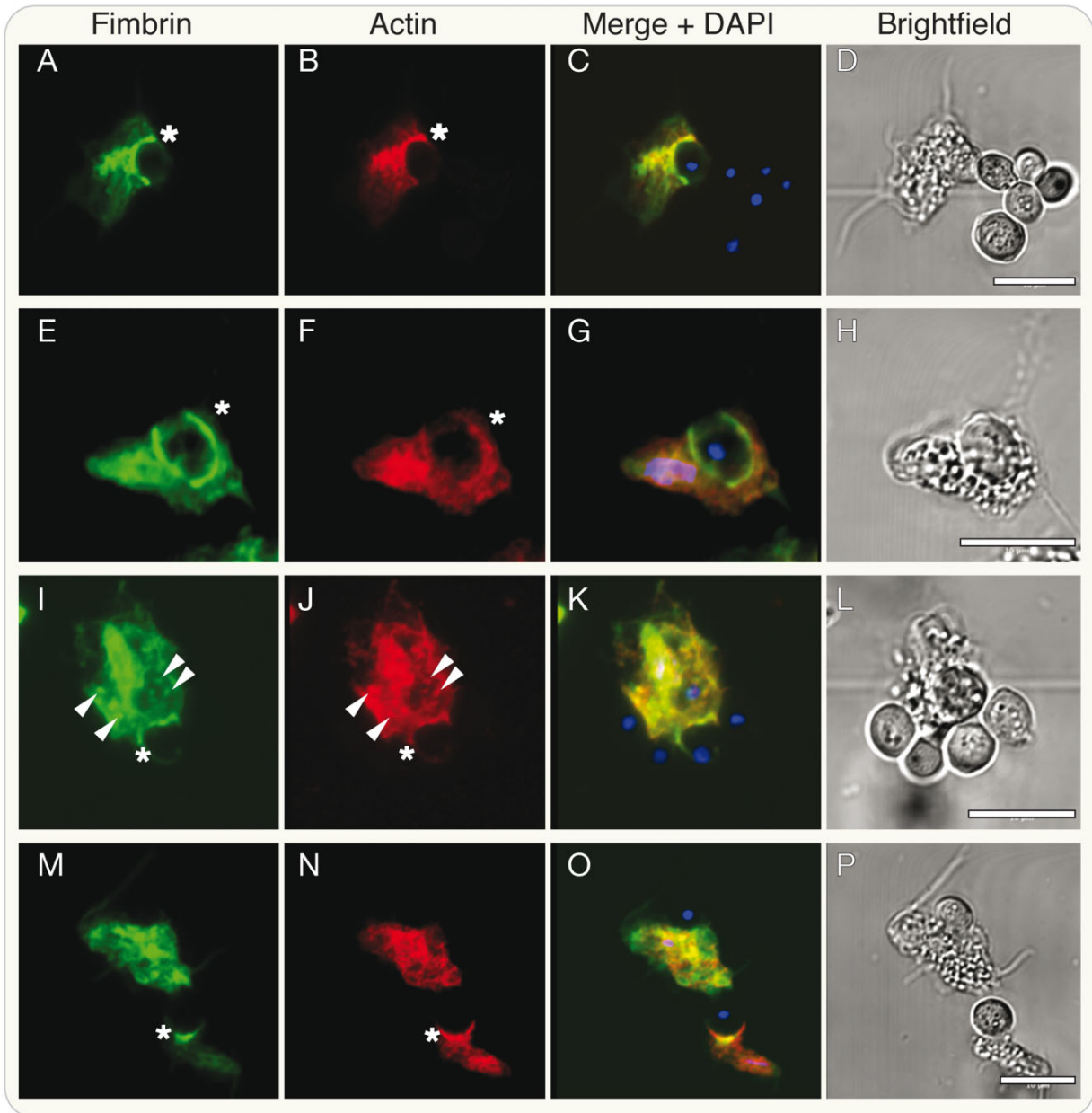


Fig. 7. Fimbrin and actin localization during phagocytosis. During phagocytosis of yeast cells actin and its bundling protein fimbrin colocalize and predominantly cluster around bordering phagocytic vesicles (asterisks). Other structures, in parts similar to those observed upon contact with host tissue, are also apparent, including smaller vesicular particles (arrow heads). Scale: 10 μ m.

analysis in *Trichomonas* is in general impaired by the many gene duplicates present.

The majority of known core components of actin regulation are encoded by *Trichomonas* (Fig. 1), and some gene families have been significantly expanded in the course of genome duplication (Carlton *et al.*, 2007). Paralogous copies of many genes are present and simultaneously expressed. Proteins of the *Trichomonas* actin

family are encoded in 29 copies and expression evidence exists for 24 of them at TrichDB. Only two genes encode proteins of the actin-bundling fimbrin family (TVAG_351310 and TVAG_116370) and the second, TvFim2, appears to be a pseudogene, a 5' truncated paralogous copy of TvFim1. During or after gene duplication, TvFim2 lost the N-terminal EF-hand – the characteristic and regulatory module of fimbrin proteins – and with

it its function. We could neither detect expression of TvFim2 on RNA level through RT-PCR, supporting the lack of expression evidence on TrichDB, nor on protein level through Western blot analysis using our fimbrin antibody. The epitope of TvFim1 differs from that of TvFim2 by only one amino acid at position 185, which suggests the antibody should recognize both proteins. However, only one band of approximately 68 kDa and none of 50 kDa was detected (Fig. S3), corresponding to the predicted sizes of TvFim1 and TvFim2 respectively.

TvFim1 increases the polymerization rate of free *Acanthamoeba* actin, albeit weakly, and served as a potent filament cross-linker, bundling rabbit F-actin in a parallel and antiparallel manner (Fig. 4). Generally, fimbrin is not considered to be a canonical actin nucleator, such as formin or the ARP complex (Butler and Cooper, 2009). Our results are not yet able to determine the exact actin polymerization potency of TvFim1; however, we can exclude 'false positive polymerization' observed through scattering due to the settings in the pyrene assays. Experiments with purified *Trichomonas* actin are currently hindered, because it does not express well in any system tested. How, specifically, fimbrin increases actin polymerization is unknown for any model system, but independent experiments on homologues from yeast and *Arabidopsis* suggest that fimbrin might lower the necessary critical concentration for actin to polymerize (Cheng *et al.*, 1999; Kovar *et al.*, 2000). Intriguingly, only the polymerization of *Acanthamoeba* G-actin was significantly increased, but not of rabbit G-actin, although rabbit F-actin was clearly bundled by TvFim1. Amino acid residues that have been implicated in the interaction with actin are as well conserved in *Trichomonas* as in other eukaryotes (Fig. S1). Furthermore, the predicted tertiary structure of TvFim1 is almost identical to that of yeast (Fig. 3), and thereby fails to offer obvious clues as to the structural basis of the differential behaviour of TvFim1 to the different actins investigated here. However, our results provide further indirect evidence that fimbrin might be a weak nucleator, too, as the polymerization increase observed was not just due to a side-effect of the bundling of filamentous actin.

Hence, *T. vaginalis* expresses only one protein of the fimbrin family that, as shown here, fulfils all functions attributed to the fimbrin family and relocates during morphogenesis and phagocytosis. A knockout or knockdown of TvFim1 to underpin its importance during infection was at this point not successful. Knockouts could not be established – which only twice has been successful for *T. vaginalis* and only once reported to reveal a noticeable phenotype (Land *et al.*, 2004; Pereira-Brás *et al.*, 2013) – and expressing TvFim1 antisense RNA did not lead to a decrease of TvFim1 transcript (Fig. S7). Yet, modifications of fimbrin seemed to influence the transfected cells:

(i) the HA-tagged line duplicated with only about half the speed, (ii) in a HaloTag line we could not detect any protein, albeit the construct was identified on RNA level (Fig. S6), suggesting either transcriptional inhibition or immediate post-translational degradation, (iii) the HA-tagged copy was also observed to show a degradation product not observed for the endogenous copy (Fig. S3B). In summary, fimbrin seems to fulfil pleiotropic roles, as indicated by its colocalization with actin during adherence to and migration across host cells and during phagocytosis of yeast, and a knockout or knockdown appears to influence the parasites viability.

The behaviour of TvFim1 during infection suggests the protein to be predominantly active during the amoeboid stage and most likely during phase transition from flagellate to the amoeboid. Clustering, specific aggregation and more defined localization of the protein were only observed in the adherent-amoeboid form of *Trichomonas* (Figs 5, 6 and S4) or during phagocytosis (Fig. 7). The signals that trigger morphogenesis and surface attachment are unknown, but TvFim1/actin re-localization and TvFim1-dependent actin polymerization provide new tools towards their investigation. *Trichomonas* can spontaneously adhere to glass and plastic surfaces and the amount of adherent cells increases when the surfaces are coated with fibronectin (Brugerolle *et al.*, 1996). Only on VECs however have we observed that the vast majority of parasite cells undergo transition from a motile to an adherent stage over time. This suggests multiple signals and a complex cascade involving the correct recognition of host tissue. This is further supported by the fact that the parasite is also able to completely engulf and phagocytose other eukaryotes such as yeast – a process also involving a co-ordinated, but different re-localization of actin and fimbrin (Fig. 7). Phagocytosis must hence require a different downstream behaviour after prey recognition in comparison to adhering to VECs, and further unknown regulatory units orchestrating the different processes. Whether the activation of TvFim1 requires the EF-hand to not be bound to Ca⁺⁺ (Namba *et al.*, 1992; Prassler *et al.*, 1997), or whether it is regulated through the phosphorylation of certain serine residues as shown for the human L-plastin (Shinomiya, 2012), remains to be investigated. TvFim1 has about 20 potential phosphorylation sites as predicted by NetPhos alone.

However, the different patterns of actin and fimbrin localization within the parasite during different morphological shifts reflect the complexity of actin and fimbrin dynamics in *Trichomonas*. In the amoeboid form actin and fimbrin colocalize in patches opposing the trailing end of the amoeba (Fig. 5M–O, Fig. 6, Fig. S4A–C), and with fibrous structures similar to those previously observed (Bricheux *et al.*, 2000). The latter could represent actin

cable-like structures. These were also observed during phagocytosis of yeast to a minor degree, albeit here the clustering around the phagocytic vacuole was the predominant structure (Fig. 7). The punctate localization of fimbrin observed during phagocytosis (Fig. 7I–K) might represent endocytic vesicles that are known to associate with fimbrin-bundled actin filaments in mammals and yeast (Hagiwara *et al.*, 2011; Skau *et al.*, 2011). This would then furthermore suggest an interaction of TvFim1 with proteins of the Rab family, as observed for the mammalian Rab5 and fimbrin during endocytosis (Hagiwara *et al.*, 2011).

The bundling of actin through fimbrin furthermore reflects only a fraction of the proteins likely associated with infection- and phagocytosis-related actin remodelling. Coronin has been localized to 'various dynamic sub-cortical zones' of the parasite, too, and suggested to play an important role during morphogenesis (Bricheux *et al.*, 2000). In other eukaryotic systems actin-regulating proteins are known to together orchestrate a variety of processes, in particular those associated with motion (Golsteyn *et al.*, 1997; Eichinger *et al.*, 1999; dos Remedios *et al.*, 2003; Xue *et al.*, 2010). As mentioned earlier, the majority of these proteins are encoded by *T. vaginalis* (Fig. 1). One must predict that only their concerted effort allows the different responses observed upon contact with host tissue and yeast cells (Figs 5/6 and 7 respectively), and the elaborate patterns of motion revealed, which include swift migration across host tissue and a sweeping of the substrate through the constricted apical tip of the parasite (Movies S4–S10). The latter also supports previous findings, which suggested a flagella-localized tetraspanin (TvTSP6) to serve sensory reception (de Miguel *et al.*, 2012). *T. vaginalis* might therefore offer an alternative system to study the actin dynamics of a pathogenic protist.

In summary our results demonstrate that fimbrin likely assists actin bundling across a multitude of different processes and it appears that many of these functions are conserved among a diverse range of evolutionary distant eukaryotes. In the parasite *T. vaginalis* one of the primary function of the early actin-based – and fimbrin accompanied – morphogenesis includes phenotypic plasticity during host cell attachment: the increase of surface interactions with VECs to scavenge substrate while gliding across host tissue. Another function could include the feeding on, and defence against, other microorganisms of the vaginal flora the parasite likely encounters during infection. Our results confirm the actin machinery to accompany phagocytosis and the parasite has been shown to phagocytose a broad range of prokaryotic and eukaryotic cells (Street *et al.*, 1984; Benchimol and de Souza, 1995; Rendon-Maldonado *et al.*, 1998; Pereira-Neves and Benchimol, 2007). These observa-

tions shift the way we view parasite–host tissue interaction and offer a system and molecular proxy to study cytoskeletal actin dynamics in a protozoan parasite during infection-associated phenotypic morphogenesis.

Experimental procedures

Cultures

Trichomonas vaginalis strains T016, FMV1 and T1 were cultivated in tryptone–yeast extract maltose medium {2.22% (w/v) tryptose, 1.11% (w/v) yeast extract, 15 mM maltose, 9.16 mM L-cysteine, 1.25 mM L(+)-ascorbic acid, 0.77 mM KH₂PO₄, 3.86 mM K₂HPO₄, 10% (v/v) horse serum, 0.71% (v/v) iron solution [= 1% (w/v) Fe(NH₄)₂(SO₄) × 6H₂O, 0.1% (w/v) 5-sulfosalicylic acid]} at 37°C and 5% CO₂ in a Galaxy 48R (Eppendorf, Germany). Immortalized VECs (VECs MS-74) were cultivated in 45% DMEM (Invitrogen, #31885), 45% Keratinocyte-SFM (Invitrogen, #37010022) and 10% fetal calf serum (FCS) in standard cell culture flasks (75 cm²) and at the same conditions as *T. vaginalis*. At high confluency, cells were washed twice with Dulbecco's PBS (PAA, #H15-001), digested with trypsin (Invitrogen, #25300-054) for 5 min, before inactivation with FCS. Cells were then pelleted and resuspended in fresh media and split 1:10 into new flasks and medium. To prevent bacterial contamination a penicillin/streptomycin mix was added to a final concentration of 100 µg ml⁻¹ to both media.

Database screening and structure prediction

We compiled a list of 77 actin and tubulin associated genes, mainly from *Homo sapiens* (Table S1). These query sequences were matched against the complete genomes of 25 eukaryotes representing the different eukaryotic phyla (Table S2). The proteomes were obtained from RefSeq (Pruitt *et al.*, 2007) except *Bigeloviella natans* (<http://genome.jgi-psf.org/Bigna1/>), *Cyanidioschyzon merolae* (Matsuzaki *et al.*, 2004) and *Cyanophora paradoxa* (Price *et al.*, 2012), which were downloaded from the corresponding genome project homepages. For five species without complete genomic sequences, ESTs were obtained downloaded from dbEST (Boguski *et al.*, 1993; see also Table S2). Clusters of homologous proteins were reconstructed from a total of 453 696 proteins encoded within the eukaryotic chromosomes. A BLAST (Altschul *et al.*, 1997) search analysis yielded 1497 best BLAST hits. All protein pairs were globally aligned using Needleman–Wunsch algorithm with needle program (Rice *et al.*, 2000). A total of 12 758 protein pairs having global amino acids identities ≥ 20% were clustered into protein families using MCL algorithm (Enright *et al.*, 2002) with the default parameters. The query proteins within each cluster were BLASTed against ESTs from *Alexandrium tamarense*, *Oxyrrhis marina*, *Porphyra yezoensis*, *Reclinomonas americana* and *Physarum polycephalum*, and matching ESTs were translated to amino acid sequences and added to the clusters based on the highest similarity. Clusters were further refined by splitting of paralogous clusters, merging of orthologous clusters, removing of single gene clusters and obvious paralogous sequences from a cluster, based on protein annotations and examination of phylogenetic trees reconstructed using PHYML (Guindon *et al.*, 2010). This procedure yielded 62 protein families with a total of 891 genes.

The tertiary structure of TvFim1 (without the EF-hand) was generated on the basis of the known crystal structure of the *Schizosaccharomyces pombe* fimbrin PDB accession 1RT8 (Klein *et al.*, 2004), using the MODELLER software (Eswar *et al.*, 2006). A local alignment of TvFim1 to 1RT8 in MODELLER revealed an identity of 46% over an alignment length of 456 amino acids, representing the highest level of sequence similarity in the PDB database (<http://www.rcsb.org/pdb/>). In order to model the actin/ABD2 complex structure we used the structure of the F-actin-fimbrin/plastin ABD2 complex of *Homo sapiens* PDB accession 3BYH (Galkin *et al.*, 2008) together with TvFim1 and an abundantly expressed actin of *T. vaginalis* (TVAG_337240).

Gene cloning and heterologous overexpression of TvFim1

TvFim1 gene sequence was amplified by TvFim1_NdeI_FOR (5'-CTGACGCATATGGCTGTAAACGCTGCG-3') and TvFim1_BamHI_REV (5'-GACGTGGATCCTTGATCCATGGCCATAAGA GA-3') using a proof-reading polymerase and ligated into expression vector pTagvag2 for IFAs, pTvGFP [based on pTagvag2 with a *Trichomonas* codon-optimized green fluorescent protein (GFP)-tag replacing the HA] for live cell imaging of FMV1 and pETEV21a for overexpression respectively, and verified by sequencing. Thirty micrograms of the plasmid DNA was used for transfection of 2.5×10^8 *T. vaginalis* cells using standard electroporation (Delgado *et al.*, 1997). After 4 h of incubation neomycin (G418) was added to a final concentration of $100 \mu\text{g ml}^{-1}$ for selection. For heterologous overexpression *E. coli* C41(DE3) was used and transformed with pETEV21a including the gene of interest. Briefly, 1 l of culture was incubated on an orbital shaker at 37°C until an optical density of OD₆₀₀ 0.4–0.6 was reached. Overexpression was induced with 1 mM isopropyl- β -D-thiogalactopyranoside (IPTG) followed by a 4 h incubation at 37°C. Cells were pelletized, washed once with phosphate-buffered saline (PBS) and again pelletized. Cells were lysed by plotting and subsequent disruption using the OneShot disruptor (Constant Systems Limited). HIS-tagged protein was isolated using a HISTrap column (HisTrap HP 5 ml, GE Healthcare) and standard fast protein liquid chromatography on an Äkta P-920 (GE Healthcare).

Quantitative real-time PCR

All experiments were carried out using a StepOnePlus and Power SYBR Green master mix (Applied Biosystems). *T. vaginalis* RNA was isolated from biological triplets using TRIzol (Invitrogen) from the motile-flagellated and three adherent-amoeboid stages at 5, 20 and 60 min after fibronectin-induced morphogenesis (fibronectin from human plasma; Sigma F0895). RNA was transcribed into cDNA with iScript cDNA Synthesis Kit (Bio-Rad) and used as a template for real-time quantitative PCR. Primers used were: Fimbrin: TvFim1_qFOR: 5'-ACAACCTTTACGACGGCA TC-3' and TvFim1_qREV: 5'-GCTTTGTCTTGTGGCCTTC-3', 40S ribosomal protein S5: 40SRibo_qFOR: 5'-GCATTGATCA GGCTCTCTCC-3', 40SRibo_qREV: 5'-ATGCGCTCAAGTTCGT CTTT-3'. For absolute quantification, a linear relationship of Ct and log (DNA weight) was plotted for the target transcripts and used to infer its corresponding amount. As the target gene sequence was known, the copy numbers implied in the quantity

may be calculated by the molecular weight of the sequence, which led to the estimate of the transcript number in the unknown sample (Lu *et al.*, 2012).

Western blotting and immunofluorescence

Protein samples were separated through standard SDS-PAGE and blotted onto nitrocellulose membrane. Membranes were blocked in 5% milk powder in Tris-buffered saline pH 7 (blocking buffer) for 30 min. Blots were incubated with the primary antibody at a dilution of 1:1000 or 1:5000 in blocking buffer either overnight at 4°C or for 1 h at room temperature (RT) and then washed 3× with TBS-T (TBS + 0.1% Tween 20), followed by the incubation with the secondary antibody (1:2000 or 1:10 000) and identical subsequent washes. Detection of the chemiluminescence signal was performed through the SuperSignal West Pico Chemiluminescent Substrate Kit (Thermo Scientific) according to the manufacturer's protocol.

For immunofluorescent labelling all wells of CultureSlides (BD Falcon, #354114) were loaded with about 1.5×10^5 VECs 48 h prior to fixation and incubated at 37°C and 5% CO₂. Medium was discarded, followed by the inoculation with 1.5×10^6 *T. vaginalis* cells for a minimum of 15 min at 37°C and 5% CO₂. Supernatant was discarded, adhesive cells washed gently with PBS and fixed and permeabilized with a solution containing 4% PFA and 0.1% Triton X-100 in PBS for 15 min at RT. After discarding the solution, cells were blocked (1% BSA, 0.25% Gelatine, 0.05% Tween 20 in PBS) for 30 min at RT. Slides were incubated with the primary antibody (1:50 to 1:500 in blocking buffer) for 1 h at RT, supernatant discarded and slides washed at least twice with PBS before incubation with the secondary antibody at 1:2500 to 1:5000 for 1 h at RT in the dark. Cells were then mounted in Fluoroshield with DAPI (Sigma #F6057). Samples were stored at 4°C, dark, until imaging using a Zeiss LSM 710 confocal microscope. Primary antibodies: polyclonal peptide antibody TvFim produced in rat (Eurogentec) against amino acid sequence CRK-FVGPREIVKGNQR, monoclonal HA-antibody (Sigma #H9658), monoclonal actin-antibody (Sigma #A4700) and mouse anti-GFP (Invitrogen #332600). Secondary antibodies: AlexaFluor488-anti-mouse IgG (Invitrogen #A11001), AlexaFluor594-anti-rat IgG (Invitrogen #A11007), ImmunoPure Goat Anti-Mouse IgG and Anti-Rat IgG (Pierce #31430 and #31470 respectively). Additional actin staining through TexasRed-X phalloidin (Invitrogen #T7471).

Yeast powder (RUF Lebensmittelwerk KG, Quakenbrück, Germany) was resuspended in sterile water (37°C) and washed three times with 0.1 M PBS. The cells were then resuspended in TYM-medium without serum at a concentration of 5×10^7 cells ml⁻¹ and immediately used for the phagocytosis assay. *T. vaginalis* and yeast were mixed with a ratio of 1:50 in TYM without serum and incubated on CultureSlides (BD Falcon, #354114) for 1 h at 37°C and at 5% CO₂. Immunofluorescent labelling and microscopy was identical to the steps described above.

Live imaging

For live imaging μ -Slide VI 0.4 (Ibidi, #80606) was used. All six slide chambers were pre-loaded with 9×10^3 VECs, 48 h prior to infection and incubated at 37°C and 5% CO₂. Medium was

discarded, followed by the inoculation with 9×10^4 *T. vaginalis* cells and immediate microscopy, using a Zeiss AxioObserver.Z1 microscope with AxioCam MRm (real-time) and AxioCam ICC1 (time-lapse) at 37°C and 5% CO₂.

Actin polymerization assays

TIRF microscopy was carried out using the Zeiss Laser TIRF3 microscope at the Summer School on Actin Dynamics in Regensburg (DFG program SPP 1464). Actin polymerization was performed using 1.5 µM rabbit muscle actin alone as a control and in the presence of 3 and 5 µM TvFim1 in FPLC elution buffer (50 mM Tris-HCl pH 7.5, 300 mM NaCl, 500 mM imidazole, 1 mM NaN₃). Protein mixtures were diluted in freshly prepared fluorescence buffer containing 10 mM imidazole-HCl (pH 7.8), 50 mM KCl, 1 mM MgCl₂, 100 mM dithiothreitol, 3 mg ml⁻¹ glucose, 20 µg ml⁻¹ catalase, 100 mg ml⁻¹ glucose oxidase and 0.5% methylcellulose to induce actin polymerization. Actin polymerization was induced in a solution containing 1.5 µM actin monomers (7% labelled with Alexa 568) and 3 or 5 µM TvFim1 respectively. Actin polymerization assays were monitored using the microplate reader Infinite200 PRO (Tecan Group, Ltd) and a Nunclon 96 flat black well plate. G-actin mix [4 µM unlabelled *Acanthamoeba* actin, 5% pyrene-actin in G-buffer (2 mM Tris pH 8.0, 0.5 mM DTT, 0.2 mM ATP, 0.1 mM CaCl₂, 0.01% NaN₃)] was measured alone and including 4 µM of TvFim1 with the following parameters: excitation = 365 nm, emission = 410 nm, Z position = 17 000, gain = 135, read every 3 s for 30 min.

Acknowledgements

We thank G. Landan and S. Nelson-Sathi for help with MatLab, all supervisors of the Actin Summer School of 2011 at the Universitätsklinikum Regensburg (Germany), in particular Margot Quinlan (University of California, USA) for helpful discussions. We thank J. Alderete and M. Benchimol for sharing *T. vaginalis* strains. This work was funded by a DFG grant (GO 1825/3-1) and the support of the 'Stiftung zur Erforschung infektiös-immunologischer Erkrankungen' to S.B.G.

References

- Addis, M.F., Rappelli, P., Delogu, G., Carta, F., Cappuccinelli, P., and Fiori, P.L. (1998) Cloning and molecular characterization of a cDNA clone coding for *Trichomonas vaginalis* alpha-actinin and intracellular localization of the protein. *Infect Immun* **66**: 4924–4931.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.
- Arroyo, R., González Robles, A., Martínez Palomo, A., and Alderete, J. (1993) Signalling of *Trichomonas vaginalis* for amoeboid transformation and adhesin synthesis follows cytoadherence. *Mol Microbiol* **7**: 299–309.
- Aurrecoechea, C., Brestelli, J., Brunk, B.P., Carlton, J.M., Dommer, J., Fischer, S., et al. (2009) GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*. *Nucleic Acids Res* **37**: D526–D530.
- Baum, J., Tonkin, C.J., Paul, A.S., Rug, M., Smith, B.J., Gould, S.B., et al. (2008) A Malaria parasite formin regulates actin polymerization and localizes to the parasite-erythrocyte moving junction during invasion. *Cell Host Microbe* **3**: 188–198.
- Benchimol, M. (2004) Trichomonads under microscopy. *Microsc Microanal* **10**: 528–550.
- Benchimol, M. (2008) The hydrogenosome as a drug target. *Curr Pharm Des* **14**: 872–881.
- Benchimol, M., and de Souza, W. (1995) Carbohydrate involvement in the association of a prokaryotic cell with *Trichomonas vaginalis* and *Tritrichomonas foetus*. *Parasitol Res* **81**: 459–464.
- Boguski, M.S., Lowe, T.M., and Tolstoshev, C.M. (1993) dbEST-database for 'expressed sequence tags. *Nat Genet* **4**: 332–333.
- Bricheux, G., Coffe, G., Pradel, N., and Brugerolle, G. (1998) Evidence for an uncommon alpha-actinin protein in *Trichomonas vaginalis*. *Mol Biochem Parasitol* **95**: 241–249.
- Bricheux, G., Coffe, G., Bayle, D., and Brugerolle, G. (2000) Characterization, cloning and immunolocalization of a coronin homologue in *Trichomonas vaginalis*. *Eur J Cell Biol* **79**: 413–422.
- Brugerolle, G., Bricheux, G., and Coffe, G. (1996) Actin cytoskeleton demonstration in *Trichomonas vaginalis* and in other trichomonads. *Biol Cell* **88**: 29–36.
- Butler, B., and Cooper, J.A. (2009) Distinct roles for the actin nucleators Arp2/3 and hDia1 during NK-mediated cytotoxicity. *Curr Biol* **19**: 1886–1896.
- Carlton, J.M., Hirt, R.P., Silva, J.C., Delcher, A.L., Schatz, M., Zhao, Q., et al. (2007) Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science* **315**: 207–212.
- Cheng, D., Marner, J., and Rubenstein, P.A. (1999) Interaction *in vivo* and *in vitro* between the yeast fimbrin, SAC6P, and a polymerization-defective yeast actin (V266G and L267G). *J Biol Chem* **274**: 35873–35880.
- Delgadillo, M.G., Liston, D.R., Niazi, K., and Johnson, P.J. (1997) Transient and selectable transformation of the parasitic protist *Trichomonas vaginalis*. *Proc Natl Acad Sci USA* **94**: 4716–4720.
- Eichinger, L., Lee, S.S., and Schleicher, M. (1999) *Dictyostelium* as model system for studies of the actin cytoskeleton by molecular genetics. *Microsc Res Tech* **47**: 124–134.
- Elmendorf, H.G., Hayes, R.D., Srivastava, S., and Johnson, P.J. (2010) New insights into the composition and function of the cytoskeleton in *Giardia lamblia* and *Trichomonas vaginalis*. In *Anaerobic Parasitic Protozoa: Genomics and Molecular Biology*. Clark, C.G., Johnson, P.J., and Adam, R.D. (eds). Norfolk: Caister Academic Press, pp. 119–156.
- Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**: 1575–1584.
- Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.Y., et al. (2006) Comparative protein structure modeling using Modeller. *Curr Protoc Bioinformatics* **5**: Unit 5.6.

- Fritz-Laylin, L.K., Prochnik, S.E., Ginger, M.L., Dacks, J.B., Carpenter, M.L., Field, M.C., *et al.* (2010a) The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. *Cell* **140**: 631–642.
- Fritz-Laylin, L.K., Assaf, Z.J., Chen, S., and Cande, W.Z. (2010b) *Naegleria gruberi* de novo basal body assembly occurs via stepwise incorporation of conserved proteins. *Eukaryot Cell* **9**: 860–865.
- Fukui, Y. (2002) Mechanistic of amoeboid locomotion: signal to forces. *Cell Biol Int* **26**: 933–944.
- Fulton, C. (1993) *Naegleria* – a research partner for cell and developmental biology. *J Euk Microbiol* **40**: 520–532.
- Galkin, V.E., Orlova, A., Cherepanova, O., Lebart, M.-C., and Egelman, E.H. (2008) High-resolution cryo-EM structure of the F-actin-fimbrin/plastin ABD2 complex. *Proc Natl Acad Sci USA* **105**: 1494–1498.
- Gold, D., and Ofek, I. (1992) Adhesion of *Trichomonas vaginalis* to plastic surfaces: requirement for energy and serum constituents. *Parasitology* **105**: 55–62.
- Golsteyn, R.M., Louvard, D., and Friederich, E. (1997) The role of actin binding proteins in epithelial morphogenesis: models based upon *Listeria* movement. *Biophys Chem* **68**: 73–82.
- Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307–321.
- Hagiwara, M., Shinomiya, H., Kashihara, M., Kobayashi, K., Tadokoro, T., and Yamamoto, Y. (2011) Interaction of activated Rab5 with actin-bundling proteins, L- and T-plastin and its relevance to endocytic functions in mammalian cells. *Biochem Biophys Res Comm* **407**: 615–619.
- Jansen, S., Collins, A., Yang, C., Rebowski, G., Svitkina, T., and Dominguez, R. (2011) Mechanism of actin filament bundling by fascin. *J Biol Chem* **286**: 30087–30096.
- Klein, M.G., Shi, W., Ramagopal, U., Tseng, Y., Wirtz, D., Kovar, D.R., *et al.* (2004) Structure of the actin crosslinking core of fimbrin. *Structure* **12**: 999–1013.
- Korenbaum, E., and Rivero, F. (2002) Calponin homology domains at a glance. *J Cell Sci* **115**: 3543–3545.
- Kovar, D., Staiger, C., Weaver, E., and McCurdy, D. (2000) AfFim1 is an actin filament crosslinking protein from *Arabidopsis thaliana*. *Plant J* **24**: 625–636.
- Kulda, J. (1999) Trichomonads, hydrogenosomes and drug resistance. *Int J Parasitol* **29**: 199–212.
- Lal, K., Noël, C.J., Field, M.C., Goulding, D., and Hirt, R.P. (2006) Dramatic reorganisation of *Trichomonas* endomembranes during amoebal transformation: a possible role for G-proteins. *Mol Biochem Parasitol* **148**: 99–102.
- Land, K.M., Delgado-Correa, M.G., Tachezy, J., Vanacova, S., Hsieh, C.L., Sutak, R., and Johnson, P.J. (2004) Targeted gene replacement of a ferredoxin gene in *Trichomonas vaginalis* does not lead to metronidazole resistance. *Mol Microbiol* **51**: 115–122.
- Lu, Y., Xie, L., and Chen, J. (2012) A novel procedure for absolute real-time quantification of gene expression patterns. *Plant Methods* **8**: 9.
- Martincová, E., Voleman, L., Najdrová, V., De Napoli, M., Eshar, S., Gualdrón, M., *et al.* (2012) Live imaging of mitochondria and hydrogenosomes by HaloTag technology. *PLoS ONE* **7**: e36314.
- Matsuzaki, M., Misumi, O., Shin, I.T., Maruyama, S., Takahara, M., Miyagishima, S.Y., *et al.* (2004) Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature* **428**: 653–657.
- de Miguel, N., Riestra, A., and Johnson, P.J. (2012) Reversible association of tetraspanin with *Trichomonas vaginalis* flagella upon adherence to host cells. *Cell Microbiol* **14**: 1797–1807.
- Namba, Y., Ito, M., Zu, Y., Shigesada, K., and Maruyama, K. (1992) Human T cell L-plastin bundles actin filaments in a calcium-dependent manner. *J Biochem* **112**: 503–507.
- Noël, C.J., Diaz, N., Sicheritz-Ponten, T., Safarikova, L., Tachezy, J., Tang, P., *et al.* (2010) *Trichomonas vaginalis* vast BspA-like gene family: evidence for functional diversity from structural organisation and transcriptomics. *BMC Genomics* **11**: 99.
- Paredes, A.R., Assaf, Z.J., Sept, D., Timofejeva, L., Dawson, S.C., Wang, C.J.R., and Cande, W. (2011) An actin cytoskeleton with evolutionarily conserved functions in the absence of canonical actin-binding proteins. *Proc Natl Acad Sci USA* **108**: 6151–6156.
- Pereira-Brás, X., Zimorski, V., Bolte, K., Maier, U.-G., Martin, W.F., and Gould, S.B. (2013) Knockout of the abundant *Trichomonas vaginalis* hydrogenosomal membrane protein Tvhmp23 increases hydrogenosome size but induces no compensatory up-regulation of paralogous copies. *FEBS Lett* (in press). doi: 10.1016/j.febslet.2013.03.001.
- Pereira-Neves, A., and Benchimol, M. (2007) Phagocytosis by *Trichomonas vaginalis*: new insights. *Biol Cell* **99**: 87–101.
- Pereira-Neves, A., and Benchimol, M. (2009) *Tritrichomonas foetus*: budding from multinucleated pseudocysts. *Protist* **160**: 536–551.
- Petrin, D., Delgaty, K., Bhatt, R., and Garber, G. (1998) Clinical and microbiological aspects of *Trichomonas vaginalis*. *Clin Microbiol Rev* **11**: 300–317.
- Prassler, J., Stocker, S., Marriott, G., Heidecker, M., Kellermann, J., and Gerisch, G. (1997) Interaction of a *Dictyostelium* member of the plastin/fimbrin family with actin filaments and actin-myosin complexes. *Mol Biol Cell* **8**: 83–95.
- Price, D.C., Chan, C.X., Yoon, H.S., Yang, E.C., Qiu, H., Weber, A.P.M., *et al.* (2012) *Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. *Science* **335**: 843–847.
- Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**: D61–D65.
- dos Remedios, C.G., Chhabra, D., Kekic, M., Dedova, I.V., Tsubakihara, M., Berry, D.A., and Nosworthy, N.J. (2003) Actin binding proteins: regulation of cytoskeletal microfilaments. *Physiol Rev* **83**: 433–473.
- Rendon-Maldonado, J.G., Espinosa-Cantellano, M., Gonzalez-Robles, A., and Martinez-Palomo, A. (1998) *Trichomonas vaginalis*: *in vitro* phagocytosis of lactobacilli, vaginal epithelial cells, leukocytes, and erythrocytes. *Exp Parasitol* **89**: 241–250.

- Rice, P., Longden, I., and Bleasby, A. (2000) EMBOS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276–277.
- Ryan, C.M., de Miguel, N., and Johnson, P.J. (2011) *Trichomonas vaginalis*: current understanding of host-parasite interactions. *Essays Biochem* **51**: 161–175.
- Santos, J.M., Lebrun, M., Daher, W., Soldati, D., and Dubremetz, J.-F. (2009) Apicomplexan cytoskeleton and motors: key regulators in morphogenesis, cell division, transport and motility. *Int J Parasitol* **39**: 153–162.
- Shinomiya, H. (2012) Platin family of actin-bundling proteins: its functions in leukocytes, neurons, intestines, and cancer. *Int J Cell Biol* **2012**: 213492.
- Skau, C.T., Courson, D.S., Bestul, A.J., Winkelman, J.D., Rock, R.S., Sirotkin, V., and Kovar, D.R. (2011) Actin filament bundling by fimbrin is important for endocytosis, cytokinesis, and polarization in fission yeast. *J Biol Chem* **286**: 26964–26977.
- Stark, J.R., Judson, G., Alderete, J.F., Mundodi, V., Kucknoor, A.S., Giovannucci, E.L., et al. (2009) Prospective study of *Trichomonas vaginalis* infection and prostate cancer incidence and mortality: Physicians' Health Study. *J Natl Cancer Inst* **101**: 1406–1411.
- Street, D.A., Wells, C., Taylor-Robinson, D., and Ackers, J.P. (1984) Interaction between *Trichomonas vaginalis* and other pathogenic micro-organisms of the human genital tract. *Br J Vener Dis* **60**: 31–38.
- Uproft, P., and Uproft, J.A. (2001) Drug targets and mechanisms of resistance in the anaerobic protozoa. *Clin Microbiol Rev* **14**: 150–164.
- Xue, F., Janzen, D.M., and Knecht, D.A. (2010) Contribution of filopodia to cell migration: a mechanical link between protrusion and contraction. *Int J Cell Biol* **2010**: 507821.
- Yusof, A., and Kumar, S. (2011) Ultrastructural changes during asexual multiple reproduction in *Trichomonas vaginalis*. *Parasitol Res* **110**: 1823–1828.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Table S1. Sources of seed sequences used in Fig. S1. The '*' indicate genes that were removed during the procedure of finding clusters of homologues, because they either had no homologues or were too similar to other gene families part of the list.

Table S2. Source of databases used to search for actin and tubulin associated genes.

Table S3. Observed migration speeds of *Trichomonas vaginalis* T016 exposed to the vaginal epithelial cell line MS74.

Fig. S1. Alignment of TvFim1 and its homologues from a range of eukaryotes is shown in (A). Known alpha-helical structures of the corresponding calponin homology domain (CH) are indicated above the alignment, whereas amino acids thought to be conserved residues for F-actin or suppressor residues are marked with a dot. Characterized actin binding sites are marked as a line below the alignment (based on Klein et al., 2004). (B) TvFim1 includes both, the EF-hand and four calponin homology domains (CH). TvFim2 is N-terminally truncated. For comparison, the human fimbrin protein (*HsFim1*, L-plastin) is only 16 amino acids

longer and shares an overall sequence identity of 43% with TvFim1. The alignment was generated using the CLC Workbench and default settings and the following protein accessions: O.c., *Oryctolagus cuniculus* (XP_002720283.1); A.c., *Acanthamoeba castellanii* (ELR12888.1); S.p., *Schizosaccharomyces pombe* (NP596289.1); M.m., *Mus musculus* (AAH05459.1); H.s., *Homo sapiens* (AAB02845.1); G.g., *Gallus gallus* (P19179); D.m., *Drosophila melanogaster* (AAF48722.1); D.d., *Dictyostelium discoideum* (P54680); S.m., *Schistosoma mansoni* (AAC14025.1); G.p., *Gibberella pulicaris* (CAA10667.1); A.t., *Arabidopsis thaliana* (AAC39359.1).

Fig. S2. FPLC-purification of heterologously expressed TvFim1 from *Escherichia coli* strain C41. In (A) the FPLC run showing the loading, washing and elution phase with the elution peak of the HIS-tagged TvFim1 at 100 mM imidazol. (B) Coomassie-stained SDS-PAGE of 1: non-induced *E. coli* cell lysate; 2: induced *E. coli* cell lysate, 3: FPLC purified TvFim1 of 68 kDa (indicated by arrow). Marker (M) in kilodalton.

Fig. S3. Antibody controls.

A. Western blot of crude pre-immune (pIS) and immune rat sera (IS) from TvFim1 peptide immunization on protein extract from *Trichomonas vaginalis* and the *E. coli* strain expression recombinant TvFim1.

B. Western blot using purified TvFim1, HA-antibody and GFP-antibody on protein extract of the *T. vaginalis* strains expressing the corresponding constructs. Cropped bands of this blot were used for Fig. 2. Molecular marker (M) in kDa.

C. Immunofluorescence controls of the *T. vaginalis* strain T016 on fibronectin-coated slides using TvFim1-pre-immune sera and anti-actin (Sigma). Scale bar: 10 μ m.

D. shows the gradient-like colocalization of the HA-tagged TvFim1 with that of the endogenous copy towards the periphery of the cells. Scale bar: 2 μ m.

Fig. S4. Additional TvFim1 distributions in *Trichomonas vaginalis* T016 cells. (A–D) TvFim1 concentrates at the protruding edge of a parasite, while gliding across host tissue, contour highlighted by a dashed line in (D). The protein furthermore associates with structures reminiscent of actin cables (E–H). Many parasites proliferate and become multinuclear (arrow heads) during infection (I–L). Arrow heads point to exemplary multi-nuclear cells. (M–R) Further localizations of fimbrin together with the hydrogonosomal marker enzyme ASCT (acetate : succinate CoA-transferase). Scale bar: 10 μ m.

Fig. S5. Live cell imaging of *Trichomonas vaginalis* on human vaginal epithelial cells. Live imaging of the parasite exposed to human epithelial cells reveals *T. vaginalis* actively migrates across tissue, while they appear to use their flagella and apical tip as the guiding end. A clear trailing end and pseudopodia are also visible. While some rapidly divide on VECs, from what appears to be di-nuclear cells, others massively increase their size (juggernauting) and develop more than two nuclei and flagellar pockets before division. Areas of interest are marked with an asterisk in the first image of every series. For details please also refer to Movies S7–S10. Scale: 10 μ m.

Fig. S6. A TvFim1::HaloTag fusion construct was generated using the identical primers as used for TvFim1:HA construct (see Material & Methods of manuscript). The HaloTag plasmid (Martincová et al., 2012, *PLoS ONE* 7: e36314) was kindly provided by Pavel Dolezal (Charles University, Prague) and the product cloned into the plasmid through NdeI and BamHI and correct insertion verified through sequencing. RT-PCR and Western blot

analysis were carried out according to the methods described in the manuscripts main text. For live imaging of TvFim1 we fused the gene to the recently reported HaloTag (Martincová *et al.*, 2012). We obtained several clones, three of which we analysed and shown above. (A) Reverse transcriptase PCR on RNA isolated from the transfected T016 cells and using specific primers, shows the fusion gene TvFim1::HaloTag is expressed in clones 2 and 3. (B) No fusion-protein is detected by Western blot analysis using an anti-haemagglutinin antibody (an HA-tag is included in the fusion protein generated by pHaloTag). This correlates with the lack of fluorescence during microscopic analysis of the clones.

Fig. S7. A. We tried to knockdown TvFim1 using anti-sense RNA as described previously by Ong *et al.* (2007, *JBC* 282: 6716–25). The anti-sense region was amplified from genomic DNA using the primers TvFim1-as_BamHI_F: 5'-GGTGGTGGATCCCGTATGAGCCTTCTCAAGAC-3' and TvFim1-as_NdeI_R: 5'-GGTGGTCATATGCGCGTCAGCGATGCCG-3'. After sequence verification the fragment was cloned into pTagvag2 for standard expression under the control of the SCS promoter. *T. vaginalis* was then transfected with 30 µg of plasmid DNA and selected through G418 (see 'Gene cloning' section).

B. RNA from recombinant and wild-type trichomonads was transcribed into cDNA and used as a template for real-time quantitative PCR using TvFim1 specific primers (see 'Quantitative PCR' section) and biological and technical triplicates. The expression level of TvFim1 in the wild type was used as the reference (100%) against TvFim1-as, which revealed no significant upregulation or downregulation of the gene.

Movie S1. Total internal reflection microscopy movie of actin self-assembly.

Movie S2. Total internal reflection microscopy movie of actin self-assembly in the presence of 3 µM TvFim1.

Movie S3. Total internal reflection microscopy movie of actin self-assembly in the presence of 5 µM TvFim1.

Movie S4. Live imaging of TvFim1::GFP. A daughter cell can be seen budding from a multinuclear mother cell that is attached to host tissue. Fimbrin::GFP is found to cluster around contractile ring-like structure of the daughter cell. One image was taken every 3 s and run at 7 frames per second.

Movie S5. Live imaging of TvFim1::GFP. During the migration of the parasite across tissue, waves of TvFim1::GFP can clearly be seen in close proximity to the migration front and moving away from it. One image was taken every 3 s and run at 7 frames per second.

Movie S6. Live imaging of TvFim1::GFP. Clustering of TvFim1::GFP was predominately observed to occur in areas of the parasite attaching to the vaginal epithelial cells. One image was taken every 5 s and run at 7 frames per second.

Movie S7. Time lapse movie of *T. vaginalis* on vaginal epithelial cells showing concerted movement and the parasite using the apical tip as the guiding end and for flagellar sensing. 40 min into infection and one image taken every second and run at 7 frames per second.

Movie S8. Time lapse movie of a *T. vaginalis* nicely showing the pseudopods and the trailing end of the parasite while gliding along host tissue 20 min after infection. One image was taken every 0.2 s and run at 10 frames per second.

Movie S9. Time lapse movie of a *T. vaginalis* cell dividing on a vaginal epithelial cell just 5 min after infection. One image taken every second and run at 7 frames per second.

Movie S10. Time lapse movie of *T. vaginalis* on vaginal epithelial cells 70 min into infection, demonstrating some adherent cells to massively increase their overall cell mass, a process we refer to as juggernauting. One image taken every second and run at 28 frames per second.

4.3 Deep sequencing of *Trichomonas vaginalis* during the early infection of vaginal epithelial cells and amoeboid transition

Sven B. Gould, Christian Woehle, Gary Kusdian, Giddy Landan, Jan Tachezy, Verena Zimorski & William F. Martin

Der vorliegende Artikel wurde 2013 in der Fachzeitschrift *International Journal for Parasitology* (Impact Factor 3,6) veröffentlicht.

Ergänzendes Material steht zur Verfügung durch die Webseiten des Verlegers⁸.

Beitrag von Christian Wöhle, Zweitautor:

Versuchsplanung	30 %
Datenanalyse	70 %
Verfassen des Manuskripts	20 %

⁸<http://www.sciencedirect.com/science/article/pii/S0020751913001288>



Deep sequencing of *Trichomonas vaginalis* during the early infection of vaginal epithelial cells and amoeboid transition [☆]



Sven B. Gould ^{a,*}, Christian Woehle ^a, Gary Kusdian ^a, Giddy Landan ^a, Jan Tachezy ^b, Verena Zimorski ^a, William F. Martin ^a

^a Institute for Molecular Evolution, Heinrich-Heine-University, 40225 Düsseldorf, Germany

^b Laboratory of Molecular and Biochemical Parasitology, Charles University Prague, 12844 Prague, Czech Republic

ARTICLE INFO

Article history:

Received 4 December 2012
Received in revised form 8 April 2013
Accepted 9 April 2013
Available online 18 May 2013

Keywords:

Trichomonas
Infection
Cytoskeleton
Oxygen stress
Gene families

ABSTRACT

The human pathogen *Trichomonas vaginalis* has the largest protozoan genome known, potentially encoding approximately 60,000 proteins. To what degree these genes are expressed is not well known and only a few key transcription factors and promoter domains have been identified. To shed light on the expression capacity of the parasite and transcriptional regulation during phase transitions, we deep sequenced the transcriptomes of the protozoan during two environmental stimuli of the early infection process: exposure to oxygen and contact with vaginal epithelial cells. Eleven 3' fragment libraries from different time points after exposure to oxygen only and in combination with human tissue were sequenced, generating more than 150 million reads which mapped onto 33,157 protein coding genes in total and a core set of more than 20,000 genes represented within all libraries. The data uncover gene family expression regulation in this parasite and give evidence for a concentrated response to the individual stimuli. Oxygen stress primarily reveals the parasite's strategies to deal with oxygen radicals. The exposure of oxygen-adapted parasites to human epithelial cells primarily induces cytoskeletal rearrangement and proliferation, reflecting the rapid morphological transition from spindle shaped flagellates to tissue-feeding and actively dividing amoeboids.

© 2013 Australian Society for Parasitology Inc. Published by Elsevier Ltd. All rights reserved.

1. Introduction

The extracellular parasite *Trichomonas vaginalis* infects the urogenital tract of approximately 3% of the world population annually and is thus the most widespread non-viral, sexually transmitted human parasite known (Schwebke et al., 2011). Although the vast majority of *T. vaginalis* infections proceed without apparent symptoms, infection with the parasite decreases fertility, elevates the risk of prostate and cervical cancer and increases the risk of acquiring HIV (Petrin et al., 1998; Ryan et al., 2011). The more severe infections with manifest symptoms, Trichomoniasis, result in urogenital tract swelling and inflammatory discharge and are commonly treated with nitroimidazole derivatives, although resistant strains are on the rise (Kulda, 1999; Upcroft and Upcroft, 2001; Benchimol, 2008; Pal and Bandyopadhyay, 2012).

A crucial step of the infection process involves a dramatic morphological shift of the parasite. The free-swimming ovoid cells,

which resemble the familiar image of a flagellated protozoan, transform into an amoeboid form (Fig. 1) upon contact with the urogenital tract. This process commences more or less immediately upon contact with host tissue and the transformation of a cell takes only minutes to complete (Lal et al., 2006). The parasite's morphogenesis entails at least two distinct but simultaneous processes: (i) the dramatic shape transition and (ii) the adherence to host cells.

Recent studies on *Trichomonas* surface proteins that mediate host cell adherence and interactions with host extracellular matrix have identified three classes of proteins (recently reviewed by Hirt et al. (2011) and Ryan et al. (2011)). First there is the large BspA family (*Bacteroides forsythus* surface protein A), members of which have been localised to the parasite's surface, but whose exact function remains elusive (de Miguel et al., 2010; Noël et al., 2010). The second class comprises components of the thick glycocalyx, which is thought to also directly interact with the host extracellular matrix. Human Galectin-1 was identified as an interaction partner for a specific single lipophosphoglycan of the pathogen (Okumura et al., 2008). A third, and most controversial, class of suggested adhesion proteins is those that have been suggested to have dual functions, as they include enzymes of the glycolytic pathway such as glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and

[☆] Note: Nucleotide sequence data reported in this paper are available through the NCBI Single Read Archive Accession No. SRA059159.

* Corresponding author. Address: Heinrich-Heine-University Düsseldorf, Universitätstr. 1, 40225 Düsseldorf, Germany. Tel.: +49 2118113983; fax: +49 2118113554.

E-mail address: gould@hhu.de (S.B. Gould).

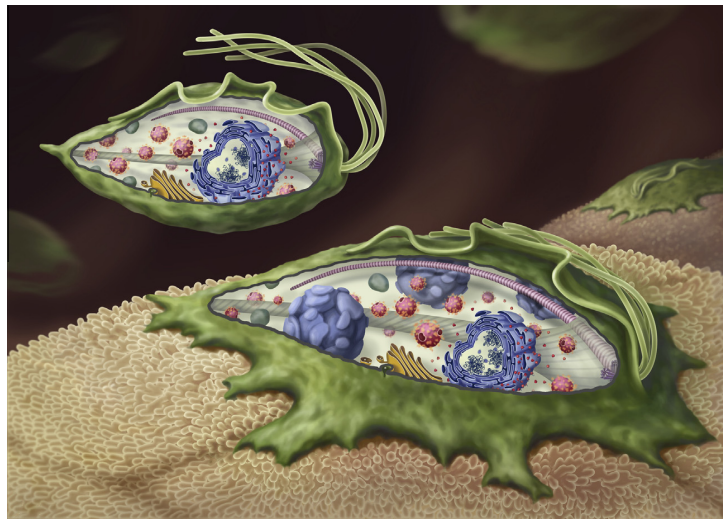


Fig. 1. The two major phenotypes of *Trichomonas vaginalis*. The human pathogen can switch between two radically different phenotypes. Phenotypic plasticity is induced through contact with tissue of the urogenital tract (light brown). After initial contact the flagellated cells become amoeboid within minutes and increase the surface area interacting with and scavenging from host tissue. Many of the amoeboid cells become multinuclear (nuclei in blue, with closely associated endoplasmic reticulum), proliferate and subsequently divide.

enzymes of pyruvate metabolism such as malic enzyme (ME), pyruvate:ferredoxin oxidoreductase (PFO) and subunits of succinyl-CoA synthetase (SCS) (Garcia and Alderete, 2007; Meza-Cervantez et al., 2011). These are common proteins with clear functions in energy metabolism (Müller et al., 2012). Their involvement in adhesion has been challenged (Addis et al., 1998; Brugerolle et al., 2000; Hirt et al., 2007; Ryan et al., 2011) and evidence for their secretion or their direct interactions with the human extracellular matrix is yet to be presented. The machinery behind trichomonad morphogenesis is not well characterised but microscopic observations and the localisation of actin and actin-associated proteins have demonstrated these cytoskeletal elements to play a crucial role (Gold and Ofek, 1992; Brugerolle et al., 1996; Bricheux et al., 1998, 2000; Kusdian et al., in press).

Trichomonas is considered a typical anaerobe. It has no oxygen requirements for growth but sometimes is designated as a microaerophile because it grows slightly better at very low oxygen tensions of $\sim 0.25 \mu\text{M}$, corresponding to about 1/1000th of ambient levels at 25°C , or $250 \mu\text{M}$ (Paget and Lloyd, 1990), while oxygen tension above $60 \mu\text{M}$ has a detrimental effect (Ellis et al., 1994). The mitochondria of *Trichomonas* are designated as hydrogenosomes because they produce hydrogen as an end product of an anaerobic, fermentative energy metabolism (Müller, 1988; Müller et al., 2012). Core enzymes of the hydrogenosomal metabolism, such as PFO and Fe–Fe hydrogenases, are highly oxygen-sensitive (Williams et al., 1990; Hrdý and Müller, 1995; Page-Sharp et al., 1996). However, the parasite typically experiences oxygen stress in its natural environment, for example during the transmission from one host to the other or with fluctuating vaginal oxygen levels during the menstruation cycle (Wagner and Levin, 1978; Ellis et al., 1992; Hill et al., 2005) and hence must possess mechanisms to avoid inactivation of oxygen-sensitive enzymes and to remove reactive oxygen species (ROS). Cytosolic NADH and NADPH oxidases are among the parasite's most important oxygen scavenging enzymes (Linstead and Bradley, 1988). Glutathione, a widespread antioxidant among eukaryotes, is absent in *T. vaginalis*, with cysteine possibly acting in its place (Ellis et al., 1994). Individual proteins shown to be up-regulated during oxygen stress include superoxide dismutase (SOD) (Ellis et al., 1994; Rasoloson et al.,

2001) and peroxiredoxins (Coombs et al., 2004), ubiquitous enzymes that are thought to be central to defenses against ROS (Gretes et al., 2012). Thioredoxin reductases are also present in *T. vaginalis* as a component of the hydrogenosomal thioredoxin-linked antioxidant system (Mentel et al., 2008). More recently, flavodiiron class A protein was shown to function as the major oxygen reductase responsible for respiration of hydrogenosomes (Smutná et al., 2009). This enzyme reduces oxygen to water in a four-electron reaction, while production of hydrogen peroxide was not detected. Thus, the parasite seems to utilise a sophisticated system to buffer oxygen stress, which includes a cytosolic and a hydrogenosomal antioxidant system in combination (Coombs et al., 2004; Mentel et al., 2008).

With approximately 160 Mbp, the genome of this protist is the largest of any protozoan genome currently available (Carlton et al., 2007). The original genome annotation identified approximately 60,000 potential protein encoding genes, which was reduced to approximately 46,000 through subtracting genes that appear to be only fragments of others (Smith and Johnson, 2011). A large portion of this redundancy of annotated genes is due to an unknown amount of duplications the genome has experienced. Approximately 65% of the genome sequence consists of repetitive elements and there are currently still $>25,000$ scaffolds that cannot be assembled. To what degree the parasite expresses this large arsenal of encoded genes is not known; expression evidence exists only for 10,470 genes at TrichDB (<http://trichdb.org/trichdb/>). Transcriptional regulation in *T. vaginalis* and in protists in general is not well characterised. Three core promoter elements have been identified in *T. vaginalis*, the initiator element (Inr) and the M3 and M5 elements (Schumacher et al., 2003; Smith et al., 2011), and one which appears to be the reverse sequence of the M3 motif and linked to the 5' untranslated region (UTR) of histone encoding genes (Cong et al., 2010). Two additional motifs, M2 and M4, were not further characterised as the M2 showed no conserved localisation in comparison with the start codon and the M4 was in too close proximity to it (Smith et al., 2011).

In order to better understand gene regulatory mechanisms underlying the initial steps of human vaginal epithelial cell (VEC) infection by *T. vaginalis* and to identify potential core changes,

we investigated 11 transcriptomes of *T. vaginalis* under varying conditions and at different time points. To be able to distinguish between the responses to oxygen itself and contact with VECs, which grow at 15% oxygen (approximately 200 μ M), we sequenced 3' fragment libraries of anaerobically cultured cells (i) after exposure to oxygen alone, (ii) after exposure to oxygen in combination with VECs and (iii) after exposure of oxygen-adapted *Trichomonas* cells to VECs, which identified transcriptional responses specific to contact with host tissue.

2. Materials and methods

2.1. Cultures

Trichomonas vaginalis T016 was cultured in tryptone-yeast extract maltose (TYM)-medium (2.22% (w/v) tryptose, 1.11% (w/v) yeast extract, 15 mM maltose, 9.16 mM L-cysteine, 1.25 mM L(+)-ascorbic acid, 0.77 mM KH₂PO₄, 3.86 mM K₂HPO₄, 10% (v/v) horse serum, 0.71% (v/v) iron solution (1% (w/v) Fe(NH₄)₂(SO₄)₂ × 6H₂O, 0.1% (w/v) 5-sulfosalicylic acid)) in 15 ml tubes in the absence of oxygen at 37 °C. To prevent bacterial contamination a penicillin/streptomycin mix was added to a final concentration of 100 μ g/ml to the media. Oxygen-adapted *T. vaginalis* were also grown in TYM medium but in cell culture flasks identical to those used for the human VEC line, MS-74 (see below), at 37 °C, 5% CO₂ and 15% O₂ in a Galaxy 48R (Eppendorf, Germany). At a cell density of approximately 9 × 10⁷ cells/ml, 1 ml was transferred into 12 ml of fresh TYM medium for continuous culturing. 'Immortalised' VECs (MS-74) were cultured in 45% DMEM (Invitrogen, Germany), 45% Keratinocyte-SFM (Invitrogen) and 10% FCS in cell culture flasks with vented lids (75 cm², VWR, Germany) at 37 °C, 5% CO₂ and 15% O₂ in a Galaxy 48R (Eppendorf). For passaging, cells were washed twice with Dulbecco's PBS (PAA, Germany), digested with trypsin (Invitrogen) for 10 min at 37 °C, before inactivation with FCS. Cells were centrifuged and resuspended in fresh media and split 1:10 into new flasks and medium.

2.2. RNA isolation

For the oxygen stress assay (*AnOx*), 50 ml of *T. vaginalis* T016 grown in sealed tubes were transferred into cell culture flasks with vented lids and incubated at 37 °C, 5% CO₂ and 15% O₂. RNA was isolated using TRIzol[®] and following the manufacturer's protocol (Invitrogen) at four time points (0, 5, 30 and 120 min of oxygen exposure), whereas the RNA isolated at 0 min served as a baseline. For the oxygen stress and infection assay (*AnOxInf*) 50 ml of T016, grown identically to those used for *AnOx*, were transferred onto a monolayer of MS74 cells grown in a cell culture flask (75 cm²) with a vented lid. Cells were harvested with a cell scraper (BD Falcon, Germany) and cell pellets produced by centrifugation at 3,500g for 1 min at 8 °C. RNA was then again isolated using TRIzol[®] at three time points (5, 30 and 120 min after infection). For the infection assay (*AdInf*), 50 ml of oxygen-adapted T016 (T016 cultured for at least five continuous days in the CO₂ incubator at 15% O₂) were exposed to a monolayer of VECs, then RNA was isolated at 5, 30 and 120 min after exposure. All experiments were performed twice (biological duplicates) and RNA duplicates pooled before DNase treatment (Fermentas, Germany) to avoid DNA contamination. Frozen RNA was then sent to Eurofins MWG (Ebersberg, Germany) for 3' fragment library preparation and deep sequencing. For quantitative real time PCR the same RNA samples that were used for 3' transcript library sequencing, were used for these experiments. RNA was transcribed into cDNA using the iScript[™] cDNA Synthesis Kit (Bio-Rad, Germany), and served as a template. Primers are listed in Supplementary Table S1. All experiments were

carried out in technical triplicates using a StepOnePlus[™] and Power SYBR[®] Green master mix (Applied Biosystems, Germany).

2.3. Transcriptome analyses

2.3.1. Reference genomes

Genomic scaffolds of *T. vaginalis*, sequences of annotated genes and genomic features were downloaded from TrichDB V1.3 (Aurrecoechea et al., 2009), including the information as to whether they encode a signal peptide or transmembrane domains (TrichDB category "Protein Features"). The nuclear genomic sequences of *Homo sapiens* (version GRCh37.p5) and its mitochondrial genome were downloaded from the National Center for Bioinformatics Information (NCBI) website (<http://www.ncbi.nlm.nih.gov>). Genomic scaffolds of *Trichomonas* smaller than 1,000 nucleotides and repeated genes were discarded (Carlton et al., 2007). Genomic sequences were reduced to the predicted mRNA locations plus 100 bp of downstream sequence to cover potential 3' UTRs. Genes labeled as "hypothetical" were manually annotated using RefSeq (as of August 2012; Pruitt et al., 2007) and NCBI BLAST (Altschul et al., 1997).

2.4. Transcriptome sequences

A total of 153,137,205 reads of a typical length of 100 bp for 11 different combinations of growth conditions and time points were obtained. The reads were deposited at the NCBI Single Read Archive (accession SRA059159). Each read was mapped onto the genomic sequences of *H. sapiens* and *T. vaginalis* using a pipeline consisting of Bowtie2 (2.0.0-beta6; Langmead and Salzberg, 2012), SAMtools (Li et al., 2009) and BEDTools (Quinlan and Hall, 2010). Non-coding sequence and RNA genes were discarded.

For comparisons of expression levels among the 11 samples, the raw read counts were normalised by the total number of reads for each sample, and scaled to match the total hit count from sample *An0*. It was noted that for expression profiling using a 3' fragment cDNA library such as ours, there is no need to normalise for transcript length ('FPKM'; fragments per kb per million sequenced reads). To cluster highly similar genes (Table 1, bottom row), transcript sequences were aligned pairwise using NEEDLE (EMBOSS; Rice et al., 2000) and MCL (Enright et al., 2002) was used to generate clusters with 90% identity. Relative up- and down-regulation values for individual genes were calculated as the ratio of the scaled read counts of the gene at the time point and in the reference sample ("*Ad0*" for "*AdInf*", and "*An0*" for "*AnOx*" and "*AnOxInf*"). *P* values and false discovery rates (FDRs), as deposited in Supplementary Tables S2–S4, were calculated using the edgeR package (Robinson et al., 2010). Due to the lack of technical replicates for the individual time points, a fixed biological coefficient of variation (BCV) of 0.1 and a significance level of 5% for the FDR were used as an indicator for differential expression (www.bioconductor.org; edgeR user guide chapter 2.9, option 2).

Identified protein coding sequences were functionally classified using the clusters of orthologous groups (COG) database (Tatusov et al., 2003), using BLASTP (*E*-value cutoff 10⁻¹⁰). Word clouds were generated using wordle.net and advanced settings. Protein tags were assigned manually by inspecting gene annotations with a known function or domain for the top 100 genes (see Supplementary Tables S2–S4).

For the identification of promoter sequences the 60 bp upstream regions from the ATG start codon of expressed genes were screened for the previously identified sequence motifs (Smith et al., 2011). The Inr, M3 and M5 motifs were based on the description in the original text, while the M2 and M4 were extracted from the sequence logo in a figure.

Table 1
Overview of the *Trichomonas vaginalis* transcriptomes analyzed. Eleven transcriptomes of the *T. vaginalis* strain T016 were sequenced in total. The transcriptomes of one culture grown under anaerobic conditions (*An0*) and one, in which the cells had been adapted to 15% CO₂ (*Ad0*) served as the base transcriptomes (time points 0). Three individual conditions were induced: oxygen stress (*AnOx*), oxygen stress and exposure to vaginal epithelial cells (VECs, *AnOxInf*) and exposure of oxygen-adapted T016 to VECs (*AdInf*). RNA was isolated at 5, 30 and 120 min after the exposure to the individual conditions.

Condition	Anaerobic			Oxygen Stress			Oxygen Stress & Infection of VECs			15% O ₂			Infection of VECs		
	0	5	30	5	30	120	5	30	120	0	5	30	5	30	120
Time point (min)	<i>An0</i>	<i>AnOx5</i>	<i>AnOx30</i>	<i>AnOx5</i>	<i>AnOx30</i>	<i>AnOx120</i>	<i>AnOxInf5</i>	<i>AnOxInf30</i>	<i>AnOxInf120</i>	<i>Ad0</i>	<i>AdInf5</i>	<i>AdInf30</i>	<i>AdInf5</i>	<i>AdInf30</i>	<i>AdInf120</i>
Total Reads	15,483,280.00	12,513,182.00	12,311,601.00	10,395,563.00	14,752,720.00	15,050,307.00	16,320,707.00	18,000,848.00	13,613,335.00	11,296,593.00	11,296,593.00	13,399,069.00	13,399,069.00	11,296,593.00	13,399,069.00
% Mapped ^a	99.29	99.18	99.20	98.88	98.89	99.11	98.86	97.42	98.36	97.42	98.14	98.66	98.66	98.14	98.66
Protein Hits ^b	25,109.00	24,910.00	24,949.00	24,161.00	26,098.00	26,053.00	26,625.00	29,294.00	24,319.00	23,891.00	23,891.00	25,852.00	25,852.00	23,891.00	25,852.00
≥ 100 hits ^c	8,056.00	8,978.00	7,889.00	7,808.00	11,083.00	10,215.00	11,092.00	14,416.00	5,043.00	5,043.00	5,372.00	6,878.00	6,878.00	5,372.00	6,878.00
Clustered 90% ^d	23,731.00	23,605.00	23,597.00	22,922.00	24,626.00	24,570.00	24,936.00	26,339.00	22,737.00	22,163.00	22,163.00	23,764.00	23,764.00	22,163.00	23,764.00

^a Percentage of reads that uniquely mapped onto the *T. vaginalis* genome.

^b Fraction of protein coding genes identified/identity ("potential paralogs").

^c Fraction of protein coding genes identified with at least 100 hits.

^d Fraction of protein coding genes with hits after clustering those with 90% sequence.

Gene families are based on the orthologous groups deposited at TrichDB. Gene pairwise co-expression was calculated as the Pearson correlation of the within sample read count ranks. Family-wise co-expression was calculated as the median of all pairwise co-expressions within the family. Nucleotide identities were derived from pairwise global alignments of annotated transcripts using the NEEDLE program (Rice et al., 2000) and the median identity for each gene family calculated.

3. Results

3.1. Transcriptional capacity

To obtain different sets of transcriptomic libraries of the parasite, we combined the infectious *T. vaginalis* strain, T016, and the human VEC line, MS74, in the following, referred to as T016 and MS74, respectively. T016 was chosen as it represents a highly virulent isolate of *T. vaginalis* (Pereira-Neves and Benchimol, 2007). In total we sequenced 3' fragment cDNA library transcriptomes of T016 from 11 individual conditions (Table 1), and each from biological duplicates. Altogether 153,137,205 reads were obtained, of which 143,767,773 (93.9%) mapped onto protein-coding genes of *T. vaginalis*. Only 3,887,899 reads from the libraries also containing mRNA from MS74 (*AnOxInf* and *AdInf*) mapped with their best hit onto human genes. Altogether the reads mapped onto 33,157 individual protein-coding genes of the parasite, of which 20,392 genes were expressed under all conditions and 23,879 genes with a minimum of 10 mapped transcripts under any condition. As *T. vaginalis* encodes many paralogous gene copies due to the duplication of large parts of the genome (Carlton et al., 2007), the transcript sequences of identified expressed proteins with 90% global identity were also clustered, leaving 27,719 individual protein-encoding genes being expressed when considering all libraries combined (Table 1). A set of actin genes (TVAG_337240, TVAG_054030 and TVAG_485210) was overall the highest expressed in the anaerobically grown trophozoite library (*An0*), followed by genes encoding proteins of the core carbon energy metabolism, namely pyruvate-flavodoxin oxidoreductase (TVAG_198110), phosphoenolpyruvate carboxykinase (TVAG_479540), malate dehydrogenase (TVAG_204360) and glucose-6-phosphate isomerase (TVAG_061930).

In the expression patterns presented for the individual experiments below, only those genes were considered for which at least 100 transcripts were present in each compared expression set after normalisation. Among the three conditions tested, statistical support for 94% of the top 200 up- and down-regulated genes was found (Supplementary Tables S2–S4). Further, quantitative real-time PCR on a set of exemplary genes such as rubrerythrin, cysteine protease or an ankyrin repeat-containing protein for example, further supported the deep-sequencing results (Supplementary Fig. S1).

3.2. Oxygen stress

After 5 min of oxygen stress (*AnOx5*), 33 genes were found to be up-regulated at least two-fold but none of them could directly be linked to the enzymatic machinery scavenging ROS. Genes up-regulated were for four transcription factors of the MYB family, other DNA- and RNA-binding proteins, 12 of unknown function and eight with homology to molecular switches and messengers such as a diverse range of kinases, calmodulin and a ubiquitin-conjugating enzyme (Supplementary Table S2). The expression pattern changed significantly after 30 min (Fig. 2) with 202 genes up-regulated at least two-fold. Although two MYB genes were still among the top 100, they represented different MYB genes from those observed

of transcript level when compared with *An0*. Except for one ferredoxin gene, all others mentioned here experienced a further increase. These, together with another peroxiredoxin, two thioredoxin and two rubrerythrin genes that were additionally up-regulated, totalled more than a dozen potential radical scavengers among the top 100 genes with increased expression after 120 min of oxygen stress (Supplementary Table S2). Two Nfu genes, just as the two *IscA* genes, potentially part of the machinery providing iron-sulfur clusters, were also among the top 100 after 120 min of oxygen stress. HydE and HydG, two of the three known iron-hydrogenase maturases (Pütz et al., 2006), were also found to be up-regulated. The overall increase in the transcript level compared with *An0* was higher after 120 min than after 30 min of oxygen stress and six of the nine genes that were up-regulated by more than 10 times after 120 min, were involved in the defense of ROS or in the supply of the proteins' active centers: peroxiredoxins, SODs, rubrerythrin and enzymes of the iron-sulfur cluster assembly machinery.

The down-regulation of genes under oxygen stress was less strong and peaked at 30 min by a gene encoding a cysteine protease that was down-regulated 21-fold, and which was displaced by a MYB transcription factor after 120 min, down nearly nine-fold. Next to the cysteine protease-encoding genes, another prominent gene family that was down-regulated included those encoding heat shock proteins (Hsps), in particular those of the Hsp70 family (Fig. 2). In general it was noteworthy in comparison to the up-regulation of genes, which still slightly increased after 2 h, that the down-regulation effect subsided after the 1 h.

3.3. Exposure to human epithelial cells

To identify major changes in the transcriptome of *T. vaginalis* upon contact with human vaginal epithelial tissue, and be able to separate them from the changes induced through oxygen stress

alone, two additional experiments were performed. In the first experiment, anaerobically grown parasites, identical to those used for the oxygen-stress only experiment, were exposed to a monolayer of MS74 (*AnOxInf*). In the second experiment the parasite was first adapted to oxygen by growing the cells for 5 days in the same CO₂ incubator as the human cell line, before exposing them to MS74 (*AdInf*). The infection of MS74 through T016 was monitored through light microscopy and in both experimental sets, adhesion to human cells was observed to occur instantly, i.e. a large proportion of the VECs presented with an adherent parasite after 5 min. Again transcriptomes of parasites were sequenced at 5, 30 and 120 min time points.

In the *AnOxInf* experiments, in which MS74 was infected with anaerobically grown parasites, an overall similar pattern of up-regulated genes was observed among the Eukaryotic Orthologous Groups (KOG) categories as in the oxygen-stress experiment (Fig. 3). At all three time points the top positions were held by the same genes as in the oxygen-only experiments, a MYB transcription factor (TVAG_076270) after 5 min, and a peroxiredoxin (TVAG_455310) after 30 and 120 min. Many other genes shared identical positions among the set of genes with increased expression, and the intensity of regulation was comparable. Generally the up-regulated factors were again dominated by proteins involved in dealing with oxygen stress, also including the enzymes of the iron-sulfur cluster generating machinery (Supplementary Table S3). Among the top 200 results, 31.5% of the genes whose expression was increased, were identical between just oxygen stress and oxygen stress combined with infection, and 40.1% of the down-regulated genes (Fig. 5B). In comparison, the *AdOxInf* set shared only 0.7% among the up-regulated genes with the *AnInf* set (see below), and 9.2% of the down-regulated genes. One obvious difference was observed among the down-regulated genes: a noticeable amount of genes associated with the core-carbon metabolism, in particular ME, malate dehydrogenase and fruc-

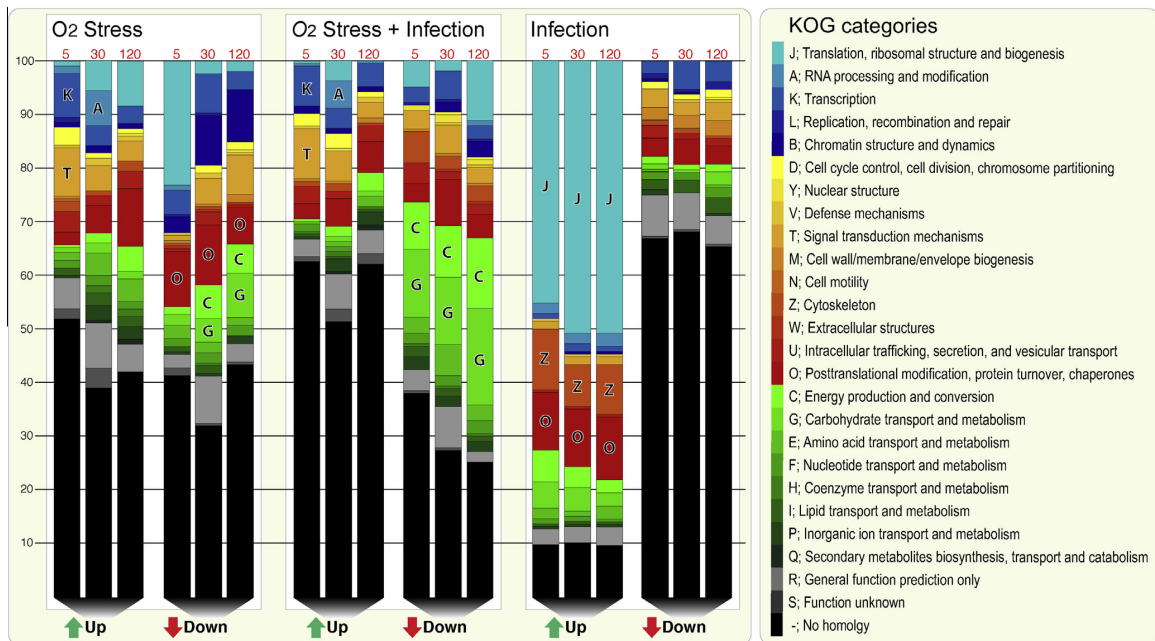


Fig. 3. Transcriptional shift of eukaryotic orthologous group categories. Regulated genes were mapped to the eukaryotic orthologous group database and each gene assigned to one of the 25 categories according to the best Basic Local Alignment Search Tool (BLAST) hit. In blue, information storage and processing; in yellow to red, cellular processes and signaling; in green, metabolism; and in grey to black, general or unknown function. Categories of particular interest are highlighted by the corresponding letter of the individual eukaryotic orthologous group category (shown on the right).

(Fig. 4). Noteworthy, the set of up-regulated genes during the first 2 h of the infection was extremely stable in comparison with the other experimental sets (Fig 5A), i.e. across all three time points approximately 200 of the genes were identical.

Next the *AdInf* data set was screened for gene families previously suggested to play a role when infection is established. For example, the large family of BspA-like proteins are thought to be important for various aspects of *T. vaginalis* pathobiology at the host-pathogen interface (Noël et al., 2010). We found expression evidence for 721 of the reported 911 putative BspA genes, albeit the vast majority was represented by a low number of mapped transcripts, with a median of 3.7 and an average of 37.2 reads/gene (max: 5446, min: 1) and compared with a general median of 19 and an average of 457 reads/gene when considering the entire data

in this study. More BspA genes were up-regulated during oxygen stress (*AnOx*), and in particular when combined with exposure to VECs (*AnOxInf*), then when oxygen-adapted parasites were exposed to VECs (*AdInf*). In the latter set the vast majority of BspA genes were observed to be down-regulated and only seven were up-regulated more than two-fold (Supplementary Table S5). We searched for further potential host-interacting and surface bound proteins of *Trichomonas* by screening for up-regulated proteins with signal peptides or a single transmembrane domain. These included previously suggested and potentially important protein families such as the leishmanolysin-like, legume-like or polymorphic membrane protein-like proteins (Carlton et al., 2007). However, apart from a few individual cases (e.g. TVAG_265530; saposin, TVAG_388060; a polymorphic membrane protein, TVAG_140850) no striking

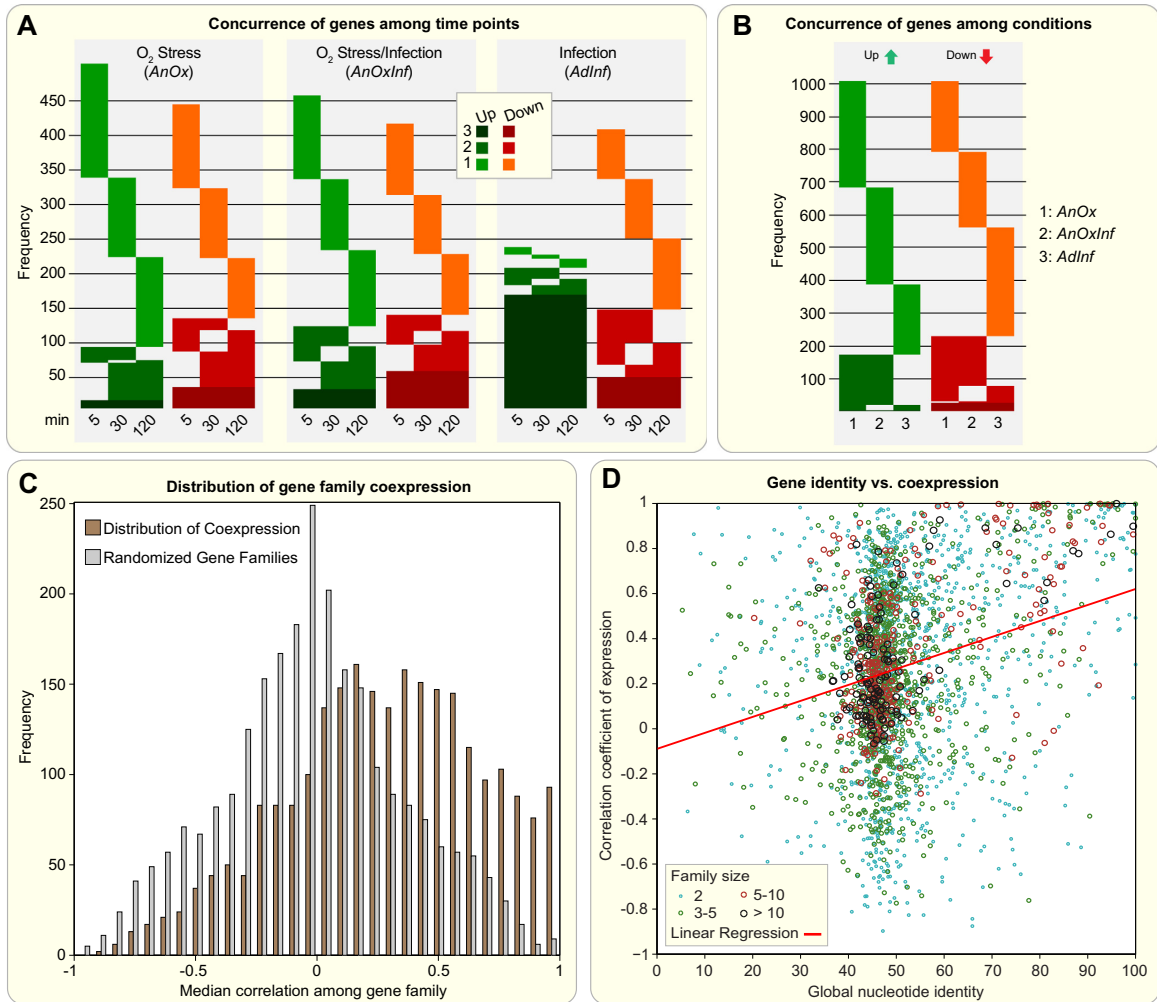


Fig. 5. Concurrence and correlated expression of *Trichomonas vaginalis* genes and gene families. (A) Concurrence of regulated genes among the individual expression sets is indicated by a color gradient. If a gene is up-regulated within only one time point this is indicated by light green; if a gene is up-regulated at all three time points, it is indicated by dark green. The highest amount of concurrence was observed among the up-regulated genes induced through the infection of MS74 with oxygen-adapted parasites (*AdInf*). (B) Co-occurrence of identical genes among the three conditions (*AnOx*, *AnOxInf*, *AdInf*), which demonstrates that oxygen alone (*AnOx*) and in combination with vaginal epithelial cells (*AnOxInf*) leads to the identical up- and down-regulation of approximately 200 genes. They differ significantly from the transcriptomes of the parasites adapted to oxygen before their exposure to vaginal epithelial cells (*AdInf*). (C) Displayed are the median expression correlations of gene family members (brown) and as a reference compared with a randomized set of gene families (grey). Expression levels of members of the same gene family correlate to a higher degree than expected by chance (Kolmogorov–Smirnow test P value <0.05). (D) Median co-expression per gene family plotted against median nucleotide identity. The linear regression line is shown in red ($r = \sim 0.25$; P value <0.05).

up-regulations were detected, especially not for entire families (Supplementary Table S6). In contrast, a protein of the tetraspanin family, TvTSP6, involved in parasite migration and potential sensor reception (de Miguel et al., 2012), was up-regulated in all of our infection libraries, as well as some hydrogenosomal proteins and proteins of the glycolytic pathway, which were reported to have a moonlighting function and serve as potential adhesion proteins, too, as mentioned in the Introduction (see also Supplementary Table S7). A surface proteome analysis revealed that 11 proteins were found more commonly present on the surface of more adherent strains (de Miguel et al., 2010). All of these genes were found expressed in T016 (a strain that was not part of the proteome analysis), with TVAG_239650 (unknown function) having the highest absolute expression. The transcriptomic response of the 11 genes is mixed and to a certain degree mirrors that of the BspA family. Genes up- or down-regulated during infection of VECs with parasites not adapted to oxygen (*AnOxInf*) did not necessarily match those up- or down-regulated during infection of VECs with oxygen-adapted parasites (Supplementary Table S8). It is noteworthy that oxygen stress alone (*AnOx*) had very little to no effect, whereas

the presence of VECs in both experimental sets (*AnOxInf* and *AdInf*) induced more significant transcriptional changes.

3.4. Regulation and co-expression of gene families

One interesting aspect of trichomonad biology is the massive expansion of gene families, whose coordinated expression has not been examined hitherto in detail due to the lack of available sequencing data. The correlation of expression among 2,509 gene families was analysed, for which we found at least two members to be identified by 100 mapped reads. A higher correlation of regulation between the members of gene families was observed when compared with randomised samples (Fig. 5C), but for many – including important gene families – the patterns were non-uniform. Whereas in some families a considerable correlation between the regulations of expression was observed, in others individual copies could be very differentially expressed. For example, for ribosomal protein-encoding genes or the actin gene family, all genes are generally expressed at high levels, the co-expression coefficient was high and expression variance was low. On the

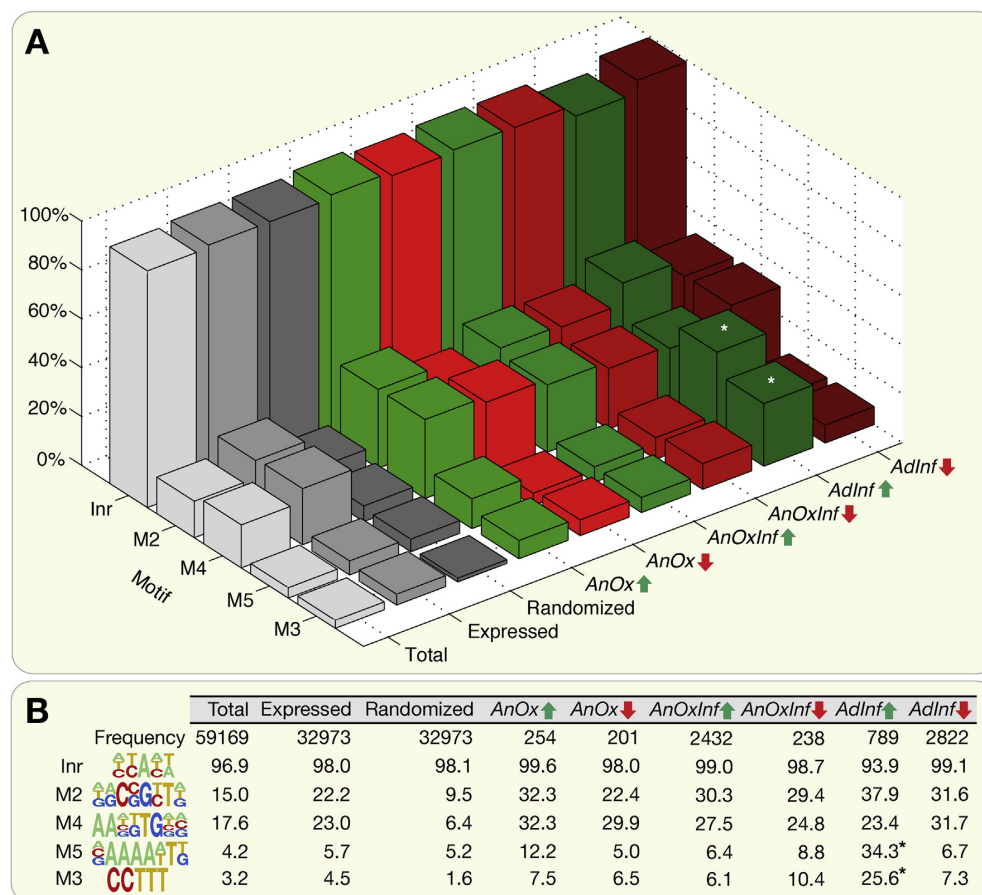


Fig. 6. *Trichomonas vaginalis* core promoter distribution. We screened the 60 bp upstream region from the start codon, for the five core promoters identified in *Trichomonas*. (A) In light grey (Total) the distribution of the promoter motifs among all genes annotated at TrichDB and medium grey (Expressed) the distribution among all genes, for which there is expression evidence. The Inr motif was found to occur at the same rate among scrambled upstream sequences (Randomized), albeit this did not take into account the distance to the start codon. The occurrence of the M3 and M5 motif is slightly increased among the genes up-regulated \geq two-fold during oxygen stress, but even more significantly among the *AdInf* set (asterisk). (B) Table of the values from (A) with the frequency of analyzed sequences given in the first row and the percentage of the individual motifs (shown in the second column) identified among the sets analyzed.

contrary, we found expression evidence for 30 genes of the Hsp70 family, within which one gene copy (TVAG_092490) was represented in total by 216,000 normalised reads, another copy (TVAG_130280) by only two. The same extreme difference was observed, but was not limited to, a glucosylceramidase and a synaptogamin family (Supplementary Table S9). Importantly, the expression correlation within an individual gene family was not random but was correlated with sequence identity, i.e. the more similar the sequences of two genes, the greater the chance their expression significantly correlates (Fig. 5D).

To elucidate whether certain promoter elements might be predominantly associated with one of the introduced environmental changes, we extracted the 5' UTRs of all annotated genes, screened them for the known core promoter regions and compared them with those genes whose expression within our libraries was regulated at least two-fold and for which at least 100 reads were identified. First, to generally elucidate promoter motif significance, we randomised the 60 bp upstream regions of all expressed genes and noticed that only the Inr motif ((A/T/C)(T/C)A(A/T/C)(T/A)), as defined by Smith et al. (2011), to be as frequently encoded among the scrambled sequences as it was among true promoter regions (Fig. 6). However, the latter does not take into account the distance to the start codon of the motif and which is likely of importance.

All five promoter motifs are enriched among the >33,000 expressed genes in comparison with the approximately 60,000 annotated at TrichDB. While the Inr motif, which mediates correct transcription initiation, was present in 98% of all the upstream regions of genes expressed, the M3 was found within only 4.5% and the M5 in 5.7% of the sequences (Fig. 6). The Inr motif experienced the lowest amount of change, the most significant being a decrease to 93.4% among the up-regulated genes of the infection set. The amount of M3 and M5 motifs changed significantly among all sets of genes but most significantly among those up-regulated during the 2 h of exposure to VECs. Twenty-five point five percent of the upstream regions of all genes up-regulated harbored a M3 motif and 34% a M5 motif, which approximately corresponded to a five- and six-fold increase, respectively. Even after subtraction of ribosomal subunit encoding genes, which are known to harbor these motifs (Smith et al., 2011), the elevation of M3 and M5 remains high at 14.9% and 7.5%, respectively.

4. Discussion

Eukaryotic genomes vary in size and protein-coding capacity ranges from approximately 2,000 in *Encephalitozoon cuniculi* to possibly even 60,000 in *T. vaginalis* (Katinka et al., 2001; Carlton et al., 2007). In the parabasal parasite, the massive coding capacity is the result of the duplication of at least large sections of the genome and additional, partly massive, expansion of some gene families (Carlton et al., 2007). However, even if it is considered that the duplication of genes can also lead to the expression of redundant copies and proteins with identical functions, the fact that almost 28,000 remain after clustering genes with 90% nucleotide identity, demonstrates the large arsenal of proteins that the parasite can tap. Our transcriptomes encompass >30,000 expressed genes and cover 98% of the 714 genes recently discussed in the microarray-based analysis of Horváthová et al. (2012) that focused on iron-regulated genes in *Trichomonas*. The difference demonstrates that deep-sequencing transcriptomes from different conditions will expand this list even further. The overall response of *T. vaginalis* to oxygen stress and upon contact with human tissue was extensive, with hundreds of at least two-fold up- and down-regulated genes, and with individual genes experiencing a 100-fold increase in gene expression.

It has been suggested that the expansion of certain gene families is not random, but rather favors those that aid the special lifestyle of *Trichomonas* (Carlton et al., 2007). However, these gene families can only then act on the parasite, if they are all expressed. By comparing the correlation of expression within gene families, we found evidence that indeed many hundreds of the expanded families, including paralogous groups of Tyrosine Kinase-Like (TKL) kinases, actin, rabs or tubulin for example, are co-regulated (Fig. 5C). What is puzzling is why so many gene families cluster around the 45–50% identity mark when plotting global nucleotide identity versus expression correlation (Fig. 5C). They do not only correspond to a limited set of certain gene families, nor do they correlate with the size of gene families. This 'cloud' might represent a footprint of an evolutionary event (genome duplication?), or maybe at this level of identity a certain threshold level is reached where any two genes that are compared can have any type of co-regulation imaginable; i.e. it could also be random. It indicates that the extension of specific gene families – or the retention thereof in comparison with others – might be related to their biological significance for the parasite's specialised lifestyle. Yet, the expansion of a gene family does not always result in high expression, as can be seen for example from the BspA family, and not always does a low or high expression level of an individual gene directly translate into a high or low amount of protein (Mair et al., 2006; Gry et al., 2009). Horváthová and colleagues (2012) also noted that among genes regulated under different iron concentrations, not all genes of a family experienced identical changes. In some cases all or the majority of paralogous copies were simultaneously regulated, in others it was only one gene of a family. Our results are congruent with those findings, for a larger sample of genes, different strains and different conditions, suggesting that this trend is general for *Trichomonas*.

The results for some genes such as those of the BspA family and for example for some of the surface-associated genes identified by de Miguel et al. (2010) are peculiar. For instance, our results for the entire BspA family confirm earlier findings for a few BspA genes, that they are up-regulated during amoeboid transition and simultaneous oxygen exposure (Noël et al., 2010). However, we found that oxygen-adapted cells do not up-regulate BspA genes upon exposure to VECs, which suggests either an oxygen-dependent component might be involved in the up-regulation of this, and perhaps other, gene families or that the strong up-regulation of, in particular, the translation machinery, in our case 'masks' less strongly regulated genes.

The many different strains of *T. vaginalis* isolated and examined to date disclose very different behaviors upon environmental stimuli (Ellis et al., 1992; Pereira-Neves and Benchimol, 2007; de Miguel et al., 2012). As the different behaviors result from different gene expression – demonstrated for example by the increased expression rate of TSP in highly-adherent strains (de Miguel et al., 2012) – this needs to be considered when comparing the sets of expression data from different *Trichomonas* strains. Also, albeit quantitative real time PCRs on individual genes can mirror patterns observed among large-scale generated data, these two techniques cannot always be directly compared. Due to the required normalisation of deep-sequencing data, genes with a very high expression will tend to mask the expression values of those with very low expression.

The parasite's core energy metabolism harbors many proteins with iron-sulfur clusters and is hence very sensitive to oxygen. High levels of oxygen therefore directly hamper substrate level phosphorylation, the only pathway for ATP synthesis within the hydrogenosomes of *Trichomonas*. However, the natural habitat of *Trichomonas* is never absolutely oxygen free and, during transmission and menstruation, the parasite experiences high levels of oxygen fluctuation (Ellis et al., 1994; Hill et al., 2005). Therefore the

parasite requires an elaborate and fast system in order to deal with oxygen stress, which our data highlights. This response can be separated into two main categories: (i) synthesizing ROS scavengers and (ii) re-synthesizing proteins, which have been damaged through ROS (Supplementary Fig. S2). Among the large set of radical scavengers that were up-regulated, one particular peroxiredoxin (TVAG_455310) appeared to have the lead role after 30 and 120 min of oxygen stress and its expression remains among the highest of all expressed genes in the oxygen-adapted parasites grown in the CO₂ incubator. The suggestion that cysteine replaces glutathione in *Trichomonas* as an antioxidant (Ellis et al., 1994) is supported by the up to nine-fold increase of two cysteine synthases (TVAG_040090 and TVAG_387920) during oxygen stress.

The response to oxygen encompasses a broad set of genes, with many members of the many radical-scavenging protein families being up-regulated at once. A large proportion of the transcriptional response appears to be dealing with the degradation of toxic ROS and shutting down energy consuming processes, in particular protein translation. Although the parasite is able to consume oxygen from the medium (Ellis et al., 1994; Chapman et al., 1999), it does not need to completely eradicate it before continuing with proliferation.

We observed a similar response in cells challenged with oxygen-stress alone and oxygen-stress combined with exposure to MS74. The extreme response to oxygen masks and possibly alters the expression shifts induced upon contact with MS74. This is underpinned by the many proteins that are down-regulated during oxygen stress, which in contrast were up-regulated during the infection of MS74 with parasites adapted in the CO₂ incubator, such as many cysteine proteases, Hsp and ribosomal subunits (Figs. 2 and 4). It shows that the cascade of events induced by oxygen stress needs to be clearly distinguished from those of host contact adhesion. Intriguingly, the oxygen stress first entails the up-regulation of transcription factors and other signaling molecules, a trend not observed during the exposure of oxygen-adapted T016 to MS74. In this latter case the translation machinery and proteins of the actin cytoskeleton are immediately among the top up-regulated genes after 5 min and this pattern does not change significantly over the 2 h analysed. It suggests that the up-regulation of these genes either requires transcription factors different from the MYB family – maybe with a WD40 repeat, as proteins with such repeats were found to be up-regulated across all three time points (Fig. 4) – or their up-regulation is independent of the up-regulation of the responsible transcription factors. That might also explain why we observed a significant increase in transcription in genes with the M3 promoter motif (Fig. 6), but no up-regulation of a recently identified M3 binding protein (TVAG_225940; Smith et al., 2011).

Trichomonas responds to oxygen with a dramatic down-regulation of the translation machinery (KOG category J, Fig. 3) and many genes encoding proteins of the KOG category O (posttranslational modification, turnover and chaperones), in particular Hsps (Figs. 2 and 3). Approximately 60% of a cell's ATP turnover is dedicated to protein synthesis, folding and posttranslational modification (Stouthamer, 1973). For *Trichomonas*, oxygen stress in direct consequence most likely also results in a depletion of the ATP pool, as core components of the hydrogenosomal metabolism are direct oxygen targets. Therefore the parasite first battles oxygen stress in a concerted effort, including the decrease of ATP consumption, before switching the cell's metabolism back to normal. The down-regulation of chaperones such as the ATPase Hsp70 perhaps reflects not only a lower demand for protein folding due to a decrease in protein synthesis, but additional means to save ATP. In accordance with this, we see that parasites that have been adapted to the oxygen conditions (*AdInf*) do the contrary when exposed to MS74: they

massively up-regulate the translational machinery, including the chaperones required for subsequent protein folding (Fig. 4).

There have been reports that several common proteins of carbon and energy metabolism moonlight as surface proteins that help the parasite to adhere to VECs (Garcia and Alderete, 2007; Mundodi et al., 2008; Meza-Cervantez et al., 2011). Specifically, the proteins ME, PFO, two subunits of SCS, GAPDH and enolase have been implicated in this function and were designated as adhesion proteins. In the oxygen-adapted and human epithelia-challenged cells (*AdInf* set), we did not observe a specific induction of transcripts beyond what was observed for enzymes of energy metabolism in general (Supplementary Table S7). Thus, if those enzymes are moonlighting for adhesion, which has been questioned (Addis et al., 1998; Brugerolle et al., 2000; Hirt et al., 2007; Ryan et al., 2011), then it does not entail a specific increase in their transcripts upon contact with VECs. In our view their up-regulation is better explained by them being part of the general up-regulation of the ATP-synthesis machinery, in particular for protein synthesis as mentioned above, rather than those enzymes acting as potential adhesion proteins on the cell surface.

The transcriptomic response of *Trichomonas* to contact with VECs underpins three fundamental processes thought to be essential for the parasite to successfully establish an infection and this is increasing the expression of genes involved in (i) protein synthesis (proliferation), (ii) phenotypic plasticity and (iii) host cell degradation (Figs. 3 and 4). Several cysteine protease-encoding genes were up-regulated across all three time points among the *AdInf* libraries – most dominantly TVAG_355480 with an up-regulation of approximately 90-fold – supporting their suspected central role regarding virulence, cytoadherence, hemolysis and cytotoxicity of *T. vaginalis* (Sommer et al., 2005; Hirt et al., 2011; Ramón-Luing et al., 2011). The major up-regulation of actin and actin-associated genes provides further evidence for this part of the cytoskeleton playing a lead role during the amoeboid transition as suggested previously (Brugerolle et al., 1996; Bricheux et al., 2000; Kusdian et al., in press). The importance of the cytoskeletal transformation of *Trichomonas* during infection is in general reflected by the increase in genes of the KOG categories N (cell motility) and Z (cytoskeleton) (Fig. 3). Flagellate-amoeboid morphogenesis is rare in nature but also occurs in another human parasite, *Naegleria fowleri* (Fulton, 1993). Finally, *Trichomonas* has been observed to not only perform binary fission, which is widely thought to be the norm, but to also become multinuclear (Yusof and Kumar, 2011); possibly allowing a faster rate of proliferation. We noticed that 60 min into the infection, some cells significantly increase their overall mass and become multinuclear, a phenotype we refer to as a juggernaut, and which is reminiscent of the schizont stage of apicomplexan parasites. This type of rapid proliferation would require massive protein synthesis and would help to explain the strong increase in genes needed for protein translation, fueled by the availability of rich substrate scavenged from the host tissue.

To summarise, *Trichomonas* expresses a core set of >24,000 genes and responds to the different environmental stimuli by the concerted regulation of hundreds of genes, with the transcript level of some increasing more than 100-fold and others decreasing more than 50-fold. The expression of the many gene families is co-regulated and the further a gene copy has diverged within a family, the less its expression correlates with that of the original “mother copy”. The massive up-regulation of ROS buffering proteins allows the parasite to survive elevated oxygen levels that allows the parasite to overcome periods of higher-than-normal oxygen levels, which are naturally part of its lifestyle. Upon contact with VECs the parasite switches to an actin-cytoskeleton driven amoeboid and induces proliferation, which is likely fueled by scavenging host tissue. The ease and speed with which morphogenesis can be

induced in *T. vaginalis* offers opportunities to study locomotion shifts from tubulin-based swimming to actin-based gliding.

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, Germany) grants (GO 1825/3-1 and MA1426/19-1) to S.B.G. and W.F.M. and the support of the “Stiftung zur Erforschung infektiös-immunologischer Erkrankungen” to S.B.G.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.ijpara.2013.04.002>.

References

- Addis, M.F., Rappelli, P., Delogu, G., Carta, F., Cappuccinelli, P., Fiori, P.L., 1998. Cloning and molecular characterization of a cDNA clone coding for *Trichomonas vaginalis* alpha-actinin and intracellular localization of the protein. *Infect. Immun.* 66, 4924–4931.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.
- Aurrecochea, C., Brestelli, J., Brunk, B.P., Carlton, J.M., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O.S., Heiges, M., Innamorato, F., Iodice, J., Kissinger, J.C., Kraemer, E., Li, W., Miller, J.A., Morrison, H.G., Nayak, V., Pennington, C., Pinney, D.F., Roos, D.S., Ross, C., Stoerckert Jr., C.J., Sullivan, S., Treatman, C., Wang, H., 2009. GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*. *Nucleic Acids Res.* 37, D526–D530.
- Benchimol, M., 2008. The hydrogenosome as a drug target. *Curr. Pharm. Des.* 14, 872–881.
- Bricheux, G., Coffe, G., Pradel, N., Brugerolle, G., 1998. Evidence for an uncommon alpha-actinin protein in *Trichomonas vaginalis*. *Mol. Biochem. Parasitol.* 95, 241–249.
- Bricheux, G., Coffe, G., Bayle, D., Brugerolle, G., 2000. Characterization, cloning and immunolocalization of a coronin homologue in *Trichomonas vaginalis*. *Eur. J. Cell Biol.* 79, 413–422.
- Brugerolle, G., Bricheux, G., Coffe, G., 1996. Actin cytoskeleton demonstration in *Trichomonas vaginalis* and in other trichomonads. *Biol. Cell* 88, 29–36.
- Brugerolle, G., Bricheux, G., Coffe, G., 2000. Immunolocalization of two hydrogenosomal enzymes of *Trichomonas vaginalis*. *Parasitol. Res.* 86, 30–35.
- Carlton, J.M., Hirt, R.P., Silva, J.C., Delcher, A.L., Schatz, M., Zhao, Q., Wortman, J.R., Bidwell, S.L., Alsmark, U.C., Besteiro, S., Sicheritz-Ponten, T., Noel, C.J., Dacks, J.B., Foster, P.G., Simillion, C., Van de Peer, Y., Miranda-Saavedra, D., Barton, G.J., Westrop, G.D., Müller, S., Dessi, D., Fiori, P.L., Ren, Q., Paulsen, I., Zhang, H., Bastida-Corcuera, F.D., Simoes-Barbosa, A., Brown, M.T., Hayes, R.D., Mukherjee, M., Okumura, C.Y., Schneider, R., Smith, A.J., Vanacova, S., Villalvazo, M., Haas, B.J., Perete, M., Feldblyum, T.V., Utterback, T.R., Shu, C.L., Osoegawa, K., de Jong, P.J., Hrdy, I., Horvathova, L., Zubacova, Z., Dolezal, P., Malik, S.B., Logsdon Jr., J.M., Henze, K., Gupta, A., Wang, C.C., Dunne, R.L., Upcroft, J.A., Upcroft, P., White, O., Salzberg, S.L., Tang, P., Chiu, C.H., Lee, Y.S., Embley, T.M., Coombs, G.H., Mottram, J.C., Tachezy, J., Fraser-Liggett, C.M., Johnson, P.J., 2007. Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science* 315, 207–212.
- Chapman, A., Linstead, D.J., Lloyd, D., 1999. Hydrogen peroxide is a product of oxygen consumption by *Trichomonas vaginalis*. *J. Biosci.* 24, 339–344.
- Cong, P., Luo, Y., Bao, W., Hu, S., 2010. Genomic organization and promoter analysis of the *Trichomonas vaginalis* core histone gene families. *Parasitol. Int.* 59, 29–34.
- Coombs, G.H., Westrop, G.D., Suchan, P., Puzova, G., Hirt, R.P., Embley, T.M., Mottram, J.C., Müller, S., 2004. The mitochondriate eukaryote *Trichomonas vaginalis* contains a divergent thioredoxin-linked peroxiredoxin antioxidant system. *J. Biol. Chem.* 279, 5249–5256.
- de Miguel, N., Lustig, G., Twu, O., Chattopadhyay, A., Wohlschlegel, J.A., Johnson, P.J., 2010. Proteome analysis of the surface of *Trichomonas vaginalis* reveals novel proteins and strain-dependent differential expression. *Mol. Cell. Proteomics* 9, 1554–1566.
- de Miguel, N., Riestra, A., Johnson, P.J., 2012. Reversible association of tetraspanin with *Trichomonas vaginalis* flagella upon adherence to host cells. *Cell. Microbiol.* 14, 1797–1807.
- Ellis, J.E., Cole, D., Lloyd, D., 1992. Influence of oxygen on the fermentative metabolism of metronidazole-sensitive and resistant strains of *Trichomonas vaginalis*. *Mol. Biochem. Parasitol.* 56, 79–88.
- Ellis, J.E., Yarlott, N., Cole, D., Humphreys, M.J., Lloyd, D., 1994. Antioxidant defences in the microaerophilic protozoan *Trichomonas vaginalis*: comparison of metronidazole-resistant and sensitive strains. *Microbiology* 140, 2489–2494.
- Enright, A.J., Van Dongen, S., Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584.
- Fulton, C., 1993. *Naegleria* – a research partner for cell and developmental biology. *J. Eukaryot. Microbiol.* 40, 520–532.
- Garcia, A.F., Alderete, J., 2007. Characterization of the *Trichomonas vaginalis* surface-associated AP65 and binding domain interacting with trichomonads and host cells. *BMC Microbiol.* 7, 116.
- Gold, D., Ofek, I., 1992. Adhesion of *Trichomonas vaginalis* to plastic surfaces: requirement for energy and serum constituents. *Parasitology* 105, 55–62.
- Gretes, M.C., Poole, L.B., Karplus, P.A., 2012. Peroxiredoxins in parasites. *Antioxid. Redox Signaling* 17, 608–633.
- Gry, M., Rimini, R., Strömberg, S., Asplund, A., Pontén, F., Uhlén, M., Nilsson, P., 2009. Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics* 10, 365.
- Hill, D.R., Brunner, M.E., Schmitz, D.C., Davis, C.C., Flood, J.A., Schlievert, P.M., Wang-Weigand, S.Z., Osborn, T.W., 2005. In vivo assessment of human vaginal oxygen and carbon dioxide levels during and post menses. *J. Appl. Physiol.* 99, 1582–1591.
- Hirt, R., Noel, C.J., Sicheritz-Ponten, T., Tachezy, J., Fiori, P., 2007. *Trichomonas vaginalis* surface proteins: a view from the genome. *Trends Parasitol.* 23, 540–547.
- Hirt, R., de Miguel, N., Nakjang, S., Dessi, D., Liu, Y.C., Diaz, N., Rappelli, P., Acosta-Serrano, A., Fiori, P.L., Mottram, J.C., 2011. *Trichomonas vaginalis* pathobiology new insights from the genome sequence. *Adv. Parasitol.* 77, 87–140.
- Horvathova, L., Šafariková, L., Basler, M., Hrdy, I., Campo, N.B., Shin, J.W., Huang, K.Y., Huang, P.J., Lin, R., Tang, P., Tachezy, J., 2012. Transcriptomic identification of iron-regulated and iron-independent gene copies within the heavily duplicated *Trichomonas vaginalis* genome. *Genome Biol. Evol.* 4, 1017–1029.
- Hrdy, I., Müller, M., 1995. Primary structure of the hydrogenosomal malic enzyme of *Trichomonas vaginalis* and its relationship to homologous enzymes. *J. Eukaryot. Microbiol.* 42, 593–603.
- Katinka, M.D., Duprat, S., Cornillot, E., Méténier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretailade, E., Brottier, P., Wincker, P., Delbac, F., El Alaoui, H., Peyret, P., Saurin, W., Gouy, M., Weissenbach, J., Vivarès, C.P., 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* 414, 450–453.
- Kulda, J., 1999. Trichomonads, hydrogenosomes and drug resistance. *Int. J. Parasitol.* 29, 199–212.
- Kusdian, G., Woehle, C., Martin, W.F., Gould, S.B., in press. The actin-based machinery of *Trichomonas vaginalis* mediates flagellate-amoeboid transition and migration across host tissue. *Cell. Microbiol.* doi:10.1111/cmi.12144.
- Lal, K., Noel, C., Field, M., Goulding, D., Hirt, R., 2006. Dramatic reorganisation of *Trichomonas* endomembranes during amoebal transformation: a possible role for G-proteins. *Mol. Biochem. Parasitol.* 148, 99–102.
- Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Genome Project Data Processing Subgroup, 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079.
- Linstead, D., Bradley, S., 1988. The purification and properties of two soluble reduced nicotinamide: acceptor oxidoreductases from *Trichomonas vaginalis*. *Mol. Biochem. Parasitol.* 27, 125–133.
- Mair, G.R., Braks, J.A., Garver, L.S., Wiegant, J.C., Hall, N., Dirks, R.W., Khan, S.M., Dimopoulos, G., Janse, C.J., Waters, A.P., 2006. Regulation of sexual development of *Plasmodium* by translational repression. *Science* 313, 667–669.
- Mentel, M., Zimorski, V., Haferkamp, P., Martin, W., Henze, K., 2008. Protein import into hydrogenosomes of *Trichomonas vaginalis* involves both N-terminal and internal targeting signals: a case study of thioredoxin reductases. *Eukaryot. Cell* 7, 1750–1757.
- Meza-Cervantez, P., González-Robles, A., Cárdenas-Guerra, R.E., Ortega-López, J., Saavedra, E., Pineda, E., Arroyo, R., 2011. Pyruvate:ferredoxin oxidoreductase (PFO) is a surface-associated cell-binding protein in *Trichomonas vaginalis* and is involved in trichomonal adherence to host cells. *Microbiology* 157, 3469–3482.
- Müller, M., 1988. Energy metabolism of protozoa without mitochondria. *Annu. Rev. Microbiol.* 42, 465–488.
- Müller, M., Mentel, M., van Hellemond, J.J., Henze, K., Woehle, C., Gould, S.B., Yu, R.Y., van der Giezen, M., Tielens, A.G., Martin, W.F., 2012. Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiol. Mol. Biol. Rev.* 76, 444–495.
- Mundodi, V., Kucknoor, A.S., Alderete, J.F., 2008. Immunogenic and plasminogen-binding surface-associated alpha-enolase of *Trichomonas vaginalis*. *Infect. Immun.* 76, 523–531.
- Noël, C.J., Diaz, N., Sicheritz-Ponten, T., Šafariková, L., Tachezy, J., Tang, P., Fiori, P.L., Hirt, R.P., 2010. *Trichomonas vaginalis* vast BspA-like gene family: evidence for functional diversity from structural organisation and transcriptomics. *BMC Genomics* 11, 99.
- Okumura, C.Y., Baum, L.G., Johnson, P.J., 2008. Galectin-1 on cervical epithelial cells is a receptor for the sexually transmitted human parasite *Trichomonas vaginalis*. *Cell. Microbiol.* 10, 2078–2090.
- Page-Sharp, M., Behm, C.A., Smith, G.D., 1996. *Tritrichomonas foetus* and *Trichomonas vaginalis*: the pattern of inactivation of hydrogenase activity by oxygen and activities of catalase and ascorbate peroxidase. *Microbiology* 142, 207–211.
- Paget, T.A., Lloyd, D., 1990. *Trichomonas vaginalis* requires traces of oxygen and high concentrations of carbon dioxide for optimal growth. *Mol. Biochem. Parasitol.* 41, 65–72.
- Pal, C., Bandyopadhyay, U., 2012. Redox-active antiparasitic drugs. *Antioxid. Redox Signal.* 17, 555–582.

- Pereira-Neves, A., Benchimol, M., 2007. Phagocytosis by *Trichomonas vaginalis*: new insights. *Biol. Cell* 99, 87–101.
- Petrin, D., Delgaty, K., Bhatt, R., Garber, G., 1998. Clinical and microbiological aspects of *Trichomonas vaginalis*. *Clin. Microbiol. Rev.* 11, 300–317.
- Pruitt, K.D., Tatusova, T., Maglott, D.R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35, D61–D65.
- Pütz, S., Dolezal, P., Gelius-Dietrich, G., Bohacova, L., Tachezy, J., Henze, K., 2006. Fe-hydrogenase maturases in the hydrogenosomes of *Trichomonas vaginalis*. *Eukaryot. Cell* 5, 579–586.
- Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842.
- Ramón-Luig, L.D., Rendón-Gandarilla, F.J., Puente-Rivera, J., Ávila-González, L., Arroyo, R., 2011. Identification and characterization of the immunogenic cytotoxic TvCP39 proteinase gene of *Trichomonas vaginalis*. *Int. J. Biochem. Cell Biol.* 43, 1500–1511.
- Rasoloson, D., Tomková, E., Cammack, R., Kulda, J., Tachezy, J., 2001. Metronidazole-resistant strains of *Trichomonas vaginalis* display increased susceptibility to oxygen. *Parasitology* 123, 45–56.
- Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* 16, 276–277.
- Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. EdgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140.
- Ryan, C.M., de Miguel, N., Johnson, P.J., 2011. *Trichomonas vaginalis*: current understanding of host-parasite interactions. *Essays Biochem.* 51, 161–175.
- Schumacher, M.A., Lau, A.O., Johnson, P.J., 2003. Structural basis of core promoter recognition in a primitive eukaryote. *Cell* 115, 413–424.
- Schwebke, J.R., Hobbs, M.M., Taylor, S.N., Sena, A.C., Catania, M.G., Weinbaum, B.S., Johnson, A.D., Getman, D.K., Gaydos, C.A., 2011. Molecular testing for *Trichomonas vaginalis* in women: results from a prospective US clinical trial. *J. Clin. Microbiol.* 49, 4106–4111.
- Smith, A., Johnson, P., 2011. Gene expression in the unicellular eukaryote *Trichomonas vaginalis*. *Res. Microbiol.* 162, 646–654.
- Smith, A.J., Chudnovsky, L., Simoes-Barbosa, A., Delgadillo-Correa, M.G., Jonsson, Z.O., Wohlschlegel, J.A., Johnson, P.J., 2011. Novel core promoter elements and a cognate transcription factor in the divergent unicellular eukaryote *Trichomonas vaginalis*. *Mol. Cell. Biol.* 31, 1444–1458.
- Smutná, T., Gonçalves, V.L., Saraiva, L.M., Tachezy, J., Teixeira, M., Hrdy, I., 2009. Flavodiiron protein from *Trichomonas vaginalis* hydrogenosomes: the terminal oxygen reductase. *Eukaryot. Cell* 8, 47–55.
- Sommer, U., Costello, C.E., Hayes, G.R., Beach, D.H., Gilbert, R.O., Lucas, J.J., Singh, B.N., 2005. Identification of *Trichomonas vaginalis* cysteine proteases that induce apoptosis in human vaginal epithelial cells. *J. Biol. Chem.* 280, 23853–23860.
- Stouthamer, A.H., 1973. A theoretical study on the amount of ATP required for synthesis of microbial cell material. *Antonie Van Leeuwenhoek* 39, 545–565.
- Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., Natale, D.A., 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4, 41.
- Upcroft, P., Upcroft, J.A., 2001. Drug targets and mechanisms of resistance in the anaerobic protozoa. *Clin. Microbiol. Rev.* 14, 150–164.
- Wagner, G., Levin, R., 1978. Oxygen tension of the vaginal surface during sexual stimulation in the human. *Fertil. Steril.* 30, 50–53.
- Williams, K.P., Leadlay, P.F., Lowe, P.N., 1990. Inhibition of pyruvate:ferredoxin oxidoreductase from *Trichomonas vaginalis* by pyruvate and its analogues. Comparison with the pyruvate decarboxylase component of the pyruvate dehydrogenase complex. *Biochem. J.* 268, 69–75.
- Yusof, A., Kumar, S., 2011. Ultrastructural changes during asexual multiple reproduction in *Trichomonas vaginalis*. *Parasitol. Res.* 110, 1823–1828.

4.4 The excavate parasite *Trichomonas vaginalis* expresses thousands of pseudogenes and long non-coding RNAs independently from neighboring genes

Christian Woehle, Gary Kusdian, Claudia Radine, Dan Graur, Giddy Landan & Sven B. Gould

Der vorliegende Artikel wurde im Februar 2014 bei der Fachzeitschrift *BMC Genomics* (Impact Factor 4,4) eingereicht.

Beitrag von Christian Wöhle, Erstautor:

Versuchsplanung	50 %
Datenanalyse	80 %
Verfassen des Manuskripts	50 %

Bestätigung über das Einreichen des Manuskripts:

Von: **BioMed Central Editorial** editorial@biomedcentral.com
Betreff: 6657326671217899 The excavate parasite *Trichomonas vaginalis* expresses thousands of pseudogenes and long non-coding RNAs independently from neighboring genes.
Datum: February 19, 2014 at 1:20 PM
An: Mr Christian Woehle Christian.Woehle@hhu.de

Article title: The excavate parasite *Trichomonas vaginalis* expresses thousands of pseudogenes and long non-coding RNAs independently from neighboring genes.

MS ID : 6657326671217899

Authors : Christian Woehle, Gary Kusdian, Claudia Radine, Dan Graur, Giddy Landan and Sven B Gould

Journal : BMC Genomics

Dear Mr Woehle

Thank you for submitting your article. This acknowledgement and any queries below are for the contact author. This e-mail has also been copied to each author on the paper, as well as the person submitting. Please bear in mind that all queries regarding the paper should be made through the contact author.

...

Best wishes,

The BMC Genomics Editorial Team

Tel: +44 (0) 20 3192 2013

e-mail: editorial@biomedcentral.com

Web: <http://www.biomedcentral.com/>

The excavate parasite *Trichomonas vaginalis* expresses thousands of pseudogenes and long non-coding RNAs independently from neighboring genes

Christian Woehle¹, Gary Kusdian¹, Claudia Radine¹, Dan Graur², Giddy Landan³, Sven B. Gould^{1*}

¹Institute of Molecular Evolution, Heinrich-Heine-University Düsseldorf, Germany

²Department of Biology and Biochemistry, University of Houston, Texas, USA

³Institute of Microbiology, Christian-Albrechts-University of Kiel, Germany

*Corresponding author

Keywords: *Trichomonas*, non-coding RNA, pseudogenes, gene families, stop codon suppression

Background: The human pathogen *Trichomonas vaginalis* is a parabasal flagellate that is estimated to infect 3% of the world's population annually. With a 160 megabase genome and up to 60,000 genes residing in six chromosomes, the parasite has the largest genome among those of sequenced protists. Although it is thought that the genome size and unusual large coding capacity is owed to genome duplication events, the exact reason and its consequences are less well studied.

Results: Among transcriptome data we found thousands of instances in which reads mapped onto genomic loci not annotated as genes, some reaching up to several kilobases in length. At first sight these appear to only represent long non-coding RNAs (lncRNAs), however, about half of these lncRNAs have significant similarities to other genomic loci annotated as protein-coding genes, providing evidence for the transcription of hundreds of pseudogenes in the parasite. In *Trichomonas*, not only the conventional lncRNAs, but also the pseudogenes are expressed through their own transcription start sites, and independently from flanking genes. Expression of several representative lncRNAs was verified through reverse-transcriptase PCR in different *T. vaginalis* strains, and case studies

exclude the use of alternative start codons or stop codon suppression for the genes analyzed.

Conclusion: Our results demonstrate that *T. vaginalis* expresses thousands of intergenic loci, including numerous transcribed pseudogenes, and in contrast to yeast these are expressed independent from neighboring genes. Our results illustrate the effect genome duplication events can have on the transcriptome of a protist. The genome is in a steady state of changing, and we hypothesize that the numerous lncRNAs could offer a large pool for potential innovation, from which novel proteins or regulatory RNA units could stem.

Background

The parabasal flagellate *Trichomonas vaginalis* is a unique human parasite causing the most common sexually transmitted disease (STD) trichomoniasis [1]. The anaerobic protist possesses the ability to rapidly shift between an amoeboid and flagellated phenotype [2, 3] and was once considered to represent an early-branching eukaryote [4]. At least 46,000 genes, and potentially up to 60,000, are encoded on six chromosomes, representing one of the highest coding capacities known [5, 6]. Exhaustive coding capacity analyses in *Trichomonas* are generally hampered through the extensive presence of repeats and transposable elements that are thought to make up 45% of the genome [7]. The expansion of the genome appears recent [5] and might coincide with the colonization of new host habitats. The genome enlargement was further fueled by a, for eukaryotes, unusual high amount of lateral gene transfer events [5, 8] and the massive expansion of some gene families [9, 10]. It has been suggested that the frequency of pseudogenes in *T. vaginalis* is at least 5% and that unstable gene families that underwent many gene duplication events – thereby producing pseudogenes on the way – further contributed to the large genome of *T. vaginalis* [11].

The transcriptome of *T. vaginalis* and its many known strains is not well characterized, but some classes of non-coding RNAs (ncRNA) have been described. Genome annotations of *T. vaginalis* include 668 ribosomal RNAs (rRNA) genes of three types and 468 transfer RNAs (tRNA) genes of 48 types [5, 7]. RNA subunits of the ribonucleoproteins RNase P and MRP were also identified [12, 13]. Furthermore, small regulatory RNAs (sRNA) have been discovered including potential microRNAs (miRNA) [14-17], small nuclear RNAs (snRNA) [18] and small nucleolar RNAs (snoRNAs) [12, 14]. Genes of the Argonaute (AGO) and Dicer-like family are encoded by *Trichomonas* and hence

suggest the existence of functional RNA interference mechanisms [5, 14], although other studies question the functionality of the miRNAs [19]. Mentioned regulatory RNAs are mostly small (<200 nucleotides), but recent reports of longer regulatory RNAs are accumulating [20-27]. Only recently deep-sequencing of the parasite's transcriptome shed light on the expression potential of the genome and provided evidence for the expression of about 30,000 genes and correlated co-expression of gene families upon different stimuli [10, 28].

Long non-coding RNAs (lncRNAs) are often defined as transcribed but not translated RNA segments larger than sRNAs (>200 nucleotides) [29]. lncRNAs affect chromosomal dynamics, the telomeres and structural organization [20, 21, 23]. Their expression can be regulated and restricted to certain developmental stages and tissues [20, 22, 24]. Some are recognized by known transcription factors [30] and their promoters can show evidence of purifying selection [26]. However, the functionality of the majority of lncRNAs is unknown, and many are thought to represent "junk" RNA or transcriptional noise attributable to the promiscuity of RNA polymerase II [31]. It has been proposed that every euchromatic nucleotide in the human genome could be transcribed [32], albeit this does obviously not necessarily translate into every expressed nucleotide having a biological function [33]. Most lncRNA studies focus on metazoan organisms [25] with yeasts representing a rare exception [27, 34-36]. Although several thousand lncRNAs have been predicted to be functional [22, 25, 37], the number of experimentally validated functional lncRNA is low, about two hundred, mainly from studies in *Homo sapiens* [38, 39]. Most lncRNAs contain only short open reading frames [39]. Still, for yeast it has been demonstrated that more than a thousand short open reading frames are translated [40].

Pseudogenes, like lncRNAs, do not encode functional proteins but can be identified through their homology to protein-coding genes from which they stem. Some are expressed and translated, but most resemble non-processed genomic remnants [41-43]. For *T. vaginalis* 1354 (~2% of predicted proteins) pseudogenes are currently annotated, but based on gene family analysis it was estimated that a minimum of 5% of the genes represent pseudogenes, as for instance half of the transmembrane cyclase family in *Trichomonas* appears to comprise pseudogenes [11]. Expressed pseudogenes are essentially a sub-group of lncRNA, and for some a biological function has been identified [42, 44]. Antisense pseudogene transcripts can be processed into small regulatory RNAs [45, 46] or complementary bind to functional counterparts and influence their expression [47, 48]. One of the best-studied functional lncRNAs that participates in X chromosome inactivation in mammals, the *Xist* RNA, is a lncRNA that originates from the pseudogenization of a protein-coding gene [49].

Here we identified and characterized lncRNAs of the parabasal parasite *T. vaginalis* by screening available transcriptional data and 270 million RNA-Seq reads we generated ourselves. We found that almost one fifth of the transcripts originate from intergenic regions of the parasite. We have characterized these transcripts in terms of their potential coding capacity, flanking genomic regions and similarity to annotated genes, in order to elucidate their origin and determine what drives their expression.

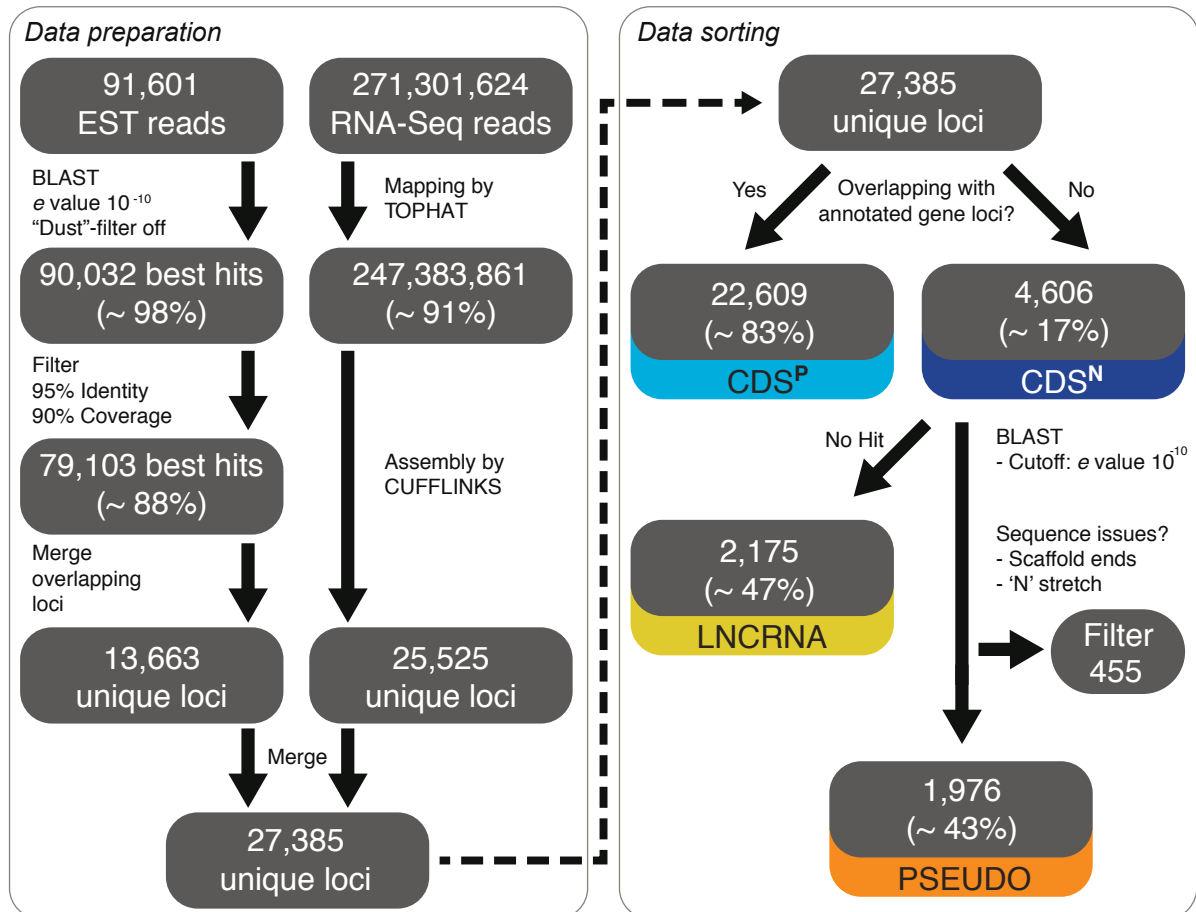


Figure 1 - Schematic workflow of data management. Sequenced reads and expressed sequence tags (ESTs) of *Trichomonas vaginalis* were mapped onto the genome as shown and sorted into the categories presented according to their best BLAST hits.

Results and discussion

General transcript mapping and homology

For our analysis we used 91,601 expressed sequence tags (ESTs) downloaded from TrichDB [7] and combined those with 271.3 million raw reads from our own RNA-Seq data. After assembling and merging the two data sets, we in total mapped 27,385 unique transcript contigs onto the genome of *Trichomonas vaginalis*. From those 22,609 (83%) mapped onto regions encoding annotated genes and 4,606 (17%) did not – we refer to these datasets as

CDS^P and CDS^N, respectively (Figure 1). The CDS^P set overlapped with 24,950 protein-coding genes, representing only 42% of annotated genes, less than half of what was found for other protists [50-53]. Yet, these transcripts represent 93% of the gene families identified in *Trichomonas* [54], indicating that (a) sequencing depth appears to be sufficient and that the numbers are not likely to change much with more sequencing data becoming available, and that (b) most of the functional proteome the genome encodes is expressed, but not all members of a gene family.

We next examined the CDS^N transcripts for their homology to annotated genes. About half (2175; 47%) had no significant similarity to any annotated genes, hence representing lncRNAs of non-recognizable origin. The remainders of the CDS^N transcripts (2431; 53%) were found to be significantly similar to annotated genes, and were thus classified as expressed pseudogenes with functional homologous genes in the same genome. These were additionally filtered to exclude contigs that mapped to the very proximal regions of genomic scaffolds and those with bad sequencing resolution, that is stretches of 'N'. 455 such contigs were identified. We termed the remaining identified set PSEUDO, and those loci without significant homologies LNCRNA (Figure 1).

The PSEUDO set includes 7% of all transcripts analyzed, and represents a lower bound on the pseudogene content of *T. vaginalis*, as this set does not include non-expressed pseudogenes, unitary pseudogenes, or pseudogenes erroneously annotated as functional genes. It has previously been estimated that at least 5% of the annotated genes of *T. vaginalis* could represent mis-annotated pseudogenes, and for one large gene family it has been shown that about half of its members could qualify as pseudogenes [55]. For the human genome it is estimated that 8 to 20% of all pseudogenes are expressed [41, 43], and if that is also true for *Trichomonas*, the parasite could potentially harbor between 10,000 and 25,000 pseudogenes. In order to estimate the number of non-expressed pseudogenes in *T. vaginalis* we performed BLASTN searches (*e* value cutoff 10^{-10}) with annotated proteins to intergenic regions lacking expression evidence. This revealed approximately 50,000 intergenic loci, for which no expression evidence exists, but with a significant homology to annotated (and likely functional) genes. Although the absolute number is much higher, the value is comparable to that from human where the amount of pseudogenes (up to 20,000) almost reaches that for the coding genes [44]. Generally, high abundances of pseudogenes are known for mammals, but their number in less complex organisms is usually smaller [56, 57]. This would support a recent hypothesis that this protist's genome (and maybe even proteome) faces constantly emerging and disappearing paralogs, and is in a steady state of changing [11].

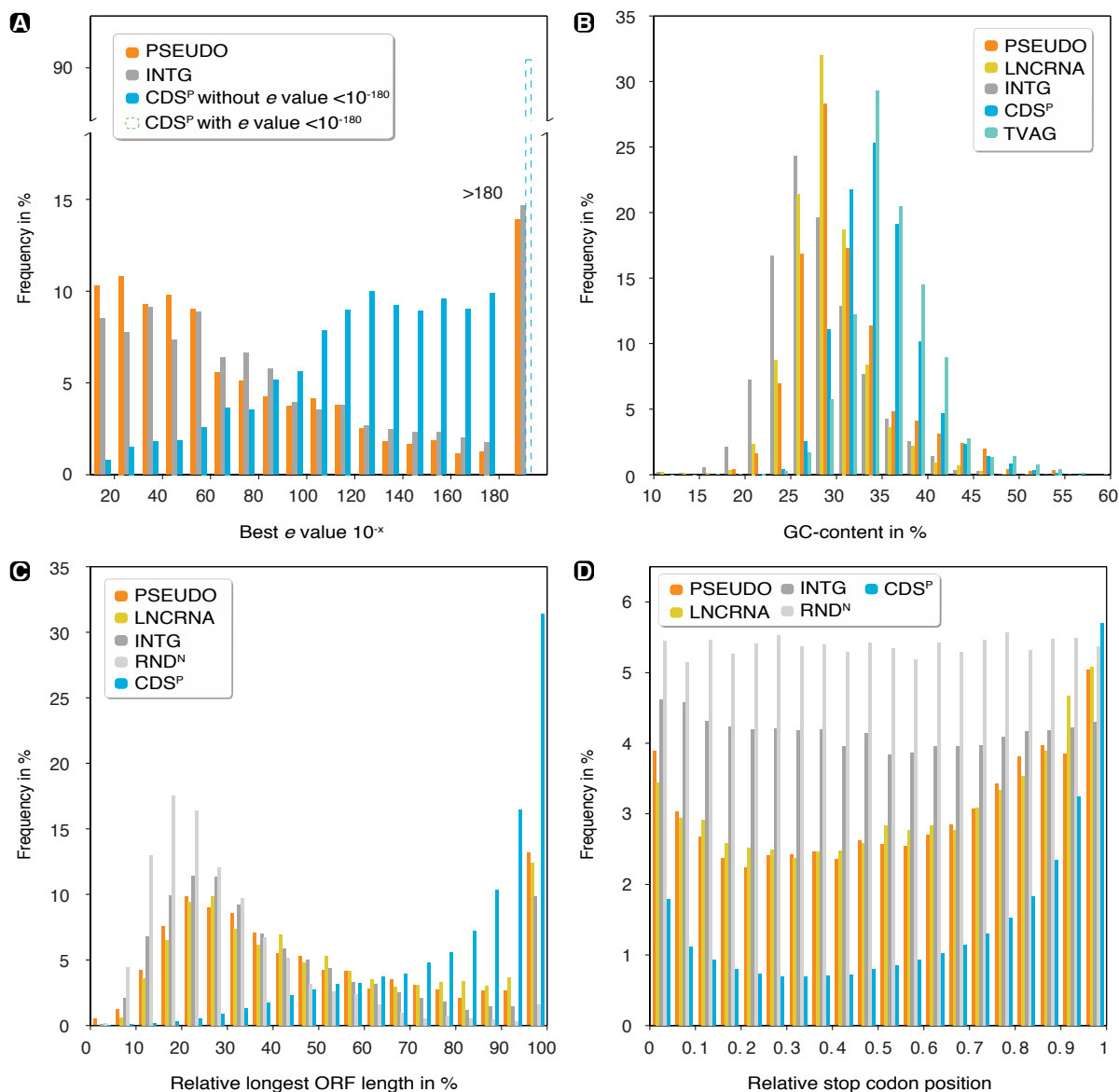


Figure 2 - Comparison of potential coding capacities for the different sets of transcripts identified. (A) shows proportions of BLASTN hits with a given e value to annotated genes of *Trichomonas vaginalis*. Relative frequencies of CDS^P were calculated excluding those e values lower 10^{-180} (the dashed bar illustrates the relation compared to all CDS^P hits). (B) Distribution of the GC-contents in percent, showing that CDS^P behaves nearly identical to TVAG (TVAG representing annotated genes of the parasite). (C) Distribution of the sequence lengths of the longest ORFs relative to the corresponding full-length sequence. The ORFs of CDS^P distribute very differently in comparison to the remaining datasets, while the intergenic regions behave similar to the PSEUDO and LNCRNA sets. (D) Distribution of stop codons over the relative positions in the full sequence of the reading frame showing the lowest number of stop codons. Counts were normalized according to total letters per bin.

In order to compare homologies of PSEUDO, CDS^P and intergenic regions (INTG) to annotated genes, we examined the distributions of their best BLASTN hit e values (Figure 2A). All compared sets differed significantly (Kolmogorov-Smirnov test; P value < 0.05 ;

Additional file 3: Table S1), with the INTG behaving similar to the CDS^N set. The BLASTN hits of PSEUDO show higher *e* values compared to CDS^P suggesting these homologies are less conserved and only partially map onto the annotated gene sequences. The several cases of pseudogenes that retrieve hits with small *e* values – indicating full sequence hits – most likely represent novel pseudogenes, that are evolutionary more recent gene duplications, and not falsely annotated genes.

17% of the transcripts did not map to any annotated genes of *T. vaginalis*. In contrast, data for the human genome suggests that half of the transcriptome consists of lncRNAs [22] and in mouse 28,000 ncRNAs were identified [37]. With more data for the parasite becoming available one will be able to determine whether this difference is simply due to the sequencing depth, which we think is unlikely, or biological differences. In any case, most will resemble transcriptional noise [31] and random expression caused for instance by sequences mimicking transcriptional promoters (see below), with only a few representing expressed and functional lncRNAs. We experimentally validated the expression of a random set of lncRNAs in the most frequently used laboratory strain T1, and the virulent T016 and highly virulent FMV1 strains. For all six cases, we could verify expression in the three *T. vaginalis* strains tested (Figure 3), which demonstrates lncRNA expression to generally be conserved across the different strains tested.

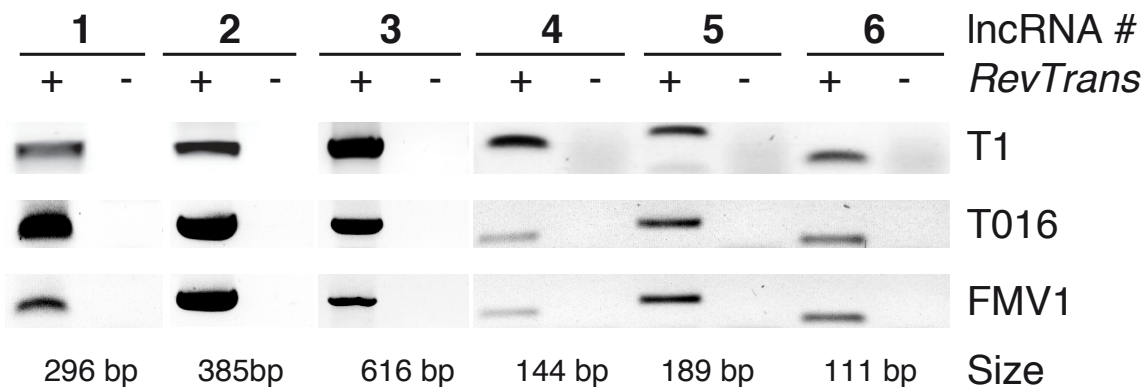


Figure 3 - Expression of lncRNAs is conserved among different *T. vaginalis* strains. Expression analysis of lncRNA candidates. Reverse transcriptase (RT-)PCR was performed on complementary DNA (cDNA) generated from wildtype RNA of *T. vaginalis* strains T1, T016 and FMV1 in the presence (+RT) or, as control, absence of the reverse transcriptase enzyme (-RT). All six lncRNAs candidates that were randomly chosen were found expressed in the three different strains analyzed.

Transcript coding capacity of CDS^N

PSEUDO, LNCRNA and CDS^P were compared in regards to their potential protein-coding capacities. Three control sets were generated: the first was based on randomized CDS^N

sequences (RND^N), the second on randomly picked intergenic loci, but with the same length distribution as the CDS^N set (INTG) and the third simply comprised all annotated *T. vaginalis* genes, that includes also those lacking expression evidence (TVAG; Table 1 and Figure 2B-D). We found that the PSEUDO and LNCRNA sets exhibit similar distributions of the several measures, and are placed in between the protein-coding CDS^P and the randomized CDS^N sets. Differences between all datasets, except PSEUDO and LNCRNA in Figure 2D, were statistically supported (Kolmogorov-Smirnov test; *P* value <0.05; Additional file 3: Table S1), where the *P* values show tendencies that CDS^P differs the most. As expected for CDS^P, this set's GC-content was found to be very similar to the GC-content described for annotated genes (34.6% versus 35%, respectively), while the GC-content of CDS^N (30.5%) was found closer that of the non-expressed intergenic sequences (28.8%). PSEUDO and LNCRNA – subsets of CDS^N – alone overall differ only slightly from the total CDS^N set, with the PSEUDO set showing only a marginal tendency towards protein-coding gene sequences (Table 1). This suggests that the PSEUDO set does not contain many, if any, genes that are not yet annotated.

Table 1 - Protein coding sequence features of the various sets analyzed.

Category	CDS ^N					
	TVAG ⁽¹⁾	CDS ^P	PSEUDO	LNCRNA	INTG ⁽²⁾	RND ^{N(3)}
Number	59672	22609	1976	2175	4606	4606
Median longest ORF length	636	1002	195	165	156	120
Mean longest ORF length	917.64	1320.23	286.64	262.63	199.45	127.05
Median relative longest ORF	99.58%	89.19%	42.11%	44.69%	34.31%	24.52%
Longest ORF ≥ 50 aa	99.59%	98.92%	64.83%	55.82%	53.58%	26.90%
Proportion of stop codons⁽⁴⁾	0.29%	1.45%	3.02%	3.08%	4.16%	5.38%
GC-Content	35.49%	34.62%	31.07%	29.42%	27.82%	30.52%

⁽¹⁾Annotated protein-coding genes

⁽²⁾Intergenic regions without expression evidence randomly selected in size of CDS^N

⁽³⁾Order of nucleotides randomized per sequence

⁽⁴⁾In reading frame with lowest number of stop codons

The relatively high amount of lncRNAs with longer open reading frames (ORFs; 55-65% ≥50 amino acids) is noteworthy. Similarities of lncRNAs to protein-coding genes have been described before and a high density of ORFs among lncRNA noted [26, 39]. We found a median ORF length of 177 nucleotides among the CDS^N set, which is lower than the median of 250 nucleotides reported for mammalian lncRNAs [39]. As expected the PSEUDO

and LNCRNA sets show a significantly lower coding capacity when compared to the CDS^P set, which demonstrates that CDS^N does not just represent erroneous protein-coding gene annotations, but largely non-coding transcripts similar to the non-expressed intergenic regions.

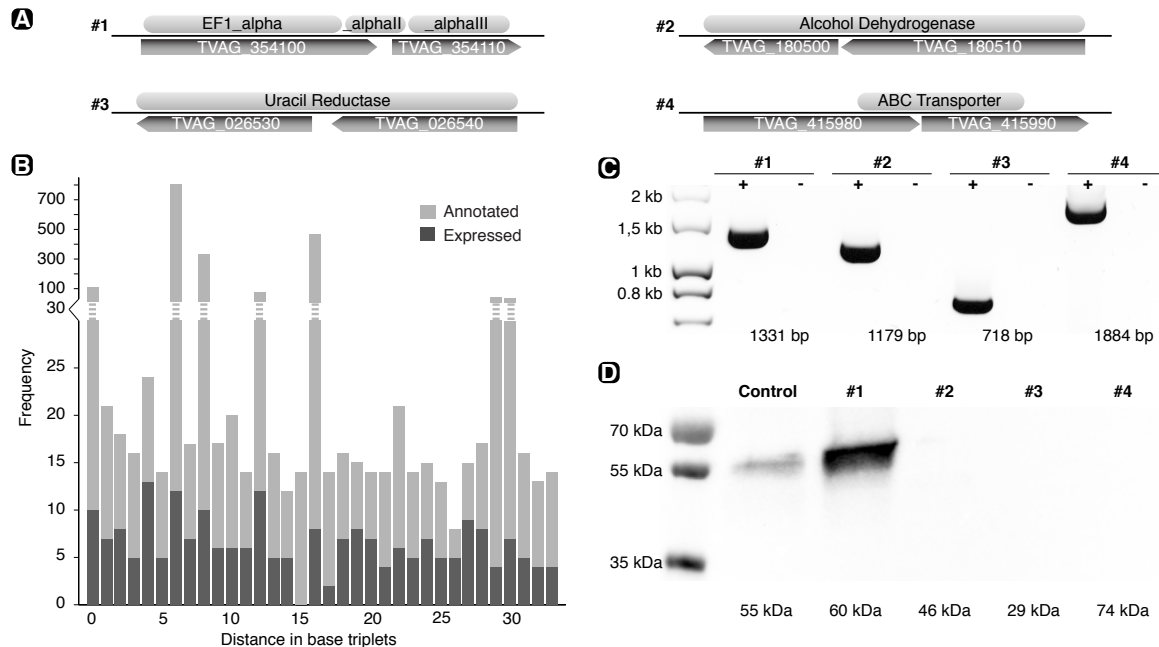


Figure 4 - No evidence for stop codon suppression in *Trichomonas vaginalis*. (A) Illustration of four selected candidates, in which two adjacent genes share the same reading frame and in combination match to a single BLAST hit. (B) Bar diagram of the frequency of annotated gene pairs and their distances in base triplets (light grey). Dark grey bars indicate the gene pairs, for which expression evidence exists. Note that the most abundant distances originate from highly conserved and large gene families. (C) RT-PCR demonstrates the full-length transcription of the gene pairs including the C-terminal HA-tag. RNA was isolated from transfected trichomonads, transcribed into complementary DNA and served as template for the PCR using specific forward and HA-reverse primers (+). RNA served as a negative control (-). (D) Western blot analysis of the same candidates shows only candidate #1 is translated. 50 μ g of protein extract loaded. TVAG_386160::HA served as a positive control.

Cui and colleagues [55] suggested stop codon read-through could explain the high number of pseudogenes in *T. vaginalis*, and which are nearly identical to their evolutionary predecessors and functional counterparts. This would mean a massive number of genes were missed during genome annotation. For a single candidate of the ABC transporter family, tentative evidence exists for stop codon suppression to occur in *Trichomonas* [58]. However, Western blot evidence for the translation of the full-length protein including its hemagglutinin (HA)-tag was not shown and the authors concluded: "...further experimental work would be required to substantiate this". In the current *T. vaginalis* genome annotation we found 2,293 cases, in which two annotated genes on the same strand are separated by

up to 33 codons (Figure 4B). For 219 of them we found expression evidence existing across their combined length. These could represent misannotations, expressed pseudogenes, or cases of stop codon suppression leading to non-interrupted translation.









We selected four candidate loci and fused the two adjacent genes to a C-terminal HA-tag (Figure 4) and checked for the transcription and translation of the fusion constructs in transfected cells. For one case (TVAG_354100 and TVAG_354110; together encoding the full-length elongation factor 1 α) the mRNA reads we obtained, and our PCR amplification product, suggested an error in the genome assembly and an incorrect annotation, as the stop codon annotated between the two genes could not be verified. This construct served as an additional control next to the expression of TVAG_386160::HA. In all cases tested we found evidence for the expression of the full-length constructs, but not for their translation (Figure 4). Only the control and the TVAG_354100::TVAG_354110 construct were translated and detectable through the C-terminal HA-tag. Alternative start codons do not appear to be used by the parasite either (Additional file 1: Figure S1A) and although the TAA stop codon is the most frequently encoded (64%), the other two – as expected – are functional (Additional file 1: Figure S1B). Hence, in summary our results confirm a conservative codon usage by the parasite, and that should stop codon suppression exist, it must be very rare and has yet to be experimentally verified.

Distribution of CDS^N relative to flanking genes

For yeast it has been reported that the expression of lncRNAs is associated with the expression of genes encoded in flanking regions [59, 60]. We analyzed the expression of the PSEUDO and LNCRNA sets of *Trichomonas vaginalis* depending on the four possible orientations to neighboring genes: divergent (\leftarrow CDS^N \rightarrow), convergent (\rightarrow CDS^N \leftarrow), co-oriented (\rightarrow CDS^N \rightarrow) and anti-oriented (\leftarrow CDS^N \leftarrow). Distances and distributions of the orientations between PSEUDO and LNCRNA did show differences (see Table 2). The distance between PSEUDO loci and flanking genes was found to be larger compared to the LNCRNA set, while the LNCRNA loci were found in divergent orientations more frequently than a convergent one. Expression of PSEUDO and LNCRNA together with flanking genes in close proximity could indicate co-expression or even the expression as one RNA molecule. To statistically test the association of co-expression with upstream or downstream genes, we performed Yates' chi-squared tests (Additional file 4: Table S2). All of the orientations tested, both for PSEUDO and LNCRNA, did not pass the false discovery rate (FDR; P value <0.05; Table 2) demonstrating no statistically significant correlation regarding the expression of these sets together with their flanking genes.

The mean intergenic distance between annotated genes in *T. vaginalis* was found to be 1165.4 [5]. The mean distances to neighboring genes for PSEUDO and LNCRNA range between 1100 and 1700 nucleotides (Table 2), being quite similar to that of the annotated genes. Overall the CDS^N, PSEUDO and LNCRNA sets behaved “autonomously” and appear independently scattered when compared to flanking, annotated gene orientation and distance. Taken together, the results indicate that these transcripts are expressed independently from their neighboring functional genes.

Table 2 - PSEUDO and LNCRNA sets are expressed in no statistically significance in correspondence to flanking genes.

Dataset	Orientation	Frequency		Mean distance		Statistics	
		Absolute	%	Upstream	Downstream	P value	FDR
PSEUDO	Convergent 	265	24.6	1419.4	1665.3	0.29	0.29
	Divergent 	260	24.2	1485.3	1543.3	0.21	0.29
	Co-oriented 	295	27.4	1286.8	1511.5	0.22	0.29
	Anti-oriented 	256	23.8	1459.9	1508.0	0.03	0.10
LNCRNA	Convergent 	233	17.5	1266.9	1207.4	0.42	0.55
	Divergent 	434	32.6	1250.0	1162.6	0.13	0.34
	Co-oriented 	329	24.7	1145.1	1283.7	0.69	0.69
	Anti-oriented 	334	25.1	1430.1	1106.4	0.17	0.34

PSEUDO and LNCRNA are transcribed, but lack obvious translation start motifs

Several promoter motifs including the DNA initiator motif (Inr) have been identified in *T. vaginalis* [61], and some are linked to the expression of gene subsets induced through changing environmental conditions [10]. We screened the upstream regions of the expressed intergenic loci we identified for overrepresented motifs (Figure 5). A motif similar to the Inr motif of the CDS^P (that is canonical annotated and expressed protein-encoding genes) was well represented among upstream sequences of all expressed loci (PSEUDO, LNCRNA). With 16.8% for LNCRNA and 15.5% for PSEUDO, the frequency of the most prominent Inr motif was comparable to the 19.9% of the CDS^P set (Additional file 2: Figure S2). Among all loci we identified one non-functional pattern recently described as the M2 motif (AAAGTGAC) [61], but only among the CDS^P set did we identify the translation-associated M4 motif (AAAAT[T/G]) and other translation start motifs containing methionine codons (Figure 5). PSEUDO and LNCRNA show approximately the same abundances of known transcription-associated motifs, while lacking any evidence for translation-associated motifs. INTG

sequences, for which we found no expression evidence, do not encode any of the previously described motifs, except M2, but with very low frequency.

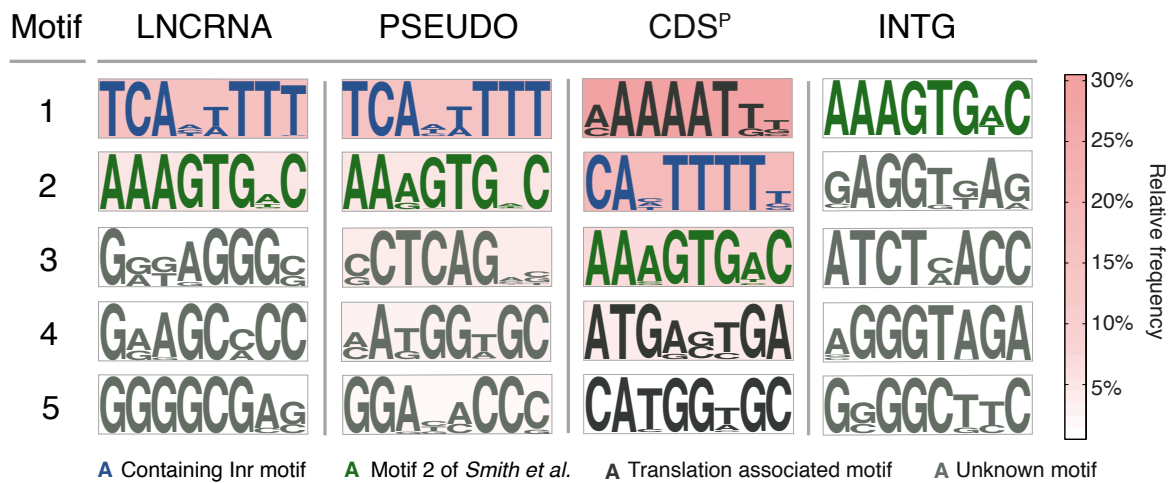


Figure 5 - Promoter sequence distribution. Shown are pictograms and scores for the five best motifs (sorted by motif abundances) of the PSEUDO, LNCRNA, CDS^P and INTG sets. Background color gradient indicates the frequency with which the motifs were identified. Note that the Inr motif of the CDS^P set misses the initial ‘T’; manual inspection revealed that 64% did however encode it. Translation initiation motifs containing an ‘ATG’ are only found among the CDS^P set.

Taken together this demonstrates that lncRNAs and pseudogenes in the parabasal parasite are not expressed as byproducts and in dependence to neighboring genes as found for other model organisms [60], but because of their own transcriptional initiator motifs. As suggested by Carvunis and colleagues [40], and supported by our data, it is possible that the LNCRNA loci only represent an intermediate and transient form of genetic elements with characteristics from both functional proteins and intergenic regions. In either case, they would not just represent transcriptional noise and could serve as a sequence pool for the development of novel functional genes. This would further explain the high number of ORFs identified among the loci and the presence of fully functional promoter motifs, yet it is too early to tell whether any of these fulfill an actual biological function.

Conclusion

By far the most information available for lncRNA is from mammals [38]. Apart from yeast [27, 35] no analysis dedicated to the characterization of lncRNA or pseudogene expression in protists is available. Our results provide insight into the expression of lncRNAs of a representative of the not well-studied eukaryotic kingdom of excavates. The expression of lncRNAs and pseudogenes in the parabasal parasite *Trichomonas vaginalis* is extensive.

Almost one-fifth of the transcripts mapped onto non-coding genomic loci, of which only half appeared to have no homology to annotated genes of the protist. They do not encode for canonical proteins, but are clearly distinct from random sequences we analyzed as control sets. Intriguingly and in contrast to yeast [59], the expression of intergenic DNA does not appear to be associated with annotated neighboring genes, but rather driven by transcription start signals mimicking those of coding genes. The fact that half of the lncRNAs expressed are pseudogenes reflects the dynamic nature of the *Trichomonas* genome, characterized by an unknown amount of duplications of at least parts of the genome and large gene families.

Material and Methods

Culture, RNA Isolation and cDNA synthesis

Trichomonas vaginalis strains T1, T016 and FMV1 were cultivated in tryptone-yeast extract maltose-medium (2.22% (w/v) tryptose, 1.11% (w/v) yeast extract, 15 mM maltose, 9.16 mM L-cysteine, 1.25 mM L(+)-ascorbic acid, 0.77 mM KH₂PO₄, 3.86 mM K₂HPO₄, 10% (v/v) horse serum, 0.71% (v/v) iron solution (= 1% (w/v) Fe(NH₄)₂(SO₄)x6H₂O, 0.1% (w/v) 5-sulfosalicylic acid)) at 37°C in Falcon tubes. To prevent bacterial contamination a penicillin/streptomycin mix was added to a final concentration of 100 µg/ml to media. Approximately 2.5x10⁸ cells were pelletized at 1,000x g for 10 min at 8 °C and total RNA isolated using TRIzol[®] (Invitrogen) according to the manufacturer's protocol. RNA was additionally digested with DNase (DNase I, RNase-free, Thermo Scientific). 1 µg of DNase digested RNA was transcribed into cDNA using the "SuperScript III First-Strand Synthesis System for RT-PCR Kit" (Invitrogen) with specific primers as stated below or the iScript Select cDNA Synthesis Kit (Bio-Rad) using its random primer mix according to manufacturer's protocol. The synthesized cDNA was used as template for test-PCRs using specific primers (Additional file 5: Table S3). Amplification products were sequenced for verification.

Sequencing, mapping and assembly

RNA-Seq reads were produced by Illumina sequencing of *Trichomonas vaginalis* under different conditions (infection and/or oxygen stress at several time points). *T. vaginalis* was cultured and RNA isolated as described in [10] and deep-sequencing was performed by Eurofins MWG (Ebersberg, Germany). Two sequencing approaches had been used: 100 basepairs paired-end reads. The filtered and trimmed reads used here are deposited in Sequence Read Archive (SRA) [62] under Accession SRA059159 (3'-library) and SRA129698 (paired-end reads).

Genomic scaffolds of *Trichomonas vaginalis*, sequences of annotated genes, genomic features (General Feature Format), orthologous gene clusters and additional EST sequences were downloaded from TrichDB V1.3 [7, 54]. The reads of both RNA-Seq sequencings were mapped separately to the draft genome and the corresponding genome annotations of *Trichomonas vaginalis* using TOPHAT2 [63]. Assembly of overlapping reads was performed by CUFFLINKS [64] and the results of the two samples were merged by CUFFMERGE [64]. We supplemented the RNA-Seq with additional ESTs from TrichDB. ESTs were matched to the *T. vaginalis* scaffolds using BLASTN [65] with disabled filtering. Best BLASTN hits with an identity of at least 95% and query coverage of at least 90% were extracted, and overlapping hits were merged to unique loci and combined with overlapping loci from the RNA-Seq experiments using BEDTOOLS [66]. Transcribed loci on smaller scaffolds (<1000 nucleotides) were discarded due to missing gene annotations [5].

Classification of transcribed loci

From downloaded genome features of *Trichomonas vaginalis* only “gene” entries were retained and used to search for overlapping between our transcribed loci and the gene annotations. Overlapping regions were classified as CDS^P while the remaining ones were referred to as CDS^N. We created two datasets to serve as controls. For the intergenic dataset (INTG) we extracted all sequences longer than 1000 basepairs from the *T. vaginalis* scaffolds that were not annotated as genes, not identified through mapped transcripts (CDS^N and CDS^P) and located in close proximity to scaffold ends. From these we randomly sampled sequences of the same lengths as those in CDS^N, thus ensuring an identical length distribution. As a second control set we subjected CDS^N sequences to a random permutation of nucleotide order (RND^N). Homologies to annotated *T. vaginalis* genes were inferred by BLASTN searches of CDS^N, CDS^P and INTG against the annotated gene sequences, with an *e* value cutoff of 10^{-10} . CDS^N loci without hits were classified as LNCRNA. Loci of CDS^N with hits were removed if either the hit or the query sequence included undetermined nucleotides (“N”), or was prematurely terminated due to scaffold termination. Remaining CDS^N loci were classified as PSEUDO. Estimates for non-expressed pseudogenes were produced by taking all BLAST hits of annotated genes to intergenic regions with an *e* value cutoff of 10^{-10} and merging those with overlapping locations into single entries.

Information on which strand transcribed loci are encoded were inferred by counting TOPHAT hits of the 3'-libraries, that are overlapping with the corresponding gene locations. An orientation was assigned, if at least 90% of the matching hits lead to the same orientation. A control with CDS^P and the corresponding genes, for which orientations are known,

revealed that for 86% of them a unique orientation was identified and 95.4% of them were congruent with overlapping annotations. For CDS^N we were able to assign orientations for 79% of the loci.

Protein-coding capacities were examined by two different methods. The length of the longest ORFs was defined as the longest peptide sequence in any reading frame beginning with the start of the sequence or a methionine and ending at the next stop codon or the end of the sequence. We defined the frequency of stop codons as minimum count found inspecting all six reading frames separately.

Flanking regions and stop codon read-through

For motif search upstream regions of transcribed loci were extracted -60 to 40 basepairs relative to the start position. Resulting sequences were clustered using CDHIT [67] with a cutoff of 90%. A search for the most overrepresented motifs was conducted using the MEME software V4.7 [68] with window size of 6-8 and zero or one occurrences per sequence. Orientations and distances of transcribed loci to surrounding annotated genes were extracted from genome annotations of scaffolds using their locations.

Candidates for stop codon read-through were determined by examining locations of genome features. We searched for gene pairs on the same strand with a distance from 0 to 33 full codons (99 nucleotides). Transcription of connected genes was determined by using CUFFLINKS results for the paired-end libraries only. Assembled transcripts had to span at least from the stop codon of the one gene to the start codon of the other.

Cloning and transfection

All fragments were cloned into expression vector pTagvag2; for primer sequences refer to Additional file 5: Table S3. For IncRNA_ATG the artificial SCS promoter of pTagvag2 [69] was replaced by the putative, endogenous promoter region of the candidate (309 bp upstream of open reading frame). To check if all three classical stop codons are valid in *T. vaginalis*, we altered the stop codon of the HA-tag (TAA) into TGA and TAG and checked the length of the translation of the actin derivative TVAG_054030 (Additional file 1: Figure S1B). To identify potential stop codon suppression, pairs of adjacent genes, for which combined expression evidence was found based on our RNA-Seq data, fragments were amplified with the 5' oligonucleotide binding to the start codon of first gene and the 3' oligonucleotide replacing the stop codon of the adjacent gene with an HA-tag (Additional file 5: Table S3). All gene sequences were amplified using a proof-reading polymerase and verified through sequencing. 30 μ g of the plasmid DNA was used for transfection of roughly

2.5×10^8 *T. vaginalis* cells using standard electroporation [70]. After four hours of incubation neomycine (G418) was added to a final concentration of 100 $\mu\text{g/ml}$ for selection.

Protein samples were separated through standard SDS-PAGE and blotted onto nitrocellulose membrane. Membranes were blocked in 5% milk powder in Tris-buffered saline pH7 (blocking buffer) for 30 min. Blots were incubated with the primary antibody at a dilution of 1:5,000 in blocking buffer either overnight (ON) at 4°C or for 1h at room temperature (RT) and then washed 3x with TBS-T (TBS + 0.1% Tween 20), followed by the incubation with the secondary antibody (1:10,000) and identical subsequent washes. Detection of the chemiluminescence signal was performed through the SuperSignal West Pico Chemiluminescent Substrate Kit (Thermo Scientific) according to the manufactures protocol. Antibodies used: monoclonal HA-antibody (Sigma H9658) and ImmunoPure Goat Anti-Mouse IgG (Pierce 31430).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SGB conceived the study. DG and GL participated in design of experiments and interpretation of results. CW performed the bioinformatic analyses and GK and CR carried out the laboratory experiments. SBG, GK and CW drafted the manuscript.

Acknowledgments

This work was funded by a Deutsche Forschungsgemeinschaft grant (DFG GO1825/3-1) to SBG, a German-Israeli Foundation grant (I-1207-264.13/2012) to SBG, and a DFG grant to William Martin (DFG MA1426/19-1).

References

1. Petrin D, Delgaty K, Bhatt R, Garber G: **Clinical and microbiological aspects of *Trichomonas vaginalis***. *Clin Microbiol Rev* 1998, **11**:300-317.
2. Lal K, Noel CJ, Field MC, Goulding D, Hirt RP: **Dramatic reorganisation of *Trichomonas* endomembranes during amoebal transformation: A possible role for G-proteins**. *Mol Biochem Parasitol* 2006, **148**:99-102.

3. Kusdian G, Woehle C, Martin WF, Gould SB: **The actin-based machinery of *Trichomonas vaginalis* mediates flagellate-amoeboid transition and migration across host tissue.** *Cell Microbiol* 2013, **15**:1707-1721.
4. Embley TM, Hirt RP: **Early branching eukaryotes?** *Curr Opin Genet Dev* 1998, **8**:624-629.
5. Carlton JM, Hirt RP, Silva JC, Delcher AL, Schatz M, Zhao Q, Wortman JR, Bidwell SL, Alsmark UC, Besteiro S, Sicheritz-Ponten T, Noel CJ, Dacks JB, Foster PG, Simillion C, Van de Peer Y, Miranda-Saavedra D, Barton GJ, Westrop GD, Müller S, Dessi D, Fiori PL, Ren Q, Paulsen I, Zhang H, Bastida-Corcuera FD, Simoes-Barbosa A, Brown MT, Hayes RD, Mukherjee M, *et al.*: **Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*.** *Science* 2007, **315**:207-212.
6. Smith A, Johnson P: **Gene expression in the unicellular eukaryote *Trichomonas vaginalis*.** *Res Microbiol* 2011, **162**:646-654.
7. Aurrecochea C, Brestelli J, Brunk BP, Carlton JM, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Innamorato F, Iodice J, Kissinger JC, Kraemer E, Li W, Miller JA, Morrison HG, Nayak V, Pennington C, Pinney DF, Roos DS, Ross C, Stoeckert CJ Jr, Sullivan S, Treatman C, Wang H: **GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*.** *Nucleic Acids Res* 2009, **37**:D526-530.
8. Alsmark UC, Sicheritz-Ponten T, Foster PG, Hirt RP, Embley TM: **Horizontal gene transfer in eukaryotic parasites: a case study of *Entamoeba histolytica* and *Trichomonas vaginalis*.** *Methods Mol Biol* 2009, **532**:489-500.
9. Noël CJ, Diaz N, Sicheritz-Ponten T, Safarikova L, Tachezy J, Tang P, Fiori P-L, Hirt RP: ***Trichomonas vaginalis* vast BspA-like gene family: evidence for functional diversity from structural organisation and transcriptomics.** *BMC Genomics* 2010, **11**:99.
10. Gould SB, Woehle C, Kusdian G, Landan G, Tachezy J, Zimorski V, Martin WF: **Deep sequencing of *Trichomonas vaginalis* during the early infection of vaginal epithelial cells and amoeboid transition.** *Int J Parasitol* 2013, **43**:707-719.
11. Cui J, Das S, Smith TF, Samuelson J: ***Trichomonas* transmembrane cyclases result from massive gene duplication and concomitant development of pseudogenes.** *PLoS Negl Trop Dis* 2010, **4**:e782.

12. Chen XS, Penny D, Collins LJ: **Characterization of RNase MRP RNA and novel snoRNAs from *Giardia intestinalis* and *Trichomonas vaginalis*.** *BMC Genomics* 2011, **12**:550.
13. Piccinelli P, Rosenblad MA, Samuelsson T: **Identification and analysis of ribonuclease P and MRP RNA in a broad range of eukaryotes.** *Nucleic Acids Res* 2005, **33**:4485-4495.
14. Chen XS, Collins LJ, Biggs PJ, Penny D: **High throughput genome-wide survey of small RNAs from the parasitic protists *Giardia intestinalis* and *Trichomonas vaginalis*.** *Genome Biol Evol* 2009, **1**:165-175.
15. Lin WC, Huang KY, Chen SC, Huang TY, Chen SJ, Huang PJ, Tang P: **Malate dehydrogenase is negatively regulated by miR-1 in *Trichomonas vaginalis*.** *Parasitol Res* 2009, **105**:1683-1689.
16. Lin WC, Li SC, Lin WC, Shin JW, Hu SN, Yu XM, Huang TY, Chen SC, Chen HC, Chen SJ, Huang PJ, Gan RR, Chiu CH, Tang P: **Identification of microRNA in the protist *Trichomonas vaginalis*.** *Genomics* 2009, **93**:487-493.
17. Huang PJ, Lin WC, Chen SC, Lin YH, Sun CH, Lyu PC, Tang P: **Identification of putative miRNAs from the deep-branching unicellular flagellates.** *Genomics* 2012, **99**:101-107.
18. Simoes-Barbosa A, Meloni D, Wohlschlegel JA, Konarska MM, Johnson PJ: **Spliceosomal snRNAs in the unicellular eukaryote *Trichomonas vaginalis* are structurally conserved but lack a 5'-cap structure.** *RNA* 2008, **14**:1617-1631.
19. Tarver JE, Donoghue PCJ, Peterson KJ: **Do miRNAs have a deep evolutionary history?** *Bioessays* 2012, **34**:857-866.
20. Amaral PP, Mattick JS: **Noncoding RNA in development.** *Mamm Genome* 2008, **19**:454-492.
21. Hu W, Alvarez-Dominguez JR, Lodish HF: **Regulation of mammalian cell differentiation by long non-coding RNAs.** *EMBO Rep* 2012, **13**:971-983.
22. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras TR: **RNA maps reveal new RNA classes and a possible function for pervasive transcription.** *Science* 2007, **316**:1484-1488.
23. Mercer TR, Dinger ME, Mattick JS: **Long non-coding RNAs: insights into functions.** *Nat Rev Genet* 2009, **10**:155-159.

24. Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS: **Specific expression of long noncoding RNAs in the mouse brain.** *Proc Natl Acad Sci U S A* 2008, **105**:716-721.
25. Nam JW, Bartel DP: **Long noncoding RNAs in *C. elegans*.** *Genome Res* 2012, **22**:2529-2540.
26. Ponjavic J, Ponting CP, Lunter G: **Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs.** *Genome Res* 2007, **17**:556-565.
27. van Werven FJ, Neuert G, Hendrick N, Lardenois A, Buratowski S, van Oudenaarden A, Primig M, Amon A: **Transcription of two long noncoding RNAs mediates mating-type control of gametogenesis in budding yeast.** *Cell* 2012, **150**:1170-1181.
28. Huang KY, Chen YY, Fang YK, Cheng WH, Cheng CC, Chen YC, Wu TE, Ku FM, Chen SC, Lin R, Tang P: **Adaptive responses to glucose restriction enhance cell survival, antioxidant capability, and autophagy of the protozoan parasite *Trichomonas vaginalis*.** *Biochim Biophys Acta* 2014, **1840**:53-64.
29. Collins LJ: **Characterizing ncRNAs in human pathogenic protists using high-throughput sequencing technology.** *Front Genet* 2011, **2**:96.
30. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Wong B, Drenkow J, Yamanaka M, Patel S, Brubaker S, Tammana H, Helt G, Struhl K, Gingeras TR: **Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs.** *Cell* 2004, **116**:499-509.
31. Struhl K: **Transcriptional noise and the fidelity of initiation by RNA polymerase II.** *Nat Struct Mol Biol* 2007, **14**:103-105.
32. The ENCODE Project Consortium: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447**:799-816.
33. Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E: **On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE.** *Genome Biol Evol* 2013, **5**:578-590.
34. Dutrow N, Nix DA, Holt D, Milash B, Dalley B, Westbrook E, Parnell TJ, Cairns BR: **Dynamic transcriptome of *Schizosaccharomyces pombe* shown by RNA-DNA hybrid mapping.** *Nat Genet* 2008, **40**:977-986.

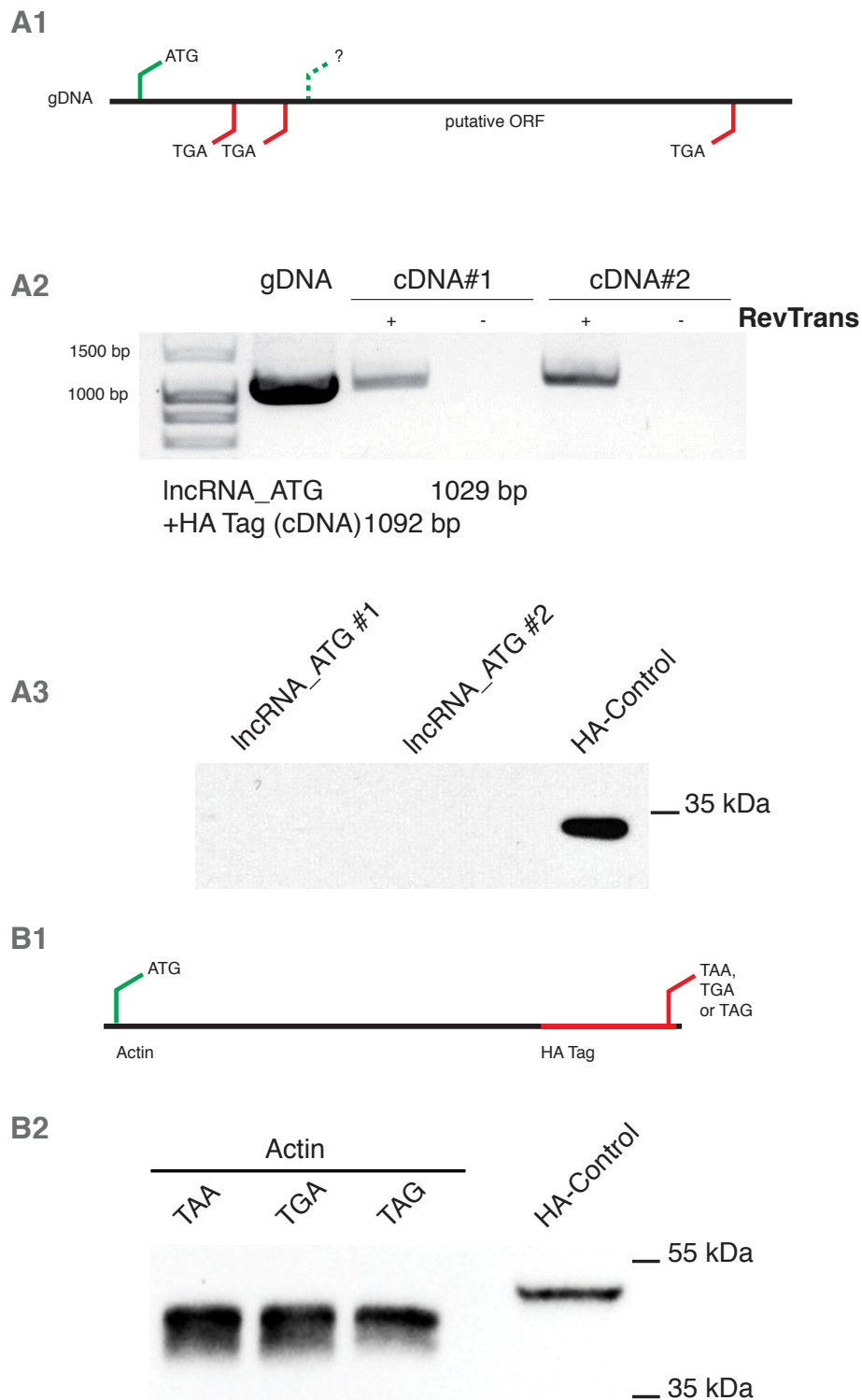
35. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: **The transcriptional landscape of the yeast genome defined by RNA sequencing.** *Science* 2008, **320**:1344-1349.
36. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bähler J: **Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution.** *Nature* 2008, **453**:1239-1243.
37. Liu J, Gough J, Rost B: **Distinguishing protein-coding from non-coding RNAs through support vector machines.** *PLoS Genet* 2006, **2**:e29.
38. Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS: **IncRNAdb: a reference database for long noncoding RNAs.** *Nucleic Acids Res* 2011, **39**:D146-151.
39. Niazi F, Valadkhan S: **Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3' UTRs.** *RNA* 2012, **18**:825-843.
40. Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charlotheaux B, Hidalgo CA, Barbette J, Santhanam B, Brar GA, Weissman JS, Regev A, Thierry-Mieg N, Cusick ME, Vidal M: **Proto-genes and de novo gene birth.** *Nature* 2012, **487**:370-374.
41. Zheng D, Frankish A, Baertsch R, Kapranov P, Reymond A, Choo SW, Lu Y, Denoeud F, Antonarakis SE, Snyder M, Ruan Y, Wei CL, Gingeras TR, Guigó R, Harrow J, Gerstein MB: **Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution.** *Genome Res* 2007, **17**:839-851.
42. Poliseno L: **Pseudogenes: newly discovered players in human cancer.** *Sci Signal* 2012, **5**:re5.
43. The ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**:57-74.
44. Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, Carter DR: **Pseudogenes: pseudo-functional or key regulators in health and disease?** *RNA* 2011, **17**:792-798.
45. Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM, Hannon GJ: **Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes.** *Nature* 2008, **453**:534-538.

46. Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T, Surani MA, Sakaki Y, Sasaki H: **Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes.** *Nature* 2008, **453**:539-543.
47. Hawkins PG, Morris KV: **Transcriptional regulation of Oct4 by a long non-coding RNA antisense to Oct4-pseudogene 5.** *Transcription* 2010, **1**:165-175.
48. Korneev SA, Park JH, O'Shea M: **Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene.** *J Neurosci* 1999, **19**:7711-7720.
49. Duret L, Chureau C, Samain S, Weissenbach J, Avner P: **The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene.** *Science* 2006, **312**:1653-1655.
50. Kolev NG, Franklin JB, Carmi S, Shi HF, Michaeli S, Tschudi C: **The Transcriptome of the Human Pathogen *Trypanosoma brucei* at Single-Nucleotide Resolution.** *PloS Pathog* 2010, **6**:e1001090.
51. Nookaew I, Papini M, Pornputtpong N, Scalcinati G, Fagerberg L, Uhlén M, Nielsen J: **A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*.** *Nucleic Acids Res* 2012, **40**:10084-10097.
52. Xiong J, Lu XY, Zhou ZM, Chang Y, Yuan DX, Tian M, Zhou ZG, Wang L, Fu CJ, Orias E, Miao W: **Transcriptome analysis of the model protozoan, *Tetrahymena thermophila*, using deep RNA sequencing.** *PloS One* 2012, **7**:e30630.
53. Dyhrman ST, Jenkins BD, Rynearson TA, Saito MA, Mercier ML, Alexander H, Whitney LP, Drzewianowski A, Bulygin VV, Bertrand EM, Wu Z, Benitez-Nelson C, Heithoff A: **The transcriptome and proteome of the diatom *Thalassiosira pseudonana* reveal a diverse phosphorus stress response.** *PLoS One* 2012, **7**:e33768.
54. Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS: **OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups.** *Nucleic Acids Res* 2006, **34**:D363-368.
55. Cui J, Smith TF, Samuelson J: **Gene expansion in *Trichomonas vaginalis*: a case study on transmembrane cyclases.** *Genome Inform* 2007, **18**:35-43.
56. Zhang Z, Gerstein M: **Large-scale analysis of pseudogenes in the human genome.** *Curr Opin Genet Dev* 2004, **14**:328-335.

57. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, Böhme U, Hannick L, Aslett MA, Shallom J, Marcello L, Hou L, Wickstead B, Alsmark UC, Arrowsmith C, Atkin RJ, Barron AJ, Bringaud F, Brooks K, Carrington M, Cherevach I, Chillingworth TJ, Churcher C, Clark LN, Corton CH, Cronin A, *et al.*: **The genome of the African trypanosome *Trypanosoma brucei***. *Science* 2005, **309**:416-422.
58. Kay C, Woodward KD, Lawler K, Self TJ, Dyall SD, Kerr ID: **The ATP-binding cassette proteins of the deep-branching protozoan parasite *Trichomonas vaginalis***. *PLoS Negl Trop Dis* 2012, **6**:e1693.
59. Wang GZ, Lercher MJ, Hurst LD: **Transcriptional coupling of neighboring genes and gene expression noise: evidence that gene orientation and noncoding transcripts are modulators of noise**. *Genome Biol Evol* 2011, **3**:320-331.
60. Ebisuya M, Yamamoto T, Nakajima M, Nishida E: **Ripples from neighbouring transcription**. *Nat Cell Biol* 2008, **10**:1106-1113.
61. Smith AJ, Chudnovsky L, Simoes-Barbosa A, Delgadillo-Correa MG, Jonsson ZO, Wohlschlegel JA, Johnson PJ: **Novel core promoter elements and a cognate transcription factor in the divergent unicellular eukaryote *Trichomonas vaginalis***. *Mol Cell Biol* 2011, **31**:1444-1458.
62. Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database Collaboration: **The Sequence Read Archive: explosive growth of sequencing data**. *Nucleic Acids Res* 2012, **40**:D54-56.
63. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009, **25**:1105-1111.
64. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nat Biotechnol* 2010, **28**:511-515.
65. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**:3389-3402.
66. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010, **26**:841-842.
67. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences**. *Bioinformatics* 2006, **22**:1658-1659.

68. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** *Proc Int Conf Intell Syst Mol Biol* 1994, **2**:28-36.
69. Hrdy I, Hirt RP, Dolezal P, Bardonová L, Foster PG, Tachezy J, Embley TM: ***Trichomonas* hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I.** *Nature* 2004, **432**:618-622.
70. Delgadillo MG, Liston DR, Niazi K, Johnson PJ: **Transient and selectable transformation of the parasitic protist *Trichomonas vaginalis*.** *Proc Natl Acad Sci U S A* 1997, **94**:4716-4720.

Additional files



Additional file 1: Figure S1. Expression and Western blot analysis of IncRNA_ATG and stop codon analysis (A1) Illustration of IncRNA_ATG consisting out of start codon followed by two stop codons and a putative open reading frame without an obvious start codon. LncRNA_ATG::HA is transcribed in two clones of transfected trichomonads shown by reverse transcriptase PCR and specific primers (A2), but not translated as shown by western analysis (A3). (B1) Illustration and Western (B2) of stop codon analysis on Actin (TVAG_054030, 42 kDa).

Motif	LNCRNA		PSEUDO		CDS ^P		INTG		Relative frequency
	Sides	e value	Sides	e value	Sides	e value	Sides	e value	
1	16.8%	1.0e-19	15.5%	8.8e-16	29.1%	2.3e-0611	2.7%	7.2e-43	
2	5.5%	1.4e-49	5.7%	4.8e-22	19.9%	1.1e-1232	2.2%	1.0e-17	
3	1.2%	3.7e-02	4.7%	1.7e-22	7.5%	7.4e-1156	1.6%	6.6e-18	
4	0.9%	3.9e-01	3.2%	1.8e-11	4.0%	2.9e-0245	1.0%	6.9e-09	
5	0.4%	8.0e+01	2.6%	2.9e-16	0.8%	3.7e-0075	0.6%	4.5e-06	

Additional file 2: Figure S2. Relative frequencies and *e* values of motifs shown in Figure 4. The background colors indicate relative frequencies in the corresponding datasets.

Additional file 3: Table S1. Kolmogorov–Smirnov test *P* values of datasets in Figure 2.

Figure 2A	Dataset	PSEUDO	INTG	CDS ^P		
	PSEUDO	1	2.45E-07	0		
	INTG	2.45E-07	1	0		
	CDS ^P	0	0	1		
Figure 2B	Dataset	PSEUDO	LNCRNA	INTG	CDS ^P	TVAG
	PSEUDO	1	8.36E-16	5.46E-79	1.99E-264	0
	LNCRNA	8.36E-16	1	1.06E-44	0	0
	INTG	5.46E-79	1.06E-44	1	0	0
	CDS ^P	1.99E-264	0	0	1	0
	TVAG	0	0	0	0	1
Figure 2C	Dataset	PSEUDO	LNCRNA	INTG	RND ^N	CDS ^P
	PSEUDO	1	0.02	2.73E-15	8.22E-129	0
	LNCRNA	0.02	1	8.77E-29	2.14E-174	0
	INTG	2.73E-15	8.77E-29	1	1.55E-102	0
	RND ^N	8.22E-129	2.14E-174	1.55E-102	1	0
	CDS ^P	0	0	0	0	1
Figure 2D	Dataset	PSEUDO	LNCRNA	INTG	RND ^N	CDS ^P
	PSEUDO	1	0.26	1.66E-61	1.29E-61	0
	LNCRNA	0.26	1	6.76E-56	1.58E-56	1.52E-289
	INTG	1.66E-61	6.76E-56	1	2.04E-05	0
	RND ^N	1.29E-61	1.58E-56	2.04E-05	1	0
	CDS ^P	0	1.52E-289	0	0	1

5 ZUSAMMENFASSUNG DER ERGEBNISSE

Die Fortschritte bei der Sequenzierung von Genomen und Transkriptomen waren in den letzten Jahren massiv. Nun lassen sich mit relativ geringem Kostenaufwand große Datenmengen generieren (siehe Abschnitt 3.2.1). Diese lassen sich nicht nur einsetzen, um genomische Eigenschaften und kodierte Gene zu entschlüsseln, zusätzlich ist es auch möglich Expressionsmuster als Reaktion auf physiologische Einflüsse und in verschiedenen Gewebetypen anhand von quantitativen Analysen an Transkriptomen zu charakterisieren (siehe Abschnitt 3.3). Ein Vorteil dieser neuen Technologien ist, dass sie auch auf weniger erforschte Organismen, ohne vorliegende Genomsequenzen, angewandt werden können (Lowe *et al.*, 2011; Pallavicini *et al.*, 2013; Vera *et al.*, 2008). Dies kommt besonders bei Protisten zum tragen. Denn hier gibt es, abgesehen von wenigen Ausnahmen in Form von Modellorganismen und pathogenen Arten, wenig sequenzierte Genome. Dabei zeigt diese Gruppierung ein besonders hohes Maß an Diversität und ist gerade dadurch eine besondere Herausforderung für aktuelle Analysen (siehe Abschnitt 3.1).

Im Rahmen dieser Arbeit wurden in mehreren unabhängigen Studien ganz verschiedene Aspekte der Sequenzierung von Transkriptomen an Protisten aufgezeigt, welche sich zu einem aussagekräftigen Gesamtbild zusammenfügen lassen. Sie bestehen aus einer *De-novo*-Sequenzierung, um exprimierte Gene zu ermitteln und phylogenetische Analysen durchzuführen (Woehle *et al.*, 2011), der Nutzung von Transkriptsequenzen, um homologe Gene zu identifizieren (Kusdian *et al.*, 2013), der Charakterisierung der Reaktion auf physiologische Veränderungen anhand von sequenzierten Transkripten (Gould *et al.*, 2013) und der Identifizierung bisher unbekannter nicht-kodierender RNAs (Woehle *et al.*, 2014). So geben diese Studien Aufschluss über unterschiedliche Anwendungsmöglichkeiten der relativ jungen Methodik der neuen Sequenziertechniken.

Wie das Genom, trägt auch das Transkriptom Information über den Aufbau der Gene. Dabei zeigten verschiedene Studien, dass der Großteil aller Gene durch Sequenzierungen des Transkriptoms abgedeckt werden können (Dyhrman *et al.*, 2012; Wilhelm *et al.*, 2008; Xiong *et al.*, 2012). Aktuelle Analysen können mit ihren massiven Datenmengen und ständig schrumpfenden Kosten wertvolle Einblicke in den funktionellen Umfang des zugrunde liegenden Genoms geben. Besonders

interessant wird diese Methodik dadurch auch bei der Anwendung an weniger erforschten Protisten. In Woehle *et al.* (2011) konnte auf diese Weise die Expression tausender Gene der kürzlich beschriebenen Alge *Chromera velia* nachgewiesen werden. Diese waren, ähnlich wie genomisch kodierte Gene, für evolutionäre Analysen verwendbar und gaben Einblicke in die phylogenetische Einordnung dieser Alge.

Transkripte geben Auskunft über den exprimierten Teil der vorhandenen Gene. So können sie genutzt werden um auf Homologe in bekannten Genfamilien zu prüfen. Da veröffentlichte Daten auch anderen zugänglich sind, können diese auch ergänzend zu unabhängigen Analysen verwertet werden, bei denen die Transkriptome nicht im Mittelpunkt stehen (Ginger *et al.*, 2010; Müller *et al.*, 2012). Auf diese Weise wurden EST-Daten genutzt um in Kusdian *et al.* (2013) Organismen ohne sequenzierte Genome auf vorhandene Gene des Zytoskeletts zu untersuchen. Dabei beschäftigt sich diese Veröffentlichung in ist erster Linie mit der Morphogenese von *Trichomonas vaginalis* vom flagellierten zum amöboiden Zustand. Demnach zeigt diese Nutzung der Daten aus Transkriptomsequenzierungen, wie sie auch als Ressource für andere Studien verwendet werden können.

Verschiedene Methoden der Transkriptomsequenzierungen ermöglichen es die Menge der Transkripte und damit die Expression der zugehörigen Gene zu bestimmen (siehe Abschnitt 3.3). So lassen sich Expressionsmuster von bekannten Annotationen oder ganzen Genombereichen beschreiben. In Gould *et al.* (2013) wurde eine spezielle Sequenzierung von 3'-Fragmenten durchgeführt und auf bekannte Gene im Genom übertragen. Das Ergebnis war eine Auswertung der umfassenden Expression am Protisten *T. vaginalis*. Dabei konnte beschrieben werden, inwieweit die Menge exprimierter Gene mit den Annotationen im Genom übereinstimmt und darüber hinaus wurden unterschiedliche Expressionsmuster erkannt und verglichen. So konnte beispielsweise gezeigt werden, dass die Reaktion von *Trichomonas vaginalis* auf Sauerstoff über Stimuli, induziert durch Wirtszellen, dominiert. Dies verdeutlicht die Notwendigkeit für *T. vaginalis* auf Veränderungen der Sauerstoffkonzentration reagieren zu können. Die erfolgreiche Verwendung dieser Expressionsdaten bestätigt, dass sie eine geeignete Anwendung zur Charakterisierung von Genexpressionsmustern darstellen.

Umfassende Analysen am Transkriptom fanden schon vor der Entwicklung der NGS-Systeme eine hohe Zahl bisher unbekannter nicht-kodierender Transkripte (siehe Abschnitt 3.4), deren Funktion nur für wenige Beispiele beschrieben werden konnte (Amaral *et al.*, 2010; Niazi & Valadkhan, 2012). Mit der Etablierung der

neuen Sequenziermethoden zog dieses Phänomen immer mehr Aufmerksamkeit auf sich. Beim Menschen konnte kürzlich die Expression von zwei Dritteln des Genoms beschrieben werden (Djebali *et al.*, 2012), was gegenüber dem geringen Prozentsatz der Protein-kodierenden Gene eine immense Menge darstellt (Taft *et al.*, 2007). Abgesehen von einigen Hefen gibt es dazu nur wenige Studien an Protisten (Kolev *et al.*, 2010; Nagalakshmi *et al.*, 2008; Wilhelm *et al.*, 2008). In Woehle *et al.* (2014) wurde *T. vaginalis* auf die umfassende Expression in intergenischen Bereichen untersucht. Dabei konnten tausende exprimierter Transkripte beschrieben werden. Ein großer Teil davon zeigte Sequenzähnlichkeiten zu Protein-kodierenden Genen, zudem waren vorwiegend transkriptionelle Promotorsequenzen vertreten. Damit konnte die umfassende Expression des Genoms auch für den Protisten *T. vaginalis* beschrieben werden und darüber hinaus wurden Hinweise auf eine selbst-induzierte Expression dieser Transkripte erfasst. Die Erforschung dieses Phänomens steht erst am Anfang, aber es ist wahrscheinlich, dass solche Analysen an Protisten helfen werden seine Ursprünge näher zu verstehen.

Die beschriebenen Analysen geben Einblicke in das umfangreiche Anwendungspotential von Transkriptomsequenzierungen an Protisten. Schon andere Veröffentlichungen konnten die Nutzung neuer Sequenziermethoden an verschiedenen Gruppen der Eukaryoten beschreiben (Dyhrman *et al.*, 2012; Kolev *et al.*, 2010; Lowe *et al.*, 2011; Wilhelm *et al.*, 2008; Xiong *et al.*, 2012). In dieser Arbeit wurde bestätigt, dass diese Methoden bei Protisten ein umfangreiches und hilfreiches Analyseverfahren darstellen. Die Möglichkeit Gene zu charakterisieren und zusätzlich differentielle Expression untersuchen zu können, eröffnet den Weg für viele neue Anwendungsmöglichkeiten. Mit sinkenden Preisen und steigenden Datendurchsatzraten ist es nur eine Frage der Zeit, bis Transkriptomsequenzierungen bei allen Eukaryoten zum Standard gehören, um einen Einblick in die Zusammensetzung des Genoms und der Reaktionen auf physiologische Einflüsse zu bekommen.

LITERATUR

- Adams MD** (1996). Serial analysis of gene expression: ESTs get smaller. *Bioessays*, 18:261–262.
- Adl SM, Simpson AGB, Farmer MA, Andersen RA, Anderson OR, Barta JR, Bowser SS, Brugerolle G, Fensome RA, Fredericq S et al.** (2005). The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J Eukaryot Microbiol*, 52:399–451.
- Adl SM, Simpson AGB, Lane CE, Lukeš J, Bass D, Bowser SS, Brown MW, Burki F, Dunthorn M, Hampl V et al.** (2012). The revised classification of eukaryotes. *J Eukaryot Microbiol*, 59:429–493.
- Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, Miller W & Lipman D** (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389–3402.
- Amaral PP, Clark MB, Gascoigne DK, Dinger ME & Mattick JS** (2010). lncRNADB: a reference database for long noncoding RNAs. *Nucleic Acids Res*, 39:D146–D151.
- Amaral PP & Mattick JS** (2008). Noncoding RNA in development. *Mamm Genome*, 19:454–492.
- Ansorge W, Sproat BS, Stegemann J & Schwager C** (1986). A non-radioactive automated method for DNA sequence determination. *J Biochem Biophys Methods*, 13:315–323.
- Ansorge W, Sproat B, Stegemann J, Schwager C & Zenke M** (1987). Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. *Nucleic Acids Res*, 15:4593–4602.
- Archibald JM** (2009). The puzzle of plastid evolution. *Curr Biol*, 19:R81–R88.
- Au PCK, Zhu Q, Dennis ES & Wang M** (2011). Long non-coding RNA-mediated mechanisms independent of the RNAi pathway in animals and plants. *RNA Biol*, 8:404–414.
- Aurrecochea C, Brestelli J, Brunk BP, Carlton JM, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G et al.** (2009). GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*. *Nucleic Acids Res*, 37:D526–D530.
- Axtell MJ, Westholm JO & Lai EC** (2011). Vive la différence: biogenesis and evolution of microRNAs in plants and animals. *Genome Biol*, 12:221.
- Bashir A, Klammer AA, Robins WP, Chin C, Webster D, Paxinos E, Hsu D, Ashby M, Wang S, Peluso P et al.** (2012). A hybrid approach for the automated finishing of bacterial genomes. *Nat Biotechnol*, 30:701–707.

- Baurain D, Brinkmann H, Petersen J, Rodríguez-Ezpeleta N, Stechmann A, Demoulin V, Roger AJ, Burger G, Lang BF & Philippe H (2010). Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. *Mol Biol Evol*, 27:1698–1709.
- Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456:53–59.
- Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B *et al.* (2005). The genome of the African trypanosome *Trypanosoma brucei*. *Science*, 309:416–422.
- Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S *et al.* (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science*, 306:2242–2246.
- Blouin NA, Brodie JA, Grossman AC, Xu P & Brawley SH (2011). *Porphyra*: a marine crop shaped by stress. *Trends Plant Sci*, 16:29–37.
- Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X *et al.* (2008). The potential and challenges of nanopore sequencing. *Nat Biotechnol*, 26:1146–1153.
- Brosius J, Dull TJ, Sleeter DD & Noller HF (1981). Gene organization and primary structure of a ribosomal RNA operon from *Escherichia coli*. *J Mol Biol*, 148:107–127.
- Burki F, Flegontov P, Oborník M, Cihlář J, Pain A, Lukeš J & Keeling PJ (2012). Re-evaluating the green versus red signal in eukaryotes with secondary plastid of red algal origin. *Genome Biol Evol*, 4:626–635.
- Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A & Rinn JL (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*, 25:1915–1927.
- Carlton JM, Hirt RP, Silva JC, Delcher AL, Schatz M, Zhao Q, Wortman JR, Bidwell SL, Alsmark UCM, Besteiro S *et al.* (2007). Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science*, 315:207–212.
- Carthew RW & Sontheimer EJ (2009). Origins and Mechanisms of miRNAs and siRNAs. *Cell*, 136:642–655.
- Carvunis A, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charlotiaux B, Hidalgo CA, Barbette J, Santhanam B *et al.* (2012). Proto-genes and *de novo* gene birth. *Nature*, 487:1–5.
- Cavalier-Smith T (1987). The origin of eukaryotic and archaeobacterial cells. *Ann N Y Acad Sci*, 503:17–54.
- Cavalier-Smith T (1999). Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J Eukaryot Microbiol*, 46:347–366.
- Cavalier-Smith T (2002). The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *Int J Syst Evol Microbiol*, 52:297–354.

- Cerutti H & Casas-Mollano JA** (2006). On the origin and functions of RNA-mediated silencing: from protists to man. *Curr Genet*, 50:81–99.
- Chapman EJ & Carrington JC** (2007). Specialization and evolution of endogenous small RNA pathways. *Nat Rev Genet*, 8:884–896.
- Chen F, Mackey AJ, Stoeckert CJ & Roos DS** (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res*, 34:D363–D368.
- Chen XS, Collins LJ, Biggs PJ & Penny D** (2009). High throughput genome-wide survey of small RNAs from the parasitic protists *Giardia intestinalis* and *Trichomonas vaginalis*. *Genome Biol Evol*, 1:165–175.
- Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G et al.** (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods*, 5:613–619.
- Cock PJA, Fields CJ, Goto N, Heuer ML & Rice PM** (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res*, 38:1767–1771.
- Costa V, Angelini C, De Feis I & Ciccodicola A** (2010). Uncovering the complexity of transcriptomes with RNA-Seq. *J Biomed Biotechnol*, 2010:853916.
- Cotch MF, Pastorek JG, Nugent RP, Hillier SL, Gibbs RS, Martin DH, Eschenbach DA, Edelman R, Carey JC, Regan JA et al.** (1997). *Trichomonas vaginalis* associated with low birth weight and preterm delivery. *Sex Transm Dis*, 24:353–360.
- Cui J, Das S, Smith TF & Samuelson J** (2010). *Trichomonas* transmembrane cyclases result from massive gene duplication and concomitant development of pseudogenes. *PLoS Negl Trop Dis*, 4:e782.
- Curtis BA, Tanifuji G, Burki F, Gruber A, Irimia M, Maruyama S, Arias MC, Ball SG, Gile GH, Hirakawa Y et al.** (2012). Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature*, 492:59–65.
- Dagan T & Martin W** (2009). Seeing green and red in diatom genomes. *Science*, 324:1651–1652.
- Dame JB, Arnot DE, Bourke PF, Chakrabarti D, Christodoulou Z, Coppel RL, Cowman AF, Craig AG, Fischer K, Foster J et al.** (1996). Current status of the *Plasmodium falciparum* genome project. *Mol Biochem Parasitol*, 79:1–12.
- Dark MJ** (2013). Whole-genome sequencing in bacteriology: state of the art. *Infect Drug Resist*, 6:115–123.
- Delcher AL, Harmon D, Kasif S, White O & Salzberg SL** (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res*, 27:4636–4641.
- Denoed F, Aury J, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C et al.** (2008). Annotating genomes with massive-scale RNA sequencing. *Genome Biol*, 9:R175.
- Derti A, Garrett-Engle P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM & Babak T** (2012). A quantitative atlas of polyadenylation in five mammals. *Genome Res*, 22:1173–1183.

- Diermann N, Matoušek J, Junge M, Riesner D & Steger G (2010). Characterization of plant miRNAs and small RNAs derived from potato spindle tuber viroid (PSTVd) in infected tomato. *Biol Chem*, 391:1379–1390.
- Diguistini S, Liao NY, Platt D, Robertson G, Seidel M, Chan SK, Docking TR, Birol I, Holt RA, Hirst M *et al.* (2009). *De novo* genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biol*, 10:R94.
- Dinger ME, Amaral PP, Mercer TR, Pang KC, Bruce SJ, Gardiner BB, Askarian-Amiri ME, Ru K, Soldà G, Simons C *et al.* (2008). Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res*, 18:1433–1445.
- Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F *et al.* (2012). Landscape of transcription in human cells. *Nature*, 489:101–108.
- Dyrhman ST, Jenkins BD, Ryneerson TA, Saito MA, Mercier ML, Alexander H, Whitney LP, Drzewianowski A, Bulygin VV, Bertrand EM *et al.* (2012). The transcriptome and proteome of the diatom *Thalassiosira pseudonana* reveal a diverse phosphorus stress response. *PLoS One*, 7:e33768.
- Ebisuya M, Yamamoto T, Nakajima M & Nishida E (2008). Ripples from neighbouring transcription. *Nat Cell Biol*, 10:1106–1113.
- Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B *et al.* (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, 323:133–138.
- Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, Wortman JR, Badger JH, Ren Q, Amedeo P, Jones KM *et al.* (2006). Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol*, 184:e286.
- El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran A, Ghedin E, Worthey EA, Delcher AL, Blandin G *et al.* (2005). The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science*, 309:409–415.
- Embley TM & Martin W (2006). Eukaryotic evolution, changes and challenges. *Nature*, 440:623–630.
- Embley TM, van der Giezen M, Horner DS, Dyal PL & Foster P (2003). Mitochondria and hydrogenosomes are two forms of the same fundamental organelle. *Philos Trans R Soc Lond B Biol Sci*, 358:191–202.
- ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447:799–816.
- Faghihi MA & Wahlestedt C (2009). Regulatory roles of natural antisense transcripts. *Nat Rev Mol Cell Biol*, 10:637–643.
- Fox-Walsh K, Davis-Turak J, Zhou Y, Li H & Fu X (2011). A multiplex RNA-seq strategy to profile poly(A+) RNA: application to analysis of transcription response and 3' end formation. *Genomics*, 98:266–271.

- Fritz-Laylin LK, Prochnik SE, Ginger ML, Dacks JB, Carpenter ML, Field MC, Kuo A, Paredez A, Chapman J, Pham J *et al.* (2010). The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. *Cell*, 140:631–642.
- Fromont-Racine M, Senger B, Saveanu C & Fasiolo F (2003). Ribosome assembly in eukaryotes. *Gene*, 313:17–42.
- Furuichi Y, LaFiandra A & Shatkin AJ (1977). 5'-Terminal structure and mRNA stability. *Nature*, 266:235–239.
- Garcia-Silva MR, Frugier M, Tosar JP, Correa-Dominguez A, Ronalte-Alves L, Parodi-Talice A, Rovira C, Robello C, Goldenberg S & Cayota A (2010). A population of tRNA-derived small RNAs is actively produced in *Trypanosoma cruzi* and recruited to specific cytoplasmic granules. *Mol Biochem Parasitol*, 171:64–73.
- Ginger ML, Fritz-Laylin LK, Fulton C, Cande WZ & Dawson SC (2010). Intermediary metabolism in protists: A sequence-based view of facultative anaerobic metabolism in evolutionarily diverse eukaryotes. *Protist*, 161:642–671.
- Glücksman E, Snell EA, Berney C, Chao EE, Bass D & Cavalier-Smith T (2011). The novel marine gliding zooflagellate genus *Mantamonas* (Mantamonadida ord. n.: Apusozoa). *Protist*, 162:207–221.
- Golden DE, Gerbasi VR & Sontheimer EJ (2008). An inside job for siRNAs. *Mol Cell*, 31:309–312.
- Gonatopoulos-Pournatzis T & Cowling VH (2014). Cap-binding complex (CBC). *Biochem J*, 457:231–242.
- Gould SB, Tham W, Cowman AF, McFadden GI & Waller RF (2008). Alveolins, a new family of cortical proteins that define the protist infrakingdom Alveolata. *Mol Biol Evol*, 25:1219–1230.
- Gould SB, Waller RF & McFadden GI (2008). Plastid evolution. *Annu Rev Plant Biol*, 59:491–517.
- Gould SB, Woehle C, Kusdian G, Landan G, Tachezy J, Zimorski V & Martin WF (2013). Deep sequencing of *Trichomonas vaginalis* during the early infection of vaginal epithelial cells and amoeboid transition. *Int J Parasitol*, 43:707–719.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*, 29:644–652.
- Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP *et al.* (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458:223–227.
- Hackett JD, Anderson DM, Erdner DL & Bhattacharya D (2004). Dinoflagellates: a remarkable evolutionary experiment. *Am J Bot*, 91:1523–1534.
- Hallick RB, Hong L, Drager RG, Favreau MR, Monfort A, Orsat B, Spielmann A & Stutz E (1993). Complete sequence of *Euglena gracilis* chloroplast DNA. *Nucleic Acids Res*, 21:3537–3544.

- Hampl V, Horner DS, Dyal P, Kulda J, Flegr J, Foster PG & Embley TM** (2005). Inference of the phylogenetic position of oxymonads based on nine genes: support for Metamonada and Excavata. *Mol Biol Evol*, 22:2508–2518.
- Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AGB & Roger AJ** (2009). Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic “supergroups”. *Proc Natl Acad Sci U S A*, 106:3859–3864.
- Harbers M & Carninci P** (2005). Tag-based approaches for transcriptome research and genome annotation. *Nat Methods*, 2:495–502.
- Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S et al.** (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*, 22:1760–1774.
- Hashimoto T, Nakamura Y, Nakamura F, Shirakura T, Adachi J, Goto N, Okamoto K & Hasegawa M** (1994). Protein phylogeny gives a robust estimation for early divergences of eukaryotes: phylogenetic place of a mitochondria-lacking protozoan, *Giardia lamblia*. *Mol Biol Evol*, 11:65–71.
- Hawkins PG & Morris KV** (2010). Transcriptional regulation of Oct4 by a long non-coding RNA antisense to Oct4-pseudogene 5. *Transcription*, 1:165–175.
- He S, Wurtzel O, Singh K, Froula JL, Yilmaz S, Tringe SG, Wang Z, Chen F, Lindquist EA, Sorek R et al.** (2010). Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat Methods*, 7:807–812.
- He Y, Vogelstein B, Velculescu VE, Papadopoulos N & Kinzler KW** (2008). The antisense transcriptomes of human cells. *Science*, 322:1855–1857.
- Holley CL & Topkara VK** (2011). An introduction to small non-coding RNAs: miRNA and snoRNA. *Cardiovasc Drugs Ther*, 25:151–159.
- Hoque M, Ji Z, Zheng D, Luo W, Li W, You B, Park JY, Yehia G & Tian B** (2013). Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat Methods*, 10:133–139.
- Howe CJ, Nisbet RER & Barbrook AC** (2008). The remarkable chloroplast genome of dinoflagellates. *J Exp Bot*, 59:1035–1045.
- Hu W, Alvarez-Dominguez JR & Lodish HF** (2012). Regulation of mammalian cell differentiation by long non-coding RNAs. *EMBO Rep*, 13:971–983.
- Huang K, Huang P, Ku F, Lin R, Alderete JF & Tang P** (2012). Comparative transcriptomic and proteomic analyses of *Trichomonas vaginalis* following adherence to fibronectin. *Infect Immun*, 80:3900–3911.
- Huez G, Bruck C & Cleuter Y** (1981). Translational stability of native and deadenylylated rabbit globin mRNA injected into HeLa cells. *Proc Natl Acad Sci U S A*, 78:908–911.
- Hutvagner G & Simard MJ** (2008). Argonaute proteins: key players in RNA silencing. *Nat Rev Mol Cell Biol*, 9:22–32.
- Jacquier A** (2009). The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat Rev Genet*, 10:833–844.

- Jan CH, Friedman RC, Ruby JG & Bartel DP** (2011). Formation, regulation and evolution of *Caenorhabditis elegans* 3'UTRs. *Nature*, 469:97–101.
- Janouškovec J, Horák A, Oborník M, Lukeš J & Keeling PJ** (2010). A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proc Natl Acad Sci U S A*, 107:10949–10954.
- Jenjaroenpun P, Kremenska Y, Nair VM, Kremenskoy M, Joseph B & Kurochkin IV** (2013). Characterization of RNA in exosomes secreted by human breast cancer cell lines using next-generation sequencing. *PeerJ*, 1:e201.
- Ji P, Diederichs S, Wang W, Böing S, Metzger R, Schneider PM, Tidow N, Brandt B, Buerger H, Bulk E et al.** (2003). *MALAT-1*, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene*, 22:8031–8041.
- Jones-Rhoades MW, Bartel DP & Bartel B** (2006). MicroRNAs and their regulatory roles in plants. *Annu Rev Plant Biol*, 57:19–53.
- Kapranov P, Cheng J, Dike S, Nix DA, Dutttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL et al.** (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, 316:1484–1488.
- Karpinets TV, Greenwood DJ, Sams CE & Ammons JT** (2006). RNA:protein ratio of the unicellular organism as a characteristic of phosphorous and nitrogen stoichiometry and of the cellular requirement of ribosomes for protein synthesis. *BMC Biol*, 4:30.
- Kasianowicz JJ, Brandin E, Branton D & Deamer DW** (1996). Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci U S A*, 93:13770–13773.
- Kim KM, Park J, Bhattacharya D & Yoon HS** (2014). Applications of next-generation sequencing to unravelling the evolutionary history of algae. *Int J Syst Evol Microbiol*, 64:333–345.
- Kiss T** (2002). Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell*, 109:145–148.
- Kodama Y, Shumway M, Leinonen R & International Nucleotide Sequence Database Collaboration** (2012). The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res*, 40:D54–D56.
- Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M et al.** (2006). CAGE: cap analysis of gene expression. *Nat Methods*, 3:211–222.
- Kolev NG, Franklin JB, Carmi S, Shi H, Michaeli S & Tschudi C** (2010). The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. *PLoS Pathog*, 6:e1001090.
- Koonin EV** (2010). The incredible expanding ancestor of eukaryotes. *Cell*, 140:606–608.
- Koonin EV** (2010). The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome Biol*, 11:209.

- Koonin EV & Wolf YI** (2010). Constraints and plasticity in genome and molecular-phenome evolution. *Nat Rev Genet*, 11:487–498.
- Korneev SA, Park JH & O’Shea M** (1999). Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. *J Neurosci*, 19:7711–7720.
- Kozak M** (1983). Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. *Microbiol Rev*, 47:1–45.
- Kozomara A & Griffiths-Jones S** (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*, 42:D68–D73.
- Kung JTY, Colognori D & Lee JT** (2013). Long noncoding RNAs: past, present, and future. *Genetics*, 193:651–669.
- Kuo C & Ochman H** (2010). The extinction dynamics of bacterial pseudogenes. *PLoS Genet*, 6:e1001050.
- Kusdian G, Woehle C, Martin WF & Gould SB** (2013). The actin-based machinery of *Trichomonas vaginalis* mediates flagellate-amoeboid transition and migration across host tissue. *Cell Microbiol*, 15:1707–1721.
- Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, Odom DT & Marques AC** (2012). Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet*, 8:e1002841.
- Larsson C, Grundberg I, Söderberg O & Nilsson M** (2010). *In situ* detection and genotyping of individual mRNA molecules. *Nat Methods*, 7:395–397.
- Lawrence JG, Hendrix RW & Casjens S** (2001). Where are the pseudogenes in bacterial genomes? *Trends Microbiol*, 9:535–540.
- Lee YS, Shibata Y, Malhotra A & Dutta A** (2009). A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev*, 23:2639–2649.
- Li B & Dewey CN** (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12:323.
- Li B, Ruotti V, Stewart RM, Thomson JA & Dewey CN** (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26:493–500.
- Li WH, Gojobori T & Nei M** (1981). Pseudogenes as a paradigm of neutral evolution. *Nature*, 292:237–239.
- Lindberg J & Lundeberg J** (2010). The plasticity of the mammalian transcriptome. *Genomics*, 95:1–6.
- Lippman Z & Martienssen R** (2004). The role of RNA interference in heterochromatic silencing. *Nature*, 431:364–370.
- Lister R, O’Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH & Ecker JR** (2008). Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell*, 133:523–536.

- Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L & Law M (2012). Comparison of next-generation sequencing systems. *J Biomed Biotechnol*, 2012:251364.
- Loman NJ, Constantinidou C, Chan JZM, Halachev M, Sergeant M, Penn CW, Robinson ER & Pallen MJ (2012). High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat Rev Microbiol*, 10:599–606.
- Louro R, Smirnova AS & Verjovski-Almeida S (2009). Long intronic noncoding RNA transcription: expression noise or expression choice? *Genomics*, 93:291–298.
- Lowe CD, Mello LV, Samatar N, Martin LE, Montagnes DJS & Watts PC (2011). The transcriptome of the novel dinoflagellate *Oxyrrhis marina* (Alveolata: Dinophyceae): response to salinity examined by 454 sequencing. *BMC Genomics*, 12:519.
- Lu C, Tej SS, Luo S, Haudenschild CD, Meyers BC & Green PJ (2005). Elucidation of the small RNA component of the transcriptome. *Science*, 309:1567–1569.
- Lynch M (2007). The evolution of genetic networks by non-adaptive processes. *Nat Rev Genet*, 8:803–813.
- Maniatis T & Reed R (1987). The role of small nuclear ribonucleoprotein particles in pre-mRNA splicing. *Nature*, 325:673–678.
- Marck C & Grosjean H (2002). tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA*, 8:1189–1232.
- Mardis ER (2008). Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*, 9:387–402.
- Marguerat S & Bähler J (2010). RNA-seq: from technology to biology. *Cell Mol Life Sci*, 67:569–579.
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bembien LA, Berka J, Braverman MS, Chen Y, Chen Z *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380.
- Marioni JC, Mason CE, Mane SM, Stephens M & Gilad Y (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 18:1509–1517.
- Martin JA & Wang Z (2011). Next-generation transcriptome assembly. *Nat Rev Genet*, 12:671–682.
- Matsumura H, Ito A, Saitoh H, Winter P, Kahl G, Reuter M, Krüger DH & Terauchi R (2005). SuperSAGE. *Cell Microbiol*, 7:11–18.
- Matsumura H, Reich S, Ito A, Saitoh H, Kamoun S, Winter P, Kahl G, Reuter M, Krüger DH & Terauchi R (2003). Gene expression analysis of plant host-pathogen interactions by SuperSAGE. *Proc Natl Acad Sci U S A*, 100:15718–15723.
- Matsuzaki M, Misumi O, Shin-i T, Maruyama S, Takahara M, Miyagishima S, Mori T, Nishida K, Yagisawa F, Nishida K *et al.* (2004). Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature*, 428:653–657.

- Mattick JS & Makunin IV** (2005). Small regulatory RNAs in mammals. *Hum Mol Genet*, 14:R121–R132.
- Maxam AM & Gilbert W** (1977). A new method for sequencing DNA. *Proc Natl Acad Sci U S A*, 74:560–564.
- McCormick KP, Willmann MR & Meyers BC** (2011). Experimental design, preprocessing, normalization and differential expression analysis of small RNA sequencing experiments. *Silence*, 2:2.
- McFadden GI** (2011). The apicoplast. *Protoplasma*, 248:641–650.
- McGinn S & Gut IG** (2013). DNA sequencing – spanning the generations. *N Biotechnol*, 30:366–372.
- McInerney JO, O’Connell MJ & Pisani D** (2014). The hybrid nature of the Eukaryota and a consilient view of life on Earth. *Nat Rev Microbiol*, 12:449–455.
- McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC et al.** (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res*, 19:1527–1541.
- Meister G** (2013). Argonaute proteins: functional insights and emerging roles. *Nat Rev Genet*, 14:447–459.
- Mercer TR, Dinger ME & Mattick JS** (2009). Long non-coding RNAs: insights into functions. *Nat Rev Genet*, 10:155–159.
- Moore RB, Oborník M, Janouškovec J, Chrudimský T, Vancová M, Green DH, Wright SW, Davies NW, Bolch CJS, Heimann K et al.** (2008). A photosynthetic alveolate closely related to apicomplexan parasites. *Nature*, 451:959–963.
- Moran NA** (2002). Microbial minimalism: genome reduction in bacterial pathogens. *Cell*, 108:583–586.
- Moreira D, von der Heyden S, Bass D, López-García P, Chao E & Cavalier-Smith T** (2007). Global eukaryote phylogeny: Combined small- and large-subunit ribosomal DNA trees support monophyly of Rhizaria, Retaria and Excavata. *Mol Phylogenet Evol*, 44:255–266.
- Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S & Marra M** (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques*, 45:81–94.
- Mortazavi A, Williams BA, McCue K, Schaeffer L & Wold B** (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, 5:621–628.
- Moustafa A, Beszteri B, Maier UG, Bowler C, Valentin K & Bhattacharya D** (2009). Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science*, 24:1724–1726.
- Müller M, Mentel M, van Hellemond JJ, Henze K, Woehle C, Gould SB, Yu R, van der Giezen M, Tielens AGM & Martin WF** (2012). Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiol Mol Biol Rev*, 76:444–495.

- Murphy FV & Ramakrishnan V** (2004). Structure of a purine-purine wobble base pair in the decoding center of the ribosome. *Nat Struct Mol Biol*, 11:1251–1252.
- Mutz K, Heilkenbrinker A, Lönne M, Walter J & Stahl F** (2013). Transcriptome analysis using next-generation sequencing. *Curr Opin Biotechnol*, 24:22–30.
- Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M & Snyder M** (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320:1344–1349.
- Nagaraj SH, Gasser RB & Ranganathan S** (2007). A hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief Bioinform*, 8:6–21.
- Niazi F & Valadkhan S** (2012). Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3' UTRs. *RNA*, 18:825–843.
- Oborník M, Modrý D, Lukeš M, Černotíková Stříbrná E, Cihlár J, Tesařová M, Kotabová E, Vancová M, Prášil Or & Lukeš J** (2012). Morphology, ultrastructure and life cycle of *Vitrella brassicaformis* n. sp., n. gen., a novel chromerid from the Great Barrier Reef. *Protist*, 163:306–323.
- Ohhata T, Hoki Y, Sasaki H & Sado T** (2008). Crucial role of antisense transcription across the *Xist* promoter in *Tsix*-mediated *Xist* chromatin modification. *Development*, 135:227–235.
- Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H et al.** (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, 420:563–573.
- Okoniewski MJ & Miller CJ** (2006). Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC bioinformatics*, 7:276.
- O'Malley MA, Simpson AGB & Roger AJ** (2012). The other eukaryotes in light of evolutionary protistology. *Biol Philos*, 28:299–330.
- Pagani I, Liolios K, Jansson J, Chen IA, Smirnova T, Nosrat B, Markowitz VM & Kyrpides NC** (2012). The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res*, 40:D571–D579.
- Pallavicini A, Canapa A, Barucca M, Alf Ldi J, Biscotti MA, Buonocore F, De Moro G, Di Palma F, Fausto AM, Forconi M et al.** (2013). Analysis of the transcriptome of the Indonesian coelacanth *Latimeria menadoensis*. *BMC Genomics*, 14:538.
- Pareek CS, Smoczynski R & Tretyn A** (2011). Sequencing technologies and genome sequencing. *J Appl Genet*, 52:413–435.
- Parfrey LW, Grant J, Tekle YI, Lasek-Nesselquist E, Morrison HG, Sogin ML, Patterson DJ & Katz LA** (2010). Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. *Syst Biol*, 59:518–533.
- Passalacqua KD, Varadarajan A, Ondov BD, Okou DT, Zwick ME & Bergman NH** (2009). Structure and complexity of a bacterial transcriptome. *J Bacteriol*, 191:3203–3211.

- Patel RK & Jain M (2011). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One*, 7:e30619.
- Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J *et al.* (2012). Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature*, 492:423–427.
- Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer A, Diekhans M *et al.* (2012). The GENCODE pseudogene resource. *Genome Biol*, 13:R51.
- Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, Yu L, Assefa SA, He M, Croucher NJ, Pickard DJ *et al.* (2009). A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. *PLoS Genet*, 5:e1000569.
- Petrin D, Delgaty K, Bhatt R & Garber G (1998). Clinical and microbiological aspects of *Trichomonas vaginalis*. *Clin Microbiol Rev*, 11:300–317.
- Phizicky EM & Hopper AK (2010). tRNA biology charges to the front. *Genes Dev*, 24:1832–1860.
- Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L & Carter DRF (2011). Pseudogenes: pseudo-functional or key regulators in health and disease? *RNA*, 17:792–798.
- Poliseno L (2012). Pseudogenes: newly discovered players in human cancer. *Sci Signal*, 5:re5.
- Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ & Pandolfi PP (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, 465:1033–1038.
- Ponjavic J, Ponting CP & Lunter G (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res*, 17:556–565.
- Price DC, Chan CX, Yoon HS, Yang EC, Qiu H, Weber APM, Schwacke R, Gross J, Blouin NA, Lane C *et al.* (2012). *Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. *Science*, 335:843–847.
- Qi Y, He X, Wang X, Kohany O, Jurka J & Hannon GJ (2006). Distinct catalytic and non-catalytic roles of ARGONAUTE4 in RNA-directed DNA methylation. *Nature*, 443:1008–1012.
- Quinn EM, Cormican P, Kenny EM, Hill M, Anney R, Gill M, Corvin AP & Morris DW (2013). Development of strategies for SNP detection in RNA-seq data: application to lymphoblastoid cell lines and evaluation using 1000 Genomes data. *PLoS One*, 8:e58815.
- Rearick D, Prakash A, McSweeney A, Shepard SS, Fedorova L & Fedorov A (2011). Critical association of ncRNA with introns. *Nucleic Acids Res*, 39:2357–2366.
- Reinhardt JA, Baltrus DA, Nishimura MT, Jeck WR, Jones CD & Dangl JL (2009). *De novo* assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae*. *Genome Res*, 19:294–305.

- Reyes-Prieto A, Moustafa A & Bhattacharya D** (2008). Multiple genes of apparent algal origin suggest ciliates may once have been photosynthetic. *Curr Biol*, 18:956–962.
- Rich A & RajBhandary UL** (1976). Transfer RNA: molecular structure, sequence, and properties. *Annu Rev Biochem*, 45:805–860.
- Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E et al.** (2007). Functional demarcation of active and silent chromatin domains in human *HOX* loci by noncoding RNAs. *Cell*, 129:1311–1323.
- Roger AJ, Svärd SG, Tovar J, Clark CG, Smith MW, Gillin FD & Sogin ML** (1998). A mitochondrial-like chaperonin 60 gene in *Giardia lamblia*: evidence that diplomonads once harbored an endosymbiont related to the progenitor of mitochondria. *Proc Natl Acad Sci U S A*, 95:229–234.
- Ronaghi M, Karamohamed S, Pettersson B, Uhlén M & Nyrén P** (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem*, 242:84–89.
- Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M et al.** (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475:348–352.
- Royce TE, Rozowsky JS & Gerstein MB** (2007). Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification. *Nucleic Acids Res*, 35:e99.
- Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H & Bartel DP** (2006). Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell*, 127:1193–1207.
- Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW & Velculescu VE** (2002). Using the transcriptome to annotate the genome. *Nat Biotechnol*, 20:508–512.
- Sanger F, Nicklen S & Coulson AR** (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74:5463–5467.
- Schnepf E & Elbrächter M** (1999). Dinophyte chloroplasts and phylogeny – A review. *Grana*, 38:81–97.
- Schwebke JR & Burgess D** (2004). Trichomoniasis. *Clin Microbiol Rev*, 17:794–803.
- Sharp PA** (2009). The Centrality of RNA. *Cell*, 136:577–580.
- Shepard PJ, Choi E, Lu J, Flanagan LA, Hertel KJ & Shi Y** (2011). Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA*, 17:761–772.
- Shiflett AM & Johnson PJ** (2010). Mitochondrion-related organelles in eukaryotic protists. *Annu Rev Microbiol*, 64:409–429.
- Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T et al.** (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A*, 100:15776–15781.

- Simpson AGB** (2003). Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota). *Int J Syst Evol Microbiol*, 53:1759–1777.
- Simpson AGB, Inagaki Y & Roger AJ** (2006). Comprehensive multigene phylogenies of excavate protists reveal the evolutionary positions of “primitive” eukaryotes. *Mol Biol Evol*, 23:615–625.
- Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM & Birol I** (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res*, 19:1117–1123.
- Siomi MC, Sato K, Pezic D & Aravin AA** (2011). PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol*, 12:246–258.
- Smith A & Johnson P** (2011). Gene expression in the unicellular eukaryote *Trichomonas vaginalis*. *Res Microbiol*, 162:646–654.
- Sogin ML** (1989). Evolution of eukaryotic microorganisms and their small subunit ribosomal RNAs. *Amer Zool*, 29:487–499.
- Sogin ML** (1991). Early evolution and the origin of eukaryotes. *Curr Opin Genet Dev*, 1:457–463.
- Sogin ML, Gunderson JH, Elwood HJ, Alonso RA & Peattie DA** (1989). Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from *Giardia lamblia*. *Science*, 243:75–77.
- Sorge J, Gross E, West C & Beutler E** (1990). High level transcription of the glucocerebrosidase pseudogene in normal subjects and patients with Gaucher disease. *J Clin Invest*, 86:1137–1141.
- Sorvillo F, Kovacs A, Kerndt P, Stek A, Muderspach L & Sanchez-Keeland L** (1998). Risk factors for trichomoniasis among women with human immunodeficiency virus (HIV) infection at a public clinic in Los Angeles County, California: implications for HIV prevention. *Am J Trop Med Hyg*, 58:495–500.
- Sorvillo F, Smith L, Kerndt P & Ash L** (2001). *Trichomonas vaginalis*, HIV, and African-Americans. *Emerg Infect Dis*, 7:927–932.
- Stiller JW, Huang J, Ding Q, Tian J & Goodwillie C** (2009). Are algal genes in nonphotosynthetic protists evidence of historical plastid endosymbioses? *BMC Genomics*, 10:484.
- Struhl K** (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct Mol Biol*, 14:103–105.
- Sugarbaker DJ, Richards WG, Gordon GJ, Dong L, De Rienzo A, Maulik G, Glickman JN, Chirieac LR, Hartman M, Taillon BE et al.** (2008). Transcriptome sequencing of malignant pleural mesothelioma tumors. *Proc Natl Acad Sci U S A*, 105:3521–3526.
- Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D et al.** (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321:956–960.

- Taft RJ, Pheasant M & Mattick JS (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays*, 29:288–299.
- Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM *et al.* (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, 453:534–538.
- Tarver JE, Donoghue PCJ & Peterson KJ (2012). Do miRNAs have a deep evolutionary history? *Bioessays*, 34:857–866.
- Teixeira SM, de Paiva RMC, Kangussu-Marcolino MM & Darocha WD (2012). Trypanosomatid comparative genomics: Contributions to the study of parasite biology and different parasitic diseases. *Genet Mol Biol*, 35:1–17.
- Tian B, Hu J, Zhang H & Lutz CS (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res*, 33:201–212.
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ & Pachter L (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28:511–515.
- Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, Freier SM, Bennett CF, Sharma A, Bubulya PA *et al.* (2010). The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell*, 39:925–938.
- Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RHY, Aerts A, Arredondo FD, Baxter L, Bensasson D, Beynon JL *et al.* (2006). *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science*, 313:1261–1266.
- Ulitsky I, Shkumatava A, Jan CH, Sive H & Bartel DP (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*, 147:1537–1550.
- Vazquez F, Vaucheret H, Rajagopalan R, Lepers C, Gascioli V, Mallory AC, Hilbert J, Bartel DP & Cr  t   P (2004). Endogenous trans-acting siRNAs regulate the accumulation of *Arabidopsis* mRNAs. *Mol Cell*, 16:69–79.
- Velculescu VE, Zhang L, Vogelstein B & Kinzler KW (1995). Serial analysis of gene expression. *Science*, 270:484–487.
- Venema J & Tollervey D (1999). Ribosome synthesis in *Saccharomyces cerevisiae*. *Annu Rev Genet*, 33:261–311.
- Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I & Marden JH (2008). Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol Ecol*, 17:1636–1647.
- Wagner GP, Kin K & Lynch VJ (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci*, 131:281–285.
- Wang G, Lercher MJ & Hurst LD (2011). Transcriptional coupling of neighboring genes and gene expression noise: evidence that gene orientation and noncoding transcripts are modulators of noise. *Genome Biol Evol*, 3:320–331.

- Wang Y, Yu Y, Pan B, Hao P, Li Y, Shao Z, Xu X & Li X (2012). Optimizing hybrid assembly of next-generation sequence data from *Enterococcus faecium*: a microbe with highly divergent genome. *BMC Syst Biol*, 6:S21.
- Wang Z, Gerstein M & Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10:57–63.
- Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T *et al.* (2008). Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature*, 453:539–543.
- Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J & Bähler J (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453:1239–1243.
- Wilkening S, Pelechano V, Järvelin AI, Tekkedil MM, Anders S, Benes V & Steinmetz LM (2013). An efficient method for genome-wide polyadenylation site mapping and RNA quantification. *Nucleic Acids Res*, 41:e65.
- Woehle C, Dagan T, Martin WF & Gould SB (2011). Red and problematic green phylogenetic signals among thousands of nuclear genes from the photosynthetic and apicomplexa-related *Chromera velia*. *Genome Biol Evol*, 3:1220–1230.
- Woehle C, Kusdian G, Radine C, Graur D, Landan G & Gould SB (2014). The excavate parasite *Trichomonas vaginalis* expresses thousands of pseudogenes and long non-coding RNAs independently from neighboring genes. *BMC Genomics*, Eingereicht.
- Wolf YI & Koonin EV (2013). Genome reduction as the dominant mode of evolution. *Bioessays*, 35:829–837.
- Xiong J, Lu X, Zhou Z, Chang Y, Yuan D, Tian M, Zhou Z, Wang L, Fu C, Orias E *et al.* (2012). Transcriptome analysis of the model protozoan, *Tetrahymena thermophila* using deep RNA sequencing. *PLoS One*, 7:e30630.
- Xu X, Zeng L, Tao Y, Vuong T, Wan J, Boerma R, Noe J, Li Z, Finnerty S, Pathan SM *et al.* (2013). Pinpointing genes underlying the quantitative trait loci for root-knot nematode resistance in palaeopolyploid soybean by whole genome resequencing. *Proc Natl Acad Sci U S A*, 110:13469–13474.
- Yabuki A, Ishida K & Cavalier-Smith T (2013). *Rigifila ramosa* n. gen., n. sp., a filose apusozoan with a distinctive pellicle, is related to *Micronuclearia*. *Protist*, 164:75–88.
- Yamamoto A, Hashimoto T, Asaga E, Hasegawa M & Goto N (1997). Phylogenetic position of the mitochondrion-lacking protozoan *Trichomonas tenax*, based on amino acid sequences of elongation factors 1 α and 2. *J Mol Evol*, 44:98–105.
- Yano Y, Saito R, Yoshida N, Yoshiki A, Wynshaw-Boris A, Tomita M & Hirotsune S (2004). A new role for expressed pseudogenes as ncRNA: regulation of mRNA stability of its homologous coding gene. *J Mol Med (Berl)*, 82:414–422.
- Yigit E, Batista PJ, Bei Y, Pang KM, Chen CG, Tolia NH, Joshua-Tor L, Mitani S, Simard MJ & Mello CC (2006). Analysis of the *C. elegans* Argonaute family reveals that distinct Argonautes act sequentially during RNAi. *Cell*, 127:747–757.

-
- Yoon OK & Brem RB** (2010). Noncanonical transcript forms in yeast and their regulation during environmental stress. *RNA*, 16:1256–1267.
- Zeevi M, Nevins JR & Darnell JE** (1982). Newly formed mRNA lacking polyadenylic acid enters the cytoplasm and the polyribosomes but has a shorter half-life in the absence of polyadenylic acid. *Mol Cell Biol*, 2:517–525.
- Zeller G, Henz SR, Widmer CK, Sachsenberg T, Rättsch G, Weigel D & Laubinger S** (2009). Stress-induced changes in the *Arabidopsis thaliana* transcriptome analyzed using whole-genome tiling arrays. *Plant J*, 58:1068–1082.
- Zhang Z & Gerstein M** (2004). Large-scale analysis of pseudogenes in the human genome. *Curr Opin Genet Dev*, 14:328–335.
- Zhao J, Hyman L & Moore C** (1999). Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev*, 63:405–445.
- Zheng D, Frankish A, Baertsch R, Kapranov P, Reymond A, Choo SW, Lu Y, Denoeud F, Antonarakis SE, Snyder M et al.** (2007). Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. *Genome Res*, 17:839–851.
- Zheng D & Gerstein MB** (2007). The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? *Trends Genet*, 23:219–224.
- Zheng D, Zhang Z, Harrison PM, Karro J, Carriero N & Gerstein M** (2005). Integrated pseudogene annotation for human chromosome 22: evidence for transcription. *J Mol Biol*, 349:27–45.
- Zhu Z, Zhang Y, Ji Z, He S & Yang X** (2013). High-throughput DNA sequence data compression. *Brief Bioinform*, Im Druck.

DANKSAGUNG

Als erstes möchte ich mich bei Prof. Dr. William F. Martin bedanken, dafür dass er mich in seinem Institut aufgenommen hat und die Betreuung meiner Doktorarbeit übernahm. Ich habe mich stets gut gefördert gefühlt und fand bei meinen Anliegen immer ein offenes Ohr.

Als meinem Betreuer und Ansprechpartner gilt Dr. Sven B. Gould ein besonderer Dank. In allen meinen Projekten konnte ich auf seinen Rat zurückgreifen. Ihm verdanke ich meine Forschung an unterschiedlichen Themengebieten, vernetzt mit Ergebnissen aus dem Labor. Ein großer Teil dieser Arbeit liegt seiner Unterstützung zugrunde.

Bei apl. Prof. Dr. Gerhard Steger möchte ich mich bedanken für die Übernahme der Rolle des Korreferenten und dafür dass er mich, durch seine Betreuung in meiner Diplomarbeit, in die richtige Richtung lenken konnte. Prof. Dr. Tal Dagan gebührt mein Dank für ihren hilfreichen Rat und Betreuung von bioinformatischer Seite, gerade zu Beginn meiner Zeit als Doktorand und Dr. Giddy Landan für seine umfangreiche Hilfe im Bezug auf den Umgang mit Statistiken.

Des Weiteren möchte ich mich herzlich bei unserem komplettem Institut bedanken. Die Arbeitsatmosphäre war angenehm und hat mich gerne dort arbeiten lassen. Diesbezüglich bedanke ich mich besonders bei Mayo und Kathrin, mit denen ich lange mein Büro teilte und Thorsten, der diese Arbeit Gegengelesen hat. Sie waren direkte Ansprechpartner bei Problemen und immer offen für produktive Diskussionen.

Zum Schluss möchte ich mich noch bei meiner Unterstützung von familiärer Seite bedanken. Meine Familie und besonders die meiner Freundin boten mir einen Rückhalt, der sich besonders in der Endphase der Promotion als unabdingbar erwies. Der weitaus größte Dank gilt dabei meiner Lebensgefährtin Yvonne, die einen essentiellen Beitrag zum Gelingen dieser Arbeit geleistet hat.

Die vorliegende Dissertation habe ich eigenständig und ohne unerlaubte Hilfe angefertigt. Die Dissertation wurde weder in der vorgelegten noch in ähnlicher Form bei einer anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Kiel, den 15.07.2014

.....
(Christian Wöhle)