FRÉDÉRIC RAYMOND

BIO-INFORMATIQUE POUR LA GÉNOMIQUE ET LE DIAGNOSTIC DES MALADIES INFECTIEUSES

Thèse présentée
à la Faculté des études supérieures de l'Université Laval
dans le cadre du programme de Doctorat en physiologie-endocrinologie
pour l'obtention du grade de Philosophiæ doctor (Ph. D.)

DÉPARTEMENT DE BIOLOGIE MÉDICALE
FACULTÉ DE MÉDECINE
UNIVERSITÉ LAVAL
QUÉBEC

2011

Résumé

Le séquençage du génome d'un microorganisme est un jalon important dans l'étude de sa biologie. Quel que soit cet organisme, les outils bio-informatiques nécessaires pour comprendre son génome et le comparer aux autres génomes séquencés seront similaires. Dans cette thèse, l'ADN génomique de parasites et de virus est mis à profit afin de mieux comprendre ces microorganismes.

Dans un premier temps, le parasite protozoaire *Leishmania* est étudié par transcriptomique et par génomique comparative afin de mieux comprendre son infectivité, sa résistance aux antiparasitaires et son mode de vie dimorphique. Ce parasite alterne entre le stade flagellé (promastigote) et le stade intracellulaire aflagellé (amastigote). Afin de faciliter l'analyse par biopuces du transcriptome de *Leishmania,* un système de gestion et d'analyse de données de biopuces a été conçu. Quatre études utilisant ce système sont présentées sommairement et leurs implications sont discutées. Deuxièmement, le génome de l'espèce *Leishmania (sauroleishmania) tarentolae,* qui n'est pas pathogène pour l'humain, a été séquencé et comparé à trois espèces infectant l'homme. Cette étude a montré que, même si peu de gènes différencient les espèces, *L. tarentolae* possède moins de gènes associés au stade amastigote que les autres espèces. Deux familles de gènes ont été trouvées en nombre de copies élevées chez *L. tarentolae :* GP63 et PSA31C. Ces résultats permettent une meilleure compréhension de la biologie de *L. tarentolae* et de la virulence des autres espèces de *Leishmania.*

Dans un deuxième temps, les séquences des génomes de virus respiratoires disponibles dans les bases de données publiques ont été analysées pour créer un test diagnostique permettant la détection et l'identification de 25 types de virus respiratoires, dont la grippe A (H1N1) responsable de la pandémie de 2009 et la grippe aviaire A (H5N1). Le test a été validé avec des échantillons de laboratoire et avec des échantillons cliniques.

Même si l'étude du parasite *Leishmania* était indépendante de celle des virus respiratoires, les approches utilisées pour ces deux projets étaient similaires. Ainsi, la bio-informatique est un outil essentiel en microbiologie, car elle est indispensable pour résoudre des problèmes de diverses natures chez des organismes différents.

Abstract

Sequencing a genome is a milestone in the study of an organism. Bioinformatics allow both to better understand single organisms and to compare them to related species through comparative genomics. This thesis centers on the idea that genome sequence of parasites and viruses can be used in various ways to better understand these microorganisms.

Transcriptomics and comparative genomics were used to study the protozoan parasite *Leishmania* in order to better understand its virulence, its resistance to antiparasitic drugs, and its dimorphic life-cycle, which includes a flagellated free form named promastigote and an aflagellate intracellular form named amastigote. In order to study gene expression in *Leishmania,* an integrated management and analysis system was created, along with protocols designed for *Leishmania* microarrays analysis. Four studies using this system are briefly described. In another study, the genome of *Leishmania (sauroleishmania) tarentolae,* a lizard parasite, was sequenced and compared to human pathogenic *Leishmania* species. This study showed little difference between the *Leishmania* species, although *L. tarentolae* seems to contain less genes associated to the amastigote life-cycle, including the amastin gene. Two gene families were highly expanded in *L. tarentolae:* the surface metalloprotease GP63 and the promastigote antigen protein PSA31C. These results provide a better understanding of *L. tarentolae* biology and give insights on the genes involved in virulence in pathogenic *Leishmania* species.

The second part of this thesis concerns the creation of a molecular diagnostic assay for the detection and identification of 25 respiratory virus types, including the influenza A/H1N1 pandemic strain and the avian influenza A/H5N1 strain. This assay was created by analyzing genome sequences available from public repositories and it was afterwards tested on laboratory and clinical virus strains.

Although *Leishmania* and respiratory viruses are distantly related, the approaches used in both projects were similar. Thus, bioinformatics is an essential and ubiquitous science that allows to solve problems in different areas (“omics”) of biology.

Avant-propos

En 1990, la population de la planète est décimée à 99,4 % par un supervirus échappé d'un centre de recherche de l'armée américaine. Ce moment, même s'il est fictif, a fait naître en moi une grande curiosité pour les maladies infectieuses. Depuis ma première lecture du roman *Le Fléau,* de Stephen King, j'ai su que j'orienterais ma carrière vers la recherche en microbiologie. Au fil des ans, d'autres œuvres de fiction ont modulé mon choix de carrière, notamment les bandes dessinées américaines X-Men, avec leurs constantes références à la génétique. Des projets comme ceux auxquels j'ai eu la chance de participer au cours de mes travaux de doctorat m'ont permis, par leur nature à la pointe des technologies biomédicales, d'étudier les maladies infectieuses d'un point de vue aussi fascinant que celui des docteurs Henry McCoy, Nathaniel Essex et Moïra McTaggert, qui ont habité mon imaginaire depuis que je sais lire. Sans Stephen King et Chris Claremont, j'aurais peut-être choisi une autre voie.

De retour à la réalité, je remercie Jacques Corbeil, mon directeur de recherche, qui m'a permis de participer à tous ces projets fascinants et qui, avec son enthousiasme et son esprit visionnaire, m'a dirigé pendant mon doctorat. La liberté qu'il m'a laissée pour mener mes projets m'a permis de m'épanouir pendant mes études de troisième cycle.

Merci à Lynda Robitaille, présence indispensable dans le laboratoire, sans qui beaucoup de projets auraient été très compliqués. Merci à Sébastien Boisvert, pour ses talents de bio-informaticien, qui a grandement facilité plusieurs des travaux présentés dans cette thèse. Merci à Nancy Boucher, qui m'a beaucoup aidé dans le laboratoire et qui m'a fait rire. Merci à Jessyka Fortin pour sa connaissance approfondie de la PCR en temps réel. Merci à Mélissa Sirois, pour toutes les discussions. Merci à Éric Madore, pour son humour et ses compétences dans le domaine des biopuces. Merci à Thibault Varin, pour les « fucking lie » et autres musiques tonitruantes. Merci à René Paradis, pour son support informatique et sa connivence musicale. Merci à toute la gang de bio-informatique de l'époque Complan. Merci au bâtiment Complan, pour m'avoir inspiré un roman d'horreur.

Merci à Marc Ouellette et à Barbara Papadopoulou, pour leur collaboration et leurs commentaires tout au long des projets décrits dans cette thèse. Merci à Danielle Légaré,

ressource essentielle en génomique de *Leishmania*. Merci à Gaétan Roy pour son support dans le projet de séquençage de *L. tarentolae*. Merci à Jean-Michel Ubeda, pour ces projets fascinants au cours desquels nous avons collaboré, même s'ils ne sont pas tous inclus dans cette thèse. Merci à Annie Rochette, pour sa collaboration dans l'analyse de ses données. Merci aux autres leishmanieux avec qui j'ai eu la chance de travailler.

Merci à Guy Boivin et à son équipe, en particulier à Julie Carbonneau et à Marie-Ève Hamelin, pour leur support dans les projets de diagnostic moléculaire d'infections respiratoires virales.

Finalement, je remercie mes parents, qui m'ont toujours encouragé et qui m'ont souvent montré la fierté qu'ils ressentaient concernant mon choix de carrière. Merci à Caroline, mon épouse adorée, qui, en plus de m'avoir encouragé tout au long de mes études universitaires, a mis à mon service ses compétences de réviseure linguistique chaque fois que j'en avais besoin, en particulier pour réviser cette thèse. Merci à Lauriane, qui ne m'a pas toujours laissé travailler, mais qui a toujours ensoleillé mes journées par sa petite binette. Merci à la famille et aux amis qui se sont tenus au courant de mes progrès tout au long de mes études.

Je remercie les Instituts de recherche en santé du Canada pour le financement de mes études de doctorat.

Merci au parasite *Leishmania* qui m'a montré que tout sujet est intéressant si l'on pose les bonnes questions.

Contributions

Cette section présente la description de mes contributions aux différents projets présentés dans cette thèse de doctorat.

Le chapitre 4 présente la création d'un système de gestion et d'analyses de données de biopuces utilisées pour l'étude du parasite *Leishmania*. Au cours de ces travaux, j'ai créé le concept de LMMAP et rédigé les spécifications du logiciel. J'ai ensuite supervisé Sébastien Boisvert dans la programmation du logiciel. J'ai créé les protocoles de normalisation et d'analyse des biopuces *Leishmania*. Finalement, j'ai conseillé les utilisateurs de la biopuce dans la conception de leurs expériences, fait la normalisation et l'analyse statistiques des données et proposé des analyses supplémentaires pour les articles énumérés dans l'annexe 1.

Le chapitre 5 décrit le génome de *Leishmania tarentolae*. Pour ce projet, j'ai supervisé et validé l'assemblage du génome, j'ai réalisé l'annotation du génome et j'ai effectué les analyses de génomique comparative. Finalement, j'ai fouillé les résultats de génomique comparative afin de trouver les différences majeures entre *L. tarentolae* et les autres espèces de *Leishmania*. J'ai supervisé et analysé les expériences de validation (biopuce et Southern). J'ai rédigé la première version de cet article et je l'ai retravaillé sous la supervision de Marc Ouellette, Barbara Papadopoulou et Jacques Corbeil.

Les chapitres 7, 8 et 9 décrivent un test de diagnostic moléculaire permettant de détecter 25 virus respiratoires dans des échantillons cliniques. Pour ce projet, j'ai conçu le test diagnostique sur biopuce ainsi que le test de PCR en temps réel. J'ai participé à l'optimisation du test ainsi qu'à sa validation. J'ai mis à jour le test pour la détection de la grippe A (H1N1) pandémique et j'ai supervisé son optimisation et sa validation. J'ai rédigé le chapitre 7 sous la supervision de Jacques Corbeil. J'ai rédigé les deux articles portant sur ce sujet sous la supervision de Guy Boivin et Jacques Corbeil.

À mes parents

Table des matières

Liste des tableaux

Liste des figures

1. Introduction

L'avènement de la génomique et du séquençage à haut débit a engendré une nouvelle ère dans l'étude de la biologie. Aujourd'hui, la seule séquence d'un génome permet de réaliser une multitude d'analyses qui, en elles-mêmes, apportent beaucoup à la biologie de l'organisme étudié. Des problèmes de biologie qui demandaient un travail colossal peuvent maintenant être résolus en quelques expériences, parfois même sans travail de laboratoire. Ce changement de paradigme nécessite une nouvelle approche à la biologie, car les questions scientifiques ne se posent plus de la même manière.

D'un point de vue fondamental, la génomique et la bio-informatique permettent une étude plus intégrée des organismes vivants. D'un point de vue appliqué, les technologies de la génomique peuvent être adaptées au diagnostic des maladies infectieuses, permettant ainsi une meilleure compréhension et une meilleure gestion de ces maladies. Ces deux aspects ont été explorés au cours de mon doctorat. Ainsi, cette thèse sera séparée en deux sections distinctes avec, pour fil conducteur, la séquence des génomes et leur étude par la bio-informatique.

Dans la première partie de cette thèse, soit les chapitres de 2 à 5, je présenterai comment la génomique, supportée par la bio-informatique, a permis de mieux comprendre le parasite protozoaire *Leishmania*. Ainsi, l'étude de ce parasite est la source de questions à partir desquelles des approches génomiques et bio-informatiques ont été établies.

Avant de passer aux différentes réalisations faites au cours de mon doctorat (chapitres 4 et 5), je présenterai un résumé de ce qui est connu du parasite *Leishmania* (chapitre 2) ainsi qu'un survol des technologies de génomique utilisées dans cette thèse, comme le séquençage de génomes et l'analyse de l'expression des gènes (chapitre 3).

Au chapitre 4, je décrirai l'implantation d'un système de gestion et d'analyses de données tirées d'études d'expression génique par biopuces. La mise en place de ce système sera traitée des points de vue de la bio-informatique et de l'analyse des données. Je présenterai aussi, dans cette section, les résultats publiés dans des études utilisant ces systèmes. Je

discuterai des avantages et des inconvénients de cette approche en m'appuyant sur différentes publications pour lesquelles j'ai réalisé les analyses bio-informatiques.

Au chapitre 5, je présenterai le séquençage du génome du parasite *Leishmania (sauroleishmania) tarentolae* et je comparerai cette séquence avec celle des génomes de *Leishmania* déjà séquencés, soit ceux de *L. braziliensis,* de *L. infantum* et de *L. major*. Dans cette section, je décrirai les caractères génomiques qui différencient une espèce de *Leishmania* non pathogène pour l'humain d'espèces pathogènes. Je montrerai aussi comment il est possible d'intégrer les données expérimentales de génomique à haut débit pour mieux comprendre des organismes dont le génome est peu connu ou peu annoté. L'ensemble de ces analyses permet une meilleure compréhension du parasite et procure des données qui s'avéreront utiles dans la création de nouveaux outils pour lutter contre la leishmaniose.

Dans la deuxième partie de cette thèse, je montrerai comment les technologies de la génomique peuvent être mises au service du diagnostic des maladies infectieuses, en particulier des infections respiratoires. Au chapitre 6, je présenterai un aperçu des infections respiratoires virales ainsi que des méthodes de diagnostic. Au chapitre 7, j'expliquerai les rudiments de la conception d'un test de diagnostic moléculaire. Au chapitre 8, je présenterai un test de diagnostic moléculaire des virus respiratoires sur biopuce qui a été adapté à un appareil de diagnostic automatisé conçu pour les laboratoires cliniques. Ensuite, au chapitre 9, je décrirai comment ce test diagnostique a été modifié pour permettre la détection de la grippe A (H1N1) d'origine porcine.

Toutes ces réalisations démontrent la puissance de la génomique et de la bio-informatique, qui permettent des avancées scientifiques et techniques utiles dans des domaines qui, par le passé, étaient exclusifs. Elles montrent aussi comment il est important de repenser les problèmes biologiques afin d'innover. Ces techniques permettent aussi de poser de nouvelles questions.

Objectifs

L'objectif général de mes travaux de doctorat était d'utiliser la bio-informatique afin de favoriser l'avancement des connaissances dans plusieurs domaines de l'étude des maladies infectieuses. Des objectifs particuliers ont été établis pour chacun des projets décrits dans cette thèse.

L'étude du parasite *Leishmania* avait pour objectif général de mieux comprendre le génome de ce parasite par le biais de la génomique et de la bio-informatique. Les objectifs particuliers étaient les suivants :

- Favoriser les études de transcriptomique et de génomique comparative par hybridation sur biopuce en créant un cahier de laboratoire électronique et des protocoles d'analyse de biopuces;
- Séquencer le génome de *Leishmania (sauroleishmania) tarentolae*, l'annoter et le comparer aux espèces de *Leishmania* pathogènes pour l'humain afin de mieux comprendre :
    - le mode de vie d'un parasite de lézard,
    - la virulence des espèces pathogènes pour l'humain en déterminant quels gènes de *L. major,* de *L. infantum* et de *L. braziliensis* sont absents ou en moins grand nombre de copies chez *L. tarentolae*.

L'étude des virus respiratoires avait pour objectif général de fournir des outils aux cliniciens et aux épidémiologistes pour qu'ils puissent mieux évaluer l'importance des différents virus respiratoires observés en clinique. Les objectifs particuliers de ce projet étaient de

- créer un outil diagnostic permettant l'identification des virus respiratoires les plus communs en un seul test automatisé,
- valider ce test avec des échantillons de laboratoire et des échantillons cliniques,

- d’étudier la distribution des virus respiratoires infectant les enfants en effectuant une étude clinique rétrospective.

2. Le parasite *Leishmania*

L'Organisation mondiale de la santé (OMS) a inclus la leishmaniose dans sa liste de priorité des maladies à combattre. Cela n'est pas surprenant, car, même si la maladie est peu connue du public nord-américain, elle touche 88 pays et met en jeu la santé de plus de 350 millions d'individus (1). Environ 14 millions de personnes souffrent de cette maladie et plus de 2 millions de nouveaux cas sont déclarés chaque année.

2.1. La répartition territoriale de la leishmaniose

D'un point de vue historique, la maladie assaille les habitants de certaines régions depuis plus de 4 millénaires, puisque des traces de leishmanioses viscérale et cutanée[1] ont été observées chez des momies égyptiennes datant d'environ 2000 av. J.-C. (2). Depuis quelques années, la co-infection par le virus d'immunodéficience humaine (VIH) et par *Leishmania* est en émergence, causant de graves problèmes chez les malades immunosupprimés (3). Le parasite *Leishmania* peut aussi infecter des animaux, par exemple le chien, ce qui fait de la leishmaniose une zoonose.

Le parasite *Leishmania* est endémique aux régions équatoriales du globe, notamment en Amérique centrale, en Amérique du Sud, en Afrique, au sud de l'Europe et au Moyen-Orient (en particulier dans le bassin méditerranéen) et en Asie. Selon l'OMS, plus de 90 % des 500 000 cas annuels de leishmaniose viscérale se situent au Bangladesh, au Brésil, en Inde, au Népal et au Soudan (1). Plus de 50 000 de ces derniers sont mortels. Parmi les 1 500 000 cas de leishmaniose cutanée répertoriés annuellement, plus de 90 % se retrouvent en Afghanistan, en Algérie, en Arabie saoudite, au Brésil, au Pérou, en République islamique d'Iran et au Soudan. Aussi, vu la situation politique mondiale et l'implication du Canada et des États-Unis en Afghanistan et en Iraq, la maladie représente une menace pour de nombreux soldats nord-américains (4).

Depuis quelques années, des cas de leishmaniose ont été diagnostiqués dans des régions qui n'étaient habituellement pas touchées, ou très peu, par cette maladie, comme le nord de l'Italie (5) et le sud de l'Allemagne (6). Les changements climatiques observés au cours des

[1] Voir la section 2.2 pour les définitions des types de leishmaniose.

dernières années pourraient être responsables d'une telle modification de la distribution de la maladie. En effet, ils permettent une meilleure survie des vecteurs de la maladie, les moustiques, dans des régions où ils ne pouvaient pas, dans le passé, résister à l'hiver (7). Une modification de la distribution géographique des espèces de *Leishmania* peut aussi être influencée par la migration de certaines espèces animales servant de réservoir à la maladie, par exemple le chien, qui peut aussi être infecté par le parasite (8).

2.2. Les types de leishmaniose

Il existe plusieurs types de leishmaniose, entre autres types les leishmanioses cutanée, viscérale et mucocutanée. La pathologie est associée à l'espèce de parasite infectant l'individu, mais certaines espèces peuvent causer différentes formes de leishmanioses, ce qui dépend probablement aussi de la réponse immunitaire de l'hôte (6, 9).

La leishmaniose cutanée est définie par le développement d'ulcères de la peau. L'infection peut prendre plusieurs mois à guérir et défigure parfois le patient. La leishmaniose cutanée est surtout causée par *L. major,* par *L. tropica* et par *L. aethiopica* dans les vieux continents et par *L. mexicana,* par *L. amazonensis* et par *L. braziliensis* en Amérique.

La leishmaniose viscérale, aussi appelée « kala-azar » ou « fièvre noire », est caractérisée par de la fièvre, des pertes de poids, des désordres chroniques des fonctions hématologiques et hépatospléniques, c'est-à-dire des désordres relatifs au sang, au foie et à la rate. Cette leishmaniose est mortelle si elle n'est pas traitée et elle est causée notamment par *L. infantum* et par *L. donovani*.

L. donovani cause aussi la leishmaniose post-kala-azar (PKDL), une forme secondaire de leishmaniose qui se caractérise par l'apparition de lésions sur le visage suivie de la dissémination de ces dernières sur tout le corps.

La leishmaniose mucocutanée, finalement, est une forme intermédiaire de leishmaniose pouvant entraîner la destruction des muqueuses du nez, de la bouche et de la trachée, défigurant ainsi le patient. Cette forme de leishmaniose est causée par *L. braziliensis*.

2.3. Le parasite *Leishmania*

Leishmania est un parasite protozoaire de l'ordre des *Kinetoplastidae* et de la famille des *Tripanosomatidae*. La diversité des espèces de *Leishmania* reflète bien la diversité des formes cliniques de la maladie (9-11). Il a été confirmé qu'une vingtaine d'espèces de *Leishmania* peuvent infecter l'humain.

Les *Leishmania* ont deux types de classification : selon le sous-genre et selon la géographie.

Sous-genre :

- *Leishmania (leishmania),*
- *Leishmania (vianna),*
- *Leishmania (sauroleishmania).*

Géographie :

- *Leishmania* du Nouveau Monde (New World), si le parasite est retrouvé en Amérique. Son vecteur est la mouche des sables du genre *Lutzomyia,*
- *Leishmania* des Vieux Continents (Old World), si le parasite est retrouvé en Europe, en Afrique ou en Asie. Il est transmis par la mouche du genre *Phlebotomus.*

Le parasite *Leishmania* a un cycle de vie dimorphique (6, 10). *Leishmania* vit sous sa forme flagellée, aussi appelée « promastigote », dans le tractus digestif de la mouche des sables, son vecteur. Les parasites sont transmis par une piqûre au cours du repas sanguin de l'insecte. Au moment de son injection, le parasite est sous la forme promastigote et il sera phagocyté par le macrophage. Une fois dans le macrophage, le parasite se différencie en amastigote, forme capable de survivre dans les conditions extrêmes du phagolysosome des macrophages. L'infection par *Leishmania* débute au site de la piqûre pour ensuite progresser en diverses pathologies selon l'espèce infectante et la réaction de l'hôte. Dans la

forme cutanée de la leishmaniose, les macrophages infectés restent confinés aux lésions de la peau alors que, dans la forme viscérale, les sites majeurs d'infections sont le foie, la rate et la moelle épinière (10). Cependant, les macrophages infectés doivent aussi être présents sous la peau afin de favoriser la retransmission du parasite par le vecteur (12). Le cycle de *Leishmania* sera discuté plus en détail dans la section 2.5.

2.3.1. Les *Leishmania (leishmania)*

Le sous-genre *Leishmania (leishmania)* a été retrouvé en Amérique et dans les Vieux Continents (13). Le complexe *L. donovani* inclut *L. donovani* et *L. archibaldi*. Le complexe *L. infantum* inclut *L. infantum* et *L. chagasi.* Les complexes *L. tropica, L. killicki, L. aethiotropica, L. major* et *L. mexicana* comprennent uniquement les espèces éponymes. Le complexe *L. amazonensis* inclut les espèces *L. amazonensis* et *L. aristidesi.* Tous ces complexes sont retrouvés dans les Vieux Continents, à l'exception des complexes *L. mexicana* et *L. amazonensis,* qui sont retrouvés en Amérique. Les espèces *L. infantum, L. tropica, L. killicki* et *L. major* sont endémiques au bassin méditerranéen.

L. major, L. tropica et *L. killicki* causent une leishmaniose cutanée, aussi nommée « oriental sore ». *L. mexicana* et *L. amazonensis* causent aussi une leishmaniose cutanée. *L. aethiotropica* cause aussi une leishmaniose cutanée, qui peut être diffuse. *L. donovani* et *L. infantum* causent une leishmaniose viscérale, aussi appelée « kala-azar » ou « fièvre noire ».

2.3.2. Les *Leishmania (vianna)*

Le sous-genre *Leishmania (vianna)* est endémique en Amérique. Le complexe *L. braziliensis* inclut *L. braziliensis* et *L. peruviana.* Le complexe *L. guyanensis* inclut *L. guyanensis, L. panamensis* et *L. shawi.* Les complexes *L. naiffi* et *L. lainsoni* incluent uniquement les espèces éponymes.

À l'exception de *L. braziliensis,* qui peut aussi causer une leishmaniose mucocutanée, ces espèces causent une leishmaniose cutanée. Quant à *L. guyanensis,* il peut causer une forme systémique et diffuse de la maladie (13). De fait, la forme viscérale est rare en Amérique (14).

2.3.3. Les *Leishmania (sauroleishmania)*

Leishmania tarentolae est le parasite le plus étudié des *Leishmania (sauroleishmania).* Il a été isolé des lézards *Tarentola mauritanica* et de *T. annularis* (15). Pendant un certain temps, son appartenance au genre *Leishmania* a été mise en doute, mais cela a été infirmé et l'espèce est maintenant confirmée comme un membre des leishmanies (16, 17). L'espèce est répandue au Moyen-Orient. D'ailleurs, une étude qui devait évaluer la proportion de phlébotomes positifs pour *Leishmania* a été réalisée dans une base militaire en Iraq. Cette étude a identifié 91,9 % de 284 phlébotomes positifs pour *L. tarentolae* (4). Il a été rapporté que, chez le lézard, *L. tarentolae* causait une faible parasitémie, d'environ un leucocyte infecté par champ de microscope à 40X, et que ces leucocytes contenaient de 3 à 9 amastigotes intracellulaires (15). Une autre étude d'observation de lézards, cette fois au Turkménistan, a rapporté une faible parasitémie chez le lézard, mais aucune *Leishmania* sous forme amastigote (18).

Il a été montré que *L. tarentolae* pouvait être phagocyté par des macrophages humains ou murins et qu'ils pouvaient s'y différencier en amastigotes. La même étude a montré que *L. tarentolae* pouvait activer la maturation des cellules dendritiques, induire la prolifération des cellules T et la production d'interférons gamma, et induire une réponse immune de type Th1 (cellulaire) (19). *L. tarentolae* est rapidement éliminé par le système immunitaire humain ou murin (18). Comme le parasite n'est pas dangereux pour l'humain et qu'il entraîne une protection immunitaire croisée avec *L. donovani* chez la souris, il est considéré comme un candidat vaccinal potentiel (19). *L. tarentolae* a aussi été considéré comme un vecteur intéressant pour la vaccination contre le virus d'immunodéficience humaine (20).

2.3.4. Les autres espèces de trypanosomatides

Tout comme *Leishmania,* les parasites du genre *Trypanosoma* appartiennent à l'ordre de *Kinetoplastidae.* Les deux principaux pathogènes de ce genre sont *Trypanosoma brucei* et *T. cruzi.* Il existe aussi plusieurs espèces de trypanosomes non pathogènes pour l'humain, mais qui s'attaquent à d'autres mammifères ou à des reptiles.

T. brucei est responsable de la maladie du sommeil, aussi appelée « trypanosomiase africaine humaine ». Cette maladie cause environ un demi-million de cas par année et

environ 70 000 morts au centre de l'Afrique (21). La maladie du sommeil causée par *T. brucei* est fatale si elle n'est pas traitée. *T. brucei* est transmis par la piqûre de la mouche tsé-tsé. Le parasite survit dans le sang du mammifère qui lui sert d'hôte. À mesure que l'infection se diffuse, le parasite modifie ses protéines de surface afin d'éviter son élimination par le système immunitaire. Finalement, quand l'infection est bien établie, le parasite passe de sa forme flagellée à une forme non flagellée qui se divise moins rapidement. Cette transformation prolonge la vie de l'hôte et favorise la transmission du parasite. *T. brucei gambiese* cause une maladie chronique à l'ouest de l'Afrique subsaharienne, *T. brucei rhodesiense* cause une maladie aiguë à l'est et au sud de cette même région et, finalement, *T. brucei brucei* n'est pas pathogène pour l'humain (22).

Le parasite *Trypanosoma cruzi,* qui cause la maladie de Chagas, est retrouvé en Amérique du Sud (22). Environ 10 millions d'individus sont infectés par *T. cruzi.* Le parasite est transmis par la défécation du vecteur pendant son repas sanguin et par la contamination subséquente de la plaie par les fèces. Les trypomastigotes métacycliques transmis par le vecteur pénètrent ensuite différents types de cellules, notamment les cellules du cœur, et se transforment en amastigotes. Les amastigotes se transforment ensuite en trypomastigotes et font exploser les cellules, ce qui permet leur libération dans le sang et la contamination de nouvelles cellules et de nouveaux vecteurs. Les parasites prennent la forme d'épimastigotes dans l'intestin moyen (*mid-gut*) des vecteurs (22). Le taxon *T. cruzi* se divise en deux groupes, *T. cruzi* I et *T. cruzi* II. À son tour, le second groupe peut être séparé en cinq sous-groupes (de *T. cruzi* IIa à *T. cruzi* IIe). Le groupe I s'attaque aux marsupiaux alors que le groupe II s'attaque aux mammifères (23).

2.4. Les antiparasitaires et la résistance

L'arsenal thérapeutique pour lutter contre *Leishmania* est limité. Même si plusieurs antiparasitaires sont offerts pour contrer la leishmaniose, la résistance du parasite à ces médicaments est fréquente (24). Le médicament classique pour lutter contre la maladie est l'antimoine, utilisé depuis plus de 70 ans (25). Cependant, certaines régions montrent un taux de résistance élevé et quasi universel. C'est le cas de l'Inde (26). Les médicaments de seconde ligne sont la pentamidine et l'amphothéricine B. Les antifolates comme le méthotrexate ont beaucoup été étudiés en laboratoire, mais ils n'ont pas été utilisés sur le

terrain. La miltéfosine, récemment introduite sur le marché, est en phase III d'essais cliniques. Il s'agit du premier traitement oral de la leishmaniose (27).

L'antimoine est principalement offert sous deux formes : le stibogluconate de sodium et l'antimoniate de méglumine, commercialisé sous le nom de Glucantime. Ces médicaments contiennent de l'antimoine sous sa forme pentavalente, le Sb(V). Afin d'être actif, l'antimoine doit être réduit sous sa forme trivalente, le Sb(III). Un doute subsiste à savoir si la réduction de l'antimoine est faite exclusivement dans le macrophage ou encore dans le parasite (28). Cependant, il a été montré que *Leishmania* possédait une enzyme capable de réduire le Sb[V] (29). Une fois à l'intérieur du parasite, l'antimoine trivalent a un effet sur le métabolisme du glutathion et du trypanothion. Il a été montré que le Sb(III) inhibait in vitro la trypanothione réductase (30) et la glutathione synthétase (31). La présence de Sb(III) dans la cellule entraîne aussi un efflux du trypanothion et du glutathion (32). Dans tous les cas, cela diminue le potentiel d'oxydoréduction du parasite, ce qui entraîne sa mort (33).

Plusieurs modes de résistance à l'antimoine sont connus et peuvent agir sur le médicament de différentes façons :

- Blocage de l'entrée de l'antimoine par la diminution du nombre de transporteurs à la surface de la cellule ou par la mutation de ces derniers (33);
- Perte d'activité réductase du parasite qui entraîne une diminution de la réduction du Sb(V) en Sb(III) (33);
- Augmentation du niveau de trypanothion par l'amplification de l'enzyme glutathione synthétase (GSH1) (34) ou la surexpression de l'ornithine décarboxylase (ODC) (35);
- Augmentation de l'efflux du médicament par des transporteurs comme MRPA (*Multidrug resistance protein A,* anciennement p-glycoprotéine A ou PGPA) ou d'autres pompes d'efflux. La PGPA est un transporteur ABC intracellulaire qui entraîne l'accumulation des conjugués d'antimoine dans des vacuoles intracellulaires (36).

La combinaison de plusieurs de ces modes de résistance est parfois nécessaire au parasite pour survivre à l'antimoine. Par exemple, l'amplification seule de la GSH1 n'est pas suffisante et requiert aussi l'augmentation de la quantité de transporteurs ABC (34, 35).

L'amphothéricine B est le traitement de seconde ligne contre la leishmaniose. Cet antiparasitaire agit sur l'ergostérol, qui compose la majeure partie de la membrane de plusieurs espèces d'eucaryotes unicellulaires, dont *Leishmania, T. cruzi* et certains champignons. L'effet de ce médicament peut varier d'une espèce à l'autre selon la proportion d'ergostérol composant leur membrane. Même si la résistance à l'amphothéricine B n'a pas été observée fréquemment, quelques mécanismes de résistance ont été décrits, incluant la modification de la structure des lipides membranaires par l'utilisation d'un précurseur de l'ergostérol dans la construction des membranes ou par la modification de leur méthylation par inactivation d'un gène (37). Finalement, des amplifications géniques causant la résistance ont été observées chez *L. tarentolae* (38).

Depuis le début des années 2000, la miltéfosine est utilisée pour traiter la leishmaniose. Il s'agit d'un analogue des phospholipides. Les différentes espèces de *Leishmania* ont des sensibilités différentes à cet antiparasitaire, *L. donovani* étant la plus sensible (39). D'un autre côté, *L. braziliensis* et *L. guyanensis* ne semblent pas sensibles à la miltéfosine. La résistance à cet antiparasitaire est principalement causée par une diminution de l'accumulation de la molécule dans la cellule (40). Ainsi, des mutations ponctuelles dans des transporteurs diminuent l'entrée dans la cellule. L'augmentation de la quantité de certains transporteurs ABC, permettant un meilleur efflux du médicament, a été décrite (41).

Même si la pentamidine a été longuement utilisée comme médicament de seconde ligne, son utilisation actuelle se résume au traitement de la leishmaniose cutanée, car le traitement de la leishmaniose viscérale semble induire une résistance assez rapide contre ce médicament. Le mode d'action de cet antiparasitaire est peu connu, mais il inclut probablement une inhibition de la synthèse des polyamines, une liaison au petit sillon de l'ADN et une modification du potentiel d'oxydoréduction de la mitochondrie (42). Les modes de résistance observés sont surtout une diminution de l'entrée de la molécule et une

augmentation de son efflux (43). L'accumulation de pentamidine dans la mitochondrie a été décrite chez des cellules résistantes et elle semblait diminuer l'efflux de la molécule (44).

Les antifolates ont été utilisés pour lutter efficacement contre plusieurs maladies, notamment contre les parasites des genres *Toxoplasma* et *Plasmodium.* Cependant, comme *Leishmania* est auxotrophe aux folates, c'est-à-dire qu'il ne les synthétise pas *de novo,* ces médicaments ne sont pas utiles contre ce parasite. Malgré cela, plusieurs groupes ont étudié l'effet des antifolates sur *Leishmania,* ce qui a permis de mieux connaître les enzymes participant à la récupération dans le milieu de ces molécules essentielles à la survie du parasite (45). Le méthotrexate (MTX) est un médicament modèle utilisé contre *Leishmania.* Le principal mode de résistance contre ce médicament est l'amplification génique de la région H, qui contient, spécialement, des transporteurs de molécules et la ptéridine réductase (PTR1). Aussi, des transporteurs comme BT1 (*biopterin transporter 1*), qui sont plus efficaces pour le transport des folates que pour le transport du MTX, peuvent mener à la résistance dans un milieu riche en folates (46).

2.5. Le cycle de vie et l'infection

Le parasite *Leishmania* vit sous trois formes successives qui lui permettent de survivre dans des conditions différentes. Chacune de ces formes a une fonction propre.

- La forme **promastigote procyclique** est adaptée à l'épithélium du tractus digestif de la mouche des sables. Cette forme n'est pas infectieuse, mais elle a la capacité de se diviser.

- La forme **promastigote métacyclique** est infectieuse, mais incapable de se diviser. La métacyclogénèse prépare le parasite à l'invasion de l'hôte vertébré et transforme le promastigote en sa forme métacyclique, qui lui permet de résister aux défenses de l'organisme et d'infecter le macrophage.

- La forme **amastigote** permet au parasite de survivre dans le phagolysosome des cellules infectées, notamment le macrophage. La forme amastigote est infectieuse et peut se diviser.

Le passage entre ces formes requiert la production de nombreuses protéines membranaires ainsi que des changements dans le métabolisme du parasite.

Selon les stades de vie de *Leishmania,* différentes molécules favorisent la survie et l'infectivité du parasite. Les gènes d'intérêt impliqués dans la virulence peuvent être classés en deux catégories : les déterminants invasifs/évasifs et les déterminants de la pathogenèse (6). Les déterminants invasifs/évasifs sont nécessaires à l'infection, mais ils n'ont pas d'impact sur la forme de la maladie. Ils incluent les molécules qui permettent l'adhésion de *Leishmania* au macrophage, l'entrée du parasite dans le macrophage, la survie de *Leishmania* à l'intérieur du macrophage et la différenciation des stades du parasite. De leur côté, les déterminants de la pathogenèse ont un impact direct sur la forme clinique de la maladie et sur sa sévérité. Plusieurs études visent à découvrir quels sont les gènes impliqués dans chacun de ces processus.

2.5.1. L'environnement et la différenciation du parasite

Les deux environnements extrêmes que visite *Leishmania* au cours de son cycle de vie sont le tractus digestif de son vecteur, la mouche des sables, et le phagolysosome de sa cellule hôte, le macrophage. Plusieurs conditions sont modifiées entre ces deux environnements et la survie du parasite nécessite une adaptation rapide à des environnements extrêmes. Le changement drastique de température au moment de la piqûre, passant de la température environnante à celle de l'hôte humain, environ 37 °C, et la variation majeure du pH, qui passe d'un pH neutre à un pH acide autour de 5,5, constituent un signal qui peut être simulé in vitro afin d'induire la différenciation des promastigotes en amastigotes dits « axéniques » (47). La quantité de glucose disponible varie aussi beaucoup entre ces deux environnements, car cette dernière est beaucoup plus élevée dans le tractus digestif du vecteur que dans le phagolysosome du macrophage. Cependant, la quantité d'acides aminés disponibles est plus élevée dans le phagolysosome (48).

Plusieurs changements dans le métabolisme de *Leishmania* au cours de son passage du stade promastigote au stade amastigote ont été décrits dans la littérature (48). Entre autres modifications, il semble y avoir une diminution de la glycolyse, qui serait compensée par une augmentation de plusieurs processus tels la gluconéogenèse, la β-oxydation, le

catabolisme des acides aminés, le cycle des acides tricarboxyliques, la chaîne respiratoire mitochondriale et les capacités de phosphorylation oxydative (48). Ainsi, le parasite utilisera le glycérol et les acides aminés présents dans le milieu pour pallier son manque de glucose.

Les éléments les plus étudiés de la différenciation du parasite sont les molécules à sa surface. De nombreuses protéines de *Leishmania* sont associées à des fonctions d'interaction avec son environnement et peuvent être décrites comme des protéines similaires aux adhésines (*adhesin-like proteins* ou ALP). Ce groupe inclut plusieurs protéines, comme l'amastine, la protéase de surface GP63, l'antigène de surface promastigote PSA2 (aussi connu sous le nom de GP46), les protéines de surface hydrophyles acétylées (*hydrophilic acetylated surface proteins* ou HASP) et les protéophosphoglycanes (PPG) (49).

Le glycolipide le plus abondant à la surface de la cellule est le lipophosphoglycane (LPG) (50). En effet, la glycocalyx de *Leishmania* peut contenir au-delà de 5 millions de molécules de LPG. Cependant, sa quantité et sa composition varient selon les stades de développement du parasite et selon les espèces de *Leishmania.* Ainsi, au cours du passage à la forme métacyclique, les LPG des promastigotes doublent leur longueur en passant de 15 à 30 unités répétitives de sucres. Cette modification aide le parasite à résister aux mécanismes de défense de l'hôte, notamment en bloquant l'action lytique du complément. De plus, dans la mouche des sables, les LPG sont responsables de l'attachement et du détachement du parasite à l'épithélium intestinal du vecteur. Une fois dans le macrophage, la quantité de LPG diminue d'un facteur trois. Pour certaines espèces de *Leishmania,* par exemple *L. major* et *L. donovani,* le LPG est essentiel à l'infection alors que pour d'autres espèces, comme *L. mexicana,* il n'est pas essentiel (51). D'autres phospholipides sont aussi importants pour le parasite, comme les protéophosphoglycanes (PPG) et les glycosylinositolphospholipides (GIPL). Ces derniers sont présents à la fois chez le promastigote et chez l'amastigote (50).

La protéine GP63, aussi appelée « leishmanolysine », est une ectométalloprotéase qui est beaucoup plus abondante chez le promastigote que chez l'amastigote. Elle pourrait jouer un rôle dans la résistance à la lyse humorale, dans l'attachement des parasites au macrophage

ou dans leur phagocytose (52, 53). Cette famille de gènes, présente sous forme de répétitions en tandem, est hautement polymorphique, même à l'intérieur d'une même espèce (6). Récemment, Gomez et ses collaborateurs ont montré que GP63 joue un rôle dans l'activation de la protéine tyrosine phosphatase SHP-1 du macrophage. En effet, la protéine GP63 induit une cascade de signalisation permettant une réduction de l'inflammation et de la réponse microbicide du macrophage (54).

L'amastine est une protéine abondante principalement exprimée chez l'amastigote. Comme GP63, cette protéine est présente en répétitions en tandem dans le génome et elle est aussi fortement polymorphique. Il existe en 3' de l'amastine une région régulatrice non codante conservée qui module son expression (55). Aussi, dans certaines régions du génome, l'amastine est trouvée en tandem avec la tuzine. La fonction de ces deux protéines reste indéterminée jusqu'à présent.

En ce qui concerne les protéines ayant un impact sur la virulence du parasite, la protéine A2 est souvent citée comme étant en partie responsable du tropisme de certaines espèces de *Leishmania* pour la forme viscérale (56, 57). Exprimée chez *L. donovani* et *L. infantum,* A2 n'est pas fonctionnelle chez *L. major.* Des études ont montré que la disparition de cette protéine chez *L. donovani* entraîne une diminution de la survie du parasite dans les viscères (58).

Comme aucune fonction n'est associée aux deux tiers des gènes de *Leishmania,* plusieurs facteurs de virulence ou protéines impliquées dans l'infection n'ont pas encore été identifiés ou décrits. Dans cette optique, plusieurs études ont criblé le génome de *L. major* afin d'y trouver des gènes qui pourraient être associés à des catégories fonctionnelles liées à la virulence ou à l'infectivité. Ces protéines ont une importance vitale pour *Leishmania,* car elles lui permettent de survivre dans son vecteur et dans son hôte. Singh et ses collaborateurs ont criblé, avec des méthodes bio-informatiques, le génome de *L. major* afin d'y trouver des protéines qui possédaient un domaine *adhesin-like* (ALP) (49). Parmi les 194 ALP potentielles, 120 avaient une ou plusieurs régions transmembranaires ou un peptide signal et 56 avaient les deux. Finalement, 6 protéines hypothétiques possédaient les trois éléments, ce qui en faisait des ALP potentielles. Néanmoins, une fonction plus précise n'a pas encore été associée à ces ALP potentielles.

Zhang et ses collaborateurs ont utilisé une approche avec un modèle animal pour tenter d'élucider la fonction de sept protéines trouvées chez *L. infantum,* mais absentes chez *L. major* (59). Ainsi, ils ont introduit ces sept protéines, clonées à partir du génome de *L. donovani,* dans *L. major* afin d'observer si elles causaient un changement de pathologie dans des souris BALB/c, en particulier si l'infection viscérale par ces parasites était augmentée. Parmi les sept gènes testés, seulement un, Li1040, a augmenté la viscéralisation de *L. major* et, étrangement, le gène était présent chez cette espèce. Leur étude suggère donc que la surexpression de Li1040 causerait ce changement de phénotype, c'est-à-dire le passage de la forme cutanée à la forme viscérale. Cette protéine est probablement impliquée dans le transport endosomal des protéines.

2.5.2. Les *Leishmania* et la réponse immunitaire

Afin d'éliminer l'infection causée par n'importe quelle espèce de *Leishmania,* un organisme doit produire une réponse immune de type cellulaire (Th1) (9, 10, 13). En effet, des expériences chez la souris ont montré qu'une réponse cellulaire élimine l'infection à *Leishmania* tandis qu'une réponse humorale (Th2) permet au parasite de survivre (10). La réponse cellulaire entraîne la production d'interleukine 2 (IL-2) et d'interféron γ (IFN-γ). L'IFN-γ active les macrophages qui produiront l'enzyme oxyde nitrique synthétase de type 2 (iNOS2), responsable de la formation d'oxyde nitrique (NO) (60). Cette réponse permet de limiter la croissance intracellulaire du parasite. De son côté, la réponse humorale entraîne la production d'interleukine 4 (IL-4), d'interleukine 5 (IL-5) et d'interleukine 10 (IL-10), qui désactivent les macrophages et entraînent la croissance intracellulaire du parasite (10). Cependant, le contact entre le parasite et le macrophage semble empêcher le macrophage de répondre à l'IFN-γ (61).

Une fois à l'abri dans le phagolysosome de la cellule infectée, le parasite utilise plusieurs stratégies pour réduire la présentation d'antigènes au système immunitaire. Par exemple, *L. donovani* réprime l'expression du complexe majeur d'histocompatibilité de classe 2 (CMH-2) (61). La présentation d'antigènes peut aussi être réprimée en interférant avec le chargement des antigènes dans le CMH-2. L'endocytose du CMH-2 suivie de sa dégradation par des protéases a été observée chez *L. amazonensis* (61).

L'étude du transcriptome de cellules dendritiques et de macrophages infectés par *L. major* ou par *L. donovani* a montré que, même si chacun des parasites générait une surexpression de gènes impliqués dans la réponse inflammatoire, l'infection par *L. major* produisait une réaction inflammatoire plus forte que celle produite par *L. donovani.* Cette réponse inflammatoire accrue pourrait être en partie responsable du confinement de *L. major* au site de l'infection (62).

Plusieurs molécules de *Leishmania,* impliquées dans la virulence de l'infection, pourraient avoir un impact sur la réponse immunitaire de l'hôte. Il s'agit notamment de glycoprotéines, comme les lipophosphoglycanes (LPG), ou de protéases, comme la leishmanolysine (GP63) (63). Il a été montré que les LPG de *L. donovani* bloquent la migration transendothéliale des monocytes et qu'ils réduisent l'expression endothéliale des protéines E-sélectine, ICAM-1 et VCAM-1 sans modifier la structure de la monocouche endothéliale (64). D'un autre point de vue, les différences entre les LPG des espèces de *Leishmania* pourraient jouer un rôle dans la compétence d'une espèce à être transmise par différentes espèces de mouches des sables (65).

Plusieurs chimiokines, de petites protéines solubles responsables du chimiotactisme des cellules immunitaires, sont impliquées dans la réponse à l'infection. Cependant, ces chimiokines dépendent de l'espèce de *Leishmania* infectante (66). Dans le cas de l'infection chronique de la souris par *L. donovani,* les cellules dendritiques murines deviennent déficientes quant à leur migration et à leur réponse aux chimiokines CCL21 et CCL19. Cette déficience serait causée par l'inhibition du récepteur de chimiokines 7 (CCR7) par l'interleukine 10 et par le facteur de nécrose tumorale alpha (*tumor necrosis factor alpha* ou TNF-α). Cette interaction pourrait jouer un rôle important dans la pathogenèse de la leishmaniose viscérale (67).

Dans le cas de la leishmaniose cutanée murine causée par *L. major,* les cytokines TNF-α, IL-1B et MIP-1α (ou CCL3) régulent le transport de *L. major* du site de l'infection jusqu'aux ganglions lymphatiques (68). Comme les cellules polymorphonucléaires (PMN) sont les premières à se rendre au site de l'infection, elles sont les premières infectées par *Leishmania.* Il a été montré que l'infection des PMN par *L. major* augmente la longévité de

ces PMN et induit une surproduction de MIP-1β, un agent chimioattracteur qui attire les macrophages au site de l'infection. Les PMN apoptotiques sont ensuite phagocytées par les macrophages, qui seront à leur tour infectés par *L. major* (69).

Les protéines CCL2 et MIP-1α, des agents chimioattracteurs, seraient responsables de la chimioattraction des macrophages dans les lésions cutanées (66). Dans des cellules dendritiques exposées à *L. major,* les récepteurs de chimiokines CCR2 et CCR5 sont réprimés tandis que le récepteur CCR7 est surexprimé. Cela diminue la chimiotaxie des cellules dendritiques à CCL2 et MIP-1α tout en augmentant la sensibilité à CCL21. Cette modulation est similaire chez des souris sensibles à *L. major* et chez des souris résistantes. De plus, il a été observé que la chimiokine CXCL10 était exprimée seulement chez les souris résistantes à *L. major* (70).

Dans le cas de la leishmaniose cutanée causée par *L. mexicana,* une forte expression de CCL2 et une expression modérée MIP-1α causent une infection localisée qui sera contenue par l'hôte tandis qu'une faible expression de CCL2 couplée à une haute expression de MIP-1α cause une dissémination de l'infection qui guérira difficilement (71).

Même si plusieurs groupes de recherche ont étudié la réaction immunitaire et les chimiokines impliquées dans l'infection de l'humain par *Leishmania,* dresser un juste portrait de ces interactions est difficile. En effet, peu d'études comparent la réaction de l'hôte selon l'espèce de *Leishmania* infectante (61). L'expression différentielle des récepteurs de chimiokines et d'autres molécules de surface jouant un rôle dans la migration des leucocytes pourrait avoir une influence importante sur la sévérité et sur la forme de leishmaniose causée par chaque espèce, notamment dans le cas de la leishmaniose viscérale.

2.6. Le génome de *Leishmania*

Les génomes de trois espèces de *Leishmania,* soit *L. major, L. infantum* et *L. braziliensis,* ont été séquencés par le Wellcome Trust Sanger Institute et des versions de ces génomes sont disponibles dans les bases de données GeneDB (72) et TriTrypDB (73). Le génome de la souche Friedlin de *L. major,* qui mesure au total 32,8 Mb, est réparti sur

36 chromosomes ayant chacun une longueur de 0,28 Mb à 2,8 Mb. Ce génome a été séquencé par le mélange de plusieurs approches, comme le séquençage complet de cosmides, le séquençage *shotgun* de chromosomes et le séquençage de chromosomes artificiels bactériens (*bacterial artificial chromosome* ou BAC) (74). Au moment d'être publié, le séquençage du génome de *L. major* était presque complet, puisqu'il ne restait que deux brèches à combler.

Les génomes des espèces *L. infantum* (clone JPCM5) et *L. braziliensis* (clone M2904) ont été séquencés par *shotgun* pour une couverture de 5 fois la longueur totale de leur génome. La séquence du génome de *L. infantum* comprend 470 contigs dont le N50, c'est-à-dire la longueur du contig pour lequel la somme des contigs plus grands correspond à la moitié de la longueur du génome, est 150 519 bases, alors que la séquence de *L. braziliensis* comprend 1031 contigs dont le N50 est 57 789 bases (75). Le génome de *L. infantum* mesure approximativement 33,5 Mb et est réparti sur 36 chromosomes. Le génome de *L. braziliensis* a approximativement la même taille que celui de *L. infantum,* mais il est réparti sur 35 chromosomes. Initialement, les chercheurs ont estimé que le génome de *L. major* contient 8370 gènes, que celui de *L. infantum* en possède 8195 et que celui de *L. braziliensis* en comprend 8312. Une première version du génome de *L. mexicana* est aussi disponible sur TriTrypDB, mais en version préliminaire.

2.6.1. La génomique comparative

L'obtention de la séquence des génomes des trypanosomatides (ou TriTryp), soit *Trypanosoma cruzi, T. brucei* et *L. major,* a permis une première incursion dans l'étude du génome de ces organismes apparentés (74, 76-78), mais ayant quand même une distance évolutive de 200 à 500 millions d'années (78). Ensuite, le séquençage de *L. infantum* et de *L. braziliensis* a permis d'évaluer avec plus de précision l'étendue des différences entre différentes espèces de *Leishmania* (75). De plus, plusieurs autres parasites protozoaires ont été séquencés, comme *Plasmodium falciparum* (75), responsable de la malaria, *Entamoeba histolytica* (79), qui cause une dysenterie, et *Dictyostelium discoideum* (80), une amibe vivant dans le sol. Enfin, plusieurs autres espèces de *Trypanosoma,* de *Leishmania,* de *Plasmodium* et d'*Entamoeba* sont à différents stades de séquençage.

2.6.1.1. La comparaison de *Leishmania* et *Trypanosoma*

La comparaison des trois premiers trypanosomatides séquencés a révélé une similitude importante entre ces organismes. Cependant, une différence majeure entre ces trois espèces est la grande taille du génome de *T. cruzi,* 55 Mb et environ 12 000 gènes, comparativement à ceux de *T. brucei,* 25 Mb et 9068 gènes, et de *L. major,* 33 Mb et 8311 gènes. Lorsque les familles de gènes (*clusters of orthologous groups* ou COG) de chacune des espèces ont été comparées entre elles, 6158 familles de gènes étaient communes entre les espèces (23). Cela correspond à environ 80 % des gènes de *T. brucei* et à 93 % des gènes de *L. major* (78). *T. cruzi* a 3736 gènes spécifiques, *T. brucei* en a 1392 et *L. major* en a seulement 910. La plupart des gènes divergents entre les espèces n'ont pas de fonction connue. Chanda et ses collaborateurs ont observé que les gènes fortement exprimés chez les trypanosomes avaient subi moins de changements dans leurs séquences protéiques que les gènes moins exprimés, en particulier chez *L. major* (81).

La synténie entre les trois espèces, c'est-à-dire la similarité entre la structure de leurs génomes et l'ordre de leurs gènes, est très élevée : 68 % et 75 % des gènes de *T. brucei* et de *L. major,* respectivement, sont dans le même ordre sur les chromosomes. Le taux de GC des trois espèces est assez différent, soit 60 % pour *L. major,* 46 % pour *T. brucei* et 51 % pour *T. cruzi.* Lorsqu'additionnés, les blocs de synténie de *T. brucei* et de *L. major* font 19,9 et 30,9 Mb, respectivement. Environ 40 % des bris de synténie sont causés par l'expansion de familles de gènes ou par des répétitions en tandem contenant des rétroéléments ou des ARN structuraux (23). De plus, 43 % des bris de synténie se produisent près des régions de changement d'orientation des groupes de gènes.

Quelques éléments distinguent les TriTryp des autres groupes d'organismes. Par exemple, le génome des TriTryp code seulement pour 45 des 61 ARNt existants. Les TriTryp ont significativement moins de facteurs de transcription que les autres organismes. Cependant, ils ont plus de domaines *CCCH-type zinc finger,* ce qui suggère un contrôle post-transcriptionnel accru. Une composante nécessaire à l'interférence via l'ARN (ARNi) est présente chez *T. brucei,* mais absente des deux autres espèces. Aucun homologue de DICER, composant essentiel du processus de ARNi, n'est présent chez les TriTryp, mais

un homologue du domaine de ribonucléase III pourrait en compenser l'absence chez *T. brucei.*

Ce sont surtout les protéines de surface qui différencient les espèces. En effet, le nombre de gènes élevés retrouvé chez *T. cruzi* correspond en majeure partie à des protéines de surface répétées en tandem, comme les protéines de surface associées à la mucine (MASPS, > 1300 copies), les trans-sialidases (1430 copies), les mucines (863 copies) et les protéases GP63 (565 copies). Le génome de *L. major* code lui aussi pour de nombreuses protéines de surface, quoiqu'elles soient présentes en moins de copies. On peut citer, par exemple, les amastines, la tuzine, les lipophosphoglycanes (LPG), les glycosylinositolphospholipides (GIPL), les protéophosphoglycanes (PPG) et les protéines ancrées à la membrane par une ancre GPI, comme GP63, PSA-2 et GP46 (74). La métalloprotéase GP63 est la protéine la plus abondante à la surface du parasite chez l'insecte (52).

Les régions subtélomériques favorisent une certaine diversité génomique chez les trypanosomes, mais pas chez *Leishmania.* Chez *Trypanosoma,* les glycoprotéines de surface à séquences variables (*variable surface glycoproteins* ou VSG) sont situées dans les régions télomériques ou subtélomériques. Ces VSG constituent l'interface entre le parasite et le système immunitaire de l'hôte. Ils le protègent contre la lyse médiée par le complément (82, 83). Le génome des *Trypanosoma* contient des centaines, voir des milliers, de VSG différentes. Selon le stade de développement du parasite, différents groupes de VSG peuvent être exprimés. Par contre, seulement une VSG à la fois sera réellement utilisée par le parasite. Les VSG sont sujets à des recombinaisons médiées ou non par la protéine RAD51, qui est aussi impliquée dans la recombinaison homologue et dans la réparation de l'ADN chez *E. coli* (84). Dans le cas de *Leishmania,* il semblerait que des gènes essentiels se retrouvent dans les régions subtélomériques et qu'ils ne soient pas sujets à de fortes recombinaisons (75). Étant donné la forte conservation de synténie entre les TriTryp, il est fort probable qu'une certaine forme de sélection maintienne la structure du génome de *Leishmania* constante dans le temps. L'organisation polycistronique (voir section 2.6.2) de ces génomes pourrait être favorisée par certains éléments de régulation encore inconnus.

2.6.1.2. La comparaison des espèces de *Leishmania*

Jusqu'ici, seulement trois espèces de *Leishmania* ont été séquencées et comparées sur la base d'un génome complet : *L. major, L. infantum* et *L. braziliensis* (75). En tout, moins de 200 gènes spécifiques à une espèce ont été trouvés chez les trois premières *Leishmania* séquencées. De plus, la synténie est conservée sur plus de 99 % des gènes des trois génomes. Ainsi, les protéines potentielles de *L. major* et de *L. infantum* ont en moyenne 92 % d'identité et la séquence nucléotidique des gènes a environ 94 % d'identité. Cependant, la comparaison de l'une ou l'autre de ces deux espèces avec *L. braziliensis* donne une identité protéique moyenne de 77 % et une identité nucléotidique qui varie de 81 à 82 %. Contrairement aux deux autres espèces de *Leishmania* qui possèdent 36 chromosomes, *L. braziliensis* en possède 35. En effet, le chromosome 20 de *L. braziliensis* constitue une fusion des chromosomes 20 et 34 trouvés chez les autres espèces. Le génome de *Leishmania* est composé de répétitions sur environ 10 % de sa séquence.

Peu de gènes ont été décrits comme étant uniques à une espèce ou à une autre. Selon Peacock et ses collaborateurs, seulement 5 gènes seraient spécifiques à *L. major,* 26 à *L. infantum* et 47 à *L. braziliensis*. De plus, environ 47 gènes sont présents chez *L. infantum* et *L. major,* mais absents chez *L. braziliensis*. Dans environ 80 % des cas, les gènes spécifiques à une espèce sont présents chez les deux autres, mais sous une forme dégénérée. Une autre portion des gènes différents résultent d'une amplification génique suivie d'une diversification. Seulement 34 % des gènes qui ne sont pas partagés par toutes les espèces ont pu être associés à une fonction. D'un autre côté, environ 8 % des gènes semblent avoir évolué à une vitesse différente entre les espèces, ce qui suggère une sélection positive de ces gènes.

À l'instar de *T. brucei, L. braziliensis* est la seule *Leishmania* chez qui la machinerie nécessaire au processus d'ARN interférence a été trouvé (75). En effet, *L. braziliensis* semble contenir un homologue de DICER (LbrM23_V2.0390) ainsi qu'un homologue d'Argonaute (AGO1). Il semble y avoir des vestiges d'AGO1 sur les chromosomes 11 de *L. major* et de *L. infantum,* mais il est probable que le gène ait été perdu.

Selon Peacock et ses collaborateurs, les génomes de *L. major* et de *L. infantum* contiennent peu d'éléments transposables, ce qui pourrait favoriser la stabilité chromosomique (75). Chez les autres trypanosomes, comme *T. brucei* et *T. cruzi,* on trouve plusieurs types de transposons, dont les rétroposons non-LTR ingi/L1Tc et SLACS/CZAR, et les rétroposons LTR VIPER. Contrairement à *L. infantum* et à *L. major,* des SLACS/CZAR sont associés aux mini-exons (voir section 2.6.2) de *L. braziliensis* et des éléments transposables associés aux télomères (*telomere associated transposable elements* ou TATE) ont été découverts près de télomères de cette espèce.

Leishmania possède aussi un autre type de rétroposon, le SIDER (*short interspersed degenerated retroposons*), qui peut être séparé en deux groupes : le SIDER1 et le SIDER2 (85). Les deux types de SIDER sont distribués de manière uniforme et synténique dans le génome de *Leishmania.* Cependant, l'association de SIDER à certaines catégories de gènes n'est similaire dans les trois espèces que dans 47 % des gènes orthologues. Bringaud et ses collaborateurs ont montré que le SIDER2 trouvé dans la région non traduite en 3' des gènes favorise leur dégradation et diminue leur niveau d'expression (86).

2.6.1.3. La comparaison entre les espèces de *Plasmodium*

La séquence complète de plusieurs parasites du genre *Plasmodium* a été déterminée par séquençage en aveugle. Entre autres parasites séquencés et comparés entre eux, on trouve *P. falciparum,* responsable de la malaria humaine, *P. yoelii yoelii,* responsable de la malaria murine, *P. vivax,* responsable de 25 à 40 % des cas de malaria humaine, et *P. knowlesi,* qui cause la malaria chez le macaque et chez l'homme. Le génome de ces espèces est réparti sur 14 chromosomes haploïdes, pour une taille de génome qui varie de 23 à 26 Mb.

Outre le fait que le génome de *Leishmania* soit diploïde et que celui de *Plasmodium* soit haploïde, plusieurs éléments distinguent ces deux parasites. En effet, *P. falciparum* est composé de 54 % de gènes qui contiennent des introns, alors que *Leishmania* n'en contient aucun. Aussi, aucun élément transposable ou rétrotransposon n'a été détecté chez *Plasmodium* (87). De plus, davantage de différences sont observées entre les espèces de *Plasmodium* qu'entre les espèces de *Leishmania.* Ainsi, environ 80 % des gènes de *P. knowlesi* ont un gène orthologue chez *P. falciparum* et *P. vivax,* ce qui est très différent

de ce qui a été observé chez les *Leishmania,* qui partagent environ 97,5 % de leurs gènes. Une partie de ces différences entre les espèces consiste en l'amplification de familles de gènes. Plus précisément, il existe chez *P. knowlesi* 5 familles de gènes uniques et distinctes dont les fonctions sont inconnues et qui contiennent de 4 à 15 paralogues.

2.6.1.4. La biologie du développement et la génomique de *Leishmania*

Le but de la biologie du développement consiste à déterminer la relation entre l'évolution sur le plan génomique et le phénotype observé au cours du développement d'un organisme. Les chercheurs de ce domaine s'intéressent particulièrement aux eucaryotes pluricellulaires, comme le ver de terre, la mouche, la souris ou l'humain. Depuis plusieurs années, le concept du gène égoïste ― dont le seul but serait de se répliquer pour survivre ― est de moins en moins prévalent dans ce domaine et de nouvelles théories sur les différences entre les espèces sont élaborées à la lumière de l'étude des génomes. Chez les eucaryotes supérieurs, la détermination des phénotypes découlerait surtout des réseaux formés par les gènes et les protéines plutôt que par les gènes seulement (88). Par exemple, la disparition d'un gène important pourrait modifier le réseau qui s'adapterait à cette situation en formant dans la population des sous-réseaux différents. Ainsi, on ne devrait pas comparer les organismes en s'appuyant seulement sur les différences nucléotidiques, qui se basent sur un changement graduel de séquence au cours du temps. En effet, on devrait aussi tenir compte du nombre de copies de gènes, du transfert horizontal de gènes et de la réorganisation des chromosomes.

La comparaison du génome de 12 espèces de drosophiles a montré que seulement 44 gènes étaient spécifiques à une espèce (89). Comme il y a peu de différences géniques entre ces drosophiles, les chercheurs étudient maintenant les différences de régulation entre les espèces : l'évolution des réseaux géniques devrait avoir plus d'impact que les gènes individuels (88). L'étude des différences entre les gènes d'une même famille entre deux espèces pourrait aussi procurer des informations importantes. Cela est résumé par la théorie de l'évolution morphologique, qui suggère que les organismes évoluent surtout par altération de l'expression de gènes dont la fonction est conservée. Cette altération est causée en grande partie par des modifications dans les régions régulant ces gènes, ce qui modifie les réseaux géniques (90).

Cette vision de l'évolution pourrait expliquer le peu de différences de gènes entre les espèces de *Leishmania,* et même entre les trypanosomes. Même si elle est capable d'échanger du matériel génique lorsque deux souches co-infectent la mouche des sables (91), *Leishmania* ne sera probablement pas exposée à beaucoup de gènes exogènes. Pour évoluer, le parasite devrait donc se contenter de modifier et de réorganiser le matériel génétique qu'il possède déjà. Comme *Leishmania* a tendance à amplifier les gènes qui lui permettent de mieux s'adapter à son environnement, la diversification de ces amplifications géniques pourrait être une approche utilisée par le parasite pour varier ses comportements.

2.6.2. La régulation de la transcription

La structure du génome de *Leishmania* est singulière, car les gènes sont organisés en groupes orientés dans la même direction (74, 77, 92, 93). Par exemple, le génome de *L. major* est organisé en 133 groupes de gènes qui ne semblent pas être réunis selon leurs fonctions (74). Ces groupes peuvent atteindre une longueur de 1259 kb. Ils sont séparés par des régions de transition de longueur variant de 0,9 kb à 1,4 kb et dont la composition en bases n'est pas traditionnelle. Il a été suggéré que ces groupes de gènes étaient transcrits par une ARN polymérase de type II (ARNPII) qui débuterait la transcription dans la région de transition et qui produirait des transcrits polycistroniques. L'analyse du génome de *L. major* a permis d'observer que certains groupes de gènes étaient entrecoupés par des ARN non codants qui seraient transcrits par une ARN polymérase de type III (ARNPIII). Cela suggère que ces groupes pourraient, en fait, en constituer plusieurs (74).

La transcription nucléaire des gènes des trypanosomes est faite par trois types d'ARN polymérase. L'étude du chromosome 3 de *L. major* avec l'alpha-amantadine, qui inhibe la ARNPII, et avec la tagétitoxine, qui inhibe sélectivement la ARNPIII, a permis de démontrer que la ARNPII était responsable de la transcription des gènes, alors que la ARNPIII était responsable de la transcription des ARNt chez *L. major* (94).

L'ARN polymérase I (ARNPI) transcrit les ARN ribosomaux (ARNr) et, chez *Trypanosoma brucei,* transcrit en plus les glycoprotéines de surface à séquences variables (VSG) responsables des variations antigéniques de ce parasite, de même que la EP-procycline. L'ARN polymérase II transcrit les ARN messagers et le *splice-leader SL*

RNA, aussi appelé « mini-exon ». L'ARN polymérase III a trois classes de promoteurs, dont une est absente chez les trypanosomatides (95). La classe 1 régule l'expression de l'ARNr 5S alors que la classe 2 régule l'expression des ARNt. Malgré l'absence de promoteurs de classe 3 chez les trypanosomes, les *small nuclear RNA* (snRNA) sont transcrits par le biais de promoteurs de classe 2.

Jusqu'à présent, aucune méthode de régulation de l'expression des gènes par régulation transcriptionnelle impliquant l'ARN polymérase II n'a été démontrée chez *Leishmania.* Plusieurs études suggèrent que la transcription spécifique débute dans la région de changement de brin entre deux unités polycistroniques convergentes (94, 96, 97). Cependant, l'ARN polymérase II des trypanosomatides peut transcrire de l'ADN exogène au parasite, comme cela a été montré par la transcription de gènes rapporteurs insérés dans des cellules de *Leishmania* ou de *Trypanosoma* à l'aide de plasmides (98).

L'organisation des gènes sur les chromosomes de *Leishmania* suggère que les gènes sont transcrits en longues unités polycistroniques sans promoteurs spécifiques aux gènes. Une fois transcrits, ces ARN messagers doivent être épissés en *trans* par l'ajout en 5' d'un mini-exon. Contrairement aux gènes transcrits par la ARNPII, le mini-exon a un promoteur particulier. De plus, le nombre de gènes codant pour le mini-exon varie de 100 à 200 copies successives par locus. Ces gènes représentent environ 6 % de la transcription cellulaire totale (95). Le mini-exon subira plusieurs modifications post-transcriptionnelles, incluant plusieurs types de méthylation et de pseudo-uridinations. Ainsi, il deviendra plus stable et davantage apte à lier différentes protéines qui participent à la traduction des ARNm, comme eIF4E. Le mini-exon ainsi modifié joue le rôle de coiffe. Il est possible que différentes structures de coiffes puissent être liées par des protéines différentes et, de ce fait, puissent avoir un profil de traduction différent (99). Les ARNm doivent aussi être polyadénylés en 3'.

Les chercheurs postulent que l'abondance des transcrits serait surtout régie par des éléments en *cis* qui affecteraient la stabilité et la modification des ARNm (100). De plus, une séquence conservée de 450 nucléotides, située en 3' des amastines de *Leishmania,* stimulerait la traduction de ces protéines de surface en réponse au choc thermique induit par l'hôte humain (101). Cette séquence appartient à une famille de rétroposons, les SIDER.

Des chercheurs ont aussi suggéré que la distribution de codons plus rares dans les gènes de *Leishmania* pourrait contribuer à réduire la traduction de certains gènes (102). Une réorganisation de la chromatine pourrait également jouer un rôle dans la sélection des régions du génome à exprimer (103). Des études d'immunoprécipitation de la chromatine sur biopuce ont d'ailleurs montré que l'acétylation de la chromatine était plus importante à certaines positions sur le génome, comme dans les régions de changement de brin. Thomas et ses collaborateurs suggèrent que ces régions servent à l'initiation de la transcription (97).

Les gènes de trypanosomatides nécessaires en quantité abondante sont souvent présents sous forme de répétitions en tandem, comme c'est le cas pour plusieurs protéines de surface (94).

2.6.3. L'étude du transcriptome de *Leishmania*

Plusieurs études décrivent l'utilisation de biopuces pour étudier la modulation des gènes de *Leishmania* selon divers facteurs.

Saxena et ses collaborateurs ont comparé l'expression de promastigotes procycliques à celle de promastigotes métacycliques chez *L. major* Freudlin en utilisant comme sondes 10 464 fragments PCR générés à partir d'une librairie de clones (104). Ils ont observé qu'environ 15 % des gènes de *L. major* étaient modulés au cours de la métacyclogénèse. La même biopuce a été utilisée pour étudier l'expression des gènes de *L. donovani* pendant la différenciation des promastigotes en amastigotes (105). Ils ont observé qu'environ 9 % des gènes de *L. donovani* étaient modulés au moment de sa différenciation. Parmi ces gènes, plusieurs étaient des protéines de choc thermique, des ubiquitines hydrolases, des protéines de liaison à l'ARN et des protéines kinases.

Salotra et ses collaborateurs ont comparé le profil d'expression de parasites *L. donovani* isolés de patients ayant des lésions de type PKDL (*post-kala azar dermal leishmaniasis*) avec ceux de patients ayant une leishmaniose viscérale. Pour ce faire, ils ont utilisé une biopuce comprenant 2268 produits PCR amplifiés provenant d'une banque d'ADN génomique de *L. donovani* (106). À terme, ils ont observé une différence d'expression dans seulement 2 % de leurs sondes, les plus modulées étant GP46, GP63 et de présumées amastines.

Holzer et ses collaborateurs ont comparé l'expression de promastigotes procycliques, d'amastigotes issus de lésions et d'amastigotes axéniques de *L. mexicana* sur une biopuce ciblant 8160 gènes. Chaque gène est ciblé par 11 oligonucléotides (60-mer) créés à partir du génome séquencé de *L. major* (107). Les chercheurs ont observé environ 3,5 % de gènes modulés entre les deux conditions. De plus, ils ont suggéré qu'une même biopuce pouvait être utilisée pour comparer différentes espèces de *Leishmania.*

Guimond et ses collaborateurs ont comparé l'expression des gènes de différentes souches de *L. tarentolae* et de *L. major* ayant différents profils de résistance aux antiparasitaires. Ils ont utilisé une biopuce contenant 44 produits PCR ciblant des gènes jouant un rôle dans la résistance aux antiparasitaires (108). Cette étude a permis d'identifier des gènes impliqués dans la résistance aux antiparasitaires chez plusieurs souches de *Leishmania.*

Leifso et ses collaborateurs ont comparé l'expression des gènes de *L. major* aux stades promastigote et amastigote en utilisant une biopuce ciblant 8160 gènes de *L. major.* Il y avait 11 sondes de 24 nucléotides par gène, chaque sonde étant associée à une sonde polymorphique (109). Les chercheurs ont conclu que 94,0 % des gènes de *L. major* étaient exprimés constitutivement entre les deux conditions. Selon leur analyse, seulement 2,9 % des gènes étaient exprimés de façon différente.

McNicoll et ses collaborateurs ont comparé les changements dans le transcriptome et le protéome de *L. infantum* au cours de son passage du stade promastigote au stade amastigote (110). L'étude transcriptomique utilisait une biopuce constituée de sondes de 70 nucléotides ciblant les protéines détectées au cours d'études protéomiques réalisées antérieurement. Les chercheurs ont observé une corrélation modérée entre le transcriptome et le protéome des amastigotes, alors qu'une corrélation faible a été observée chez les promastigotes.

Alcolea et ses collaborateurs ont comparé le profil d'expression de promastigotes procycliques et métacycliques de *L. infantum* issus d'une même culture et séparés en utilisant une méthode d'agglutination avec des lectines d'arachides (111). La biopuce utilisée dans cette expérience a été construite à partir d'une librairie dérivée de *L. infantum* MCAN/ES/98/10445. Un total de 29 952 produits PCR ont été déposés sur une biopuce.

Parmi ces produits PCR, 317 sont considérés comme statistiquement significatifs. De ceux-ci, 197 contenaient des portions de gènes, alors que 120 ne contenaient que des régions intergéniques. Les promastigotes métacycliques surexprimaient plusieurs gènes liés à la virulence du parasite, comme la cystéine peptidase A et plusieurs gènes ayant trait à la biosynthèse des lipophosphoglycanes et des glycoprotéines.

Depledge et ses collaborateurs ont comparé l'expression entre les stades promastigote procyclique, promastigote métacyclique et amastigote de *L. braziliensis* pour 785 gènes afin d'évaluer quels gènes étaient modulés pendant la différenciation de cette espèce (112). En général, seulement 9 % des gènes étudiés étaient modifiés entre les stades, ce qui est similaire aux observations faites sur d'autres espèces. Près de la moitié de ces gènes étaient surexprimés dans le stade promastigote métacyclique. Les auteurs ont aussi comparé l'expression des gènes au stade amastigote entre *L. braziliensis, L. infantum* et *L. major*. Ils observent que la majorité des gènes distribués différentiellement entre les espèces sont habituellement exprimés de façon constitutive. Ils suggèrent aussi que la plupart des gènes subissant une régulation de l'expression ne sont pas les mêmes entre les espèces. Ils arrivent à cette conclusion en appuyant leurs résultats par d'autres études publiées.

La comparaison de ces différentes études suggère que peu de gènes de *Leishmania* sont exprimés différemment entre les stades promastigote et amastigote (113). Cependant, comme les biopuces utilisées varient d'une étude à l'autre quant à leur couverture du génome de *Leishmania,* quant aux sondes utilisées et quant à l'analyse statistique effectuée, il est difficile de donner raison à une étude plutôt qu'à une autre. Les conditions de culture des parasites ainsi que les souches utilisées pourraient aussi influencer les résultats.

3. La génomique et la bio-informatique

L'étude du parasite *Leishmania* peut être faite de plusieurs points de vue, mais cette thèse concerne surtout les aspects génomiques de l'étude du parasite. Ainsi, cette section a pour objectif de familiariser le lecteur avec différents aspects du séquençage de génomes et de l'étude de l'expression génique. De même, le séquençage de génomes tel que décrit dans cette section génère la matière première qui sera nécessaire à la conception de tests de diagnostic moléculaire, et les biopuces peuvent être utilisées comme outils diagnostics.

3.1. Le séquençage de génomes

La génomique est née avec l'apparition des techniques de séquençage de l'ARN, puis de l'ADN. En 1965, Holley et ses collaborateurs ont séquencé les deux premiers acides nucléiques de l'histoire, l'ARNt de l'alanine d'*Escherichia coli* (114), puis celui de la levure (115). C'est grâce à la capacité de purifier des ARNt particuliers et à la connaissance de RNAses dont la spécificité était connue que ces premiers séquençages ont pu avoir lieu. De plus, il a été possible de déterminer la structure secondaire de l'ARNt, puisque l'hybridation entre les bases était connue à l'époque. C'est en 1971 que la première molécule d'ADN a été séquencée. Cette molécule consistait en une séquence de 12 nucléotides, soit la séquence des extrémités cohésives du phage lambda (116, 117). Ces premières séquences ont été obtenues à l'aide de réactions chimiques spécifiques, comme la dépurination. Ces méthodes permettaient d'obtenir des séquences longues de 10 à 20 nucléotides.

En 1975, Sanger et Coulson ont introduit la méthode de terminaison des chaînes pour le séquençage de l'ADN (118). En 1977, Maxam et Gilbert ont conçu une méthode similaire à celle de Sanger, mais ils utilisaient plutôt des nucléotides qui ne permettaient pas l'élongation des chaînes (119). La même année, Sanger a introduit la méthode des didéoxynucléotides, méthode qui permettait de séquencer jusqu'à 100 nucléotides. Cette technique a permis le séquençage du génome du phage PhiX, aussi publié en 1977 (120).

La grande innovation suivante dans l'histoire du séquençage a été l'automatisation des protocoles et de l'analyse (121). Cette avancée importante a permis de démocratiser le

séquençage, jusqu'à permettre le séquençage de génomes complets, dont le génome humain en février 2001 (122, 123) et le génome des TriTryp en 2005 (74, 76-78).

3.1.1. Le séquençage en aveugle

Le séquençage du premier génome bactérien, *Haemophilus influenzae,* a introduit la méthode du séquençage en aveugle (en anglais, *whole-genome shotgun*) (124). La méthode avait été proposée en 1982 par Sanger et ses collaborateurs et avait été utilisée, à plus petite échelle, pour le séquençage du génome du phage lambda (125). Cependant, les chercheurs avaient utilisé une méthode de positionnement des fragments sur le génome (BacMap) plutôt qu'un assemblage bio-informatique, comme l'ont fait Fleischmann et ses collaborateurs. La méthode du séquençage en aveugle est encore utilisée et consiste en les étapes suivantes :

1. Fragmentation de l'ADN génomique de l'organisme à séquencer;
2. Clonage de ces fragments au hasard dans un vecteur, par exemple un plasmide ou un BAC. Le vecteur est ensuite introduit dans *E. coli* afin d'obtenir une librairie de séquences;
3. Séquençage du fragment introduit dans chaque clone avec la méthode Sanger, souvent en séquençant les deux extrémités du clone, ce qui produit des séquences pairées;
4. Assemblage des séquences obtenues, par des méthodes bio-informatiques.

L'utilisation de protocoles de séquençage apparié a facilité le séquençage de génomes cellulaires, dont ceux de l'humain (122, 123) et de la souris (126). Ce type de protocole permet d'ordonner les contigs et de les organiser sous forme d'échafaudages (en anglais, *scaffolds*). Lorsque les régions flanquant les régions à séquencer sont connues, il est possible de cibler la région à séquencer par PCR, ce qui permet de finaliser l'analyse du génome.

3.1.2. Les méthodes à haut débit

Le principal avantage de ces méthodes à haut débit concerne le fait qu'elles sont hautement parallélisées, ce qui permet d'obtenir des millions de séquences différentes en une seule expérience, sans clonage de fragments génomiques dans un vecteur. La longueur des fragments séquencés, aussi appelés « reads », est par contre limitée comparativement aux méthodes de séquençage classiques, qui peuvent séquencer jusqu'à 1000 nucléotides contigus. Cette limite augmente avec l'affinement des méthodes de séquençage à haut débit, mais elle demeure une limitation des nouvelles méthodes. De plus, l'assemblage de ces données reste l'élément limitant, puisqu'il nécessite une capacité informatique beaucoup plus grande.

Malgré leurs différences méthodologiques, les techniques de séquençage à haut débit actuellement offertes suivent un protocole dont les grandes lignes sont similaires (127). Même si les méthodes utilisées à chacune des étapes diffèrent, l'ordre dans lequel elles sont effectuées est invariable :

1. Préparation d'une librairie de séquences marquées par des adaptateurs;
2. Amplification des séquences marquées de manière à ce qu'elles soient séparées spatialement;
3. Séquençage par réactions enzymatiques cycliques mesurées en temps réel.

Il existe présentement cinq technologies majeures de séquençage à haut débit. Les particularités de chacune d'elles sont décrites ci-dessous.

3.1.2.1. La technologie 454

La technologie 454 a été mise sur le marché par l'entreprise 454 Life Sciences en 2005, une filiale de Roche Diagnostics (128). Cette méthode consiste en plusieurs étapes. Après préparation de la librairie de fragments à séquencer, les fragments sont mis en contact avec des billes couvertes de sondes oligonucléotidiques complémentaires aux adaptateurs. L'étape suivante est la PCR en émulsion (emPCR), qui consiste à ségréger les billes dans des bulles qui servent de microréacteurs. Ainsi, chaque bille sera couverte par une

amplification clonale d'un seul fragment à séquencer. Après l'emPCR, les billes sont immobilisées dans des puits qui n'accommodent qu'une seule bille et qui permettent de ségréger spatialement le signal de pyroséquençage de chaque fragment. Ainsi, sur chaque bille, la séquence complémentaire du fragment amplifié sur cette bille est synthétisée à partir de nucléotides qui sont irrigués sur la surface dans un ordre prédéterminé. Chaque nucléotide ajouté à une séquence libère un pyrophosphate, qui est ensuite dégradé par une ATP sulfurylase et une luciférase. Cette réaction enzymatique produit de la lumière qui est détectée par une caméra CCD (en anglais, *charge coupled device*). La détection du signal lumineux peut ensuite être associée à l'ordre d'introduction des nucléotides, ce qui permet de déterminer la séquence de chaque fragment immobilisé sur une bille.

Dans le cas où plusieurs nucléotides identiques se suivraient, formant des homopolymères, la quantité de nucléotides à inclure dans la séquence est inférée à partir de l'intensité du signal, ce qui est une source d'erreur fréquente de la technologie 454. Une autre erreur possible est la présence résiduelle de nucléotides irrigués au cours de cycles précédents, qui est interprétée comme étant le nucléotide courant. Cela entraîne une erreur appelée « carry-forward ». Ces erreurs limitent l'utilité de cette méthode pour le séquençage de nouveaux génomes et la détection de mutations.

En général, la technologie 454 produit des fragments séquencés plus longs que les autres méthodes et elle peut produire jusqu'à 400 millions de nucléotides par expérience. Le protocole peut être facilement adapté à d'autres applications.

3.1.2.2. L'Illumina Genome Analyzer (Solexa)

Le séquenceur Illumina Genome Analyzer est aussi connu sous le nom de Solexa, le nom de la compagnie qui l'avait initialement mis au point (129). Une fois les fragments garnis d'adaptateurs, la première étape particulière au Illumina Genome Analyzer est la PCR par pont, aussi appelée « bridge PCR ». Ainsi, des sondes complémentaires aux adaptateurs sont disposées sur une surface solide et permettent l'amplification des cibles par une première hybridation à une amorce immobilisée. Ensuite, l'extrémité opposée du brin nouvellement synthétisé est hybridée à l'amorce anti-sens, qui est aussi attachée à la surface. Des cycles d'extension des amorces par une polymérase suivis d'une dénaturation

au formamide permettent de créer sur la surface des groupes d'environ 1000 amplicons représentant la même séquence. Une amorce de séquençage est ensuite hybridée aux fragments immobilisés et permet l'incorporation de nucléotides contenant un terminateur de chaîne réversible marqué par un fluorophore différent selon le nucléotide. La fluorescence est ensuite détectée sur toute la surface afin d'associer les nucléotides incorporés aux différentes positions spatiales, ce qui permet de construire la séquence de chaque fragment. Les terminateurs de chaîne sont ensuite clivés afin de permettre un nouveau cycle d'ajout de nucléotides. La longueur des fragments séquencés avec cette méthode était, en juin 2010, d'environ 100 nucléotides. L'erreur la plus fréquente avec cette technique est la substitution.

Cette méthode domine actuellement le marché et elle peut produire de 25 à 200 milliards de nucléotides par expérience, ce qui augmente la couverture de séquençage d'un génome et ainsi améliore la précision de la séquence finale.

3.1.2.3. Le SOLiD

Le protocole de séquençage du SOLiD de Applied Biosystems est similaire à celui de 454 Life Sciences, car il utilise une PCR en émulsion sur des billes paramagnétiques (130). C'est après le bris de l'émulsion que le protocole diffère. Dans le SOLiD, les billes sont triées afin de ne conserver que celles couvertes d'amplicons. Elles sont ensuite accrochées de façon désordonnée à une surface qui, contrairement à celle de la technique 454, n'est pas poreuse. Le séquençage proprement dit est tout à fait différent des autres méthodes, car il se base sur l'hybridation et la ligation d'octamères fluorescents plutôt que sur la polymérisation de l'ADN.

Premièrement, une amorce universelle est hybridée à l'adaptateur qui a été ajouté à chaque fragment pendant la création de la librairie à séquencer. Ensuite, des octamères dégénérés sur toutes les bases, sauf une (la position *n*), sont hybridés et liés à l'amorce universelle par la ligase. Chacun de ces octamères est marqué par un des quatre fluorophores, dont la couleur est associée au nucléotide à la position *n.* Après la ligation, l'octamère est clivé entre les nucléotides 5 et 6 afin d'éliminer le fluorophore. Plusieurs cycles similaires sont effectués, puis le fragment produit par ligation des octamères est dénaturé. Une nouvelle

amorce universelle est hybridée aux fragments, suivie d'une série de cycles mettant en jeu des octamères dont une position *n′* (différente de *n*) est associée à la couleur du fluorophore. L'utilisation d'octamères dont deux bases sont connues permet de diminuer les erreurs de séquences. Le SOLiD produit des séquences mesurant de 35 à 50 nucléotides. Cette méthode peut produire jusqu'à 30 milliards de nucléotides.

3.1.2.4. Le Polonator

Le Polonator est le seul séquenceur de nouvelle génération qui suit un modèle libre (en anglais, *Open Source*) sur les protocoles, les réactifs et les logiciels. Le Polonator utilise aussi un protocole incluant une PCR en émulsion et un séquençage par ligation. Sa conception est basée sur la même publication que le SOLiD (130). Cependant, la version actuelle de l'appareil permet seulement de séquencer des fragments pairés de 13 nucléotides chacun, soit 26 nucléotides au total par fragment.

3.1.2.5. L'Heliscope

Ce qui distingue la technologie Heliscope des autres séquenceurs de nouvelle génération est qu'elle ne nécessite ni PCR, ni amplification clonale. Après la fragmentation de l'ADN à séquencer et l'ajout par ligation d'une queue poly-A, les fragments sont hybridés à une sonde poly-T disposée sur une surface. Ensuite, un nucléotide fluorescent est ajouté à chaque cycle de séquençage, un peu comme pour la technologie 454. Puis, un système de détection de fluorescence très sensible interroge indépendamment chaque molécule, ce qui permet de reconstruire la séquence de chacun des fragments.

Comme pour la technologie 454, la principale erreur générée par cette méthode se trouve dans les homopolymères. Les délétions sont aussi une erreur fréquente qui se produit lorsque des bases ne sont pas marquées ou qu'elles sont peu fluorescentes. De plus, il est possible que tous les fragments ne soient pas séquencés à la même vitesse, mais cela ne cause pas d'erreur dans les séquences obtenues. Une stratégie de double séquençage est aussi disponible : elle permet de séquencer les fragments dans les deux sens.

3.1.2.6. Les approches hybrides

La technologie 454 a été utilisée pour le séquençage de portions du génome de l'orge, un génome dont certaines portions sont complexes (131). Dans cet exemple, la couverture des régions séquencées était uniforme, mais, même si l'assemblage des régions géniques a été complété avec succès, l'assemblage de régions complexes n'a pas bien réussi. Wicker et ses collaborateurs sont arrivés à la conclusion que l'inclusion de séquençage Sanger dans leur assemblage améliorait la qualité des résultats.

Goldberg et ses collaborateurs ont aussi suggéré une approche hybride qui combine le séquençage Sanger et le séquençage à haut débit utilisant la technologie 454 (132). Selon eux, un séquençage Sanger à 5.3X est suffisant pour séquencer les bactéries marines (génome < 3 Mb) sur lesquelles portait leur étude. Ils ont suggéré que le séquençage à haut débit est à privilégier pour séquencer les régions non clonables ou celles contenant des *hard stop,* alors que le séquençage Sanger est utile afin de combler les trous et de couvrir les régions répétées. Cependant, leur étude n'inclut pas de séquençage pairé qui pourrait, selon eux, résoudre certains de ces problèmes. Ils ont aussi suggéré que la réalisation de deux expériences avec la technologie 454 pourrait être suffisante si un génome informant est disponible et si l'objectif de l'étude est la génomique comparative. Il faut noter que cette étude n'utilisait pas les assembleurs de dernière génération, comme CABOG, Ray ou Newbler, qui seront discutés dans la section suivante.

Le perfectionnement des techniques de séquençage de nouvelle génération a rendu désuète la combinaison de séquençage Sanger et de séquençage à haut débit. En effet, la longueur améliorée des fragments lus, l'utilisation de fragments pairés et les nouveaux logiciels d'assemblage de génomes ont permis d'obtenir de meilleurs résultats en utilisant uniquement des méthodes de séquençage à haut débit (133). En particulier, plusieurs distances séparant des fragments pairés peuvent être utilisées pour séquencer un organisme, ce qui facilitera l'assemblage *de novo* d'un génome (134).

3.1.3. L'assemblage de génomes

La reconstruction de la séquence des génomes a toujours été un défi, même lorsque les méthodes de séquençage ne produisaient pas une quantité appréciable de résultats à

analyser. Malgré les progrès informatiques qui permettent maintenant d'assembler facilement des séquences longues et peu abondantes, les méthodes informatiques ne sont pas encore adaptées pour traiter à la perfection les données générées par un séquenceur de seconde génération. Une seule expérience utilisant les techniques de séquençage à haut débit produit une quantité phénoménale de séquences, de l'ordre des milliards de bases, d'une longueur variant de 25 à plus de 500 nucléotides. Selon la longueur des séquences produites, leur assemblage sera plus ou moins difficile. Or, des fragments appariés facilitent cette étape et permettent d'assembler un génome plus long avec des fragments plus courts (135). Cette méthode permet aussi de reconstruire un génome riche en séquences répétées.

Présentement, il existe deux approches algorithmiques pour l'assemblage. La première, basée sur l'alignement des séquences, est plus adaptée au séquençage avec la méthode de Sanger et aux longs fragments, alors que la seconde méthode, basée sur la théorie mathématique des graphes de Bruijn, a un fort potentiel pour l'assemblage de fragments plus courts. Les logiciels les plus fréquemment utilisés ou les plus représentatifs de ces deux types d'assembleurs de génomes sont décrits dans les sections suivantes.

3.1.3.1. Le CAP, un assembleur par alignement et consensus

Le logiciel Contig Assembly Program (CAP) est l'un des premiers assembleurs par alignement qui a été publié (136). Le logiciel commence par éliminer les paires de fragments qui ne peuvent pas être alignées. Ensuite, les paires de fragments restantes sont alignées entre elles et une note est calculée pour chaque alignement. Les fragments sont ensuite assemblés selon leur note avec un algorithme vorace, c'est-à-dire qui requiert beaucoup de temps et de pouvoir informatique. Ce logiciel simple résume bien les algorithmes d'assemblage par alignement et consensus.

3.1.3.2. Les assembleurs Celera et CABOG

Plus complexe que le logiciel CAP, l'assembleur Celera a été conçu pour assembler des séquences obtenues par séquençage en aveugle, comme des chercheurs l'ont démontré dans l'assemblage du génome de *Drosophila melanogaster* (137). L'algorithme utilisé par cet assembleur est de type *overlap-layout-consensus,* c'est-à-dire qu'il nécessite l'alignement

des séquences afin d'établir un consensus. En bref, cet assembleur construit un graphe dans lequel les contigs sont les nœuds, et les interactions entre ces contigs sont les arêtes.

L'assembleur Celera inclut plusieurs étapes :

1. Le *screener,* d'abord, détecte et filtre les séquences répétées, les transposons, les séquences provenant de vecteurs et les autres séquences qui pourraient nuire à l'assemblage;

2. L'*overlapper* aligne ensuite les fragments afin de trouver ceux qui peuvent être assemblés, en tolérant un certain pourcentage de différences;

3. L'*unitiger* transforme par la suite les fragments associés par l'*overlapper* en unitigs, soit en minicontigs;

4. Le *scaffolder* ordonne ensuite les unitigs selon l'information obtenue avec les fragments pairés afin de former des échafaudages (*scaffolds*) et de reconstituer des chromosomes. Puis, les séquences répétées sont résolues afin d'éviter les erreurs d'assemblage;

5. Le consensus est finalement formé à partir des échafaudages en réalignant les fragments participant à la construction de chaque échafaudage afin de vérifier chaque base selon des critères de qualité de séquençage.

Comme l'assembleur Celera n'est pas adapté à la quantité et à la longueur de fragments produits par une expérience de séquençage à haut débit, une version modifiée de cet assembleur a été publiée : Celera Assembler with the Best Overlap Graph (CABOG) (138). Cet assembleur permet l'assemblage mixte de données issues de séquençage en aveugle et de séquençage avec la technologie 454. Voici ses étapes :

1. Les fragments sont d'abord rognés afin d'éliminer les positions ayant un plus fort risque d'erreurs;

2. Les alignements exacts entre les fragments sont ensuite trouvés. Les homopolymères sont compressés à un seul caractère afin d'éliminer les erreurs

causées par une détection erronée des homopolymères par la technologie 454. CABOG recherche les k-mers, c'est-à-dire les séquences nucléotidiques de longueur *k,* afin de trouver les homologies entre les fragments. Un alignement complet des fragments est calculé lorsque nécessaire;

3. Les fragments se superposant sont ensuite représentés dans un graphe (*Best overlap graph*), chaque fragment étant évoqué par une paire de nœuds reliés entre eux, l'un représentant le début du fragment et l'autre, sa fin. Ces nœuds peuvent ensuite être reliés à d'autres fragments avec des arêtes orientées qui indiquent l'orientation relative des deux fragments;

4. Le graphe créé en 3 est ensuite utilisé afin de construire les unitigs qui sont brisés aux positions où le graphe contient des intersections multiples insolubles. Puis, CABOG utilise l'information sur les fragments appariés afin d'éliminer les erreurs d'assemblage;

5. Les modules 4 et 5 de l'assembleur Celera original, soit le *scaffolder* et le générateur de consensus, sont utilisés pour les étapes subséquentes de l'assemblage.

CABOG a été comparé aux assembleurs Newbler version 1, PCAP (version ultérieure de CAP), Euler-SR et Velvet, produisant des résultats supérieurs à ces assembleurs en matière de nombre minimal de contigs et de quantité maximale de nucléotides incorporés dans l'assemblage (138). Malgré cela, certaines erreurs d'assemblage ont été trouvées dans les assemblages faits avec CABOG.

3.1.3.3. Le premier assembleur de Bruijn : Euler

Le logiciel Euler est le premier assembleur de génomes utilisant le graphe de Bruijn (139). La première étape de cet assembleur consiste à tracer un graphe de Bruijn correspondant au problème à étudier, c'est-à-dire en se basant sur tous les fragments séquencés à partir d'un même génome. Afin de construire le graphe, tous les mots de taille *k* de chaque séquence, c'est-à-dire toutes les séquences de *k* nucléotides comprises dans la grande séquence, sont répertoriés. Les séquences de *k-1* nucléotides constitueront les nœuds du graphe. Ensuite,

les mots de taille *k-1* contigus de chaque séquence, les séquences de longueur *k* se chevauchant sur *k-1* nucléotides, sont reliés par une arête dirigée qui correspond à l'orientation 5'–3' de l'ADN. L'assemblage proprement dit est ensuite effectué en trouvant les chemins eulériens dans le graphe, c'est-à-dire en trouvant la combinaison de chemins permettant de visiter une seule fois chacune des arêtes du graphe. Ces chemins, convertis en séquences nucléotidiques, sont les contigs.

Euler peut aussi utiliser l'information de séquençage apparié afin de construire des échafaudages. Cet assembleur ne tolère pas bien les erreurs de séquençage et les régions répétées dans les génomes. Chaisson et ses collaborateurs ont publié une version plus récente de ce logiciel, baptisée Euler-USR, qui permet l'assemblage de fragments courts pairés (135). Dans ce même article, ils ont suggéré que, même si les fragments séquencés sont courts, l'utilisation de fragments pairés contrebalance le problème pendant l'assemblage.

3.1.3.4. Le logiciel Newbler

Le logiciel Newbler, de Roche, est le logiciel officiel pour l'assemblage de données obtenues avec la technologie 454 (140). La première étape de l'assemblage est effectuée par l'*overlapper*, qui compare les séquences entre elles afin de trouver celles qui ont une forte probabilité d'être assemblables. Cela est fait en comparant les données brutes obtenues par le séquenceur, aussi appelées « flowgrams », afin de déterminer les fragments ayant un potentiel d'être assemblés. Ensuite, l'*unitiger* forme des groupes de séquences qui doivent, sans équivoque, être assemblés. Ensuite, le *multialigner* aligne les fragments de chaque unitig et le transforme en une seule séquence. Les fragments formant chaque unitig sont alors comparés afin de créer un consensus, et les contigs ainsi formés sont validés par l'analyse des *flowgrams* les composant. Si le séquençage utilise des fragments appariés, ces données seront utilisées afin d'échafauder les contigs.

Miller et ses collaborateurs ont comparé CABOG à Newbler, Euler-SR, PCAP et Velvet. Newbler arrivait en deuxième place en ce qui concerne le nombre de contigs créés et la longueur de ces derniers (138). Cependant, cette étude utilisait la version 1.1.03.24, qui est

une version antérieure à la version actuelle, ce qui rend cette comparaison désuète. À ce jour, aucune autre comparaison majeure et indépendante n'a été publiée.

3.1.3.5. Le Velvet

L'assembleur Velvet utilise un algorithme basé sur les graphes de Bruijn (141). Ainsi, ce logiciel libre construit un graphe de Bruijn comme Euler le fait, mais, en plus, il annote ce graphe en y ajoutant de l'information sur les chemins correspondant aux fragments séquencés. Cette addition permet de reconstruire la séquence des fragments à partir du graphe. Le logiciel parcourt ensuite ce graphe afin d'en déterminer les fragments qui se superposent. L'assembleur Velvet inclut aussi des étapes de simplification du graphe, comme la suppression d'erreurs de séquençage, la suppression des aspérités du graphe et la suppression des bulles dans le graphe.

3.1.3.6. L'assembleur Ray

Publié en 2010, l'assembleur Ray pousse plus loin l'utilisation des graphes de Bruijn pour l'assemblage des génomes (142). La principale différence entre Ray et EULER est le fait que Ray ne recherche pas les chemins eulériens dans le graphe, mais compare plutôt la séquence des fragments séquencés avec le contig en construction pour allonger ce dernier. Cette méthode limite les erreurs d'assemblage. L'annotation du graphe avec la séquence des fragments séquencés permet cette approche. De plus, Ray utilise les fragments pairés afin d'allonger les contigs en construction. Récemment, Ray a été adapté pour l'utilisation en informatique parallèle (*message passing interface* ou MPI), c'est-à-dire avec plusieurs processeurs, ce qui permet d'assembler un génome plus rapidement. Aussi, le logiciel est optimisé pour réduire la mémoire nécessaire à son exécution et permettre l'assemblage de génomes plus longs, comme le génome humain. De plus, cet assembleur n'est pas limité à une seule technologie de séquençage.

3.1.4. L'annotation de génomes

Une fois la séquence d'un génome construite, il faut l'annoter en associant des éléments (en anglais, *features*) à des positions du génome. Ces éléments peuvent être de plusieurs natures, comme des séquences codant pour des gènes ou des ARN fonctionnels, des séquences répétées ou des régions régulatrices. Plusieurs outils sont disponibles pour

identifier ces éléments. L'annotation d'un génome ne peut être faite en utilisant qu'un seul outil, car ils ont tous leurs forces et leurs faiblesses. De plus, selon le type de génome étudié, procaryote ou eucaryote par exemple, les outils à utiliser devront tenir compte de paramètres différents. En effet, la présence ou l'absence d'introns, le nombre de répétitions, la présence de duplications géniques, de délétions et de pseudogènes entraînent des besoins bio-informatiques différents.

Le type d'élément qui a le plus d'intérêt dans un génome est le gène, c'est-à-dire la séquence nucléotidique codant pour une protéine. Lorsque l'organisme d'intérêt est phylogénétiquement proche d'un autre organisme déjà séquencé et annoté, une première étape d'annotation peut être faite par transfert d'annotation, c'est-à-dire en comparant le nouveau génome et le génome déjà annoté avec des logiciels de comparaison de séquences comme BLAST ou BLAT (143-145). Dans la plupart des cas, ce parallèle permettra de trouver la majorité des gènes contenus dans le génome. Cependant, il est possible que certaines structures génomiques contenant des régions répétées, des insertions ou des délétions ne soient pas détectables avec cette méthode. La comparaison avec des espèces plus éloignées permet aussi d'identifier des gènes. Toutefois, utiliser uniquement la génomique comparative pour identifier les gènes d'une espèce nouvellement séquencée peut entraîner l'oubli de nombreux gènes.

Il est simple de trouver tous les cadres de lecture ouverts (*open reading frame* ou ORF) d'une séquence et de postuler que les plus longs ORF ont une forte probabilité de coder pour des protéines. Cependant, cette méthode rudimentaire d'identification des gènes est inutile si l'organisme étudié possède des introns.

Afin de trouver les gènes qui seraient uniques à l'organisme séquencé, une approche simple et valide consiste à coupler la recherche des ORF et les comparaisons avec les gènes connus d'autres organismes ou avec des méthodes d'annotation *ab initio.* Par ailleurs, une des premières méthodes de détection des gènes consiste à étudier l'utilisation des codons. En effet, chez plusieurs organismes, l'utilisation des codons dans les gènes est différente de celle dans des régions intergéniques (146). Cependant, ces méthodes ont été remplacées par des modèles plus complexes qui permettent une meilleure identification des gènes selon l'espèce (147). Plusieurs de ces outils utilisent les chaînes de Markov (*hidden Markov*

models ou HMM), un modèle statistique bayésien, afin d'entraîner un modèle sur des gènes connus d'un organisme pour identifier les gènes dans des sections non annotées du même organisme ou d'un organisme apparenté. Plusieurs logiciels utilisent les HMM pour trouver les gènes d'une espèce. En général, la première étape consiste à créer un modèle HMM particulier à l'organisme étudié en utilisant des séquences de gènes déjà connues. Ensuite, la séquence du génome nouvellement séquencé est analysée à l'aide de ce modèle et les gènes y sont identifiés en leur attribuant un score. De nombreux exemples de logiciels utilisant cette méthode peuvent être cités : AUGUSTUS (148), GLIMMER (149), GENEID (150) et ARGOT (151).

Trouver les séquences codantes potentielles ne représente plus le problème le plus difficile de l'annotation, mais l'identification du premier codon d'un gène peut s'avérer difficile. En effet, plusieurs codons START sont souvent possibles au début d'un gène et il est difficile d'identifier avec certitude lequel de ces codons est le bon. La méthode la plus efficace pour identifier le début des gènes est de comparer la séquence du génome à annoter avec des marqueurs de séquence exprimée (*expressed sequence tags* ou EST) (152). La sensibilité de cette méthode très spécifique dépend de la profondeur du séquençage des EST. Il existe aussi des méthodes bio-informatiques de détection du début des gènes. Par exemple, Smith et ses collaborateurs ont utilisé des *position specific scoring matrix* (PSSM) pour identifier chez *Leishmania* environ 50 % des sites de polyadénylation dans un intervalle de 100 nucléotides et 65 % des sites de trans-épissage dans un intervalle de 25 nucléotides (153). Chez la plupart des eucaryotes, le problème le plus difficile du processus de détection de gènes est l'identification des jonctions intron-exon.

Les éléments répétés dans les génomes peuvent être recherchés dans les séquences en utilisant des outils comme RepeatMasker (154). Il existe plusieurs types de séquences répétées et des logiciels différents doivent être utilisés pour les détecter. Entre autres outils, des modèles HMM ont été bâtis pour détecter les SIDER de *Leishmania* (85).

L'association des éléments à une fonction potentielle passe par la comparaison avec des gènes dont la fonction est connue et par la recherche de motifs encodés dans des modèles HMM. En particulier, le génome d'intérêt peut être interrogé avec le logiciel HMMER

(155) afin d'y déceler des motifs connus construits à partir de protéines conservées entre plusieurs espèces, comme la base de données Pfam (156).

3.2. La transcriptomique et les biopuces

L'utilisation de techniques d'hybridation moléculaire a commencé par la mise au point des hybridations sur membranes dans les années 1970, mais c'est en 1996 que ce type d'approche a pris toute son ampleur par la miniaturisation de ces dispositifs expérimentaux (157, 158). L'avènement des biopuces a été le premier pas vers une biologie quantitative à grande échelle. Cette méthode a permis de mesurer simultanément l'expression de centaines, voire de dizaines de milliers de gènes transcrits en une seule expérience. Même s'il existe plusieurs technologies de biopuces, elles ont toutes le même principe, soit l'hybridation moléculaire d'échantillons biologiques marqués sur une sonde immobilisée à un endroit précis d'une surface.

3.2.1. Le principe des biopuces

En général, une expérience de biopuce ayant pour objectif de mesurer l'expression des gènes d'un organisme utilise comme échantillon de base les ARN messagers (ARNm) extraits d'une ou de plusieurs populations de cellules représentatives des conditions à l'étude.

Cet ARNm sera habituellement transformé en ADN complémentaire (ADNc) par transcription inverse et, par le même processus, il sera marqué par une molécule fluorescente qui pourra ensuite être détectée en utilisant un appareil approprié, en général un scanneur confocal. L'ADNc marqué sera ensuite hybridé sur une biopuce adaptée aux questions biologiques étudiées. Sur la biopuce, des sondes composées de séquences d'ADN particulières sont immobilisées sur une surface, de manière à ce que des positions distinctes de la surface puissent être associées à des cibles précises. Les sondes disponibles sur la surface s'hybrideront avec l'ADNc marqué si leurs séquences sont complémentaires. La complémentarité de l'ADN selon les principes découverts par Watson et Crick sera respectée au cours de l'hybridation, comme dans une hybridation Southern, mais il est possible qu'une sonde et une cible ayant de faibles différences de séquences forment un complexe moléculaire assez stable pour être détecté. La biopuce est ensuite lavée dans des

conditions suffisamment stringentes pour améliorer la spécificité de l'hybridation. Le signal des biopuces est ensuite quantifié par un appareil qui permet de déterminer la fluorescence sur toute la surface de la biopuce. Le signal fluorescent obtenu avec chaque sonde peut ensuite être corrélé avec la quantité de cibles présentes dans l'échantillon original, pour permettre ainsi d'estimer le niveau d'expression ou de modulation des gènes ciblés.

Les biopuces peuvent être utilisées avec un ou deux canaux, soit en hybridant un seul échantillon marqué par biopuce, soit en hybridant sur une même biopuce deux échantillons marqués par des fluorophores différents.

3.2.2. Les types de biopuces

Avant que la séquence d'un génome soit disponible, l'étude d'un organisme avec des biopuces était possible, mais demandait beaucoup de travail. Premièrement, une librairie de clones devait être construite avec l'ADNc de l'organisme à étudier. Ensuite, l'ADN de chaque clone était extrait et la séquence de l'insert était amplifiée et purifiée. Chacun des clones sélectionnés était déposé sur une surface de verre couverte de groupements amines. Cette biopuce était ensuite hybridée avec les échantillons d'intérêt. L'annotation de ce type de biopuce nécessitait le séquençage de chacun des clones sélectionnés.

Lorsque la séquence du génome de l'organisme d'intérêt est disponible, une biopuce peut être conçue par ordinateur, pour être ensuite construite par impression d'oligonucléotides préalablement synthétisés ou par photolithographie (159). Dans les deux cas, une analyse bio-informatique de la séquence du génome de l'organisme d'intérêt permet de déterminer, pour chaque gène, une ou plusieurs séquences qui permettront de mesurer précisément l'expression de ce gène. Toutes les sondes d'une biopuce doivent correspondre aux mêmes paramètres thermodynamiques, incluant le pourcentage GC de la sonde et sa température de dénaturation. En général, ces types de biopuces utilisent des oligonucléotides dont la longueur varie de 20 à 70 nucléotides, mais qui reste constante sur une biopuce.

L'impression de biopuces oligonucléotidiques requiert la construction préalable d'une banque d'oligonucléotides dans des plaques de 96 ou de 384 puits. Ces oligonucléotides sont ensuite déposés sur la surface par contact physique, par un processus piézoélectrique ou par une technologie similaire à celle trouvée dans une imprimante à jet d'encre.

Plusieurs chimies de surface permettent l'immobilisation de ces sondes, mais les groupements chimiques les plus fréquemment utilisés sont le poly-L-lysine ou simplement l'amine. Cette dernière permet la liaison électrostatique de l'ADN à la surface.

L'impression de biopuces *in situ* par photolithographie permet plus de liberté dans leur conception, car les séquences des sondes peuvent être modifiées entre chaque lot de production sans nécessiter la commande de nouveaux oligonucléotides ou la construction de nouvelles librairies. Les biopuces d'Affymetrix et d'Agilent sont les exemples les plus répandus de biopuces construites par photolithographie.

Les biopuces d'Affymetrix ont l'avantage de permettre la standardisation des plateformes pour les utilisateurs intéressés à l'étude de l'humain, de la souris ou d'autres organismes pour lesquels Affymetrix produit une biopuce particulière. Quant à la biopuce sur mesure d'Agilent, elle permet aux utilisateurs de créer une biopuce correspondant à leurs besoins sans nécessiter la commande de milliers d'oligonucléotides et en leur laissant le soin de choisir les séquences à utiliser.

Dans un futur rapproché, les biopuces seront probablement remplacées par le séquençage à haut débit des ARNm, aussi appelé « RNA-seq » (133).

3.2.3. La conception d'une expérience

Avant de concevoir une expérience de biopuce, il faut définir les questions biologiques à résoudre. Or, cette étape essentielle est souvent oubliée. Dans le cas d'expériences dans lesquelles les chercheurs comparent seulement deux conditions, la question biologique sera souvent simple : « Quels sont les gènes exprimés de manière différente entre les deux conditions? » La question se complique lorsque plus de deux conditions sont étudiées et que plusieurs comparaisons différentes sont nécessaires. Le protocole de normalisation et d'analyse devra alors être modifié en conséquence.

3.2.4. La normalisation et l'analyse statistique

La normalisation et l'analyse statistique de biopuces sont des étapes essentielles qui nécessitent à la fois une bonne connaissance de la question biologique étudiée et des méthodes statistiques à employer.

L'étape de la normalisation permet d'éliminer une partie de la variabilité expérimentale causée par des différences de traitement entre les biopuces, des différences entre le marquage des échantillons ou des différences dans la concentration des échantillons. Les méthodes de normalisation varient selon les biopuces utilisées, les questions à l'étude et le choix d'un protocole à un ou deux canaux.

L'analyse statistique des biopuces permet de déterminer les résultats significatifs, c'est-à-dire les gènes dont l'expression est significativement modifiée entre les conditions étudiées. Dans ce cas, le résultat pour une sonde ou un gène sera considéré comme statistiquement significatif si son signal varie suffisamment comparativement à la variabilité expérimentale. Plusieurs logiciels sont disponibles pour effectuer ces analyses, mais cette thèse sera concentrée sur les algorithmes de normalisation et sur les méthodes statistiques les plus fréquemment utilisés pour les biopuces à deux couleurs.

Le logiciel le plus puissant pour la normalisation et l'analyse des biopuces est probablement R, qui, avec le projet Bioconductor, permet l'utilisation de plusieurs librairies utiles pour l'analyse de biopuces (160). Entre autres logiciels, LIMMA permet l'analyse de biopuces deux couleurs (161) et AFFY l'analyse de biopuces Affymetrix (162). Parmi les autres logiciels d'intérêt, on compte la suite TM4 (163), Mayday (164), MAGMA (165), ArrayMagic (166) et ArrayPipe (167). Toutes ces plateformes permettent l'intégration avec le logiciel de programmation statistique R pour la normalisation et le calcul des statistiques.

3.2.4.1. La normalisation de biopuces

La normalisation de biopuces inclut plusieurs étapes. Pour normaliser une biopuce à deux couleurs, on compte trois étapes principales, soit la correction du bruit de fond, la normalisation entre les canaux d'une même biopuce et la normalisation entre biopuces

indépendantes. Dans le cas d'une biopuce à un seul canal, comme les biopuces d'Affymetrix, seules la première et la troisième étape sont nécessaires.

La correction du bruit de fond doit permettre d'éliminer l'impact d'une fluorescence résiduelle inégale à différentes positions de la biopuce. Plusieurs méthodes sont disponibles. Les plus simples consistent en une soustraction du signal correspondant au signal aspécifique autour d'une sonde de la biopuce. Ces méthodes peuvent cependant entraîner des problèmes si le signal obtenu après la soustraction est négatif. Cela peut être évité en attribuant aux sondes dont le signal devient négatif une valeur arbitraire, ou en leur associant une valeur dérivée du bruit de fond environnant (168).

Les méthodes les plus évoluées de correction de la fluorescence basale utilisent des distributions statistiques, par exemple une loi normale (169, 170). Dans le cas des biopuces d'Affymetrix, la méthode Robust Multi-array Average (RMA) ajuste le bruit de fond en se basant sur la distribution empirique du signal des sondes (171), alors que la méthode MAS5 calcule le bruit de fond en se basant sur les sondes dont le signal est le moins intense (172). Ce dernier sera ensuite utilisé pour corriger les intensités en fonction de l'importance du signal aspécifique. Historiquement, la correction du bruit de fond des biopuces d'Affymetrix était basée sur une sonde de séquence inexacte associée à chacune des biopuces. Cette méthode n'est pas recommandée, car elle biaise les résultats (173, 174).

La normalisation interne d'une biopuce à deux couleurs permet de mettre à niveau les deux canaux d'une biopuce afin d'équilibrer le signal si la quantité d'ADN ajouté pour les deux échantillons est différente ou si le niveau de marquage des échantillons diverge. L'objectif vise une distribution comparable de l'intensité des deux canaux d'une même biopuce. La méthode la plus simple pour y arriver est la normalisation des deux canaux selon la moyenne ou la médiane des signaux. Cependant, une normalisation se basant sur une distribution statistique, comme une normalisation par lœss, donne de meilleurs résultats. De plus, dans le cas de biopuces imprimées par contact avec plusieurs pointes d'impression, généralement 48, il est nécessaire de normaliser les sections de la biopuce ayant été imprimées avec des pointes différentes en utilisant la méthode de Lœss selon les pointes d'impression (en anglais, *print-tips loess*). Il est aussi possible de normaliser des biopuces

selon des contrôles ajoutés à l'échantillon pendant sa préparation, mais cette méthode devrait être faite conjointement à l'une des méthodes préalablement décrites.

Finalement, les biopuces doivent être normalisées entre elles. L'objectif de cette étape est de rendre comparables les signaux de toutes les biopuces d'une expérience afin de permettre l'analyse simultanée et la comparaison de ces biopuces. Aussi, cette étape peut permettre de calculer le rapport entre les différentes conditions testées sous la forme de log-ratio, c'est-à-dire la valeur logarithmique en base deux du ratio entre deux signaux. Pour ce faire, deux approches peuvent être utilisées. La première ajuste les intensités des biopuces sans modifier les log-ratios de chaque biopuce, alors que la seconde peut avoir un impact sur les log-ratios.

Les biopuces peuvent être normalisées selon les quantiles de l'un ou l'autre des deux canaux ou selon le log des produits de l'intensité des deux canaux. D'autres variantes sont aussi disponibles, comme une normalisation qui tient compte des groupes d'échantillons auxquels appartiennent les cibles. La méthode d'échelle (*scale*), mise en place dans LIMMA (161), a pour but d'ajuster les log-ratios afin que l'écart absolu de la médiane soit similaire entre les biopuces. Cette dernière méthode modifie les log-ratios au cours de la normalisation. Dans le cas des biopuces d'Affymetrix, la méthode RMA utilise une normalisation selon les quantiles, alors que la méthode MAS5 utilise une normalisation d'échelle.

3.2.4.2. L'analyse statistique

Les données normalisées peuvent ensuite être analysées par des tests statistiques. La méthode la plus simple d'analyse statistique consiste en un test de Student sur chacune des sondes. Ainsi, en comparant chaque condition test entre elles pour une même sonde, il est possible de déterminer les sondes pour lesquelles la différence entre deux conditions est statistiquement significative. Des méthodes plus complexes sont aussi disponibles. Par exemple, le protocole d'analyse statistique utilisé par le logiciel LIMMA recourt à une régression linéaire pour déterminer les log-ratios finaux, et à des statistiques de Bayes pour faire le calcul statistique, une analyse de variance (ANOVA). L'utilisation d'une méthode

de correction de la valeur p peu stricte, couplée à un seuil minimal du log-ratio est la méthode recommandée pour l'analyse de biopuces (175).

Néanmoins, et élément important, peu importe le test statistique utilisé, il est nécessaire de corriger la valeur p (en anglais, *p-value*) pour les tests multiples. En effet, lorsque plusieurs milliers de tests statistiques sont faits simultanément, il est attendu qu'une proportion de ces tests correspondant à la valeur p soient des faux positifs. Par exemple, pour une valeur p de 0.05, 5 % des gènes testés pourraient être des faux positifs, ce qui n'est pas souhaitable lorsque plusieurs milliers de gènes sont testés. Les méthodes de correction de la valeur p permettent de contrôler le nombre de faux positifs obtenus. Deux approches permettent cette correction, l'erreur d'ensemble de la famille (*familywise error rate* ou FWER), plus stricte, et le taux de fausses découvertes (*false discovery rate* ou FDR), moins stricte.

D'une part, la correction de la FWER est très stricte et diminue de beaucoup le taux de faux positifs, mais cela au détriment de vrais positifs. Utilisant cette approche, la correction de Bonferroni consiste simplement à multiplier la valeur p du test statistique par le nombre de tests statistiques effectués. Par exemple, si 1000 gènes sont testés, une valeur p de 0.05 deviendra 50, ce qui ne sera pas significatif. Ainsi, une valeur p non corrigée plus petite que 0.00005 sera nécessaire pour qu'un gène soit significatif, ce qui conduit à un taux de faux positifs presque nul. D'autre part, la méthode de Holm ressemble à la méthode de Bonferroni, mais elle est un peu moins stricte. Ces deux méthodes contrôlent la probabilité de faire une ou plusieurs erreurs dans une série de tests statistiques.

La méthode de Benjamini et Hochberg contrôle le taux de faux positifs (*false discovery rate* ou FDR). Cette méthode est beaucoup moins stricte, mais plus complexe. Si la valeur p utilisée est 0.05, alors 5 % des gènes significatifs pourraient être des faux positifs.

3.2.4.3. L'exploration des données

Il est difficile de tirer profit d'une simple liste de gènes comme celle générée par l'analyse statistique des données. Souvent, il est nécessaire d'analyser les résultats significatifs d'une expérience de biopuce avec plusieurs méthodes afin d'obtenir des résultats ayant un sens

biologique. Ces méthodes peuvent consister en des calculs uniquement mathématiques ou elles peuvent être ancrées dans la biologie.

Plusieurs méthodes de ségrégation (en anglais, *clustering*) permettent de classer les données dans différents groupes en se basant uniquement sur des calculs mathématiques. Par exemple, la ségrégation hiérarchique recourt à des méthodes de calcul similaires à celles utilisées pour la construction d'arbres phylogénétiques. Les gènes ou les échantillons sont ainsi mis en relation et ceux ayant des profils similaires seront plus proches dans l'arbre. Une autre méthode, la ségrégation par *k-means,* permet de séparer les données en différents groupes qui réagissent d'une manière similaire aux conditions étudiées. Le logiciel TMeV permet de mettre en application de nombreuses méthodes de ce type (163).

Il est essentiel d'analyser les données avec la biologie en tête. L'analyse de voies métaboliques, d'ontologies géniques et de réseaux de gènes permet une telle analyse. La visualisation de voies métaboliques mise en relation avec les résultats provenant d'expériences de biopuces permet de voir l'impact des conditions étudiées sur le métabolisme de l'organisme étudié. D'une manière plus globale, l'enrichissement ou la diminution du nombre de gènes significatifs appartenant à une catégorie d'ontologies géniques permet d'associer les conditions étudiées à des effets généraux observés dans les cellules (176). Finalement, différentes méthodes de génération de réseaux géniques, comme les analyses d'interactions protéine-protéine (177, 178) ou les analyses du Web sémantique (179), permettent d'observer les effets des conditions étudiées sur des groupes de gènes pour lesquels il y a présomption d'interaction. Ces dernières méthodes sont surtout limitées par les résultats disponibles et par le niveau d'annotation des génomes.

3.2.5. Les autres applications des biopuces

L'hybridation d'ADN génomique à une biopuce permet de poser des questions différentes de celles posées pour l'hybridation d'ARNm. En effet, des phénomènes comme les amplifications chromosomiques, l'aneuploïdie ou les amplifications géniques seront détectables au moment de la comparaison d'ADN génomique sur biopuce (180).

Les biopuces ont aussi été considérées pour le séquençage de génomes (180). Cependant, la complexité de ce type de séquençage requiert une quantité de sondes qui n'était pas

compatible avec la densité des biopuces disponibles au moment ou cela a été proposé. Le concept était de disposer, sur une surface, les sondes couvrant toutes les possibilités de séquences d'un oligonucléotide d'une longueur déterminée. L'assemblage de ces génomes utiliserait la théorie des graphes de Bruijn. Cependant, le nombre de séquences différentes nécessaires pour un oligonucléotide est 4^n, où *n* est la longueur de la sonde. Ainsi, pour une sonde de 20 nucléotides, plus d'un million de sondes sont nécessaires.

Une autre application des biopuces est le diagnostic des maladies infectieuses (159). Ce sujet sera discuté dans les chapitres 6 à 9.

4. La bio-informatique pour analyser les biopuces *Leishmania*

La biopuce est l'un des premiers outils disponibles qui ont permis de donner à la biologie un haut débit. En permettant d'évaluer semi-quantitativement l'expression de tous les gènes d'un organisme, cette technologie a beaucoup apporté à cette science tout en créant de nouveaux problèmes qui doivent être résolus par de nouvelles approches. Heureusement, l'avènement de ces technologies va de pair avec les technologies de traitement et de gestion de l'information, ce qui permet de faire croître conjointement le besoin et les ressources permettant d'y répondre. Ces problèmes peuvent aussi être associés aux autres méthodes de génomique à haut débit, comme le séquençage à haut débit ou la protéomique (181).

Plus précisément, plusieurs défis sont associés à l'utilisation de biopuces :

- Gérer les données à haut débit générées par ces expériences;
- Éviter la perte de données;
- Favoriser le passage de résultats entre les générations d'employés et d'étudiants;
- Effectuer des copies de sauvegarde;
- Traiter et normaliser les données;
- Effectuer l'analyse statistique des données;
- S'assurer de l'uniformité du traitement des données entre les expériences d'individus différents;
- Faciliter l'analyse pour les gens qui n'ont pas les compétences adéquates en analyse statistique;
- S'assurer que toutes les expériences sont analysées avec une méthode valide.

L'utilisation d'un système informatique de gestion de laboratoire (*laboratory information management system* ou LIMS) qui inclut le traitement des données permet de répondre à la majorité de ces problèmes. La création et l'implantation d'un tel système seront décrites dans ce chapitre.

4.1. La gestion des données à haut débit

Le premier problème lié à l'utilisation de biopuces est la quantité massive de données générées par cette technique. Une expérience est nécessairement composée de plusieurs biopuces, chacune composée d'une série de points auxquels sont associées plusieurs données numériques. Ainsi, plusieurs milliers de points de données sont générés pour chaque biopuce, le tout multiplié par le nombre de biopuces incluses dans l'expérience. En plus de ces données quantitatives, chaque biopuce génère un ou plusieurs fichiers d'images, qui sont les données brutes de l'expérience. Il est essentiel de conserver les données brutes afin de permettre, si nécessaire, une analyse additionnelle advenant la découverte d'une erreur dans les processus de traitement ou d'analyse informatiques, ou l'avènement de nouveaux protocoles d'analyse.

Dans le passé, il était facile de consigner par écrit, dans un cahier de laboratoire, les résultats d'une expérience qui générait une quantité limitée de données. La consignation de ces données à un cahier de laboratoire permettait la pérennité des données dans le laboratoire et favorisait le passage de l'information entre les générations d'employés ou d'étudiants. Vu la quantité de données produites par les nouvelles méthodes de génomique, comme les biopuces, il est impossible de tout consigner à un cahier de laboratoire physique. Dans le cas où chaque utilisateur serait responsable de la conservation des données brutes de ses expériences de génomique, plusieurs problèmes pourraient survenir, comme l'absence de copies de sauvegarde, une mauvaise annotation des fichiers ne permettant pas une réanalyse ou encore la perte de ces données au départ de l'utilisateur. Il devient alors essentiel pour un laboratoire qui fait usage de ces techniques de mettre sur pied un système qui favorisera la pérennité des données brutes.

Le dépôt des données dans une base de données publique comme Gene Expression Omnibus (GEO) (182) ou ArrayExpress (183, 184) permettrait la consultation ultérieure

des données. Ces bases de données sont utiles pour la communauté scientifique, mais leur nature générique ne permet pas nécessairement de noter toute l'information qui serait utile à un laboratoire particulier. Aussi, lorsque les données ne sont pas encore publiées, il n'est pas souhaitable de les rendre immédiatement accessibles à la communauté scientifique. De plus, ces bases de données ne permettent pas l'addition de tous les fichiers associés à une expérience. Le format de données MIAME (de l'anglais, *minimum information about a microarray experiment*) permet la standardisation entre les laboratoires (185), mais il n'est souvent pas suffisant pour gérer les expériences à l'intérieur d'un laboratoire particulier. Finalement, les revues scientifiques exigent que des données de biopuces publiées soient déposées dans une base de données publique.

4.2. L'analyse des données à haut débit

Le second problème associé à l'utilisation de biopuces concerne le traitement et l'analyse de ces données. Chaque biopuce produit plusieurs milliers de points de données et chacun de ces points subit plusieurs types de variations (techniques, biologiques et expérimentales). Il est donc nécessaire de prétraiter les données avant d'en commencer l'analyse afin de permettre la comparaison entre les biopuces. De plus, vu la variabilité des données, une analyse statistique est essentielle. Étant donné le nombre élevé de questions posées dans une expérience de biopuces, il est nécessaire d'ajuster le test statistique utilisé en fonction du nombre de tests effectués par une méthode de correction de la valeur p pour les tests statistiques multiples.

Il est souhaitable que, à l'intérieur d'un groupe de recherche, les données issues d'un même type d'expérience soient traitées de manière similaire. Cette uniformité fournit plusieurs avantages, notamment la facilitation de la réanalyse des expériences, de la comparaison de ces dernières, de la formation des utilisateurs et de la publication des données. Aussi, comme ces étapes demandent une connaissance approfondie du type de données analysées et des méthodes utilisées pour le faire, il est nécessaire que les protocoles de prétraitement et d'analyse soient conçus par des experts. Tous les utilisateurs ne souhaitent pas apprendre les notions de biostatistique et de bio-informatique nécessaires à la conception d'un protocole de traitement de données à haut débit, comme des données issues d'expériences

de biopuces. La conception de protocoles d'analyse par des utilisateurs moins expérimentés peut diminuer la puissance statistique de certaines expériences. Ainsi, certains protocoles de prétraitement ou d'analyse pourraient ne pas être valides dans certains contextes et mener à des conclusions erronées. Le choix du protocole d'analyse devrait aussi être guidé par la ou les questions biologiques derrière l'expérience, ce qui est souvent oublié.

4.3. La biopuce *Leishmania*

L'étude du transcriptome de *Leishmania* demandait la création d'une biopuce permettant d'évaluer l'expression des gènes de deux espèces : *L. infantum* et *L. major.* Une première génération de biopuces a été créée en 2005 en utilisant l'impression par contact, tandis qu'une deuxième génération a été créée en 2008 à partir de versions plus récentes des génomes et en utilisant une méthode d'impression *in situ*.

4.3.1. La première génération

En collaboration avec Philippe Rigault, du Centre de génomique de Québec, nous avons développé une biopuce à oligonucléotides qui cible tous les gènes de *L. major* et de *L. infantum.* Les sondes de 70 nucléotides pour chaque gène ont été choisies de manière automatisée et permettent de mesurer l'expression de chacun des gènes des deux organismes. La création de cette biopuce était principalement basée sur la séquence du génome de *L. infantum,* version 2.0. Ainsi, la biopuce composée de 8841 sondes permet de mesurer l'expression de 8173 gènes de *L. infantum,* version 2, et de 8311 gènes de *L. major,* version 5.2. Les biopuces *Leishmania* ont été imprimées par l'équipe du D^{r} Gary Hardiman, de l'University of California, San Diego (UCSD).

Sur cette biopuce, 7926 sondes ciblent *L. infantum* avec une hybridation parfaite. Parmi ces sondes, 3580 ont aussi une hybridation parfaite avec *L. major,* 3410 ont 1 ou 2 nucléotides de différence avec *L. major,* 649 ont de 3 à 7 nucléotides de différence et 297 ont 8 nucléotides ou plus de différence. Pour les gènes de *L. major* qui n'avaient pas de sondes associées ayant 2 nucléotides de différence ou moins, 915 nouvelles sondes ont été ajoutées. À ces sondes s'ajoutent 372 oligonucléotides contrôle ciblant 32 gènes différents. Pour chacun de ces gènes, les sondes contrôle incluent un contrôle de resynthèse,

3 contrôles ciblant des positions différentes du gène et 6 contrôles contenant des polymorphismes (1, 2, 3, 5, 7, 10). La biopuce contient aussi 49 sondes utilisées au cours d'études précédentes (108).

L'identifiant GEO de la biopuce *Leishmania* de première génération est GPL8804.

4.3.2. La deuxième génération

La seconde génération de biopuces *Leishmania* a aussi été conçue par Philippe Rigault, du Centre de génomique de Québec. Les sondes de 60 nucléotides ont été sélectionnées de manière automatisée et permettent de mesurer l'expression de chacun des gènes de *L. infantum* et de *L. major*. La version 3.0a du génome de *L. infantum* et la version 5.2 du génome de *L. major* ont été utilisées pour la sélection des sondes. La biopuce comprend 9173 sondes spécifiques à *Leishmania,* en plus des sondes contrôle particulières aux biopuces d'Agilent. Les sondes sont imprimées sur 8 régions distinctes d'une seule biopuce, ce qui permet de l'utiliser pour hybrider simultanément 8 échantillons. Les biopuces sont synthétisées *in situ* par le service de biopuce sur mesure d'Agilent Technologies (Mississauga, Ontario).

Sur cette biopuce, exactement 8100 sondes ciblent parfaitement *L. infantum*. Parmi ces sondes, 3972 sondes ont aussi une hybridation parfaite avec *L. major,* 3407 ont 1 ou 2 nucléotides de différence avec la séquence de *L. major,* 456 ont de 3 à 7 nucléotides de différence et 604 ont plus de 8 polymorphismes de différence. De plus, 822 sondes ont été ajoutées pour compenser les sondes ayant plus de 2 nucléotides de différence entre les deux espèces. Des sondes de contrôle sont incluses pour les mêmes 32 gènes que dans la première génération de biopuces *Leishmania.* Trois sondes ciblent des positions différentes de ces 32 gènes, alors que 8 sondes contiennent 1, 2, 3, 5, 7, 10, 12 et 20 mutations.

L'identifiant GEO de la biopuce *Leishmania* de deuxième génération est GPL11330.

4.4. Le protocole d'analyse

Plusieurs étapes de normalisation et d'analyse sont nécessaires afin de transformer les données brutes obtenues à la suite d'une expérience de biopuces en résultats ayant une

signification biologique pertinente et analysable. Voici, en termes généraux, chacune des étapes d'analyse d'une expérience de biopuces utilisant deux canaux :

1. Correction du bruit de fond;
2. Normalisation interne de chaque biopuce afin d'équilibrer les deux canaux de fluorescence;
3. Normalisation entre elles de toutes les biopuces afin de permettre leur comparaison;
4. Analyse statistique incluant une correction pour tests multiples.

Plusieurs méthodes sont disponibles pour chacune de ces étapes. Ces méthodes ont donc été sélectionnées en considérant des paramètres particuliers à la biopuce *Leishmania.* Pour l'étape de la normalisation, les méthodes ont été choisies en les testant sur des données réelles et en observant leur effet sur la distribution des données.

4.4.1. L'association entre les sondes et les gènes

La nature des génomes, notamment celui de *Leishmania,* est telle que plusieurs structures sont répétées à plusieurs endroits dans le génome, avec ou sans polymorphismes. Ainsi, il est possible qu'une même sonde s'hybride à plusieurs gènes sur le génome et que ces différents gènes aient un impact variable sur le signal de la sonde. D'un autre côté, l'utilisation d'une biopuce conçue pour étudier l'expression des gènes de plusieurs espèces proches phylogénétiquement entraîne, sur la biopuce, la présence de sondes ayant des différences de séquences avec l'organisme étudié ou de sondes qui n'ont aucune cible dans l'espèce étudiée. Finalement, comme il est nécessaire d'utiliser des protocoles de corrections de la valeur p au cours de l'analyse statistique des données, l'inclusion dans l'analyse de nombreuses sondes qui n'ont pas de cibles dans le génome étudié entraîne une diminution inutile de la puissance statistique de l'étude.

Il est donc souhaitable d'annoter les sondes pour chaque espèce étudiée afin de faciliter l'analyse de l'expérience de biopuces et d'éliminer a priori les sondes inutiles. Le fichier de

type SpotType, utilisé par la librairie LIMMA du logiciel R (161), permet de tenir compte de cette information. Il faut, pour chaque espèce étudiée, créer un fichier SpotType qui permet d'associer adéquatement les sondes de la biopuce aux gènes de l'espèce étudiée.

Il est important de noter que le fichier SpotType utilisé pour analyser la première génération de biopuces *Leishmania* a été créé avec un protocole différent de celui-ci. La première génération de biopuces *Leishmania* utilisait un fichier SpotType commun à *L. infantum* et à *L. major*. Ce fichier contenait moins d'information que le fichier SpotType utilisé pour l'étude de la seconde génération de biopuces *Leishmania.* Afin d'améliorer l'analyse de données similaires à la première biopuce *Leishmania,* il ne sera décrit ici que le protocole de création du fichier SpotType de seconde génération. Ce type de fichier sera aussi utilisé dans la description des analyses de la biopuce de première génération.

La création du fichier SpotType de seconde génération nécessite plusieurs étapes, qui sont décrites ci-dessous :

1. Obtenir la séquence des sondes de la biopuce en format FASTA;

2. Comparer la séquence des sondes avec la séquence des gènes de l'organisme d'intérêt en utilisant le logiciel Blastall, avec les résultats en format BLAST tabulaire (m = 8), ou le logiciel BLAT, avec les résultats en format BLAST tabulaire;

3. Préparer un fichier tabulaire décrivant les annotations du génome d'intérêt et contenant les quatre colonnes suivantes :

    - Identifiant du gène,
    - Position de début du gène,
    - Position de fin du gène,
    - Description du gène entre guillemets;

4. Préparer un fichier tabulaire contenant la liste de sondes à raison d'une sonde par ligne;

5. Traiter les résultats de comparaison entre les gènes et le génome complet en utilisant le script d'annotation des sondes (voir l'annexe 2). La commande à utiliser a la forme suivante : *perl AnnotateProbes-v1.pl ListeDeSondes.probelist ListeDeGènes.genelist RésultatsComparaison.blat > Fichier.spottypes.*

Ce protocole associe à chaque sonde de la biopuce un SpotType sélectionné parmi les types suivants :

- Perfect : toute la sonde hybride sur le gène;
- 1-2 : la sonde hybride sur le gène avec 1 ou 2 polymorphismes;
- 3-7 : la sonde hybride sur le gène avec de 3 à 7 polymorphismes;
- Badhit : la sonde possède plus de 8 polymorphismes avec le gène;
- NoMatch : la sonde n'est hybridée nulle part sur les gènes de l'organisme;
- ControlAgilent : la sonde est un contrôle de la biopuce d'Agilent (seulement pour la 2^{e} génération de biopuce *Leishmania*).

Ensuite, le programme associe chaque sonde avec le gène pour lequel elle a la meilleure hybridation. Le fichier contient alors les positions de début et de fin de ce gène dans le génome ainsi que la description de ce gène. Finalement, le fichier tabulé contient aussi la liste des gènes pour lesquels la sonde a une hybridation parfaite, 1 ou 2 polymorphismes ou de 3 à 7 polymorphismes. Ces dernières informations permettront, dans le cas de discordances entre des résultats de biopuces et de qRT-PCR, de déterminer si d'autres gènes pourraient avoir un impact sur le signal d'une sonde. Une partie de ces informations est utilisée par le protocole d'analyse décrit plus bas.

4.4.2. La pondération des sondes

Une pondération est associée à chaque sonde de chaque biopuce incluse dans l'analyse. Cette pondération dépend de trois facteurs : la qualité du signal de chaque sonde sur la biopuce, la qualité générale de la biopuce contenant cette sonde, et la similarité entre la séquence de la sonde et la séquence cible de l'organisme hybridé.

Au cours de la quantification de la fluorescence, la qualité de chaque point est évaluée par un logiciel et elle est confirmée par des paramètres comme la proportion de pixels saturés, la variance du point par rapport à celle du bruit de fond et la variance du point par rapport aux autres points. Les points seront alors triés comme étant à accepter ou à rejeter. Le poids des points à rejeter sera ajusté à 0, ce qui les exclura de la normalisation et de l'analyse.

L'ajustement de la pondération des sondes selon la qualité de la biopuce est réalisé selon la méthode de Ritchie et ses collaborateurs (186). La qualité de chaque biopuce est notée numériquement selon la variabilité de son signal, comparativement aux autres puces à ADN de l'expérience. Pour chaque hybridation, la pondération de chaque sonde est modifiée en multipliant cette dernière par la pondération de la biopuce. Cette étape permet d'inclure, dans l'analyse, des puces à ADN de moins bonne qualité et, ainsi, de profiter du résultat de certaines sondes tout en minimisant l'impact des observations aberrantes.

Finalement, les sondes qui ne sont pas utiles pour l'analyse, puisqu'elles ont un trop grand nombre de nucléotides divergents de la séquence de l'organisme cible, sont éliminées en multipliant leur pondération par 0. Cette étape utilise le fichier SpotType pour modifier la pondération des sondes qui ont trop de différences nucléotidiques avec l'espèce étudiée.

4.4.3. Le protocole d'analyse de la biopuce de première génération

La biopuce de première génération possède certaines particularités qui ont été tenues en compte au moment de la création du protocole d'analyse. D'abord, cette biopuce a été imprimée par contact physique avec 48 pointes. Puis, chaque sonde a été imprimée deux fois sur la biopuce. Finalement, ces biopuces ont été quantifiées par le logiciel GenePix.

Si l'on considère ces éléments, la correction du bruit de fond a été faite avec la méthode *normexp* (169), la normalisation interne avec la méthode printTiplœss et la normalisation globale avec la méthode d'échelle (en anglais, *scale*). Après la normalisation, les sondes dupliquées sont combinées en faisant leur moyenne pondérée et en conservant les données sur leur variabilité.

Le protocole d'analyse de la biopuce *Leishmania* de première génération est inclus dans l'annexe 3.

4.4.4. Le protocole d'analyse de la biopuce de seconde génération

La biopuce *Leishmania* de seconde génération a certaines particularités qui la distinguent de celle de première génération. Contrairement à la biopuce de première génération, celle de deuxième génération ne supporte pas le protocole de correction du bruit de fond *normexp*, car il augmente la variabilité des résultats (187). Ainsi, la méthode *edwards* a été utilisée (188). L'impression par photolithographie nécessite une normalisation interne moins élaborée que pour la première génération, qui était imprimée par contact. Ainsi, la normalisation interne se fait par la simple méthode lœss, comparativement à la méthode printTiplœss qui est essentielle à la normalisation de biopuces imprimées avec plus d'une pointe d'impression. La normalisation par loess a été recommandée par Zaharak et ses collaborateurs (187). La normalisation globale a été effectuée avec la méthode *Aquantile*. Comme cette biopuce ne comprend pas de duplicatas des sondes, contrairement à la biopuce de première génération qui nécessitait une étape d'analyse et de réduction des répliquas. Finalement, les fichiers de résultats produits par cette analyse incluent une association plus détaillée entre les sondes et les gènes qu'elles ciblent.

Le protocole d'analyse de la biopuce *Leishmania* de seconde génération est inclus dans l'annexe 4.

4.5. Le système informatique de gestion de laboratoire

La gestion des données issues d'études transcriptomiques par biopuces reste un défi important. Ainsi, plusieurs groupes de recherche ont géré des données de biopuces grâce au concept de cahier de laboratoire électronique, aussi appelé « système informatique de

gestion de laboratoire » (*laboratory information management system* ou LIMS) (163, 189-191).

Le système BASE est l'un des premiers outils publiés adaptés aux biopuces. Cet outil Web permet l'accès à la base de données par plusieurs utilisateurs, et ainsi le partage de l'information (189, 191). Cependant, ce système ne répond pas à nos besoins, car, notamment, l'extraction des données est difficile. Néanmoins, il a l'avantage de permettre aux utilisateurs d'avoir accès aux données à partir de n'importe quel endroit. Il existe aussi des outils comme MADAM, de TIGR, qui nécessite une application Java. Cela ne correspond pas davantage à nos besoins, car l'accès à la base de données par plusieurs utilisateurs serait plus difficile.

Comme aucun des outils existants ne correspondait à nos attentes, nous avons conclu que nous devions développer notre propre outil. Nous avons commencé par déterminer nos besoins particuliers :

- Regrouper les expériences liées au même projet;
- Faciliter le téléchargement des données complètes (brutes et analysées) pour une expérience;
- Permettre la consultation de résultats par des membres du laboratoire qui n'ont pas participé à cette expérience;
- Permettre la compréhension d'une expérience par des membres du laboratoire qui n'y ont pas participé.

En gardant ces objectifs en tête, nous avons développé un portail Web permettant d'accéder à une base de données de résultats génomiques et transcriptomiques pour le projet *Leishmania*. Cet outil a été nommé *Leishmania Microarray Management and Analysis Platform* (LMMAP). Cette application a été développée en collaboration avec Sébastien Boisvert, étudiant en bio-informatique à l'Université Laval, qui a programmé LMMAP.

Les buts de cet outil sont le stockage, la normalisation, l'analyse statistique et l'interprétation des données générées par la biopuce *Leishmania*. À long terme, cet outil regroupera les données générées par la biopuce *Leishmania* afin d'en faciliter la méta-analyse. Le site Web a deux fonctions principales : une fonction de cahier de laboratoire électronique et une fonction d'analyse de données de biopuces. L'outil permet de séparer l'information par projet. Chaque projet est alors défini par les facteurs expérimentaux étudiés et plusieurs expériences de biopuces peuvent y être associées. Par ailleurs, le système est conçu pour contenir toute l'information requise par le format de données MIAME (de l'anglais, *minimum information about a microarray experiment*) (185).

La conception du cahier de laboratoire électronique suit un modèle dont la colonne vertébrale est l'interaction entre les termes « projet », « expérience », « condition » et « facteur ».

Voici la définition de ces termes :

- Projet : Groupe d'expériences permettant de répondre à une question biologique en évaluant des facteurs particuliers à ce projet;
- Expérience : Série de conditions testées avec une méthode commune (biopuces) dans le cadre d'un projet;
- Condition : Combinaison de facteurs;
- Facteur : Paramètre qui varie entre les conditions d'une expérience et qui est particulier à un projet.

Ainsi, un projet inclut une ou plusieurs expériences au cours desquelles des facteurs sont combinés en conditions afin de comparer ces conditions dans le but de répondre à la question biologique d'intérêt. Les paragraphes suivants décrivent comment un utilisateur ajoute une expérience de biopuce à la base de données.

La création d'une expérience est conditionnelle à la création d'un projet. Les principaux attributs d'un projet sont le titre, la description générale du projet, les organismes étudiés, les facteurs étudiés et des mots clés permettant de classer les projets (Figure 4.1A).

La création d'une expérience se fait en suivant le même cheminement qu'une expérience dans le laboratoire (Figure 4.1B). Des cultures cellulaires virtuelles, appelées « samples », auxquelles sont associés un ou plusieurs organismes, sont traitées par une condition pour devenir des *processed samples.* Ces dernières sont définies par un organisme, par une condition et par un protocole expérimental. Le matériel génétique des *processed samples* est extrait, puis ces dernières deviennent des *extracts,* qui sont définis par une *processed sample* et par un protocole d'extraction. Les *extracts* sont ensuite marqués et deviennent des *labeled extracts,* qui sont définis par un *extract,* par un marqueur fluorescent et par un protocole de marquage. Finalement, les *labeled extracts* sont hybridés à une biopuce, ce qui est décrit par une *hybridization.* Cette dernière est définie par un ou plusieurs *extracts,* par une biopuce, par un protocole de préhybridation, par un protocole d'hybridation et par un protocole de posthybridation. Ainsi, chaque *hybridization* correspond à une biopuce physique, c'est-à-dire à une lame de verre dans le cas des biopuces *Leishmania.* Un cahier de laboratoire électronique plus détaillé est associé à chacune de ces étapes. Il permet de noter l'information essentielle relativement aux réactifs utilisés à chacune des étapes d'une expérience.

À chaque *hybridization* peuvent ensuite être associés des *scans,* des images produites à la lecture de la fluorescence sur les biopuces. Puis, à chaque *scan* peuvent être associés des *quantifications,* un fichier produit pendant l'évaluation de l'intensité de chacune des sondes de la biopuce. Ces fichiers de *quantifications* seront utilisés au moment de la normalisation et de l'analyse statistique. Cela permet de conserver les données brutes des expériences avec suffisamment d'information pour faciliter l'analyse des données. Ainsi, les données brutes proprement dites ne sont pas ajoutées à la base de données.

La normalisation et l'analyse statistique des données sont facilitées par une intégration, invisible à l'utilisateur, de LMMAP et de la librairie LIMMA (161) au logiciel d'analyse statistique R. De plus, des paramètres de normalisation particuliers à la biopuce *Leishmania* ont été mis en place. Un tableau synthèse regroupe les résultats de l'analyse avec la

correspondance entre les sondes et les gènes de l'organisme étudié, en plus de fournir les ontologies (192) de chaque gène ainsi que son affiliation à des voies métaboliques telles que définies dans la base de données KEGG (193, 194). À l'aide d'analyses utilisant le logiciel BLAST (195), nous avons annoté chaque gène de *L. infantum* avec les ontologies et les voies métaboliques associées à ses orthologues chez *L. major*. Il est aussi possible d'observer le signal et la modulation des gènes selon leur position sur le génome de l'organisme étudié.

Finalement, l'information décrivant l'expérience est utilisée pour normaliser les données et pour les soumettre à une analyse statistique telle que décrite dans les sections précédentes. L'utilisateur peut définir le protocole de normalisation et d'analyse statistique ou utiliser le protocole par défaut, qui est adapté à la biopuce *Leishmania.* Une fois la normalisation et l'analyse complétées, les résultats de l'analyse sont disponibles en ligne. Les différentes expériences peuvent alors être comparées.

4.6. Les forces et les faiblesses de LMMAP

L'évaluation d'un cahier de laboratoire électronique ne tient pas qu'à la qualité de sa conception, mais surtout à l'adoption du système par les membres du laboratoire. C'est à partir de ce point de vue que j'évaluerai les forces et les faiblesses de notre portail LMMAP. Cependant, même si l'évaluation est faite à partir de l'usage qui en a été fait, il est possible que les sections peu utilisées de l'outil Web le soient parce que peu de formation a été donnée concernant ces sections ou parce que la conception de ces sections est déficiente.

Après cinq ans d'utilisation, j'évalue les forces de LMMAP suivantes :

- Analyse intégrée;
- Conservation des données;
- Fonctions de base faciles d'utilisation;

- Obligation pour les utilisateurs d'utiliser le cahier de laboratoire électronique a favorisé son adoption.

En opposition, les faiblesses de notre approche sont les suivantes :

- Fonctions périphériques peu utilisées, notamment la section cahier de laboratoire supplémentaire à la description des expériences;
- Exploration de données limitée et peu utilisée, car les utilisateurs préfèrent télécharger un seul fichier tabulé résumant les résultats;
- Organisme pas toujours bien défini dans les projets et les expériences, ce qui entraîne de la confusion dans certains cas.

Au final, les options de LMMAP qui répondaient au besoin initial ont été adoptées par les utilisateurs, alors que les options qui constituaient des ajouts de fonctions supplémentaires ne répondant pas aux besoins primaires des utilisateurs n'ont presque pas été utilisées.

4.7. Les études publiées utilisant la biopuce *Leishmania*

Quatre études utilisant la biopuce *Leishmania* de première génération ont été publiées. Deux de ces projets étudiaient la résistance aux antiparasitaires alors que les deux autres étudiaient les stades de vie du parasite.

Ubeda et ses collaborateurs ont étudié l'effet de la résistance au méthotrexate (MTX) sur le niveau d'expression des gènes de *L. infantum* et de *L. major* (196). Lorsqu'un seuil de détection strict était utilisé dans l'analyse, 61 gènes d'un mutant *L. infantum* résistant au MTX et 75 gènes d'un mutant *L. major* résistant au MTX étaient modulés de manière statistiquement significative. Peu des gènes modulés étaient communs entre les deux mutants, mais un groupe de 6 gènes situés sur le chromosome 6 donnait des résultats similaires entre les deux mutants. Ces gènes incluent DHFR-TS, la cible du MTX.

Par le passé, ce gène a souvent été décrit comme étant amplifié sous forme d'amplicon circulaire chez les mutants résistants au MTX. La recherche de séquences répétées flanquant ce groupe de gènes a montré que, dans le génome des deux espèces, les gènes

étaient flanqués par des séquences répétées. Ces régions de 575 nucléotides et de 837 nucléotides étaient conservées entre les deux espèces. Le gène PTR1 était lui aussi amplifié, cette fois sous forme d'un amplicon linéaire, mais seulement dans le mutant *L. infantum.* Cette amplification linéaire est aussi possible chez *L. major,* qui possède des répétitions inverses similaires (99 % d'identité) de 578 nucléotides entre les mêmes gènes sur le chromosome 23. Les chercheurs ont aussi observé que certains chromosomes étaient aneuploïdes. Finalement, les auteurs proposent deux modèles d'amplification génique. L'amplification circulaire requiert des répétitions « sens » d'une séquence d'ADN, alors que l'amplification linéaire requiert des répétitions « antisens ».

Leprohon et ses collaborateurs ont comparé des souches de *L. infantum* résistantes à l'antimoine avec une souche sensible (197). Un total de 84 gènes distincts ont été décrits comme étant différentiellement exprimés, soit environ 1 % du génome de *L. infantum.* La particularité de cette liste de gènes concerne le fait que plusieurs d'entre eux sont situés dans des régions chromosomiques proximales, ce qui suggère une expression conjointe de certains groupes de gènes. Une étude plus approfondie de ces régions génomiques a permis aux auteurs de découvrir qu'elle était flanquée de séquences répétées longues de 1,4 kb. Le parasite résistant semble aussi avoir modulé le nombre de copies de certains chromosomes afin de répondre à la présence de l'antimoine.

Rochette et ses collaborateurs ont utilisé la biopuce *Leishmania* pour comparer des amastigotes à des promastigotes de *L. major* et de *L. infantum* (198). Un total de 274 gènes ont été associés au stade promastigote de *L. infantum* et 481 à celui de *L. major.* En ce qui concerne les gènes associés au stade amastigote, 309 gènes de *L. infantum* et 301 gènes de *L. major* ont été associés à ce stade. En tout, 12,05 % des gènes communs aux deux espèces étaient associés au stade promastigote, alors que 10,50 % des gènes étaient associés au stade amastigote.

Cependant, il faut noter une différence majeure entre la culture des amastigotes de ces deux espèces pour cette étude. En effet, les amastigotes de *L. infantum* ont été obtenus dans des cellules THP1 différenciées en macrophages, alors que les amastigotes de *L. major* ont été obtenus de lésions de souris. Cette différence de provenance pourrait expliquer la régulation différente d'une portion des gènes entre les deux espèces. Néanmoins, la

comparaison entre l'expression de plusieurs gènes dans les amastigotes de *L. major* cultivés dans la lignée cellulaire THP1 par rapport à l'expression de ceux prélevés de lésions de souris a montré une forte concordance entre les deux systèmes. De même, la comparaison par qRT-PCR de différents gènes de *L. infantum* et de *L. major* a montré une divergence dans 67 % des gènes testés. Cela suggère une différence marquée entre les gènes exprimés entre les deux espèces. Il est intéressant de noter que les gènes modulés par deux espèces semblaient distribués aléatoirement dans le génome. Ils ne semblent donc pas être modulés par amplifications extrachromosomiques.

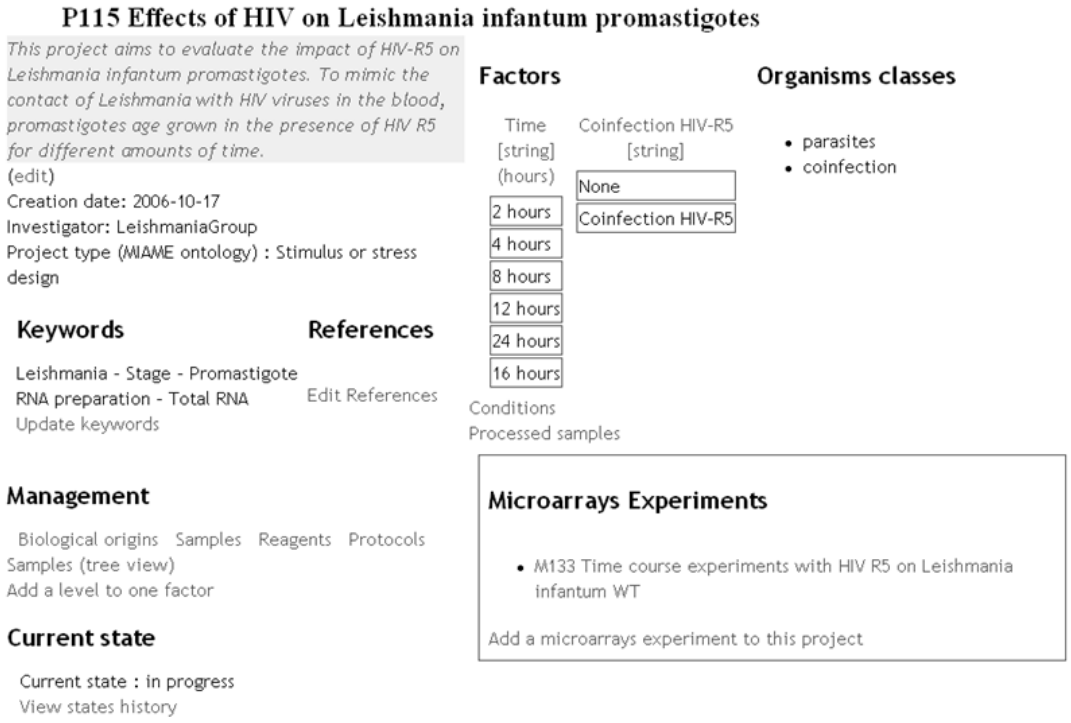
Une seconde étude sur les stades de développement de *Leishmania* a été présentée par Rochette et ses collaborateurs (47). Cette fois, les auteurs se sont intéressés plus précisément aux différences d'expression entre les formes amastigotes axéniques et amastigotes intracellulaires de *L. infantum*. Environ 40 % plus de gènes ont été modulés chez les amastigotes axéniques (518 gènes), comparativement aux amastigotes intracellulaires (309 gènes). En tout, seulement 12 % des gènes modulés étaient communs aux deux types d'amastigotes. Cette étude montre que les amastigotes axéniques devraient être utilisés avec discernement dans les études d'expression génique, car ils ne réagissent pas de manière identique aux amastigotes intracellulaires en ce qui concerne l'expression de leurs gènes.

Les quatre études résumées dans cette section ont été faites avec la biopuce *Leishmania* de première génération. En décembre 2010, un premier article utilisant la biopuce *Leishmania* de deuxième génération a été soumis à PloS Neglected Tropical Diseases (Rubens et coll., soumis). La biopuce *Leishmania* de deuxième génération a aussi été utilisée pour effectuer des hybridations comparatives de génomes afin de valider des résultats obtenus au cours du séquençage du génome de *L. tarentolae*. Ces résultats sont présentés au chapitre 5. De plus, ces six articles sont répertoriés dans l'annexe 1. Dans le futur, une troisième génération de la biopuce *Leishmania* sera créée, cette fois en considérant dans sa conception la séquence du génome de *L. tarentolae*. Cependant, comme le séquençage à haut débit devient de plus en plus accessible, il est possible que les études de transcriptomique du futur utilisent ces technologies plutôt que les biopuces.

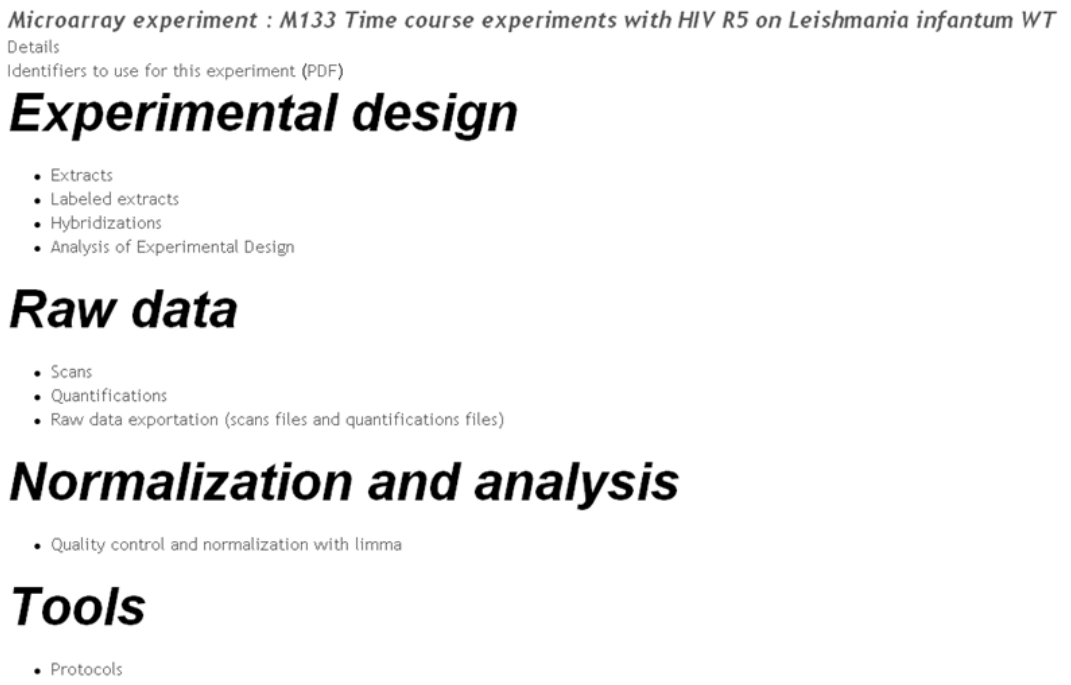
4.8. La figure

Figure 4.1. Captures d'écran du logiciel LMMAP pour une expérience fictive de biopuce. (A) Présentation générale d'un projet dans LMMAP. (B) Les étapes d'une expérience de biopuce dans LMMAP.

A.



B.



5. Le génome de *Leishmania tarentolae*

5.1. Le résumé de l'article

5.1.1. Le résumé en français

Le parasite *Leishmania (sauroleishmania) tarentolae* a été découvert chez le lézard *Tarentola mauritanica.* Cette espèce de parasite n'est pas pathogène pour l'humain, mais elle est utilisée en laboratoire pour des analyses moléculaires et pour la surproduction de protéines.

Le génome de la souche Parrot-TarII de *Leishmania tarentolae* a été séquencé à l'aide de méthodes de séquençage à haut débit pour obtenir une couverture de 23 fois la taille du génome. Il a été assemblé *de novo,* puis les contigs ont été ordonnés selon leur homologie avec le génome de *L. major* Friedlin. C'est la première fois qu'une souche de kinétoplastidé non pathogène pour l'humain est séquencée, ce qui permettra de la comparer avec les espèces de *Leishmania* pathogènes pour l'humain.

Le génome de *L. tarentolae* est fortement synténique avec celui de *L. infantum* et de *L. major,* quoiqu'il soit plus proche de ce dernier. Globalement, plus de 90 % des gènes de *L. tarentolae* sont partagés avec les autres espèces de *Leishmania* déjà séquencées. En tout, 250 gènes de *L. major* sont absents de *L. tarentolae* et une forte proportion d'entre eux sont exprimés dans le stade intracellulaire des *Leishmania* pathogènes pour l'humain. Plusieurs de ces gènes sont associés au transport vésiculaire et aux antioxydants. Cela suggère que *L. tarentolae* pourrait être moins apte à survivre à l'intérieur de cellules de mammifères. D'un autre côté, deux familles de gènes sont en nombre de copies beaucoup plus élevé chez *L. tarentolae* que chez les autres espèces de *Leishmania,* soit la leishmanolysine (GP63) et l'antigène de surface promastigote (PSA31C).

Dans l'ensemble, les gènes de *L. tarentolae* suggèrent que le parasite est plus adapté au stade promastigote qu'au stade amastigote. Cela pourrait expliquer, en partie, pourquoi *L tarentolae* est incapable de se multiplier dans des macrophages et pourrait suggérer un mode de vie promastigote dans le lézard.

5.1.2. Abstract

Leishmania (Sauroleishmania) tarentolae was first isolated in the lizard *Tarentola mauritanica*. This species is not pathogenic to humans but is often used as a model organism for molecular analyses or protein production.

The *Leishmania tarentolae* Parrot-TarII strain genome sequence was resolved to an average 16-fold mean coverage by next-generation DNA sequencing technologies. This is the first nonpathogenic to humans kinetoplastid protozoan genome to be described thus providing an opportunity for comparison with the completed genomes of pathogenic *Leishmania* species.

A high synteny was observed between all sequenced *Leishmania* species. A limited number of chromosomal regions diverged between *L. tarentolae* and *L. infantum,* while remaining syntenic to *L. major*. Globally, over 90% of the *L. tarentolae* gene content was shared with the other *Leishmania* species. We identified 95 predicted coding sequences unique to *L. tarentolae* and 250 genes that were absent from *L. tarentolae.* Interestingly, many of the latter genes were expressed in the intracellular amastigote stage of pathogenic species. In addition, genes coding for products involved in antioxidant defense or participating in vesicular-mediated protein transport were underrepresented in *L. tarentolae*. In contrast to other *Leishmania* genomes, two gene families were expanded in *L. tarentolae*, namely the leishmanolysin (*GP63*) and a gene related to the promastigote surface antigen (*PSA31C*).

Overall, *L. tarentolae'*s gene content appears better adapted to the promastigote insect stage rather than the amastigote mammalian stage. This may partly explain its inability to replicate within mammalian macrophages and its suspected preferred life style as extracellular promastigotes in the lizards.

5.2. L'article

Genome analysis of the lizard parasite *Leishmania tarentolae* reveals expansion of insect-specific and loss of mammalian-specific genes in comparison to human pathogenic *Leishmania* species

Running title: *Leishmania tarentolae* whole-genome sequence

Authors: Frédéric Raymond, Sébastien Boisvert, Gaétan Roy, Jean-François Ritt, Danielle Légaré, Mario Stanke[1], Martin Olivier[2], Michel J Tremblay, Barbara Papadopoulou*, Marc Ouellette*, Jacques Corbeil*

Centre de recherche en infectiologie du Centre de recherche du CHUL and Département de microbiologie et immunologie, Faculté de médecine, Université Laval, Québec, Canada.

[1] University of Greifswald, Greifswald, Germany.

[2] McGill University, Montreal, Canada.

* Corresponding authors: Jacques Corbeil, Marc Ouellette and Barbara Papadopoulou

5.2.1. Introduction

Leishmania is an early-branching unicellular eukaryote that belongs to the Kinetoplastida order and the family Trypanosomatidae. *Leishmania* species are transmitted by the bite of female phlebotomine sand flies as extracellular flagellated metacyclic promastigotes and replicate in mammalian macrophages as intracellular aflagellated amastigotes. *Leishmania* infections represent a global health problem with 350 million people at risk, an annual incidence of 2 million and an overall prevalence estimated at 12 million people worldwide (199). At least 20 *Leishmania* species cause a large spectrum of clinical manifestations ranging from self-resolving skin lesions (*L. major*) to mucocutaneous manifestations (*L. braziliensis*) reaching to life-threatening visceral diseases (*L. donovani/L. infantum*).

Through phylogenetic analyses, *Leishmania* was divided in three distinct subgenera: the *Leishmania,* the *Viannia* and the *Sauroleishmania* (17). The classification of lizard *Leishmania* as a distinct genus was once debated (16, 200) but the molecular evidence did not support this assumption. *Leishmania (Sauroleishmania) tarentolae* was first isolated from the lizard *Tarentola mauritanica* in 1921 (15) and is probably the most widely studied *Leishmania* (*Sauroleishmania)* species. In lizards, the parasites live predominantly as promastigotes in the lumen of the cloacae and intestine or in the bloodstream (201). Amastigotes, either free or inside monocytes, are rarely observed in lizards (16, 201), although both free promastigotes and amastigotes from the blood were reported (202). The ability of *L. tarentolae* to develop into amastigote forms in the lizard is still debated but, as for several *Leishmania* species infecting lizards (201), *L. tarentolae* is able to enter human phagocytic cells and differentiate into amastigote-like forms. However, there is no clear evidence for their efficient replication within macrophages (19, 20, 203). Because of its rapid growth in defined media and lack of pathogenicity to humans, *L. tarentolae* has been used widely as a model organism for studies on gene amplification (204-206) or RNA editing (207). Furthermore, *L. tarentolae* has been used as a platform for the production of recombinant proteins (208) and as a potential vaccine candidate (19, 20).

The genomes of two *Leishmania (Leishmania)* species, *L. major* and *L. infantum,* and one *Leishmania (Viannia)* species, *L. braziliensis,* have already been completed and annotated (74, 75), and more are being sequenced (www.tritrypdb.org). The 32.8 Mb genome of

L. major clone Friedlin spreads over 36 chromosomes and is presently the best annotated *Leishmania* genome (74). A 5-fold shotgun sequencing was provided for *L. braziliensis* clone M2904 and *L. infantum* clone JPCM5. The genomes of the various *Leishmania* species contain a similar number of genes, estimated at 8200 (Table 5.1). Despite the 20-100 million years of divergence within the *Leishmania* genus, a recent sequence comparison of the genomes of *L. major*, *L. infantum* and *L. braziliensis* has revealed a strong conservation of gene content and synteny across the genus (75). Comparative genomics of *L. major, L. infantum* and *L. braziliensis* have shown that approximately 200 genes were differentially distributed between the species and that only 78 genes were unique to one species (75). Notable differences were observed in *L. braziliensis,* which contains a putative RNA interference pathway and two types of transposons (TATES) and retroposons (SLACS) absent in the two other species.

Here we sequenced and assembled *de novo* the *Leishmania tarentolae* strain Parrot-TarII using next-generation high-throughput DNA sequencing technologies. The comparison of the *L. tarentolae* genome with the published genomes of pathogenic *Leishmania* species revealed a high degree of synteny. We identified 95 predicted coding sequences unique to *L. tarentolae* and 250 genes present in the pathogenic species but absent in *L. tarentolae*.

5.2.2. Materials and methods

5.2.2.1. Sample preparation and sequencing

Promastigotes of *Leishmania tarentolae* (strain Parrot-TarII) were grown up to the late log phase in SDM-79 (Schneider's Drosophila medium) supplemented with 5 µg/ml hemin and 5% heat-inactivated fetal calf serum (FCS) (Multicell, Wisent Inc.). High molecular weight genomic DNA was extracted from parasites after chemical lysis with SDS (1%) followed by two sequential phenol extractions, proteinase K and RNase A treatments. Chromosomal DNA was further purified on cesium chloride gradients to limit the extent of kinetoplastid DNA inclusion (209). Purified *L. tarentolae* genomic DNA was sequenced with next-generation sequencing technologies using paired (2.5 kb inserts) and unpaired GS-FLX or Titanium sequencing procedures (Roche 454). GS-FLX and Titanium sequencing were performed at the McGill University and Génome Québec Innovation Centre, Montréal,

Canada. Genomic sequences using unpaired Solexa/Illumina sequencing were performed at the Netherlands Cancer Institute. Sequencing runs are described in Table 5.2.

5.2.2.2. Assembly

Sequences produced on the GS-FLX and Titanium sequencers were assembled using Newbler 2.0 (Roche). The N_{50} scaffold length of the *L. tarentolae* assembly was 61,619 bp. Sequences produced by Solexa/Illumina were assembled with the Velvet version 0.7.55 software (141). Additional validation of the assembly was performed using our own *de novo* assembler Ray (142). The assemblies were merged using the minimus2 scaffolder found in the AMOS package version 2.0.8 (210). Chromosome-like scaffolds were created by comparing gene order in the *de novo* assembled scaffolds and contigs to the gene order in the version 5.2 of the reference *L. major* genome using BLAST and custom python scripts.

5.2.2.3. Annotation

Gene identification was performed by comparing the assembled *L. tarentolae* genome sequence to the putative genes of *L. infantum, L. major* and *L. braziliensis* using BLAST 2.2.21 (195). Sequences from the assembly were compared to the coding sequences (CDS) of the three other species and all *L. tarentolae* open reading frames (ORFs) larger than 100 nucleotides were translated and compared to the gene sequences of the three other species. Similar analysis was also performed on *T. cruzi* and *T. brucei.* Another set of genes was predicted with Augustus using evidence from protein homology to *Trypanosomatidae* and an *ab initio* model trained for *L. tarentolae* (148, 211, 212). A consensus set of genes was compiled by pooling the two annotation sets, which were filtered in order to identify the most probable genes in the *L. tarentolae* sequence. All putative genes were mined for domain structures and motifs using HMMER 2.0 and the pfam database (156, 213). Further gene ontology (GO) analyses were performed using Blast2GO (214). Additional phylogenetic analyses were conducted in MEGA4 (215). Synteny maps were drawn using the R software (http://www.r-project.org/) based on chromosome comparisons with BLAST (blastn).

5.2.2.4. Orthologous group identification

Putative *L. tarentolae* genes were compared to genes from other sequenced *Leishmania* species using the OrthoMCL 2.0 web tool (216, 217). Genes from *L. major, L. infantum* and *L. braziliensis* have already been clustered in groups of orthologs, which are publicly available in the OrthoMCL database (216). Each group of orthologous genes contains orthologs (related genes found in different species) and paralogs (related genes found within a species). Orthologous groups may include whole gene families, subfamilies or single genes, depending on the extent and variability of the gene family. Genes qualified as absent from *L. tarentolae* were validated manually by searching the sequence of this gene in the original reads and by assembling the matching reads using CAP3 (218).

5.2.2.5. Validation of copy number variation

Reads were mapped onto the *L. tarentolae* scaffolds using the BWA software (219). The number of reads corresponding to each nucleotide position was calculated and the mean read coverage for each gene was estimated. For each orthologous group (OG), the read coverages of all genes were summed to estimate the total coverage for each OG. The total read coverage of OG was compared to the number of genes in the OG. Orthologous groups for which copy number variation were observed between *L. tarentolae* and the other species were manually inspected to validate that total coverage of OG confirmed the number of genes.

5.2.2.6. Southern blots

Promastigotes of *L. tarentolae* strain Parrot-TarII, *L. tarentolae* strain S125, *L. major* strain Friedlin, *L. infantum* strain JPCM5 and *L. braziliensis* strain WHOM/BR/75/M2904 were grown at 25 °C in SDM-79 medium supplemented with 10% heat-inactivated fetal calf serum and 5 µg/ml of hemin. Total DNA from each culture was prepared with DNAzole (Invitrogen), digested with XhoI restriction enzyme (New England Biolabs, Pickering, ON, Canada) and run on standard agarose gels. Southern blot analyses with [α-^{32}P]-dCTP labelled DNA probes were performed according to standard protocols (209). All probes were generated by polymerase chain reaction using primer sets listed in Table 5.3. For each target, we generated a specific probe for each *Leishmania* species, which were

co-hybridized on a blot of total digested genomic DNAs. Equal amount of DNA was layered for each strain and monitored by hybridization with the single copy *PTR1* gene (not shown).

5.2.2.7. Comparative genomic hybridization

Leishmania whole-genome DNA microarrays used in CGH experiments, which included 8100 60-mer probes that were designed to hybridize all genes of *L. major* 5.2 and *L. infantum* 3.0a, were obtained from Agilent Technologies (Mississauga, ON, Canada). Microarray platform details and probe sequences were deposited in the GEO database under the accession GPL11330. Sample preparation, prehybridization and hybridization steps were performed as previously described (196). Normalization and data analysis were done in R with LIMMA 2.7.3. (161). Multiple testing correction was done using the false discovery rate (FDR) method and probes were considered significant when $p < 0.05$ and log-2 ratio > 2. The entire dataset was deposited in GEO under the reference number series GSE27184.

5.2.2.8. H_2O_2 IC50 assay

Promastigotes of *L. major* LV39, *L. infantum* MHOMMA#67#ITMAP-263 and *L. tarentolae* S125 were grown at pH 7.0 and 25 °C in SDM-79 medium supplemented with 10% fetal bovine serum, 5 µg/ml hemin, and 5µM biopterin (Sigma-Aldrich, St-Louis, MO). *Leishmania* promastigotes (5×10^6) were inoculated in 5 ml medium and H_2O_2 (Rougier Pharma, Mirabel, QC, Canada) was added at various concentration (100 µM, 250 µM, 500 µM, 1000 µM, 2500 µM and 5000 µM). OD^{600nm} values were taken after 72 h. Each curve was performed in triplicate. IC50 were calculated for each curve and the mean and 95% confidence intervals were calculated.

5.2.2.9. Gelatin zymography assay

Protease activity of parasite GP63 was assayed as previously described by 10% sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) incorporated with gelatin (1 mg/ml) (270). With some modification to the previous protocol, the gels were loaded with 10 µg of parasite lysates that were added to SDS-PAGE sample buffer (15.6 mM Tris pH 6.8, 2% SDS, 10% glycerol, 0.05% bromophenol blue). Electrophoresis

was performed at a constant current of 20 mA at room temperature. After electrophoresis, SDS was removed by incubation with washing buffer (2.5% Triton X-100 in 50 mM Tris pH 7.4, 5 mM CaCl2, 1 µM ZnCl2) for 1h on a rotating shaker at room temperature. Then, the gels were briefly rinsed twice with deionized water and incubated in a buffer containing 50 mM Tris pH 7.4, 5 mM CaCl2, 1 µM ZnCl2, overnight at 37 °C. After incubation, gels were stained 30 min in 0.5% Coomassie brilliant blue R-250 in 30% ethanol and 10% acetic acid, and destained few hours in a solution containing 30% ethanol and 10% acetic acid.

5.2.2.10. Western blot

Promastigotes were collected at day 7 by centrifugation and washed 3 times in PBS. They were lysed with cold buffer (Tris-HCl 50 mM, EDTA 0.1 mM, EGTA 0.1 mM, Igepal 1%, 2β-mercaptoethanol, 100 µg/ml aprotinin and 25 µg/ml leupeptin) for 45 min on ice. Proteins were dosed by Bradford assay (Bio-Rad, Mississauga, Ontario, Canada) and 50 µg were separated by SDS-PAGE (10% acrylamide), and transferred to PVDF membranes. Western blot were adapted from (258). Membranes were blocked in Tris Buffer Saline and Tween 0.1% (TBS-T) containing 5% BSA for 1h and incubated overnight with monoclonal antibody clone #253 against GP63 (271). After washing with TBS-T (2 times for 5 min), membranes were incubated 1h with anti-mouse HRP-conjugated antibody (GE Healthcare, Mississauga, ON, Canada). After washing with TBS-T (3 times for 5min), they were developed by chemiluminescence immunodetection with ECL reagents (Thermo Fisher Scientific, Rockford, IL) and autoradiography.

5.2.3. Results

5.2.3.1. *Leishmania tarentolae* genome sequencing

The genome of *L. tarentolae* strain Parrot-TarII was resolved using high-throughput sequencing technologies (Table 5.2**)** to a 16-fold mean coverage and 23-fold peak coverage. The assembled *L. tarentolae* genome contains a total of 30,440,719 bases with 95.1% of the GS-FLX and Titanium reads found in 773 scaffolds and the remaining 4.9% distributed in 2,499 contigs. Reads obtained by Illumina sequencing were also incorporated in the final sequence to assist in the assembly. After *de novo* assembly, sequences of specific

chromosomes were built using contigs and scaffolds based on their homology to *L. major*. Directed assembly on *L. major* allowed the mapping of 29,862,062 bases (98.1%) leaving 578,657 bases (1.9%) in 1315 small contigs. A summary of the sequencing statistics of the *L. tarentolae* genome and the other published *Leishmania* spp. genomes are presented in Table 5.1. The *L. tarentolae* genome has 36 chromosomes but appears smaller and has the lowest coding GC content of all sequenced *Leishmania* species (Table 5.1).

5.2.3.2. *Leishmania tarentolae* is syntenic with other *Leishmania* species

De novo assembled contigs and scaffolds were generally syntenic with both *L. major* and *L. infantum*. Most differences between *L. tarentolae* and the other species consisted in gene insertions or deletions distributed randomly or in tandem arrays throughout the genome. Although *L. tarentolae* possesses similar percent identity with *L. infantum* and *L. major* (Table 5.1), its synteny is closer to the latter. For example on chromosome 28, a stretch of 90 kb is syntenic between *L. tarentolae* and *L. major* but the gene ordering is different in *L. infantum* (Figure 5.1A). *De novo* assembled scaffolds are, in this case, long enough to confirm that *L. tarentolae* is more synthenic to *L. major*. Similar differences were found at the proximal region of chromosome 7 (Figure 5.1B) and at the distal end of chromosome 35 (Figure 5.1C), both suggesting greater synteny to *L. major* than to *L. infantum*. Other loci where *L. tarentolae de novo* assembled scaffolds were syntenic to *L. major* but not to *L. infantum* can be found on chromosomes 7, 9, 11, 12, 13, 32 and 35.

5.2.3.3. *Leishmania tarentolae* gene content

Genome annotation of *L. tarentolae* indicated a total of 8201 putative protein-coding genes, a number similar to the other sequenced *Leishmania* species (Table 5.1). Annotation was performed by comparing the *L. tarentolae* genome to other *Leishmania* and *Trypanosoma* species along with *ab initio* annotation using the Augustus software trained for *Leishmania* gene detection (148, 211, 212). Given that assembly of repeated gene clusters is more difficult, the count of genes found in *L. tarentolae* may be biased, especially for genes present in high copy number.

Using the OrthoMCL web tool, the set of *L. tarentolae* putative genes was compared to the OrthoMCL database (www.orthomcl.org) in order to assign each gene to a group of

orthologs (216). This allowed us to readily compare the gene content of the four sequenced *Leishmania* species, determine which genes are unique to a given species and calculate which orthologous group (OG) of genes vary in copy number between the different species. These results were further confirmed by interspecies comparative genomics hybridization (CGH) microarrays, read depth analysis and, in selected cases, by southern blots analyses. The gene content of *L. tarentolae* is highly similar to the three pathogenic *Leishmania* species sequenced, which contain a similar number of OG (Table 5.1).

Figure 5.2 compares the gene content of *L. tarentolae* to other *Leishmania* species with emphasis on *L. major*. Overall, 7331 OG are shared by *L. tarentolae* and *L. major.* Of these, 7225 OG (7662 genes in *L. tarentolae* and 7845 genes in *L. major)* have a similar copy number in *L. tarentolae* and at least another *Leishmania* species, 20 OG (32 genes in *L. tarentolae* and 131 genes in *L. major)* have a lower copy number in *L. tarentolae* than the three other species (Table 5.4) and 86 OG (363 genes in *L. tarentolae* and 133 in *L. major)* have a higher copy number in *L. tarentolae* than the three other species (Table 5.5). More than a third of the orthologous groups with varying copy numbers have a putative function (see Figure 5.2 and Tables 5.4 and 5.5).

A total of 250 *L. major* genes distributed between 188 orthologous groups were found to be absent from *L. tarentolae* (Figure 5.2 and Table 5.6). Of these, 83 OG were shared by the three pathogenic species, 74 OG by *L. major* and *L. infantum,* 5 OG by *L. major* and *L. braziliensis* and 26 OG were unique to *L. major* (Figure 5.2).

A total of 73 OG (95 genes) were unique to *L. tarentolae* (Table 5.7). From these, 31 OG had orthologs in other non-*Leishmania* species, including 29 in *Trypanosoma* spp. We also found 42 OG (65 genes) that were sequence orphans.

5.2.3.4. Genes absent from *L. tarentolae* or present in lower copy number compared to the pathogenic *Leishmania* spp.

L. tarentolae lacks several genes coding for proteins implicated in trafficking. Indeed, the beta1/beta2-adaptins (LmjF11.0990; LmjF36.5595), mu-adaptin (LmjF31.3035) and the epsilon-adaptin (LmjF30.1545) (Figure 5.2 and Table 5.6) were absent from *L. tarentolae.* Adaptins are involved in the formation of clathrin-associated adaptor protein (AP)

complexes, which play a key role in the transport of proteins by regulating the formation of transport vesicles as well as cargo selection between the trans-Golgi network, endosomes, lysosomes and the plasma membrane (220, 221). The calcium-dependent membrane binding proteins copines (LmjF28.1190) and Ras-like small GTP-binding proteins (LmjF36.1820), both involved in cell signaling and/or membrane trafficking/transport pathways and exocytosis, were also absent from *L. tarentolae.* The endosomal/lysosomal membrane-bound acid phosphatase (LmjF28.2650), potentially involved in intracellular trafficking (222), is also missing from *L. tarentolae*. Also, *L. tarentolae* lacks the phosphatidylinositol 3-kinase 2 gene (LmjF14.0020) and one phosphatidylinositol-4-phosphate 5-kinase gene (LmjF26.2495) (Figure 5.2 and Table 5.6) whose activities were linked to a diverse set of key cellular functions, including intracellular trafficking (223). The Tubby protein 1, that has been reported to bind phosphatidylinositol 4,5-bisphosphate on the plasma membrane and to facilitate macrophage phagocytosis (224), is not present in *L. tarentolae.*

L. tarentolae lacks three of the seven subunits of the Arp2/3 complex, notably the p40/ARPC1 (LmjF18.0920) p20/ARPC4 (LmjF02.0600) and p15/ARPC5 (LmjF05.0285) (Figure 5.2 and Table 5.6). Only one Arp2/3 complex subunit coding gene was found and annotated in the *L. tarentolae* genome (not shown). The Arp2/3 complex is required for actin polymerization and reported to associate with several other cytoskeletal components, including microtubules (225, 226). Interestingly, a number of microtubule-associated proteins (9 on chromosome 19) and a kinesin microtubule-associated protein (LmjF16.1580) are also missing from the *L. tarentolae* genome (Figure 2 and Table 5.6).

Genes reported to play a role in the resistance to oxidative stress are missing from *L. tarentolae.* These include a subtilisin (LmjF28.2380) belonging to the S8A subfamily of proteases, shown recently to process the terminal peroxidases of the trypanothione reductase system in *Leishmania* (227); a Pfpi/DJ-1-like protein (LmjF35.3910) that defends cells against reactive oxygen species and mitochondrial damage (228); and a tyrosine/dopa decarboxylase (LmjF30.2500), a precursor of catecholamines, which may act as scavengers of free radicals (229) (Figure 2 and Table 5.6). Collectively, these data suggest that *L. tarentolae* may deal less efficiently with oxidants than *L. major.* Interestingly,

L. tarentolae strains were found to be more than 3.5-fold more sensitive to hydrogen peroxide than *L. major* and *L. infantum*. Indeed, the IC50 of *L. tarentolae* to H_2O_2 was 152 ± 7 µM for *L. tarentolae* but 514 ± 115 µM for *L. major* and 574 ± 121 µM for *L. infantum.*

Gene content analysis also revealed differences in the glycoprotein content between *L. tarentolae* and the other pathogenic species. Genes involved in lipophosphoglycan (LPG) modifications were found to be either in lower copy number or absent from *L. tarentolae*. Chromosome 2 displayed important differences between *L. tarentolae* and the two other sequenced Old World species in terms of phosphoglycan transferases (Figure 5.3). This region extends on two *de novo* assembled scaffolds in *L. tarentolae,* providing reliable information on this locus. We found that phosphoglycan beta 1,3 galactosyltransferase orthologous group of genes necessary for LPG side chain addition (230) are present in 6-8 copies in other *Leishmania* species but only in 2 copies in *L. tarentolae* (Table 5.4). Within the same locus, phosphoglycan beta 1,2 arabinosyltransferase (LmjF02.0180; LmjF02.0220), known to modify LPG in *L. major* (231), is found in *L. major* and *L. infantum* but not in *L. tarentolae* (Table 5.6). *L. tarentolae* also lacks two glycosyltransferase genes (LmjF35.5250, LmjF29.2110), enzymes that catalyze the transfer of a monosaccharide unit to a glycosyl acceptor molecule. Moreover, *L. tarentolae* lacks calreticulin (LmjF31.2600), an endoplasmic reticulum (ER) chaperone ensuring the proper folding and quality control of newly synthesized glycoproteins destined for secretion or the cell surface expression (232) (Table 5.6).

Amastins, one of the largest family of surface proteins in *Leishmania* (75) shown to be expressed primarily in the intracellular stage of the parasite (55, 233), are underrepresented in *L. tarentolae* (Table 5.4). Amastins, whose function has yet to be determined, are divided into four subfamilies based on their phylogeny and genomic positioning (234). We show that *L. tarentolae* contains the amastins of the alpha, beta and gamma subfamilies that are also shared with *Trypanosoma* or *Crithidia,* but lacks all but two of the amastins of the delta subfamily (Figure 5.4A). This subfamily is expressed preferentially in the intracellular stage of the parasite (55). No new amastins were discovered in *L. tarentolae.* Interestingly,

tuzins, which are often associated on the same chromosomal locus with amastin genes in the pathogenic *Leishmania* species (55), were less diverse in *L. tarentolae* with only 4 tuzin genes compared to 6 and 28 genes in *L. infantum* and *L. major,* respectively. The amastin- and tuzin-rich region within chromosome 8 contains only two copies of the tuzin gene in *L. tarentolae* (Figure 5.4B). On *L. tarentolae* chromosome 34, amastin and tuzin genes are present in single copy as opposed to the situation in *L. major* where they are organized in tandem, suggesting co-expression (Figure 5.4C).

5.2.3.5. Gene family expansion in *L. tarentolae*

Genome comparative analysis between the *L. tarentolae* and the other sequenced *Leishmania* pathogenic species revealed that two surface-associated protein gene families were particularly enriched in the non-pathogenic *L. tarentolae.* The first family is leishmanolysin, a major surface protease, also known as GP63; which is mostly, but not exclusively, expressed in the promastigote stage of several *Leishmania* species (52, 235, 236). The GP63 orthologous group in *L. tarentolae* is highly expanded with 49 putative GP63 genes as compared to 29 in *L. braziliensis,* 7 in *L. infantum,* and 5 in *L. major* (Table 5.5). Due to the high copy number and sequence variability, assembly of the GP63 gene family was limited and resulted in several partial gene sequences. The density of read coverage for GP63 indicated portions of the gene that are present in *L. tarentolae* are well conserved, while other regions were less represented, further supporting a sequence diversity of GP63 in *L. tarentolae* (Figure 5.5A). GP63 sequence diversity suggests that the function of GP63 in *L. tarentolae* may differ from the other species. Indeed, a gelatin zymography assay showed no protease activity in *L. tarentolae,* in contrast to the pathogenic *Leishmania* species (Figure 5.6).

The second surface protein gene family that was expanded in *L. tarentolae* is distantly related to the promastigote surface antigen proteins (PSA) located on chromosome 31. This family has recently been described (237) and the *L. tarentolae* genes are orthologous to PSA31C. There are 63 paralogs in *L. tarentolae,* while only a single copy is present in *L. major* and *L. infantum* and two copies in *L. braziliensis* (Table 5.5). *L. tarentolae* PSA proteins seem to contain only the leucine-rich repeated motif (LRR) (Figure 5.5B), a

domain often implicated in the binding to other proteins or glycolipids. This gene is suspected to be involved in host-pathogen interactions (237).

Smaller copy number variations were observed for 84 additional orthologous groups, which are shown in Table 5.5.

5.2.3.6. Genes unique to *L. tarentolae*

We found 73 OG (95 genes) unique to *L. tarentolae* (Figure 5.2 and Table 5.7). Of these groups, 31 (42%) had OrthoMCL orthologs in species other than *L. infantum, L. major* and *L. braziliensis,* including *Trypanosoma* spp. (26 groups; 36%). Putative functions could be ascribed to 10 OG (14%). These include a C3HC4 finger protein, a trafficking protein particle complex subunit 2-like protein, a La domain-containing protein, a malate dehydrogenase, an OmpA family protein, a phosphoinositide kinase, two surface protein GP63 orthologous groups, a zinc finger (Ran-binding) family protein, and a poly(A) polymerase, which is different from the poly(A) polymerases found in the pathogenic *Leishmania*. The remaining OG are sequence orphans with no orthologs found in the orthoMCL database.

5.2.3.7. Experimental validation for gene content and copy number variations

We validated experimentally some of the changes highlighted from the genome sequence comparisons using comparative genomic hybridizations (CGH), southern blot and read depth analysis. Full *Leishmania* genome microarrays designed for *L. major* and *L. infantum* were used to co-hybridize *L. tarentolae* with either the *L. major* or the *L. infantum* DNA to test whether we could identify genes specific to pathogenic species or genes with altered copy number. We found that 31% to 35% of the genes absent in *L. tarentolae* showed significantly higher signals with the DNA derived from the pathogenic species, which is statistically significant when compared to the overall ratio of 8% due to better hybridization of the DNAs derived from the pathogens ($p < 0.001$, one sample test for binomial proportion). Similarly, 20% of the *L. tarentolae* genes with lower copy number were validated by the CGH experiment ($p < 0.001$) confirming in part the sequencing data. These correlations are explained by the design of the microarray that contains several probes (45%) that have more than 10 mismatches with *L. tarentolae,* and by the fact that this

microarray did not allow testing for the presence of *L. tarentolae*-specific genes. Considering these limitations, the CGH results allowed the validation of many differentially distributed genes between the *Leishmania* species. However, genes that are not validated by CGH may still be differentially distributed between the species. Genes confirmed by CGH are highlighted in Tables 5.4 and 5.6.

The correlation between sequence data and gene copy number was validated by Southern blot hybridizations for a few chosen genes (Figure 5.7). The delta-amastin subfamily is present in high copy number in *L. major* and *L. infantum* but not in *L. tarentolae* (Table 5.4). The low copy number of delta-amastins in the TarII-Parrot *L. tarentolae* strain predicted from the sequencing data and their high copy number in the pathogenic species were confirmed by southern blot hybridization (Figure 5.7A, lanes 3-5). Indeed the intense hybridization signals in the pathogenic species were due to the highly repetitive nature of a cluster of amastins giving rise to single restriction fragments (see Figure 5.7A). Interestingly, the copy number of delta-amastins was low not only in the TarII-Parrot strain (Figure 5.7A, lane 1) cultured for decades in the laboratory, but also in a more recent isolate from lizard with a limited history of in vitro culture passages (Figure 5.7A, lane 2). As a control, we used a probe recognizing the proto-delta amastin gene, which according to the sequence data is present in similar copy numbers in all species. Southern blots have indeed validated this assumption (Figure 5.7B).

In an independent set of experiments, southern blot hybridizations corroborated sequencing data for the two genes involved in LPG side chain addition and LPG modification, indicating that phosphoglycan beta 1,3 galactosyltransferase was present in lower copy number in *L. tarentolae* (Figure 5.7C) whereas phosphoglycan beta 1,2 arabinosyltransferase was absent from this species (Figure 5.7D). Also, southern blot analyses confirmed the expansion of the GP63 and PSA31C protein gene families in *L. tarentolae* as deduced from the sequencing data (Figures 5.7E and 5.7F). The heterogeneity in the hybridizing fragments gives some credence to the size diversity of GP63 (Figure 5.5A) and PSA31C (Figure 5.5B) genes, as suggested by the sequencing data. Moreover, both gene families have a high read coverage, 3413 reads per nucleotide for GP63 and 7280 reads per nucleotide for PSA31C, which supports the idea that these

genes are in high copy number (Table 5.5). Read depth analysis also supports the higher copy number of genes listed in Table 5.5 and the lower copy number of genes listed in Table 5.4. The mean read coverage for single-copy genes was 22.4 read/nucleotide.

5.2.3.8. Orthologous groups of genes varying in copy number or missing from *L. tarentolae* are developmentally regulated in pathogenic *Leishmania* spp.

We show here that some genes preferentially expressed in the amastigote stage, such as the amastins, are present in lower copy number or are absent from *L. tarentolae,* while genes reputed to have a higher expression in the insect stage (e.g. GP63 or PSA31C) are represented in higher copy number in *L. tarentolae* (Figures 5.7E and 5.7F). We performed additional bioinformatics analyses to test whether orthologous groups of genes with a variable copy number between *L. tarentolae* and the pathogenic species were stage-regulated in the later. Specific expression of these groups to *Leishmania*'s life cycle stages was based on published transcriptomics and proteomics studies (47, 105, 105, 107, 109-111, 113, 198, 238-241). Overall, 23% of the 7694 OG were associated to one or the other developmental stage of the parasite with 890 OG associated to the promastigote stage and 1013 OG associated to the amastigote stage (Table 5.8). Association to either developmental stage is indicated in Tables 5.4-5.6. Comparison of each category to overall distribution was performed using the Fisher exact test. Interestingly, 70% (16/23) of the OG with a lower copy number and ~ 34% (29/86) of the OG with a higher copy number in *L. tarentolae* were differentially expressed in either life stage of the pathogenic *Leishmania* species (Table 5.8). Remarkably, OG absent from *L. tarentolae* were mainly associated to the amastigote life stage in other *Leishmania* ($p < 0.001$). Indeed, of the 66 orthologous groups present in pathogenic *Leishmania* spp. but absent from *L. tarentolae* that were associated to one or the other developmental stage of the parasite, 18 genes were linked to the promastigote life stage, 52 to the amastigote stage, and 4 to both life stages (Table 5.8).

Examples of genes that were absent from *L. tarentolae* but preferentially expressed in the amastigote stage of other *Leishmania* species are the delta-amastins and the hydrophilic surface protein gene family HASPA1, HASPA2, HASPB1, HASPB2 (Table 5.6), known to be preferentially expressed in *L. major, L. infantum* and *L. donovani* amastigotes as determined by DNA microarray studies (105, 198). This list also includes a 3,2-trans-enoyl-

CoA isomerase, an ABC transporter-like protein, a class I nuclease-like protein, a D-isomer specific 2-hydroxyacid dehydrogenase-protein, an Ef hand-like protein, a GTPase activator protein, a pteridine transporter and a tyrosine/dopa decarboxylase. The remaining amastigote orthologous groups absent from *L. tarentolae* encode hypothetical proteins (Table 5.6).

5.2.4. Discussion

We used next-generation DNA sequencing technologies to obtain a high quality draft (242) of the genome sequence of *Leishmania tarentolae,* the first non-human pathogen trypanosomatid sequenced. Genome sequence analysis revealed that *L. tarentolae* is syntenic to the three sequenced pathogenic *Leishmania* species and that over 90% of the ~ 8200 genes are shared by all the species. However, a number of genes shown to be either important for pathogenesis or preferentially expressed in the intracellular parasitic stage in the pathogenic species or more relevant to the insect stage were either absent from *L. tarentolae* or present at variable copy number, supporting the hypothesis that some of these genes may be responsible for the reduced capacity of *L. tarentolae* to live as an intracellular parasite and its diminished pathogenic potential to humans.

Among the genes absent from *L. tarentolae,* there was a significant bias for genes expressed specifically in the amastigote intracellular stage in the pathogenic species (Table 5.8). Most of these genes correspond to hypothetical conserved or hypothetical proteins of unknown function (Table 5.6) further highlighting the gaps in our understanding of the biology of intracellular parasites. Still, well-studied genes, such as the amastin gene family, especially the delta-amastins, are in low copy number in *L. tarentolae* (only 2 found, see Figure 5.4A) whereas high numbers of these genes (12-25 members) are found in the pathogenic species. The delta subfamily contains a diverse clade that is exclusive to *Leishmania* with one locus (proto-delta-amastins) found in other trypanosomatids (234). Although the function of amastins is still unknown, most of the delta-amastins undergo a stage-specific regulation and are preferentially expressed in amastigotes (55, 101, 198, 233) and few in infective metacyclics (111). Given their specific expression in amastigotes, it has been argued that their expansion should be viewed as the adaptation of the parasite to a novel life stage after the acquisition of vertebrate parasitism (234). *L. tarentolae* also lacks

tuzins, conserved transmembrane proteins with unknown function, which are often contiguous with delta-amastins. Moreover, the majority of *L. tarentolae* genes displaying high variation in their copy number (e.g. GP63, PSA31C, phosphoglycan beta 1,3 galactosyltransferase, ama1, etc.) are developmentally regulated in the pathogenic *Leishmania.*

Two of the most investigated abundant surface constituents of *Leishmania* promastigotes are the GPI-anchored molecules lipophosphoglycan (LPG) and GP63 protease (243-245). In contrast to the mammalian parasites, lizard *Leishmania* such as *L. adleri* and *L. tarentolae* seem to lack LPG on their surface (246). Sequencing data did not point out any specific gene, absent or mutated, that could explain the lack of LPG. However, *L. tarentolae* lacks the phosphoglycan beta 1,2 arabinosyltransferase and has significantly lower copies of the phosphoglycan beta 1,3 galactosyltransferase gene compared to the pathogenic species (Figure 5.3). LPG has been implicated in many steps required for the establishment of initial macrophage infection, notably in the inhibition of phagolysosome biogenesis and in resistance to oxidants (247), but also in the development of innate immunity (248, 249). Indeed, LPG-deficient *L. major* mutants showed attenuated virulence (250, 251) although this was not the case for *L. mexicana* LPG-deficient mutants (252).

In opposition to missing genes, the major surface protease GP63, a well known virulence factor implicated in phagocytosis, parasite evasion of complement-mediated lysis, and induction of the host immune response (52, 254), is highly expanded in *L. tarentolae,* as confirmed by Southern blot (Figure 5.7E) and read coverage analyses (Figure 5.5A). In the sand fly, GP63 is thought to be implicated in nutrient acquisition and attachment of promastigotes to the gut wall (52, 253). Expansion of GP63 genes in *L. tarentolae* may thus facilitate nutrient acquisition, and in the absence of LPG, GP63 may be important for the parasite to maintain its attachment to the midgut. More recently, it was shown that during *Leishmania*-macrophage interaction, GP63 is internalized by macrophages, where it interacts and cleaves several intracellular macrophage proteins, including actin cytoskeleton regulators, protein tyrosine phosphatases, and transcription factors (54, 257-259). Collectively, these cleavage-dependent activation events lead to the down-regulation of IFN-gamma signaling and downstream macrophage activation, including NO

production (61). Interestingly, *L. tarentolae* was shown to lack protease activity (261), as also validated on the sequenced strain by gelatin zymography (Figure 5.6). While GP63 genes were difficult to assemble, read analysis showed that some portions of the gene were underrepresented or had increased sequence variability, while other portions were conserved (Figure 5.5A). It is therefore possible that GP63 sequence variations may change the tertiary structure of the protein and affect thus its anchoring to the plasma membrane. Overexpression of multiple altered GP63 copies may further interfere with its activity and act as dominant repressor curtailing the virulence of *L. tarentolae.*

In addition to GP63 genes, *L. tarentolae* has expanded the promastigote surface antigen PSA31C for which only a single copy is present in the pathogenic *Leishmania* (237). The PSA proteins are part of eight subfamilies that have usually a signal peptide, a cysteine-rich region, leucine-rich repeats, a threonine/serine-rich region and a domain possibly acting as a GPI anchor (237). PSA were found to be either membrane-bound, secreted, or soluble but little is known about their function (237). The expression of some *PSA* genes was found to be increased in metacyclic parasites (262). In *L. tarentolae,* the only recognizable feature in the PSA31C subfamily is the leucine-rich repeats domain (Figure 5.5B). The expansion of this subfamily in *L. tarentolae* (Figure 5.7F) is striking and warrants further analysis. One hypothesis is that the PSA31C gene has expanded in lizard parasites to facilitate their survival as promastigotes either in the insect vector or the lizard.

Another interesting feature of the *L. tarentolae* genome is the lack of a number of genes encoding key functions of vesicular protein trafficking. In fact, *L. tarentolae* lacks the large adaptin subunits (beta1/beta2) and the medium-sized mu-adaptins that are part of the heterotetrameric AP-1 complex, which is involved in the formation of clathrin-coated vesicles (CCVs) mediating protein transport between the trans-Golgi network and endosomes (220). *Leishmania* mutants lacking AP-1 subunits showed significant defects in Golgi structure, endocytosis, or exocytic transport and displayed reduced rates of endosome-to-lysosome transport (265). Moreover, it was shown that Sigma 1- and mu 1-adaptin homologues of *L. mexicana* are required for parasite survival in the infected host (266). Also, a *Leishmania* AP3-deficient mutant was unable of transporting membrane proteins to acidocalcisomes (267). In addition, the absence of a membrane-bound acid

phosphatase localizing to endosomal/lysosomal compartments (222) and of copines, soluble, calcium-dependent membrane binding proteins with C2 domains known to be involved in cell signaling and/or membrane trafficking pathways and exocytosis (268) further suggests that *L. tarentolae* may have a deficient vesicular protein transport. In line with this, *L. tarentolae* lacks members of the Ras-like small GTP-binding proteins that have emerged as master regulators of cellular membrane transport by interacting with C2 domains of proteins (269).

Also, *L. tarentolae* lacks three of the seven subunits of the Arp2/3 complex, which has been shown to function in cellular processes ranging from cell motility, cytokinesis and endocytosis to trafficking and cell–cell communication (272). Arp2/3 complex is required for actin polymerization and it has been reported to associate with several cytoskeletal components, including microtubules. Alterations in the exocytic and endocytic systems may impact the capacity of the parasite to communicate efficiently with its intracellular environment, the macrophage phagolysosome. Some other genes known to play a role in virulence were also missing from *L. tarentolae.* Indeed, *L. tarentolae* lacks one of the two subtilisin (SUB) proteases found to process the terminal peroxidases of the trypanothione reductase system (227). Subtilisin promotes survival of *Leishmania* amastigotes by serving as a maturase for the trypanothione reductase system, which is essential to maintain redox homeostasis in the host macrophage and to protect the parasite against oxidative damage (273). Interestingly, SUB-deficient *Leishmania* showed increased sensitivity to hydroperoxides and reduced viability in mouse infection models (227). In addition, *L. tarentolae* lacks DJ-1, a multifunctional oxidative stress response protein that defends cells against reactive oxygen species and mitochondrial damage (228). This is consistent with the observed decrease of *L. tarentolae* survival in the presence of H_2O_2.

Here, we provide a high-quality draft of the genome sequence of the lizard *L. tarentolae* using next-generation sequencing (NGS) technologies. Limitations of NGS technologies are well known (274) and similar problems were observed with the *L. tarentolae* genome assembly. The size of the *L. tarentolae* genome is somewhat smaller (~ 5%) than that of the other sequenced *Leishmania* spp. due to collapse of regions with identical nucleotide sequence. Also, repeated genes such as GP63 or PSA31C had a tendency to be fragmented

in the assembly. However, these problems were attenuated by read number analysis. Homopolymers are an important cause of sequencing errors when using the 454 platform and can lead to frame shifts that make it difficult to differentiate actual sequencing errors from pseudogenes. In most cases, comparison of these specific regions to reads obtained by Illumina sequencing validated the presence or absence of genes. Most importantly, the current sequence draft provides additional research avenues to investigate *Leishmania* pathogenesis. The absence of LPG, acid phosphatase, and delta-amastin glycoproteins on the surface of *L. tarentolae* combined to GP63 lacking most likely its protease activity could lead to more vulnerable parasites, which may be more sensitive to complement-mediated lysis with a diminished capacity of survival within the host macrophage. In addition, a possible defect in the vesicular trafficking of glycoproteins, plasma-membrane proteins and secreted proteins may have important consequences on the ability of *L. tarentolae* to survive as an intracellular parasite. Furthermore, *L. tarentolae* lacks several proteins of unknown functions (including the delta-amastins) shown to be expressed preferentially in the amastigote stage. The absence of these proteins may explain why *L. tarentolae* cannot replicate efficiently in mammalian macrophages. In summary, this study provides insights into the reconstruction of the steps leading to increase adaptation for intracellular parasitism and suggests a number of hypothesis that can be experimentally tested.

5.2.5. Acknowledgments

We thank Arno Velds and Ron Kerkhoven, Netherlands Cancer Institute, for providing Solexa runs and Dr. Ken Dewar at the McGill University and Genome Quebec Innovation Centre for generating the Roche/454 sequencing data included in the genome assembly. We thank also Prof. Jean-Pierre Dedet and Prof. François Pratlong at Université de Montpellier, France, for sending us the *L. tarentolae* S125 isolate from the *Leishmania* Biological Resource Centre and Centre national de références des *Leishmania* and for useful discussions on the lack of consensus for intracellular amastigote stage of lizard *Leishmania.* We are thankful to Prof. Steve Beverley, Washington University School of Medicine, St. Louis, Missouri, USA, for discussions on NGS and for suggestions on kinetoplast DNA removal. FR was the recipient of a CIHR studentship. JC, MJT and MO are holders of

Tier 1 Canada Research Chairs. This work was funded in part by the CIHR group grant GR14500 to MOu, BP, JC, MJT and MOl.

5.2.6. Tables

Table 5.1. Summary of sequenced *Leishmania* spp. genomes.

	L. tarentolae	***L. major (V5.2)***	***L. infantum (V3a)***	***L. braziliensis (V2)***
Strain	*Parrott-TarII*	*Friedlin*	*JPCM5*	*WHOM/BR/75/M2904*
Number of chromosomes	36	36	36	35
Genome size (bp)	30,440,719	32,816,678	32,134,935	32,005,207
Overall G+C content (%)	57.2%	59.7%	59.3%	57.8%
Coding G+C content (%)	58.4%	62.5%	62.5%	60.4%
Number of coding genes in database (time of study)	8201	8304	8216	8133
Number of orthologous groups	7449	7530	7506	7353
Mean nucleotide identity with *L. tarentolae*	-	84.9%	85.0%	79.2%
Mean amino acid identity with *L. tarentolae*	-	81.9%	82.4%	74.8%
Reference	Current study	Ivens *et al.*, 2005	Peacock *et al.*, 2007	Peacock *et al.*, 2007

Table 5.2. Description of *Leishmania tarentolae* sequencing runs with statistics.

Run #	Total reads	Total nucleotides	Mean length (nt)	Paired reads*	Method
1	363,762	76,440,590	210	No	GS-FLX
2	425,167	102,230,855	240	No	GS-FLX
3	57,726	13,668,451	237	No	GS-FLX
4	479,778	115,173,949	240	No	GS-FLX
5	4,342	978,720	225	1,597	GS-FLX Paired ends
6	342,918	68,909,190	201	47,616	GS-FLX Paired ends
7	114,301	25,227,697	221	37,752	GS-FLX Paired ends
8	1,005,839	311,111,682	309	No	Titanium
9	11,650,003	687,350,177	59	No	Solexa (enzymatic)
10	9,291,243	334,484,748	36	No	Solexa (sheared)
11	15,437,942	910,838,578	59	No	Solexa (sheared)

* Paired reads were 2500 nucleotides distant.

Table 5.3. Sequence of primers used to amplify probes for southern blot hybridizations.

Gene description (Figure number)	Probe length (bp)	Species	Gene	Forward primer	Reverse primer
Delta-Amastin (Figure 5.6A)	538	*L. major*	LmjF34.1660	GTTGTCTACGTGGTCGTGCA	ACTCCGGCTTCTCGTTC
		L. infantum	LinJ08_V3.0710	TCGGTATTATTCTCTACGCGATCC	CGTCACCCTTACCGGACT
		L. braziliensis	LbrM20_V2.1080	TATCCCGCTGTTAGTCTACGTG	CTCCGTTTGCTCCACCTGG
		L. tarentolae	LtaP34.0070	CCTTTTGGTGCTGGTGG	GTGCTTTGTTTGTCCATCCTCG
Proto-delta Amastin (Figure 5.6B)	440	*L. major*	LmjF34.0970	ACGATATGGGGCTTCAAGGA	ATGAAGCAGAGGCYCCAG
		L. infantum	LinJ34_V3.1040	ACGATATGGGGCTTCAAGGA	ATGAAGCAGAGGCYCCAG
		L. braziliensis	LbrM20_V2.2870	GACGGTGTGGGGCTTGAAG	ATGAAGCAGAGGCYCCAG
		L. tarentolae	LtaP34.0100	ACGATATGGGGCTTCAAGGA	ATGAAGCAGAGGCACCAAC
Phosphoglycan beta 1,3 galactosyltransferase (Figure 5.6C)	865	*L. major*	LmjF25.2460	GTGTTCGCGACGACATCTGC	GCAGTGTACGCGTAGTTCAGCCA
		L. infantum	LinJ25_V3.2570	GTGTTCGCGACAACATCTGC	GCAGTGTACGCGTAGTTCAGCCA
		L. braziliensis	LmjF25.2460	GTGTTCGCGACGACATCTGC	GCAGTGTACGCGTAGTTTAGCCA
		L. tarentolae	LtaP07.1340	CGTGTTCGCAAGGACATCTGC	GCAGTGTACGCGTAGTTCAGCCA
Phosphoglycan beta 1,2 arabinosyltransferase (Figure 5.6D)	695	*L. major*	LmjF02.0180	GAAACCAGACGGGTACGCAAG	TGTCACGCACAGCCGCAGTC
		L. infantum	LinJ02_V3.0190		
GP63 (Figure 5.6E)	462	*L. major*	LmjF10.0460	GCGASTTCAAGGTGCCG	GCCATGAGCTCGTCCT
		L. infantum	LinJ10_V3.0490		
		L. braziliensis	LbrM10_V2.0470		
		L. tarentolae	LtaP10.0660		
Surface antigen protein-like PSA31C (Figure 5.6F)	504	*L. major*	LmjF31.1440	TGACGTGGACTACTCGATGGT	CAGTCAGCCATACGATGCTTGGCA
		L. infantum	LinJ31_V3.1470	TGACGTGGACTACTCGCTG	CAGTCAGCCATGCGATGCTTG
		L. braziliensis	LbrM31_V2.1670	TGACGTGGACTACTCAAAGGTG	CAGTCAGCTTTGCGATGCCTTG

Table 5.4. Genes in orthologous groups with lower copy number in *Leishmania tarentolae* compared to the pathogenic *Leishmania* species.

		Number of genes in orthologous groups				Life stage association		
Orthologous group ID	**Description**	***L. major***	***L. infantum***	***L. braziliensis***	***L. tarentolae***	**Promastigote**	**Amastigote**	**Read depth**
OG4_10652	60s ribosomal protein L35, putative	2	2	2	1	Leifso *et al.* 2007		22.3
OG4_11934	Ama1 protein, putative	3*	3*	2	1		Rochette *et al.* 2008	24.6
OG4_12080	Delta-Amastin	25*	12*	15	2	Alcolea *et al.* 2009	Rochette *et al.* 2009; Alcolea *et al.* 2010; Rochette *et al.* 2008; Saxena *et al.* 2007	32.6
OG4_10729	Aminopeptidase P1, putative	2	2*	2	1			20.4
OG4_112181	Gp63, leishmanolysin	0	0	2	1			11.9
	Heat shock protein 83, Lipophosphoglycan biosynthetic protein#	18	4	4	1	Brotherton *et al.* 2010; Saxena *et al.* 2007; Leifso *et al.* 2007; Saxena *et al.* 2007; Rochette *et al.* 2008	McNicoll *et al.* 2006;Sriv	24.2
OG4_64458	Histone H2A, putative	0	3	0	1	Alcolea *et al.* 2010		23.7
OG4_10036	Histone H3	7	5	5	2	Rochette *et al.* 2008		80.9
OG4_10043	Histone H4	7	8	13	1	Alcolea *et al.* 2010	Srividia *et al.* 2007	103.1
OG4_14227	Hydrolase, alpha/beta fold family, putative	2	2	2	1		Rochette *et al.* 2008; Rochette *et al.* 2009	19.0
OG4_18410	Hypothetical protein, conserved	2	3*	2	1		Rochette *et al.* 2009	24.7
OG4_21149	Hypothetical protein, conserved	2	2	2	1		Rochette *et al.* 2009; Saxena *et al.* 2007	22.6
OG4_29287	Hypothetical protein, conserved	2	2	2	1			20.5
OG4_12208	Hypothetical protein, conserved	2	2	2	1	Rochette *et al.* 2008; Leifso *et al.* 2007		15.3
OG4_15862	Hypothetical protein, unknown function	9*	8*	7	4			89.6
OG4_10070	Long chain fatty Acyl CoA synthetase, putative	5*	5*	5	4	Rochette *et al.* 2008		85.1
OG4_13314	Phosphoglycan beta 1,3 galactosyltransferase	8*	6*	6	2	Alcolea *et al.* 2009	Rochette *et al.* 2009; Rochette *et al.* 2008; Srividia *et al.* 2007; Saxena *et al.* 2007	42.4
OG4_20578	Phosphoprotein phosphatase, putative	2	2	2	1		Rochette *et al.* 2009	23.3
OG4_47918	SHERP	3	2	0	1	Alcolea *et al.* 2010; Leifso *et al.* 2007; Rochette *et al.* 2008	Rochette *et al.* 2009; Rochette *et al.* 2008; Saxena *et al.* 2007	21.0

		Number of genes in orthologous groups				Life stage association		
Orthologous group ID	**Description**	***L. major***	***L. infantum***	***L. braziliensis***	***L. tarentolae***	**Promastigote**	**Amastigote**	**Read depth**
OG4_11186	Sucrose hydrolase-like protein;beta-fructofuranosidase-like protein	6*	6*	5	4	Rochette *et al.* 2008; Saxena *et al.* 2007; Rochette *et al.* 2008; Rochette *et al.* 2008	Rochette *et al.* 2008	88.3
OG4_50672	Surface antigen-like protein	0	0	3	1			19.2
OG4_17051	Tuzin-like protein	22*	0	0	1			18.9
OG4_10108	Uracil phosphoribosyltransferase, putative;uridine kinase-like protein	2*	2*	3	1			18.7

* Genes in higher copy number in *Leishmania* pathogenic species than in *L. tarentolae* as determined by CGH.

Table 5.5. Genes in orthologous groups with higher copy number in *Leishmania tarentolae* compared to the pathogenic *Leishmania* species.

		Number of genes in orthologous groups				Life stage association		
Orthologous group ID	**Description**	***L. major***	***L. infantum***	***L. braziliensis***	***L. tarentolae***	**Promastigote**	**Amastigote**	**Read depth**
OG4_23961	Hypothetical protein, unknown function (related to surface antigen protein)	1	1	2	63			7280.1
OG4_10176	Major surface protease gp63, putative; GP63, leishmanolysin	5	7	29	49	Leifso *et al.* 2007; Rochette *et al.* 2008; Rochette *et al.* 2008	Rochette *et al.* 2008	3413.7
OG4_10419	40s ribosomal protein S6, putative	2	2	2	3	McNicoll *et al.* 2006		64.3
OG4_10152	Acetyl-CoA synthetase, putative	2	2	2	3	Rochette *et al.* 2008	McNicoll *et al.* 2006	80.8
OG4_10915	Acetylornithine deacetylase-like protein; glutamamyl carboxypeptidase, putative	2	2	2	3	Rochette *et al.* 2008		66.3
OG4_12086	Amino acid permease, putative; amino acid transporter aATP11, putative	4	3	3	5	Alcolea *et al.* 2009; Rochette *et al.* 2008	Rochette *et al.* 2009; Rochette *et al.* 2008	239.9
OG4_25781	Amino acid transporter, putative	2	2	2	3	Sriv; Leifso *et al.* 2007; Rochette *et al.* 2008	Srividia *et al.* 2007	111.1
OG4_10581	Beta-adaptin, putative	1	1	1	2			48.5
OG4_10223	Branched-chain amino acid aminotransferase, putative	1	1	1	3			86.2
OG4_10522	cAMP specific phosphodiesterase, putative	1	2	2	3	Rochette *et al.* 2008		101.8
OG4_11071	Citrate synthase, putative	2	2	2	3			75.0
OG4_36553	DEAH-box RNA helicase, putative	1	1	1	2			39.6
OG4_17226	Endonuclease/Exonuclease/phosphatase, putative	1	1	1	2			45.8
OG4_10922	Eukaryotic translation initiation factor 2 subunit, putative	1	1	1	2			58.6
OG4_21087	Glucokinase 1-like protein	1	1	1	3			52.9
OG4_14440	Glucose transporter, putative	4	4	4	5	Alcolea *et al.* 2010; Leifso *et al.* 2007; Rochette *et al.* 2008	Rochette *et al.* 2009	376.7
OG4_26454	Hypothetical predicted transmembrane protein	3	3	3	4			118.0
OG4_17047	Hypothetical protein, conserved	4	6	6	8	Rochette *et al.* 2008		175.5
OG4_18417	Hypothetical protein, conserved	2	2	2	4			120.2
OG4_20068	Hypothetical protein, conserved	1	1	1	4	Rochette *et al.* 2008	Rochette *et al.* 2009	99.7
OG4_21810	Hypothetical protein, conserved	2	2	2	4			200.1
OG4_24865	Hypothetical protein, conserved	1	1	1	4			123.0
OG4_10296	Hypothetical protein, conserved	1	1	1	3			166.7
OG4_24371	Hypothetical protein, conserved	2	2	2	3			85.2

		Number of genes in orthologous groups				Life stage association		
Orthologous group ID	**Description**	***L. major***	***L. infantum***	***L. braziliensis***	***L. tarentolae***	**Promastigote**	**Amastigote**	**Read depth**
OG4_24872	Hypothetical protein, conserved	1	1	1	3			62.3
OG4_26114	Hypothetical protein, conserved	2	1	1	3			76.5
OG4_26901	Hypothetical protein, conserved	1	1	1	3			67.5
OG4_27476	Hypothetical protein, conserved	1	1	1	3			40.4
OG4_32453	Hypothetical protein, conserved	1	1	1	3			39.6
OG4_47864	Hypothetical protein, conserved	1	1	1	3			44.8
OG4_63581	Hypothetical protein, conserved	1	1	1	3			39.2
OG4_64304	Hypothetical protein, conserved	1	1	1	3		Saxena *et al.* 2007	47.9
OG4_11153	Hypothetical protein, conserved	1	1	1	2			40.8
OG4_12474	Hypothetical protein, conserved	1	1	1	2			40.4
OG4_17843	Hypothetical protein, conserved	1	1	1	2			71.4
OG4_18279	Hypothetical protein, conserved	1	1	1	2		Rochette *et al.* 2008	41.9
OG4_24851	Hypothetical protein, conserved	1	1	1	2			41.6
OG4_26445	Hypothetical protein, conserved	1	1	1	2			38.6
OG4_26668	Hypothetical protein, conserved	1	1	1	2			57.1
OG4_26782	Hypothetical protein, conserved	1	1	1	2			174.7
OG4_28831	Hypothetical protein, conserved	1	1	1	2			41.3
OG4_33875	Hypothetical protein, conserved	1	1	1	2			43.4
OG4_35631	Hypothetical protein, conserved	1	1	1	2			55.6
OG4_49358	Hypothetical protein, conserved	1	1	1	2	Rochette *et al.* 2008		39.7
OG4_56558	Hypothetical protein, conserved	1	1	1	2			48.9
OG4_60177	Hypothetical protein, conserved	1	1	1	2		Rochette *et al.* 2008	41.8
OG4_63448	Hypothetical protein, conserved	1	1	1	2			46.0
OG4_63485	Hypothetical protein, conserved	1	1	1	2			57.7
OG4_63777	Hypothetical protein, conserved	1	1	1	2			38.2
OG4_63872	Hypothetical protein, conserved	1	1	1	2			48.4
OG4_63939	Hypothetical protein, conserved	1	1	1	2			43.0
OG4_64172	Hypothetical protein, conserved	1	1	1	2			191.7
OG4_64291	Hypothetical protein, conserved	1	1	1	2			39.7
OG4_64344	Hypothetical protein, conserved	1	1	1	2			39.5
OG4_23608	Hypothetical protein, conserved (GO function: cyclic nucleotide biosynthetic process, intracellular signaling cascade)	2	2	2	3		Rochette *et al.* 2009; Rochette *et al.* 2008	87.9
OG4_35583	Hypothetical protein, conserved (GO function: protein folding)	1	1	1	2			63.4

		Number of genes in orthologous groups				Life stage association		
Orthologous group ID	**Description**	***L. major***	***L. infantum***	***L. braziliensis***	***L. tarentolae***	**Promastigote**	**Amastigote**	**Read depth**
OG4_23441	Hypothetical protein, conserved (GO function: protein modification process)	1	1	1	5			118.9
OG4_29790	Hypothetical protein, conserved (GO function: RNA modification, pseudouridine synthesis)	1	1	1	2			39.1
OG4_30772	Hypothetical protein, conserved (GO function: translation)	1	1	1	4		Rochette *et al.* 2008	72.2
OG4_63716	Hypothetical protein, unknown function	1	1	1	3			43.4
OG4_16980	Hypothetical protein, unknown function	1	1	1	2			43.1
OG4_57509	Hypothetical protein, unknown function	1	1	1	2			52.0
OG4_62853	Hypothetical protein, unknown function	1	1	1	2			57.0
OG4_63629	Hypothetical protein, unknown function	1	1	1	2			276.2
OG4_63883	Hypothetical protein, unknown function	1	1	1	2			83.9
OG4_64177	Hypothetical protein, unknown function	1	1	1	2	Rochette *et al.* 2008		41.3
OG4_64193	Hypothetical protein, unknown function	1	1	1	2			74.2
OG4_19604	Inhibitor of cysteine peptidase	1	1	1	4		Rochette *et al.* 2008	128.3
OG4_20563	Kinesin, putative	2	2	2	5		Rochette *et al.* 2008	120.7
OG4_23508	Kinesin, putative	2	2	2	4	Rochette *et al.* 2008		123.0
OG4_30276	Kinesin, putative	1	1	1	2	Srividia *et al.* 2007	Rochette *et al.* 2009;Srividia *et al.* 2007	47.7
OG4_20153	Membrane associated protein-like protein	1	1	1	2			42.6
OG4_24862	Mitotubule-associated protein Gb4, putative	1	1	1	2			40.5
OG4_10027	Multidrug resistance protein, putative; p-glycoprotein-like protein	7	7	6	11	Alcolea *et al.* 2010; Alcolea *et al.* 2009	Rochette *et al.* 2009; Saxena *et al.* 2007	316.6
OG4_10276	NADH:flavin oxidoreductase/NADH oxidase, putative; n-ethylmaleimide reductase-like protein	2	2	2	5		Rochette *et al.* 2009	279.4
OG4_10599	Prolyl-tRNA synthetase, putative	2	2	1	3			73.9
OG4_63781	Protein kinase, putative	1	1	1	2	Alcolea *et al.* 2009	Rochette *et al.* 2009	66.7
OG4_12706	Pteridine transporter	10	10	8	12	Rochette *et al.* 2008	Rochette *et al.* 2009; Rochette *et al.* 2008	443.4
OG4_11347	Pterin-4-alpha-carbinolamine dehydrogenase	1	1	1	3			120.5
OG4_10126	P-type ATPase, putative; vacuolar-type Ca2+-ATPase, putative	3	3	3	4			114.1
OG4_10262	Serine/threonine protein phosphatase catalytic subunit, putative	1	1	1	3			62.7

		Number of genes in orthologous groups				Life stage association		
Orthologous group ID	**Description**	***L. major***	***L. infantum***	***L. braziliensis***	***L. tarentolae***	**Promastigote**	**Amastigote**	**Read depth**
OG4_12097	Serine/threonine-protein phosphatase PP1, putative	1	1	1	2	Rochette *et al.* 2008		64.8
OG4_42636	Surface membrane protein gp46-like protein	2	1	1	3		Rochette *et al.* 2008	91.6
OG4_11890	Ubiquitin carboxyl-terminal hydrolase, putative	1	1	1	3			39.2
OG4_11391	Vacuolar ATP synthase subunit, putative	1	1	1	2	Rochette *et al.* 2008	Leifso *et al.* 2007	37.7
OG4_12221	Vacuolar-type proton translocating pyrophosphatase 1, putative	1	1	1	2	Alcolea *et al.* 2010		46.9

Table 5.6. Orthologous groups of genes absent from *Leishmania tarentolae,* sorted by distribution between the pathogenic species.

		Number of genes in orthologous groups			Life stage association	
Orthologous group ID	**Description**	***L. major***	***L. infantum***	***L. braziliensis***	**Promastigote**	**Amastigote**
Shared by *L. major, L. infantum and L. braziliensis*						
OG4_44658	Adaptin-related protein-like protein	1	1	1		
OG4_58061	Aldose 1-epimerase, putative; aldose 1-epimerase-like protein	1*	1	1	Rochette *et al.* 2008	
OG4_14254	Amastin-like protein; amastin-like surface protein, putative	16*	7*	4	Alcolea *et al.* 2009	Alcolea *et al.* 2010; Rochette *et al.* 2009; Rochette *et al.* 2008; Saxena *et al.* 2007; Srividya *et al.* 2007
OG4_38756	Amastin-like surface protein-like protein	2	2	2		
OG4_63472	Arp2/3 complex 16kDa subunit, putative	1	1	1		
OG4_11508	Arp2/3 complex subunit, putative	1*	1*	1		
OG4_12005	Arp2/3 complex subunit, putative	1	1	1		
OG4_83337	Beta-adaptin protein, putative	1	1	1		
OG4_12536	Calreticulin, putative	1*	1*	1		McNicoll *et al.* 2006
OG4_63650	Carboxypeptidase, putative	1	1	2	Depledge *et al.* 2009	
OG4_10608	Copine I-like protein	1	1	1		
OG4_63659	Fatty acid elongase, puatative	1	1	1		
OG4_11941	Glycosyl transferase, putative	1	1	1	Rochette *et al.* 2008	
OG4_63599	Hypothetical protein	1	1	1		
OG4_59030	Hypothetical predicted transmembrane protein	1*	1	1		Rochette *et al.* 2009; Rochette *et al.* 2008
OG4_15400	Hypothetical protein	1	1	1		
OG4_42632	Hypothetical protein	1	1	1		
OG4_64378	Hypothetical protein	1	1*	1		
OG4_83322	Hypothetical protein	1	1	1		
OG4_10690	Hypothetical protein, conserved	1*	1*	1		
OG4_11427	Hypothetical protein, conserved	1	1	1		
OG4_11774	Hypothetical protein, conserved	1	1	1		Rochette *et al.* 2008
OG4_12259	Hypothetical protein, conserved	1	1	1		
OG4_12833	Hypothetical protein, conserved	1	1	1		
OG4_13358	Hypothetical protein, conserved	1	1	1		
OG4_15615	Hypothetical protein, conserved	1	1	1		
OG4_16181	Hypothetical protein, conserved	1	1	1		
OG4_16997	Hypothetical protein, conserved	1	1*	1		
OG4_17568	Hypothetical protein, conserved	17*	1	1		Rochette *et al.* 2009; Rochette *et al.* 2008
OG4_20423	Hypothetical protein, conserved	1*	1*	1		Rochette *et al.* 2009; Leifso *et al.* 2007; Rochette *et al.* 2008

		Number of genes in orthologous groups			Life stage association	
Orthologous group ID	**Description**	***L. major***	***L. infantum***	***L. braziliensis***	**Promastigote**	**Amastigote**
OG4_26452	Hypothetical protein, conserved	1*	1*	2		Rochette *et al.* 2008
OG4_29012	Hypothetical protein, conserved	1	1	1		Rochette *et al.* 2009
OG4_29456	Hypothetical protein, conserved	1*	1*	1		
OG4_31853	Hypothetical protein, conserved	1*	1*	1		
OG4_38714	Hypothetical protein, conserved	1	1	1		Rochette *et al.* 2009
OG4_38771	Hypothetical protein, conserved	1*	1*	1		Saxena *et al.* 2007
OG4_42591	Hypothetical protein, conserved	1	1	1		
OG4_47797	Hypothetical protein, conserved	1*	1	1		Rochette *et al.* 2008
OG4_53571	Hypothetical protein, conserved	1	1	1		
OG4_54006	Hypothetical protein, conserved	1*	1*	2		
OG4_54445	Hypothetical protein, conserved	1	1*	1		
OG4_54446	Hypothetical protein, conserved	1	1	1		
OG4_54484	Hypothetical protein, conserved	1*	1*	1		Rochette *et al.* 2008
OG4_56438	Hypothetical protein, conserved	1	1	1		McNicoll *et al.* 2006
OG4_60268	Hypothetical protein, conserved	1*	1*	1		Rochette *et al.* 2009
OG4_63577	Hypothetical protein, conserved	1	1	1		Rochette *et al.* 2009; Rochette *et al.* 2008
OG4_64257	Hypothetical protein, conserved	1	1	1		Rochette *et al.* 2009
OG4_64295	Hypothetical protein, conserved	1	1	1		Rochette *et al.* 2009
OG4_64330	Hypothetical protein, conserved	1*	1	1		
OG4_29281	Hypothetical protein, conserved (GO function: iron-sulfur cluster assembly)	1	1	1		
OG4_62091	Hypothetical protein, conserved (GO function: metabolic process)	1*	1*	1		
OG4_83310	Hypothetical protein, conserved	1	1	1		Rochette *et al.* 2009; Rochette *et al.* 2008
OG4_14805	Hypothetical protein, unknown function	1	1	1	Rochette *et al.* 2008	
OG4_47176	Hypothetical protein, unknown function	1	1	1		
OG4_63523	Hypothetical protein, unknown function	1	1	1		
OG4_63618	Hypothetical protein, unknown function	1	1*	1		
OG4_63633	Hypothetical protein, unknown function	1	1	1		Rochette *et al.* 2009; Rochette *et al.* 2008
OG4_63691	Hypothetical protein, unknown function	1*	1*	1		Rochette *et al.* 2008
OG4_63704	Hypothetical protein, unknown function	1	1	1		Rochette *et al.* 2008
OG4_63741	Hypothetical protein, unknown function	1	1	1		
OG4_63742	Hypothetical protein, unknown function	1	1	1		
OG4_63792	Hypothetical protein, unknown function	1	1	1		
OG4_64084	Hypothetical protein, unknown function	1	1	1		
OG4_64203	Hypothetical protein, unknown function	1*	1*	1		
OG4_64329	Hypothetical protein, unknown function	1	1	1		
OG4_64371	Hypothetical protein, unknown function	1	1	1		

		Number of genes in orthologous groups			Life stage association	
Orthologous group ID	**Description**	***L. major***	***L. infantum***	***L. braziliensis***	**Promastigote**	**Amastigote**
OG4_63531	Hypothetical protein, unknown function(GO function: metabolic process)	1	1*	1		
OG4_35275	Hypothetical protein, unknown function;hypothetical protein	5*	1	2		Saxena *et al.* 2007
OG4_42644	Hypothetical protein, unknown function;hypothetical protein	1	1	1	Depledge *et al.* 2009	
OG4_30174	Hypothetical protein	1	1*	1		
OG4_31906	Kinesin, putative	1	1	1		
OG4_55710	Membrane-bound acid phosphatase, putative	1*	1*	1		Rochette *et al.* 2009; Rochette *et al.* 2008
OG4_19653	Microtubule associated protein; microtubule associated protein-like; microtubial binding protein-like protein; microtubule associated protein-like protein	9	3	2	Rochette *et al.* 2008	
OG4_64220	Mu-adaptin 4, putative;mu-adaptin 4, putative (pseudogene)	1*	1*	1		
OG4_10163	Nucleoside diphosphate kinase b	1	2	2	Alcolea *et al.* 2010; Alcolea *et al.* 2009	
OG4_64302	Oxidoreductase-like protein (pseudogene)	1*	1*	1		
OG4_15136	Pfpi/DJ-1-like protein, putative	1	1	1		
OG4_10852	Phosphatidylinositol 3-kinase 2, putative	1	1	1		
OG4_11902	Polynucleotide kinase 3'-phosphatase, putative	1	1*	1		
OG4_64396	Ras-like small GTPases, putative	1	1	1		
OG4_11833	Subtilisin-like serine peptidase	1	1	1		Rochette *et al.* 2008
OG4_51499	Tub family protein-like protein	1*	1*	1		
OG4_12283	Tyrosine/dopa decarboxylase, putative	1*	1*	1		Rochette *et al.* 2008
Shared by *L. major* and *L. infantum*						
OG4_112236	ABC transporter-like protein	1	1*	0		Alcolea *et al.* 2010; Saxena *et al.* 2007
OG4_112262	Acetylornithine deacetylase-like protein	1	1	0		
OG4_112256	Delta-Amastin	1*	1*	0		Alcolea *et al.* 2010; Rochette *et al.* 2009; Rochette *et al.* 2008; Srividya *et al.* 2007
OG4_83296	Delta-Amastin	1	2*	0		Rochette *et al.* 2009; Rochette *et al.* 2008
OG4_64460	Delta-Amastin	1*	1*	0		Rochette *et al.* 2009; Rochette *et al.* 2008
OG4_83391	Aminoacyclase, putative	1	1	0		
OG4_83351	Cytochrome b5-like protein	1	1	0	Alcolea *et al.* 2009	
OG4_83399	DNA polymerase kappa, putative	1*	1*	0		
OG4_43730	Ef hand-like protein	1	1	0		Rochette *et al.* 2009; Rochette *et al.* 2008
OG4_112253	Epsilon-adaptin (pseudogene), putative	1	1	0		
OG4_112238	Fatty acid elongase, putative	1	1	0		
OG4_80241	Glutamamyl carboxypeptidase (pseudogene), putative	1*	1*	0		

		Number of genes in orthologous groups			Life stage association	
Orthologous group ID	**Description**	***L. major***	***L. infantum***	***L. braziliensis***	**Promastigote**	**Amastigote**
OG4_11590	GTPase activator protein, putative	1	1	0		Rochette *et al.* 2008
OG4_64454	Hydrophilic surface protein 2	3*	1*	0	Alcolea *et al*. 2010	Rochette *et al.* 2008; Saxena *et al.* 2007
OG4_112251	Hypothetical protein	1	1	0		
OG4_112271	Hypothetical protein	1	1*	0		
OG4_112272	Hypothetical protein	1	1	0		
OG4_112275	Hypothetical protein	1	1*	0		
OG4_112278	Hypothetical protein	1	1*	0		
OG4_18477	Hypothetical protein	1	1	0		Rochette *et al.* 2009
OG4_83358	Hypothetical protein	1*	1	0	Rochette *et al*. 2008	
OG4_83362	Hypothetical protein	1	1	0		Rochette *et al.* 2009
OG4_83376	Hypothetical protein	1	1	0		
OG4_83377	Hypothetical protein	1	1	0		
OG4_83382	Hypothetical protein	1	1*	0		
OG4_83427	Hypothetical protein	1	1	0		
OG4_76050	Hypothetical protein (GO function: cytoskeleton organization)	1*	1	0		
OG4_112255	Hypothetical protein, conserved	1*	1*	0		Rochette *et al.* 2009
OG4_112258	Hypothetical protein, conserved	1*	1*	0		
OG4_34463	Hypothetical protein, conserved	1*	1	0		
OG4_45776	Hypothetical protein, conserved	1	1	0		
OG4_83350	Hypothetical protein, conserved	1	1	0	Rochette *et al*. 2008	
OG4_83409	Hypothetical protein, conserved	1*	1*	0		
OG4_32705	Hypothetical protein, conserved (pseudogene);hypothetical protein, conserved	1*	1*	0		
OG4_12670	Hypothetical protein, conserved (GO function: GPI anchor biosynthetic process)	1*	1*	0		
OG4_112224	Hypothetical protein, unknown function	1	1	0		
OG4_112237	Hypothetical protein, unknown function	1	1	0		
OG4_112250	Hypothetical protein, unknown function	1*	1	0		
OG4_112254	Hypothetical protein, unknown function	1	1	0		
OG4_112263	Hypothetical protein, unknown function	1	1	0		
OG4_112276	Hypothetical protein, unknown function	1	1	0		
OG4_45804	Hypothetical protein, unknown function	1	1	0	Brotherton *et al.* 2010	
OG4_45991	Hypothetical protein, unknown function	1*	1	0	Rochette *et al*. 2008	Saxena *et al.* 2007
OG4_64455	Hypothetical protein, unknown function	1	2*	0		
OG4_83346	Hypothetical protein, unknown function	1	1	0		
OG4_83359	Hypothetical protein, unknown function	1	1	0		
OG4_83360	Hypothetical protein, unknown function	1	1	0		Rochette *et al.* 2009
OG4_83369	Hypothetical protein, unknown function	1	1	0		
OG4_83375	Hypothetical protein, unknown function	1*	1*	0		
OG4_83380	Hypothetical protein, unknown function	1	1	0		Rochette *et al.* 2008
OG4_83387	Hypothetical protein, unknown function	1	1	0		

		Number of genes in orthologous groups			Life stage association	
Orthologous group ID	**Description**	***L. major***	***L. infantum***	***L. braziliensis***	**Promastigote**	**Amastigote**
OG4_83398	Hypothetical protein, unknown function	1	1	0		
OG4_83408	Hypothetical protein, unknown function	1*	1*	0		
OG4_83411	Hypothetical protein, unknown function	1*	1*	0		
OG4_83414	Hypothetical protein, unknown function	1*	1	0		
OG4_83415	Hypothetical protein, unknown function	1*	1*	0		Saxena *et al.* 2007
OG4_83421	Hypothetical protein, unknown function	1	1	0		
OG4_83422	Hypothetical protein, unknown function	1	1	0	Rochette *et al.* 2008	
OG4_83423	Hypothetical protein, unknown function	1	1	0		
OG4_83431	Hypothetical protein, unknown function	1	1	0		
OG4_83432	Hypothetical protein, unknown function	1	1	0		
OG4_83433	Hypothetical protein, unknown function	1	1	0		
OG4_112243	Hypothetical protein, unknown function (pseudogene) (GO function: metabolic process)	1	1*	0		
OG4_70346	Hypothetical protein, unknown function, pseudogene	1*	1*	0		Rochette *et al.* 2009
OG4_112232	Hypothetical protein, unknown function;hypothetical protein	1	1	0		
OG4_16114	Lectin, putative	1	1	0		
OG4_112266	Map kinase pseudogene	1*	1*	0		
OG4_83397	Phosphatidylinositol-4-phosphate 5-kinase, putative	1	1	0		
OG4_83339	Phosphoglycan beta 1,2 arabinosyltransferase, putative;phosphoglycan beta 1,2 arabinosyltransferase	2	1	0		
OG4_112239	Pteridine transporter	1	1	0		Alcolea *et al.* 2010
OG4_83425	Serine/threonine-protein phosphatase PP1, putative	1	1	0		
OG4_112282	Tartrate-sensitive acid phosphatase acp-3.2, putative	1	1*	0		
OG4_112257	Tuzin like protein;tuzin	1	1	0		Rochette *et al.* 2009; Rochette *et al.* 2008
OG4_112233	Tuzin, putative (pseudogene);tuzin, putative	1	1	0		
Shared by *L. infantum* and *L. braziliensis*						
OG4_11357	40s ribosomal protein S5	0	1	1		
OG4_47556	Flagellar calcium-binding protein, putative	0	1	1		
OG4_112197	Hypothetical protein	0	1*	1		
OG4_54447	Hypothetical protein	0	1	1		Rochette *et al.* 2008
OG4_83298	Hypothetical protein	0	1	1		
OG4_112216	Hypothetical protein, conserved (pseudogene)	0	1*	1		
OG4_45864	Protein kinase-like protein	0	1	1	Alcolea *et al.* 2009	Rochette *et al.* 2009
OG4_112191	Tuzin-like protein;tuzin like protein, putative	0	1	1		Rochette *et al.* 2009; Rochette *et al.* 2008
Only in *L. infantum*						
OG4_64457	Amastin-like protein	0	4*	0		Rochette *et al.* 2009; Rochette *et al.* 2008
OG4_12296	Cullin-like protein-like protein	0	1	0		

		Number of genes in orthologous groups			Life stage association	
Orthologous group ID	**Description**	***L. major***	***L. infantum***	***L. braziliensis***	**Promastigote**	**Amastigote**
OG4_44934	Hydrophilic surface protein	0	1*	0		Alcolea *et al.* 2010; Rochette *et al.* 2009; Rochette *et al.* 2008
OG4_112244	Hydrophilic surface protein 2	0	1	0		Rochette *et al.* 2009
OG4_112226	Hypothetical protein	0	2	0		
OG4_104453	Hypothetical protein	0	1	0		
OG4_36933	Hypothetical protein	0	1	0		
OG4_38827	Hypothetical protein, conserved	0	1	0		
OG4_40582	Hypothetical protein, conserved	0	1	0		
OG4_112225	Hypothetical protein, unknown function	0	1	0		
OG4_112235	Hypothetical protein, unknown function	0	1*	0		Rochette *et al.* 2009
OG4_39213	Hypothetical protein, unknown function	0	1	0		
OG4_42331	Hypothetical protein, unknown function	0	1*	0		
OG4_37478	Hypothetical repeat protein	0	1	0		
OG4_112234	Microtubule associated protein-like protein	0	2	0	Alcolea *et al.* 2010	
OG4_20452	Protein transport protein sec31, putative	0	1	0		
OG4_16857	Surface antigen protein 2, putative	0	1	0		
OG4_56279	Universal minicircle sequence binding protein	0	1	0		
Shared by *L. major* and *L. braziliensis*						
OG4_38786	Hypothetical protein, conserved	1	0	1		
OG4_83311	Hypothetical protein, conserved	1	0	1		
OG4_49433	Hypothetical protein, unknown function	1	0	1		
OG4_33074	Hypothetical protein, unknown function	1	0	2		
OG4_83307	Serine/threonine-protein phosphatase PP1, putative	1	0	1		
Only in *L. major*						
OG4_112264	3,2-trans-enoyl-CoA isomerase, mitochondrial precursor, putative	1*	0	0		Leifso *et al.* 2007
OG4_64465	Class I nuclease-like protein	4	0	0		Leifso *et al.* 2007; Rochette *et al.* 2008
OG4_90898	D-isomer specific 2-hydroxyacid dehydrogenase-protein	1	0	0		McNicoll *et al.* 2006
OG4_20155	Glycoprotein 96-92, putative	1*	0	0		
OG4_112289	Histone H4;histone H4, putative, pseudogene	2	0	0		
OG4_21637	Hydrophilic surface protein	2*	0	0		
OG4_112286	Hypothetical protein	1	0	0		
OG4_112288	Hypothetical protein	1*	0	0		
OG4_112291	Hypothetical protein	2	0	0		
OG4_112295	Hypothetical protein	1	0	0		
OG4_112298	Hypothetical protein	1*	0	0		
OG4_20136	Hypothetical protein	1	0	0		
OG4_83438	Hypothetical protein	3*	0	0		
OG4_112292	Hypothetical protein, conserved	1*	0	0		
OG4_75671	Hypothetical protein, conserved	1	0	0		
OG4_112301	Hypothetical protein, conserved (pseudogene)	1	0	0		
OG4_112303	Hypothetical protein, unknown function	1	0	0		

		Number of genes in orthologous groups			Life stage association	
Orthologous group ID	**Description**	***L. major***	***L. infantum***	***L. braziliensis***	**Promastigote**	**Amastigote**
OG4_51357	Hypothetical protein, unknown function	1	0	0		
OG4_112284	Microtubule associated protein-like protein	2	0	0		
OG4_83436	Microtubule associated protein-like protein	3	0	0		
OG4_112294	Phosphoglycerate kinase (pseudogene), putative	1	0	0		
OG4_54550	Promastigote surface antigen protein 2 PSA2;surface antigen protein 2 precursor	5*	0	0	Rochette *et al.* 2008	
OG4_66645	Surface antigen protein 2, putative	1	0	0		
OG4_13497	Surface antigen-like protein	1	0	0		
OG4_112283	Tuzin, putative (pseudogene)	1	0	0		
OG4_47931	Hypothetical protein	1	0	0		
Only in *L. braziliensis*						
OG4_87788	60s ribosomal protein L19, putative	0	0	1		
OG4_10661	Amino acid transporter	0	0	1		
OG4_61342	Aminophospholipid translocase, putative	0	0	1		
OG4_63547	Beta tubulin;amastin-like protein;amastin-like surface protein, putative	0	0	4		
OG4_63885	Centromere/microtubule binding protein cbf5, putative	0	0	1		
OG4_83315	Cyclophilin a	0	0	1		
OG4_73205	Cystathionine beta-synthase	0	0	1		
OG4_16940	Diacylglycerol kinase-like protein	0	0	1		
OG4_62562	Elongation factor 2	0	0	1		
OG4_36688	Eukaryotic release factor 3, putative	0	0	1		
OG4_39227	Eukaryotic translation initiation factor-like	0	0	1		
OG4_112206	Heat shock 70-related protein 1, mitochondrial precursor, putative	0	0	2		
OG4_112207	Heat shock 70-related protein 1, mitochondrial precursor, putative	0	0	1		
OG4_88964	Heat shock protein DNAJ, putative	0	0	1		
OG4_10091	Hypothetical protein	0	0	1		
OG4_16928	Hypothetical protein	0	0	1		
OG4_21813	Hypothetical protein	0	0	1		
OG4_23220	Hypothetical protein	0	0	4		
OG4_24293	Hypothetical protein	0	0	1		
OG4_29571	Hypothetical protein	0	0	1		
OG4_29958	Hypothetical protein	0	0	1		
OG4_36986	Hypothetical protein	0	0	1		
OG4_40506	Hypothetical protein	0	0	2		
OG4_42625	Hypothetical protein	0	0	1		
OG4_44035	Hypothetical protein	0	0	5		
OG4_54527	Hypothetical protein	0	0	1		
OG4_56054	Hypothetical protein	0	0	1		
OG4_83313	Hypothetical protein	0	0	1		
OG4_83323	Hypothetical protein	0	0	1		
OG4_17997	Hypothetical protein, conserved	0	0	1		
OG4_21464	Hypothetical protein, conserved	0	0	1		
OG4_33181	Hypothetical protein, conserved	0	0	1		
OG4_36775	Hypothetical protein, conserved	0	0	1		

		Number of genes in orthologous groups			Life stage association	
Orthologous group ID	**Description**	***L. major***	***L. infantum***	***L. braziliensis***	**Promastigote**	**Amastigote**
OG4_39837	Hypothetical protein, conserved	0	0	1		
OG4_56480	Hypothetical protein, conserved	0	0	1		
OG4_57646	Hypothetical protein, conserved	0	0	1		
OG4_42628	Hypothetical protein, ppg like	0	0	1		
OG4_112211	Hypothetical protein, unknown function	0	0	1		
OG4_52509	Hypothetical repeat protein	0	0	1		
OG4_93385	Iron/zinc transporter protein-like protein	0	0	2		
OG4_112213	Major surface protease gp63-like	0	0	2		
OG4_17328	Nucleobase/nucleoside transporter	0	0	1		
OG4_73411	Phosphatidic acid phosphatase, putative	0	0	2		
OG4_34533	Proteophosphoglycan ppg1	0	0	1		
OG4_112185	Pteridine transporter	0	0	2		
OG4_93445	Pyruvate kinase, putative	0	0	1		
OG4_112174	Repeat gene hypothetical protein	0	0	2		
OG4_10748	Repeat gene hypothetical protein;Retrotransposable element SLACS (GO function: RNA-dependent DNA replication)	0	0	13		
OG4_56821	Ribosomal protein l3, putative	0	0	1		
OG4_54469	RNase III domain gene	0	0	1		
OG4_71037	Serine/threonine protein phosphatase, putative	0	0	1		
OG4_112175	Slacs like gene retrotransposon element;hypothetical protein	0	0	2		
OG4_112210	Sodium stibogluconate resistance protein, putative	0	0	2		
OG4_35972	Surface antigen-like protein	0	0	1		
OG4_14950	Tate DNA Transposon (GO function: DNA integration, DNA recombination)	0	0	40		
OG4_16474	Translation elongation factor 1-beta, putative; elongation factor 1-beta	0	0	2		
OG4_112209	Triacylglycerol lipase-like protein	0	0	1		
OG4_13770	Tubulin-tyrsoine ligase-like protein	0	0	1		
OG4_83308	Tuzin-like protein;hypothetical protein, conserved	0	0	2		
OG4_66534	Tyrosine specific protein phosphatase, putative	0	0	1		
OG4_100766	Zinc-finger protein ZPR1, putative	0	0	1		
OG4_105391	Product unspecified	0	0	1		
OG4_42606	Hypothetical protein	0	0	1		
OG4_47816	Hypothetical protein	0	0	1		

* Genes absent from *L. tarentolae* as determined by CGH.

Table 5.7. Orthologous groups of genes found in *Leishmania tarentolae* but not in the other sequenced *Leishmania*.

Description	Nb genes	Other species with orthologues
C3HC4 finger protein	1	*T. cruzi, T. brucei, T. vivax, Aspergilus fumigatus*
Hypothetical protein	1	*Monosiga brevicollis, Trichoplax adhaerens*
Hypothetical protein	1	*T. brucei, Canis lupus, Gallus gallus, Homo sapiens, Monodelphis domestica, Mus musculus, Ornithorhynchus anatinus, Pan troglodytes, Rattus norvegicus, Tetraodon nigroviridis, Trichoplax adhaerens*
Hypothetical protein	1	*T. brucei, T. vivax*
Hypothetical protein	1	*T. brucei, T. vivax*
Hypothetical protein	1	*T. brucei, T. vivax*
Hypothetical protein	1	*T. brucei, T. vivax*
Hypothetical protein	1	*T. congolense*
Hypothetical protein	1	*T. cruzi, T. vivax, Tetrahymena thermophila*
Hypothetical protein similar to trafficking protein particle complex subunit 2	1	*T. brucei, T. congolense, T. cruzi, Acyrthosiphon pisum, Aedes aegypti, Anopheles gambiae, Arabidopsis thaliana, Aspergillus fumigatus, Aspergillus oryzae, Babesia bovis, Bombyx mori, Brugia malayi, Caenorhabditis briggsae, Caenorhabditis elegans, Candida glabrata, Canis lupus, Chlamydomonas reinhardtii, Ciona intestinalis, Coccidioides immitis, Coccidioides posadasii, Cryptococcus neoformans, Cryptosporidiummuris, Cryptosporidium parvum, Culex pipiens, Cyanidioschyzon merolae, Danio rerio, Debaryomyces hansenii, Dictyostelium discoideum, Drosophila melanogaster, Entamoeba dispar, Entamoeba histolytica, Entamoeba invadens, Eremothecium gossypii, Gallus gallus, Gibberella zeae, Homo sapiens, Kluyveromyces lactis, Laccaria bicolor, Monodelphis domestica, Monosigabrevicollis, Mus musculus, Nematostella vectensis, Neospora caninum, Ornithorhynchus anatinus, Oryza sativa, Ostreococcus tauri, Pan troglodytes, Pediculus humanus, Phanerochaetechrysosporium, Physcomitrella patens, Phytophthora ramorum, Pichia stipitis, Plasmodium berghei, Plasmodium chabaudi, Plasmodium falciparum, Plasmodium knowlesi, Plasmodium vivax,Plasmodium yoelii, Rattus norvegicus, Ricinus communis, Saccharomyces cerevisiae, Schistosoma mansoni, Schizosaccharomyces pombe, Takifugu rubripes, Tetrahymena thermophila, Thalassiosirapseudonana, Theileria annulata, Theileria parva, Toxoplasma gondii, Trichomonas vaginalis, Trichoplax adhaerens, Yarrowia lipolytica*
Hypothetical protein, conserved	1	*T. brucei, T. cruzi, T. vivax*

Description	Nb genes	Other species with orthologues
Hypothetical protein, conserved	1	*T. cruzi, Arabidopsis thaliana, Aspergillus fumigatus, Aspergillus oryzae, Candida glabrata, Coccidioides immitis, Coccidioides posadasii, Cryptococcus neoformans, Debaryomyces hansenii, Dictyostelium discoideum, Entamoeba dispar, Entamoebahistolytica, Entamoeba invadens, Eremothecium gossypii, Gibberella zeae, Kluyveromyces lactis, Laccaria bicolor, Neurospora crassa, Oryza sativa, Phanerochaete chrysosporium, Physcomitrella patens, Phytophthora ramorum, Pichia stipitis, Ricinus communis, Saccharomyces cerevisiae, Trichoplax adhaerens, Yarrowia lipolytica*
Hypothetical protein, conserved	1	*T. cruzi, T. brucei, T. vivax*
Hypothetical protein, conserved	1	*T. cruzi, T. brucei, T. vivax*
Hypothetical protein, conserved	1	*T. cruzi, T. brucei, T. vivax*
Hypothetical protein, conserved	1	*T. cruzi, T. brucei, T. vivax*
Hypothetical protein, conserved	1	*T. cruzi, T. brucei, T. vivax, T. congolense*
Hypothetical protein, conserved	1	*T. cruzi, T. brucei, T. vivax, T. congolense*
Hypothetical protein, conserved	1	*T. cruzi, T. brucei, T. vivax, T. congolense*
Hypothetical protein, conserved	1	*T. cruzi, T. brucei, T. vivax, T. congolense*
Hypothetical protein, conserved	1	*T. cruzi, T. brucei, T. vivax, T. congolense*
Hypothetical protein, conserved	1	*T. cruzi, T. brucei, T. vivax, T. congolense*
La domain-containing protein	1	*Arabidopsis thaliana, Oryza sativa, Ostreococcus tauri, Physcomitrella patens, Ricinus communis, Trichomonas vaginalis*
Malate dehydrogenase	1	*Culex pipiens*
OmpA family protein	1	*Agrobacterium tumefaciens, Gallus gallus, Nematostella vectensis*

Description	Nb genes	Other species with orthologues
Phosphoinositide kinase, putative	1	*Acyrthosiphon pisum, Aedes aegypti, Anopheles gambiae, Apis mellifera, Arabidopsis thaliana, Aspergillus fumigatus, Aspergillus oryzae, Bombyx mori, Brugia malayi, Caenorhabditis briggsae, Caenorhabditis elegans, Candida glabrata, Canis lupus, Chlamydomonas reinhardtii, Ciona intestinalis, Coccidioides immitis, Coccidioides posadasii, Cryptococcus neoformans, Danio rerio, Debaryomyces hansenii, Dictyostelium discoideum, Drosophila melanogaster, Entamoeba dispar, Entamoeba histolytica, Entamoeba invadens, Eremothecium gossypii, Gallus gallus, Gibberella zeae, Homo sapiens, Kluyveromyces lactis, Laccaria bicolor, Monodelphis domestica, Monosiga brevicollis, Mus musculus, Nematostella vectensis, Neurospora crassa, Ornithorhynchus anatinus, Oryzasativa, Pan troglodytes, Pediculus humanus, Phanerochaete chrysosporium, Physcomitrella patens, Phytophthora ramorum, Pichia stipitis, Plasmodium berghei, Rattus norvegicus, Ricinuscommunis, Saccharomyces cerevisiae, Schistosoma mansoni, Schizosaccharomyces pombe, Takifugu rubripes, Tetrahymena thermophila, Tetraodon nigroviridis, Thalassiosira pseudonana, Theileria parva, Trichomonas vaginalis, Trichoplax adhaerens, Yarrowia lipolytica*
Poly(A) polymerase, putative	1	*T. cruzi, T. congolense*
Surface protein GP63 (Possible pseudogene)	1	*T. cruzi*
Surface protein GP63 (Possible pseudogene)	1	*T. cruzi*
Zinc finger (Ran-binding) family protein	1	*T. brucei, T. congolense, T. cruzi, T. vivax, Arabidopsis thaliana, Oryza sativa, Ostreococcus tauri, Ricinus communis, Tetraodon nigroviridis, Thalassiosira pseudonana*
Hypothetical protein	5	Sequence orphan
Hypothetical protein	3	Sequence orphan
Hypothetical protein	3	Sequence orphan
Hypothetical protein	3	Sequence orphan
Hypothetical protein	2	Sequence orphan
Hypothetical protein	2	Sequence orphan
Hypothetical protein	2	Sequence orphan
Hypothetical protein	2	Sequence orphan
Hypothetical protein	2	Sequence orphan
Hypothetical protein	2	Sequence orphan
Hypothetical protein	2	Sequence orphan
Hypothetical protein	2	Sequence orphan
Hypothetical protein	2	Sequence orphan
Hypothetical protein	2	Sequence orphan
Hypothetical protein	2	Sequence orphan
Hypothetical protein	2	Sequence orphan
Hypothetical protein	2	Sequence orphan
Hypothetical protein	2	Sequence orphan
Hypothetical protein	2	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan

Description	Nb genes	Other species with orthologues
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan
Hypothetical protein	1	Sequence orphan

Table 5.8. Association of *Leishmania tarentolae* orthologous groups of genes with differential expression in promastigote and/or amastigote life stages in pathogenic *Leishmania* species.

	n	Promastigote	Amastigote	p-value*
All orthologous groups	*7694*	890	1013	NA
Higher copy number in *L. tarentolae*	*86*	20	20	0.75
Lower copy number in *L. tarentolae*	*23*	11	11	0.83
Orthologous groups absent from *L. tarentolae*	*278*	18	52	< 0.001

* Promastigote vs amastigote comparison using two-tailed Fisher exact test.

5.2.7. Figures

Figure 5.1. Synteny map of *L. tarentolae* (middle) compared to *L. major* (top) and *L. infantum* (bottom). Genes are grey on chromosome tracks. *L. tarentolae* contig delimitation is in black in the middle lane. Shade of synteny blocks is proportional to sequence identity, the darker the more similar are the sequences. The scale represents nucleotide position on the chromosome. (A) 5' region of chromosome 28. (B) 5' region of chromosome 7. (C) 3' end of chromosome 35.

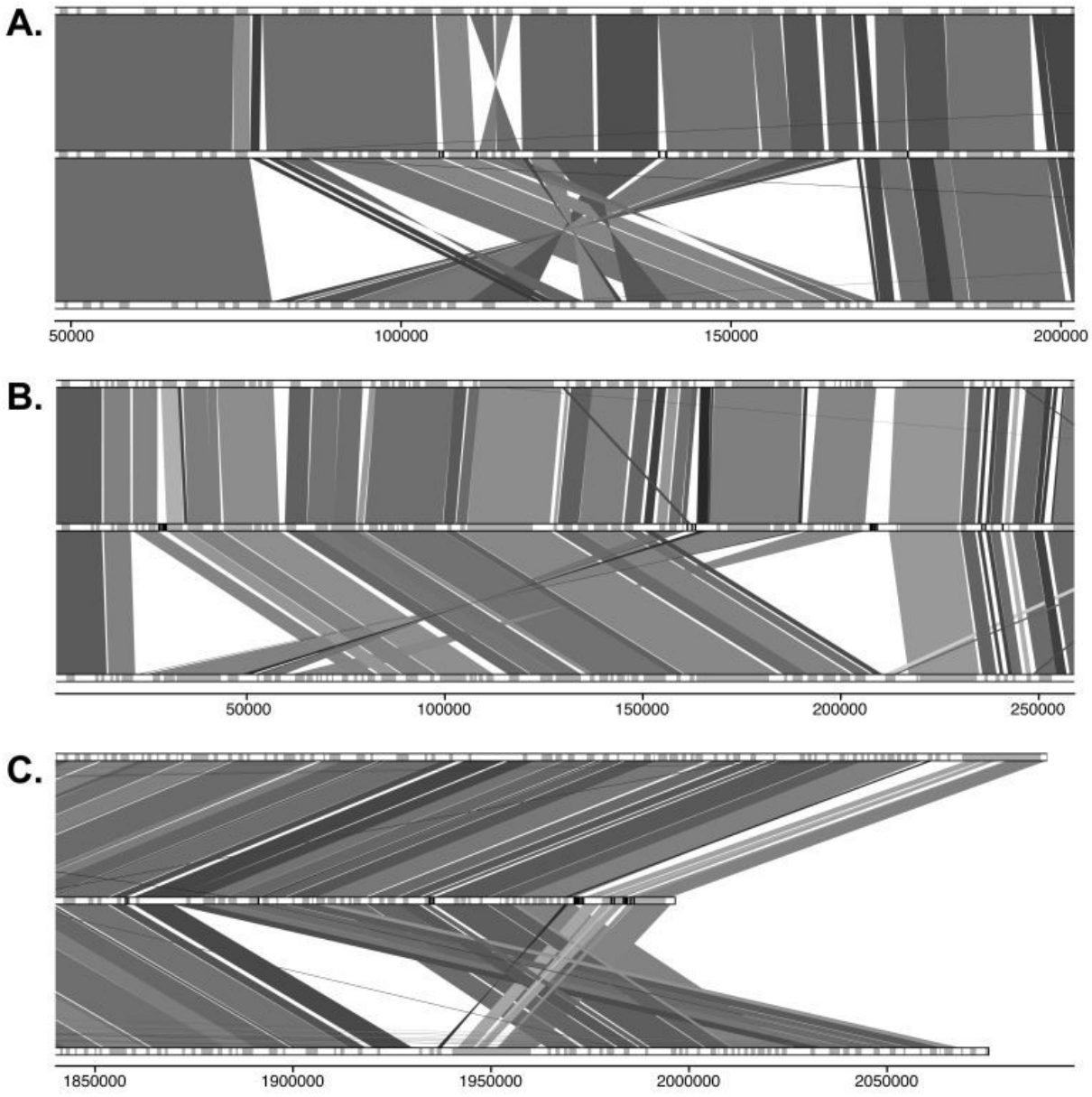


Figure 5.2. Differential distribution of genes and orthologous groups of genes between *L. tarentolae* and *L. major*. [a]Gene counts referring to *L. major*. [b]Gene counts referring to *L. tarentolae*. Lists include the description of orthologous groups (OG) of genes that have differential distribution between *L. tarentolae* and the three sequenced *Leishmania* pathogenic species. Counts of genes within the different OG are in parenthesis. The complete list of genes and orthologous group for selected categories is shown in Tables 5.4, 5.5, 5.6 and 5.7. *Genes with the highest copy number variability in *L. tarentolae*.

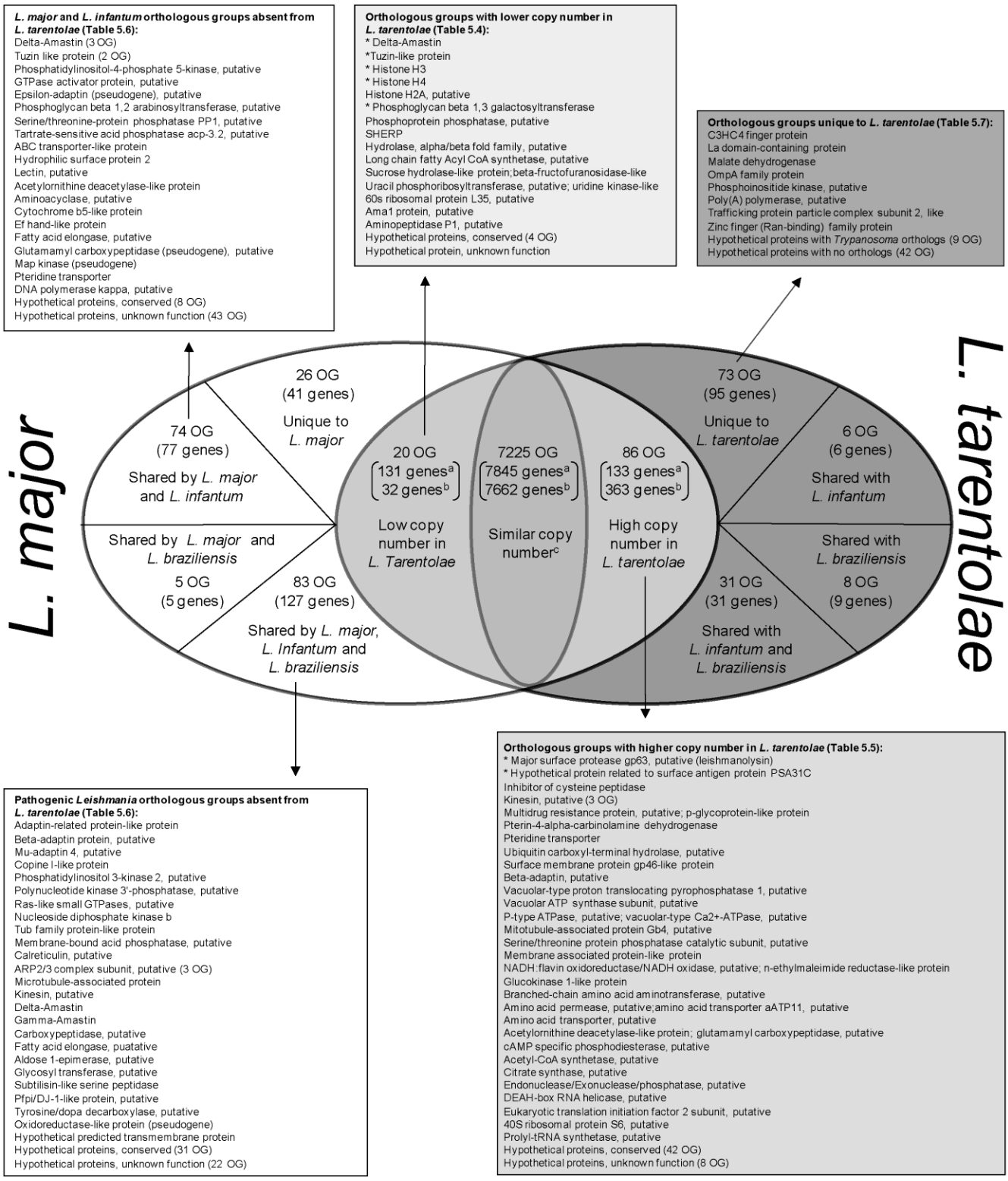

L. major and *L. infantum* orthologous groups absent from *L. tarentolae* (Table 5.6):
Delta-Amastin (3 OG)
Tuzin like protein (2 OG)
Phosphatidylinositol-4-phosphate 5-kinase, putative
GTPase activator protein, putative
Epsilon-adaptin (pseudogene), putative
Phosphoglycan beta 1,2 arabinosyltransferase, putative
Serine/threonine-protein phosphatase PP1, putative
Tartrate-sensitive acid phosphatase acp-3.2, putative
ABC transporter-like protein
Hydrophilic surface protein 2
Lectin, putative
Acetylornithine deacetylase-like protein
Aminoacyclase, putative
Cytochrome b5-like protein
Ef hand-like protein
Fatty acid elongase, putative
Glutamamyl carboxypeptidase (pseudogene), putative
Map kinase (pseudogene)
Pteridine transporter
DNA polymerase kappa, putative
Hypothetical proteins, conserved (8 OG)
Hypothetical proteins, unknown function (43 OG)
Orthologous groups with lower copy number in *L. tarentolae* (Table 5.4):
* Delta-Amastin
*Tuzin-like protein
* Histone H3
* Histone H4
Histone H2A, putative
* Phosphoglycan beta 1,3 galactosyltransferase
Phosphoprotein phosphatase, putative
SHERP
Hydrolase, alpha/beta fold family, putative
Long chain fatty Acyl CoA synthetase, putative
Sucrose hydrolase-like protein;beta-fructofuranosidase-like
Uracil phosphoribosyltransferase, putative; uridine kinase-like
60s ribosomal protein L35, putative
Ama1 protein, putative
Aminopeptidase P1, putative
Hypothetical proteins, conserved (4 OG)
Hypothetical protein, unknown function
Orthologous groups unique to *L. tarentolae* (Table 5.7):
C3HC4 finger protein
La domain-containing protein
Malate dehydrogenase
OmpA family protein
Phosphoinositide kinase, putative
Poly(A) polymerase, putative
Trafficking protein particle complex subunit 2, like
Zinc finger (Ran-binding) family protein
Hypothetical proteins with *Trypanosoma* orthologs (9 OG)
Hypothetical proteins with no orthologs (42 OG)
L. major
L. tarentolae
74 OG
(77 genes)
Shared by *L. major* and *L. infantum*
26 OG
(41 genes)
Unique to *L. major*
Shared by *L. major* and *L. braziliensis*
5 OG
(5 genes)
83 OG
(127 genes)
Shared by *L. major*, *L. Infantum* and *L. braziliensis*
20 OG
131 genes[a]
32 genes[b]
Low copy number in *L. Tarentolae*
7225 OG
7845 genes[a]
7662 genes[b]
Similar copy number[c]
86 OG
133 genes[a]
363 genes[b]
High copy number in *L. tarentolae*
73 OG
(95 genes)
Unique to *L. tarentolae*
6 OG
(6 genes)
Shared with *L. infantum*
Shared with *L. braziliensis*
8 OG
(9 genes)
31 OG
(31 genes)
Shared with *L. infantum* and *L. braziliensis*
Pathogenic *Leishmania* orthologous groups absent from *L. tarentolae* (Table 5.6):
Adaptin-related protein-like protein
Beta-adaptin protein, putative
Mu-adaptin 4, putative
Copine I-like protein
Phosphatidylinositol 3-kinase 2, putative
Polynucleotide kinase 3'-phosphatase, putative
Ras-like small GTPases, putative
Nucleoside diphosphate kinase b
Tub family protein-like protein
Membrane-bound acid phosphatase, putative
Calreticulin, putative
ARP2/3 complex subunit, putative (3 OG)
Microtubule-associated protein
Kinesin, putative
Delta-Amastin
Gamma-Amastin
Carboxypeptidase, putative
Fatty acid elongase, puatative
Aldose 1-epimerase, putative
Glycosyl transferase, putative
Subtilisin-like serine peptidase
Pfpi/DJ-1-like protein, putative
Tyrosine/dopa decarboxylase, putative
Oxidoreductase-like protein (pseudogene)
Hypothetical predicted transmembrane protein
Hypothetical proteins, conserved (31 OG)
Hypothetical proteins, unknown function (22 OG)
Orthologous groups with higher copy number in *L. tarentolae* (Table 5.5):
* Major surface protease gp63, putative (leishmanolysin)
* Hypothetical protein related to surface antigen protein PSA31C
Inhibitor of cysteine peptidase
Kinesin, putative (3 OG)
Multidrug resistance protein, putative; p-glycoprotein-like protein
Pterin-4-alpha-carbinolamine dehydrogenase
Pteridine transporter
Ubiquitin carboxyl-terminal hydrolase, putative
Surface membrane protein gp46-like protein
Beta-adaptin, putative
Vacuolar-type proton translocating pyrophosphatase 1, putative
Vacuolar ATP synthase subunit, putative
P-type ATPase, putative; vacuolar-type Ca2+-ATPase, putative
Mitotubule-associated protein Gb4, putative
Serine/threonine protein phosphatase catalytic subunit, putative
Membrane associated protein-like protein
NADH:flavin oxidoreductase/NADH oxidase, putative; n-ethylmaleimide reductase-like protein
Glucokinase 1-like protein
Branched-chain amino acid aminotransferase, putative
Amino acid permease, putative;amino acid transporter aATP11, putative
Amino acid transporter, putative
Acetylornithine deacetylase-like protein; glutamamyl carboxypeptidase, putative
cAMP specific phosphodiesterase, putative
Acetyl-CoA synthetase, putative
Citrate synthase, putative
Endonuclease/Exonuclease/phosphatase, putative
DEAH-box RNA helicase, putative
Eukaryotic translation initiation factor 2 subunit, putative
40S ribosomal protein S6, putative
Prolyl-tRNA synthetase, putative
Hypothetical proteins, conserved (42 OG)
Hypothetical proteins, unknown function (8 OG)

Figure 5.3. Differential distribution of genes involved in lipophosphoglycan and phosphoglycan modification in *L. tarentolae* (middle) as compared to *L. major* (top) and *L. infantum* (bottom). Phosphoglycan beta 1,2 arabinosyltransferase are shaded with crosses. A first group of phosphoglycan beta 1,3 galactosyltransferase are shaded with thin lines, and a second group with bold lines. Other genes are grey on chromosome tracks. *L. tarentolae* contig delimitation is in black in the middle lane. Shade of synteny blocks is proportional to sequence identity, the darker the more similar are the sequences. The scale represents nucleotide position on chromosome 2.

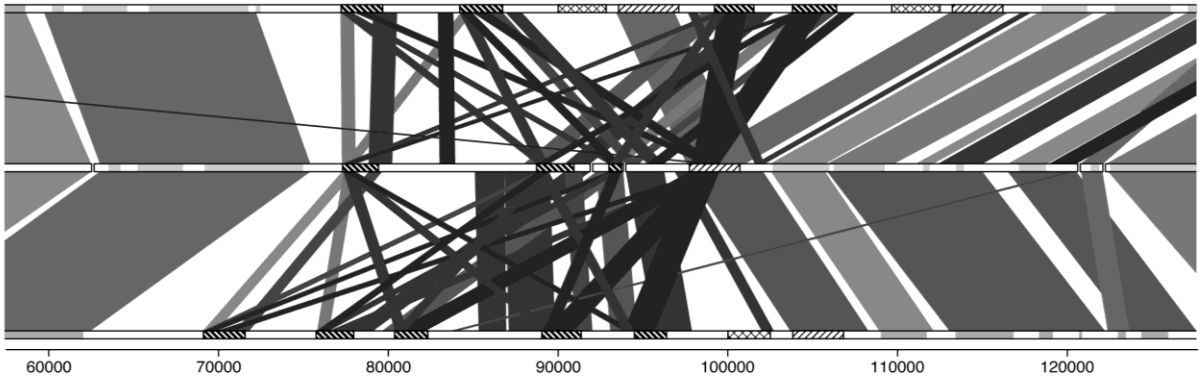


Figure 5.4. Analysis of amastin-coding genes in pathogenic and non-pathogenic *Leishmania* spp. (A) Phylogenetic tree of the amastin genes in *L. tarentolae, L. major, L. infantum* and *L. braziliensis*. Labels refer to amastin subfamilies. *L. tarentolae* amastins are in bold. The evolutionary history was inferred using the Neighbor-Joining method, with a bootstrap test of 500 replicates. Phylogenetic analyses were conducted in MEGA4 (215). (B) Synteny of *L. major* (top), *L. tarentolae* (middle) and *L. infantum* (bottom) amastin/tuzin cluster located on chromosome 8. Amastins are marked with the letter A and tuzins with the letter T. (C) Synteny of *L. major* (top), *L. tarentolae* (middle) and *L. infantum* (bottom) amastin/tuzin cluster located on chromosome 34.

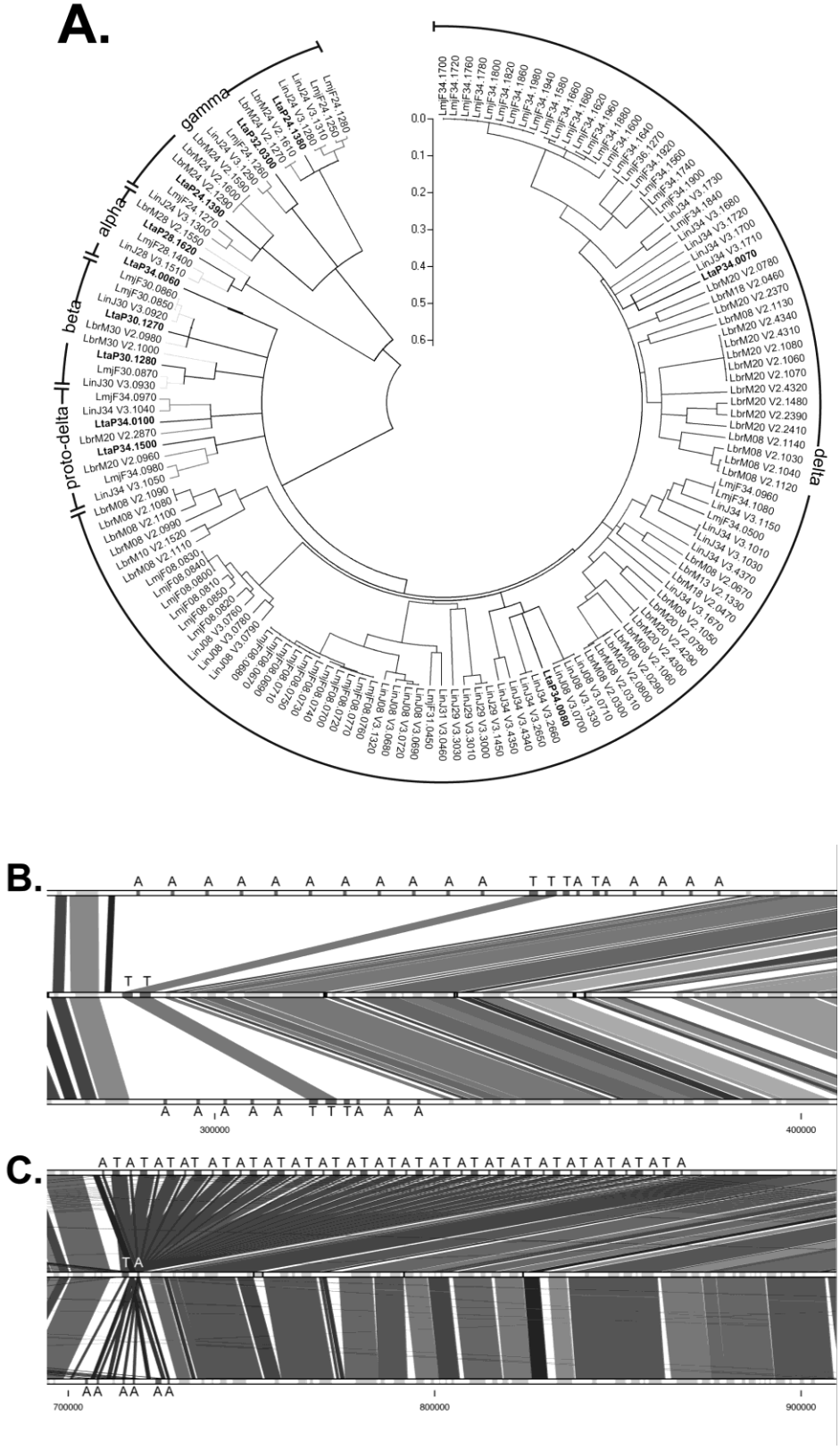

A.
gamma
alpha
beta
proto-delta
delta
0.0
0.1
0.2
0.3
0.4
0.5
0.6
B.
A A A A A A A A A A A A T T TA TA A A A A
T T
A A A A A T T TA A A
300000
400000
C.
A T A T A T A T A T A T A T A T A T A T A T A T A T A T A T A T A T A T A T A T A T A
T A
A A A A A A
700000
800000
900000

Figure 5.5. Density of read coverage for genes present in high copy number in *L. tarentolae.* For each position of the reference *L. major* genes, the number of corresponding reads were counted and plotted on the graph. Protein domains are indicated on the upper portion of each graph. (A) Leishmanolysin (GP63) gene; LmjF10.0480 is used as a reference. (B) Promastigote surface antigen PSA31C gene; LmjF31.1440 is used as a reference.

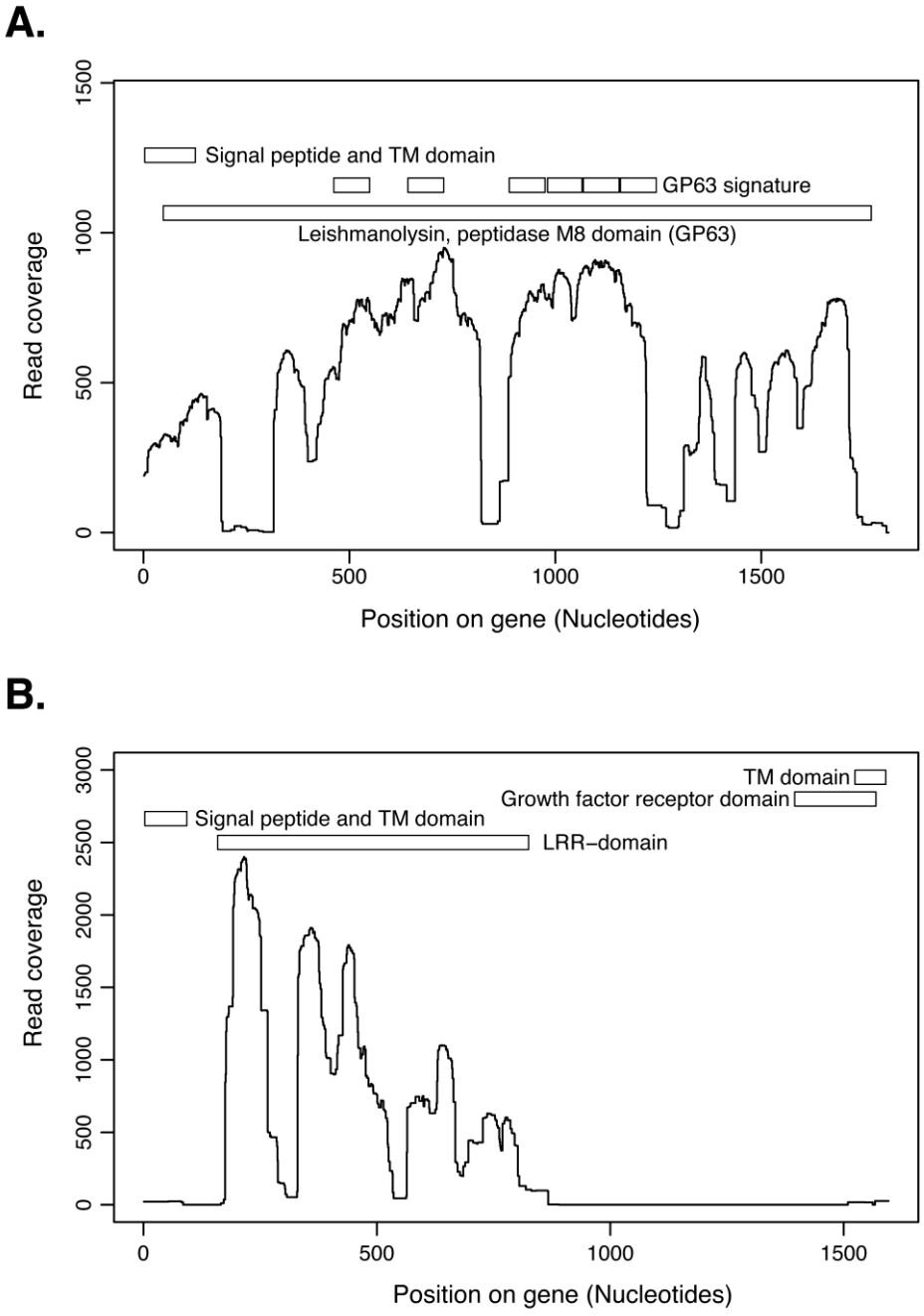


Figure 5.6. Protease activity of GP63 in six *Leishmania* species. (A) Western blot using monoclonal antibody targetting GP63 show the quantity of this protease in each sample. (B) Gelatin zymography assay determining the protease activity of GP63. No signal was observed for *L. tarentolae,* suggesting the absence of GP63 activity in this species. Lane 1. *L. mexicana;* 2. *L. major;* 3. *L. d. donovani;* 4. *L. d. infantum;* 5. *L. amazonensis;* 6. *L. tarentolae*.

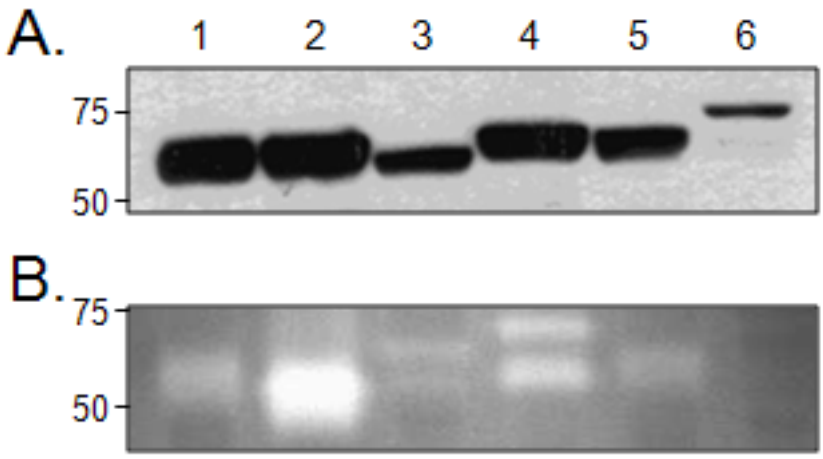
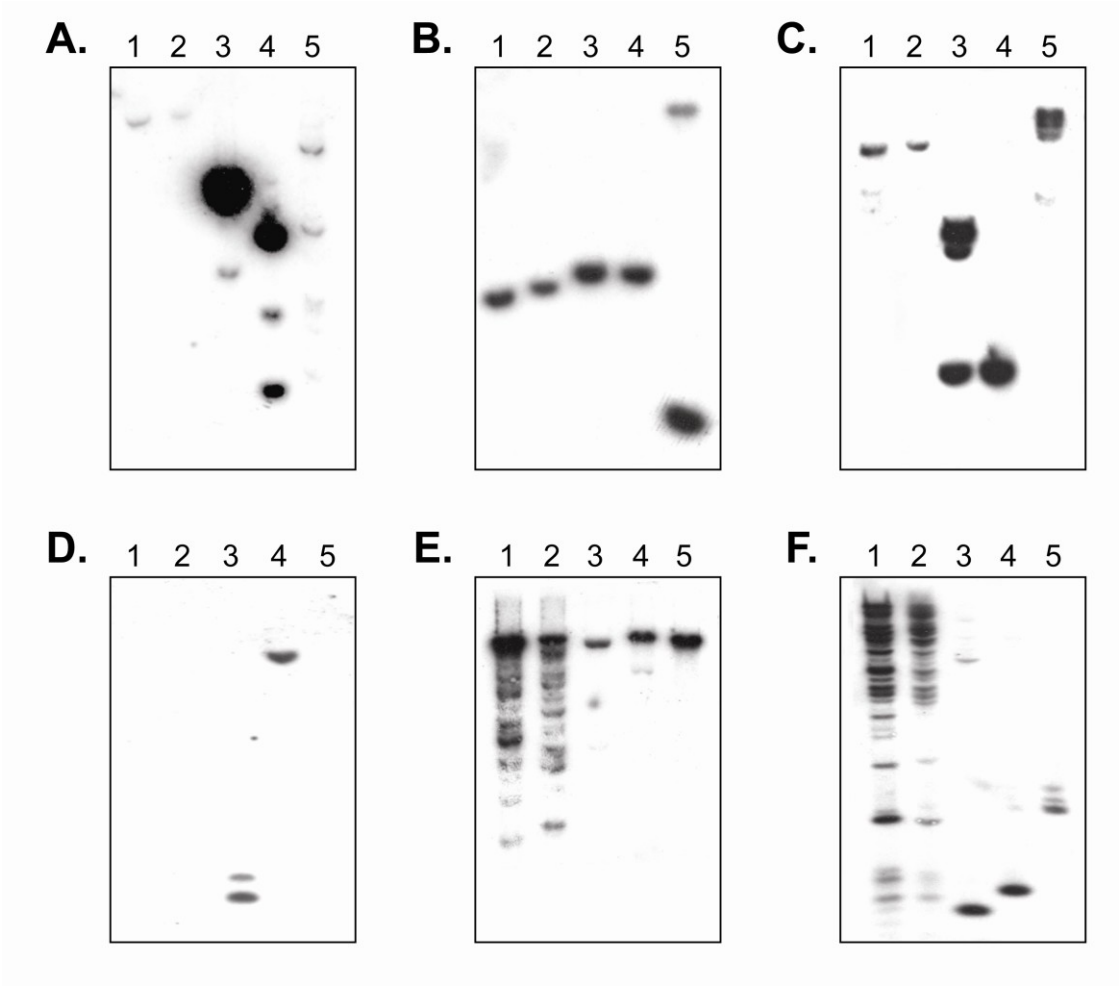


Figure 5.7. Southern blot hybridization of genes that are present in variable copy number between *L. tarentolae* and the other pathogenic species. Total genomic DNA of isolates was digested with XhoI, run on agarose gels, blotted and hybridized with a combination of PCR-specific probes derived from each species (see Table 5.3 for primer sequences and probe details): (A) Amastin delta. (B) Amastin proto-delta (shared by the four species) as a control. (C) Phosphoglycan beta 1,3 galactosyltransferase. (D) Phosphoglycan beta 1,2 arabinosyltransferase. (E) Leishmanolysin (GP63). (F) Surface antigen protein PSA31C. Lanes 1, *L. tarentolae* Parrott-TarII; 2, *L. tarentolae* S125; 3, *L. major* Friedlin; 4, *L. infantum* JPCM5; 5, *L. braziliensis* WHOM/BR/75/M2904.



6. Le diagnostic moléculaire des infections respiratoires virales

Les dernières années ont été riches en éclosions virales qui ont fait les manchettes. Le syndrome respiratoire aiguë-sévère (SRAS), la grippe aviaire et la grippe A (H1N1) d'origine porcine ont toutes trois fait couler beaucoup d'encre et ont inquiété la population à l'échelle mondiale. Elles ont aussi permis de donner une importance au diagnostic des infections respiratoires, en particulier celles d'origine virale. Cependant, l'importance à la fois clinique, sociale et économique des infections plus communes, comme l'influenza saisonnière et les virus causant des syndromes d'allure grippale, ne doit pas être sous-estimée. Le diagnostic précis de ces infections a une fonction clinique et épidémiologique importante.

6.1. Les infections respiratoires

Plusieurs pathogènes peuvent infecter les voies respiratoires et causer différents types d'infections comme le rhume, la grippe, les syndromes d'allure grippale, les bronchiolites, les otites, les sinusites et les pneumonies. Ces maladies peuvent être d'origine virale ou bactérienne. La sévérité de l'infection par un virus respiratoire est habituellement proportionnelle à la charge virale (275).

Même si la plupart des infections respiratoires d'origine virale présentent des symptômes semblables, elles peuvent être causées par plusieurs virus très différents, à la fois du point de vue de la génétique que du point de vue des antigènes. Le rhume est l'infection respiratoire la plus fréquente. Ses symptômes sont l'écoulement nasal, les maux de gorge, la toux et les éternuements. Comme ces symptômes peuvent être causés par la plupart des virus respiratoires, on ne peut se baser uniquement sur les symptômes pour identifier l'agent infectieux, ou pour les différencier de la grippe, même si cette dernière est habituellement plus virulente que le rhume.

Dans le cas d'infections virales ressemblant à la grippe, on peut aussi parler de syndrome d'allure grippale. Elles peuvent être causées, entre autres virus, par les rhinovirus A et B, certains types d'entérovirus A, B, C et D, les para-influenzavirus 1, 2, 3 et 4, les

métapneumovirus humains A et B, les virus respiratoires syncytiaux A et B, les coronavirus OC43, NL63, 229E et HKU1, et les adénovirus A, B, C, D, E et F. Plusieurs sous-types de chacun de ces virus existent, mais les différences entre les types viraux ne sont pas toujours connues. Par exemple, les symptômes de l'infection au virus para-influenza sont similaires entre les différents sérotypes. Cependant, leur distribution saisonnière varie : le sérotype 1 est observé toute l'année, le sérotype 2, à la fin de l'été et à l'automne et le sérotype 3, au printemps et au début de l'été (276).

De plus, plusieurs de ces virus sont connus depuis les années 1960, comme le coronavirus et le rhinovirus, mais ce n'est que récemment que leur importance a été relevée (277). Malgré tout, le rhinovirus est responsable de 25 à 50 % des rhumes dans la population en général (278). L'otite aiguë et la sinusite peuvent être causées par la plupart de ces virus respiratoires, alors que le croup est surtout causé par les para-influenzavirus (279), même s'il peut être causé par d'autres virus, comme le coronavirus NL63 (280).

Plus virulente, la grippe est causée par le virus de l'influenza. Il existe trois types d'influenza : l'influenza de type A, la plus fréquente et la plus variée, l'influenza de type B, et l'influenza de type C, qui reste peu étudiée. Les symptômes de l'influenza incluent la fièvre, la toux, les maux de gorge, la rhinorrhée, la congestion nasale ainsi que des symptômes systémiques comme les maux de tête ou les douleurs musculaires (277). Les souches saisonnières sont habituellement de types H1N1 ou H3N2, alors que certaines variantes peuvent apparaître ponctuellement, comme la grippe aviaire H5N1 découverte à Hong Kong en 1997 et la grippe H1N1 d'origine porcine responsable de la pandémie de 2009.

Une particularité du virus de l'influenza est son fort taux de mutation, qui lui permet de varier ses antigènes d'une année à l'autre, réinfectant ainsi des gens qui avaient une immunité pour une version antérieure de ses antigènes. La grippe peut être dangereuse pour les personnes à risque, soit les jeunes enfants, les femmes enceintes, les malades immunosupprimés et les personnes âgées (281). L'infection par l'influenza de type A entraîne de l'absentéisme au travail pouvant aller de 1,5 à 4,9 jours par épisode (282). De même, l'absentéisme de parents de jeunes enfants peut être augmenté s'ils doivent s'occuper de leurs enfants malades ou s'ils sont contaminés par ces derniers (282).

Environ 80 % des coûts indirects de l'influenza sont associés à l'absentéisme (283). Aux États-Unis seulement, l'influenza entraîne des coûts d'environ 87,1 milliards de dollars par année (284).

La bronchiolite est la plus grande cause d'hospitalisation chez les jeunes enfants. Cette infection est surtout causée par le virus respiratoire syncytial, le métapneumovirus et le coronavirus. Il semblerait que l'action lytique du virus dans les bronches soit responsable de ce type de pathogenèse chez les enfants (285). En 2000, aux États-Unis, le virus respiratoire syncytial était responsable d'environ 86 000 hospitalisations, 1,7 million de visites chez le médecin, et 402 000 visites en salle d'urgence par des enfants de moins de cinq ans, entraînant des coûts directs et indirects de 652 millions de dollars américains (286). La majorité des infections à *Bordetella pertussis,* soit 67 %, chez des enfants de moins de six mois admis à l'hôpital pour une bronchiolite étaient concomitantes avec le virus respiratoire syncytial (287). Dans la même étude, 8,5 % des 142 nouveau-nés testés étaient co-infectés par ces deux pathogènes. La bactérie *Bordetella pertussis* est surtout connue comme l'agent responsable de la coqueluche. Le virus respiratoire syncytial serait responsable de 40 à 90 % des coûts médicaux causés par les bronchiolites et jusqu'à 50 % des cas de pneumonie chez les enfants de moins de deux ans (286).

La pneumonie acquise dans la communauté est responsable de nombreux décès, à la fois chez les enfants et chez les adultes. En 2005, aux États-Unis, plus de 60 000 adultes de plus de 15 ans sont décédés des suites d'une pneumonie acquise dans la communauté (288). De 10 à 20 % des individus admis à l'hôpital pour une pneumonie devront être admis dans une unité de soins intensifs (288). Les personnes ayant reçu une transplantation forment aussi un groupe à risque pour la pneumonie, qui est le plus souvent d'origine bactérienne (289). En effet, *Streptococcus pneumoniae* est la principale cause de pneumonie acquise dans la communauté (de 20 à 60 % des cas), surtout chez les enfants de moins de 2 ans et chez les adultes de plus de 65 ans (290, 291). Il est estimé que, mondialement, environ un million d'enfants meurent annuellement d'infections par *S. pneumoniae* (290).

La bactérie *Haemophilus influenzae* ainsi que les bactéries dites « atypiques », soit *Mycoplasma pneumoniae, Chlamydia pneumoniae* et *Legionella pneumophila,* peuvent aussi être responsables de la pneumonie acquise dans la communauté (292). Finalement, ce

type de pneumonie est causé à environ 29 % par des virus respiratoires, en particulier l'influenza et le rhinovirus, qui ont tendance à causer des myalgies (293). La co-infection par *S. pneumoniae* et par le rhinovirus causerait une infection sévère (293). Des co-infections par *S. pneumoniae* et l'influenza de type B ont causé des pneumonies sévères chez des femmes adultes n'ayant aucun facteur de risque connu (294).

Pour l'instant, des antiviraux sont disponibles uniquement pour l'influenza de type A et pour le virus respiratoire syncytial.

6.2. Le diagnostic des infections respiratoires

Plusieurs raisons justifient le besoin d'un diagnostic moléculaire des infections respiratoires (295). En effet, l'identification d'un agent bactérien ou viral permettra de limiter l'usage inadéquat d'antibiotiques (296) ou d'antiviraux. Des tests diagnostiques permettant l'identification de gènes de résistance aux antimicrobiens permettront de sélectionner un traitement adéquat pour les patients avec une probabilité d'échec thérapeutique réduit (297, 298).

Dans le cas d'une infection virale, l'identification précise de l'agent infectieux, comme la différenciation de l'influenza des autres virus respiratoires, permettra une meilleure gestion des soins hospitaliers. Cela pourrait permettre d'isoler les patients atteints d'infections respiratoires à plus haut risque et d'éviter la contamination du personnel et des autres malades (295).

D'un autre point de vue, l'identification précise des microorganismes pourrait permettre une meilleure compréhension de l'épidémiologie de ces différents agents infectieux et de la distribution globale des infections respiratoires dans la population (277).

6.2.1. Les méthodes utilisant la culture ou l'immunologie

Le diagnostic des infections respiratoires d'origine virale peut être fait par diverses méthodes. Le standard historique consiste à isoler des virus par croissance sur un tapis cellulaire (295). Cependant, les méthodes traditionnelles de culture sont généralement trop longues pour être utiles cliniquement. L'utilisation de lignées cellulaires multiples et de

shell vial utilisant la centrifugation permet un protocole plus simple et plus rapide, soit de 24 à 48 heures pour obtenir des résultats (299).

Très utiles pour le diagnostic rapide en laboratoire, les tests par immunologie (en anglais, *immunoassays*) rapides permettent la détection des virus en moins de 30 minutes. Ils sont surtout utilisés pour le diagnostic de l'influenza et du virus respiratoire syncytial chez les enfants, puisque leur titre viral est plus élevé (299). Ces tests sont moins utiles chez les adultes, chez qui le titre viral est moins élevé (295). La spécificité de ces tests est limitée, surtout lorsqu'elle est comparée au RT-PCR (300, 301).

Le test de détection directe par anticorps fluorescents (*direct fluorescent antibody test* ou DFA) est une autre méthode basée sur la reconnaissance d'antigènes par des anticorps. Cette méthode est très sensible et spécifique, mais requiert des compétences techniques élevées. Même si plusieurs réactifs commerciaux sont approuvés par la FDA, le travail requis pour effectuer la DFA limite son utilisation en laboratoire (301).

Finalement, la sérologie peut aussi être utilisée pour détecter des infections virales, mais elle est peu utile pour le diagnostic des infections respiratoires virales. En effet, elle requiert la détection d'IgG ou d'IgM chez le patient pendant la phase aiguë et la convalescence (302), ce qui rend le diagnostic trop lent pour être utile (299). Dans le cas des virus respiratoires, les méthodes basées sur l'immunologie peuvent être limitées par la disponibilité des réactifs pour un sous-type particulier de virus, par exemple dans le cas de l'influenza ou des rhinovirus, qui comptent plus de 100 sérotypes.

6.2.2. Les méthodes utilisant la détection d'acides nucléiques

Les méthodes les plus sensibles pour détecter des virus respiratoires sont celles utilisant la réaction de polymérisation en chaîne (PCR). Même si ces méthodes ont nécessité un changement de paradigme avant d'être acceptées en clinique, elles ont néanmoins de plus en plus de place dans le diagnostic moléculaire des infections respiratoires. Plusieurs tests diagnostiques utilisant la PCR ont été approuvés par la Food and Drug Administration (FDA). La détection de plusieurs virus respiratoires requiert une étape de transcription inverse, c'est-à-dire de transformation de l'ARN génomique en ADN complémentaire. Ces tests peuvent avoir plusieurs niveaux de complexité.

Le test le plus simple consiste à amplifier une seule cible par PCR conventionnelle, puis de la détecter sur un gel d'agarose. L'apparition d'une bande de longueur appropriée sur un gel d'agarose permettra de dire si oui ou non l'organisme recherché est présent dans l'échantillon. Il est aussi possible de multiplexer plusieurs paires d'amorces PCR afin de détecter plusieurs cibles à la fois. La détection sur gel permettra de mesurer la longueur des fragments observés et ainsi de déterminer quelles cibles étaient présentes dans l'échantillon.

Les méthodes de diagnostic basées sur la PCR en temps réel peuvent être quantitatives et permettent d'atteindre une meilleure spécificité et une meilleure sensibilité. Ces tests utilisent des sondes dont la fluorescence est modifiée au cours de l'amplification du matériel génétique cible pendant la PCR. Cette modification de la fluorescence est mesurée en temps réel. L'analyse de la courbe de fluorescence permet de déterminer si une cible est présente et, dans certains cas, d'évaluer sa concentration. L'utilisation de fluorophores de couleurs différentes permet de détecter plusieurs cibles à la fois ou d'inclure un contrôle interne.

Plusieurs tests de ce type ont été approuvés par la FDA. Par exemple, la compagnie Gen-Probe commercialise une série de tests, nommés Prodesse, utilisant la PCR en temps réel pour détecter des virus respiratoires (303). À la fin de décembre 2010, leur menu de tests permettant la détection de virus respiratoires incluait :

- Prodesse ProFlu+, qui permet d'identifier l'influenza A, l'influenza B et le virus respiratoire syncytial;
- Prodesse ProFAST+, qui différencie les influenzas A de type H1 saisonnier, H3 saisonnier et H1 pandémique d'origine porcine;
- Prodesse ProParaflu+, qui détecte les virus para-influenza 1, 2 et 3;
- Prodesse ProhMPV+, qui détecte le métapneumovirus humain.

Ces tests sont rapides et plusieurs ont été approuvés par la FDA pour usage diagnostique. Cependant, pour étudier toutes les sources d'infections respiratoires virales probables, il faut réaliser plusieurs tests.

La nouvelle génération de tests diagnostiques moléculaires pour la détection des virus respiratoires utilise des PCR hautement multiplexées suivies par une hybridation sur biopuce ou sur des billes marquées par de la fluorescence. Les biopuces ont souvent été utilisées en recherche pour identifier des microorganismes, mais la complexité expérimentale requérant un personnel qualifié, le temps nécessaire pour réaliser une expérience et leur coût élevé limitaient leur implantation dans le diagnostic clinique (304-306).

Des méthodes plus simples que les biopuces classiques, comme les billes marquées en suspension, aussi appelées « biopuces liquides », ou des biopuces conçues pour l'utilisation en laboratoire clinique, ont été nécessaires afin que cette technologie ait un impact sur le diagnostic des maladies infectieuses. Une des premières applications commercialisées de ces nouvelles technologies a été la détection des virus respiratoires.

En général, les tests hautement multiplexés utilisent un protocole dont les étapes sont similaires d'un test à l'autre (307). La première étape du test consiste en la transcription de l'ARN viral en ADN complémentaire par transcription inverse. L'ADN complémentaire ainsi produit sera ensuite soumis à une PCR multiplexe qui permettra l'amplification de régions spécifiques à chacun des virus ciblés par le test. Dans certains cas, la transcription inverse et la PCR multiplexe peuvent être combinées pour limiter les étapes expérimentales. Les amplicons seront ensuite soumis à une étape d'amplification asymétrique, aussi appelée « primer extension », utilisant des amorces formées de deux parties : une séquence d'adressage en 5' et une séquence spécifique à l'organisme ciblé en 3' de l'amorce. Ainsi, cette PCR asymétrique permet d'amplifier des séquences spécifiques à une espèce virale ou à un sous-type de virus, tout en incorporant des nucléotides marqués. La séquence d'adressage sera ensuite hybridée à la biopuce, que ce soit une biopuce sur une surface solide ou sur des billes marquées en suspension dans un liquide. La position d'hybridation des amplicons marqués sur la biopuce ou sur l'étiquette de la bille hybridée permettra d'identifier les virus présents dans un échantillon.

À ce jour, au moins cinq tests de diagnostic moléculaire hautement multiplexés ont été commercialisés pour la détection des virus respiratoires. La majorité de ces tests utilisent la technologie de Luminex, qui consiste à hybrider des cibles sur des billes marquées par de la

fluorescence. Ces tests incluent le xTAG RVP de Luminex Molecular Diagnostics (308, 309), le Resplex II de QIAgen (310) et le Multicode-PLx RVP (311). La compagnie Nanogen a commercialisé un test utilisant l'hybridation sur une biopuce qui se sert de champs électriques pour favoriser l'hybridation, le NGEN Respiratory Virus ASR (312). Le seul test utilisant les biopuces classiques est le test RVP Plus commercialisé par AutoGenomics (313). Ce test fait partie des réalisations de mon doctorat et il sera décrit dans les chapitres 7, 8 et 9 de cette thèse.

7. La conception d'un test diagnostique

Ce chapitre a été originalement publié dans le livre *Microarray innovations: Technology and experimentation,* édité par Gary Hardiman et publié aux éditions Taylor and Francis, en 2009.

Comme ce chapitre a été rédigé pendant la conception du test diagnostique, les informations concernant la composition du test concernent une version antérieure à celle présentée dans les chapitres 8 et 9. Ainsi, certaines informations comme la liste des virus détectés ou le nombre d'amorces composant le test peuvent différer entre les chapitres 7, 8 et 9.

7.1. Le résumé du chapitre de livre

Ce chapitre décrit les étapes de la conception d'un test de diagnostic moléculaire permettant l'identification simultanée de virus respiratoires, de l'analyse des séquences jusqu'à l'optimisation et la validation du test.

7.2. Le chapitre

Development of an integrated molecular diagnostic test to identify respiratory viruses.

Short title: Respiratory virus diagnostic test development

Frédéric Raymond[1], Whei-Kuo Wu[2] and Jacques Corbeil[1]

[1] Infectious Disease Research Center of the CHUQ-CHUL and Laval University, Canada.

[2] AutoGenomics Inc., Carlsbad, CA.

7.2.1. Introduction

This chapter describes the development of an integrated molecular diagnostic assay to identify respiratory viruses. From the analysis of the problem to the design of the assay and its validation, this chapter aims to walk the reader through all the design steps of a molecular diagnostic test involving primer design, PCR, primer extension and microarray hybridization. The system chosen for the integration of this assay is the AutoGenomics Infiniti System.

7.2.1.1. Burden of respiratory virus disease

From the avian flu to the severe acute respiratory syndrome (SARS), respiratory viruses are currently the cause of great concern worldwide. Fear of an influenza pandemic is present through all strata of the population and the media devotes a substantial amount of airtime to every new case of a suspicious respiratory disease. However, these highly talked about viruses are only the tip of the iceberg when speaking of viral respiratory infections. They are, first and foremost, the main causes of the common cold. Non-influenza-related viral respiratory tract infections (VRTI) had an estimated cost of $39.5 billion for the year 2000 in just the United States, a number that includes $17 billion of direct costs and $22.5 billion of indirect costs (314). According to the same study, there are approximately 500 million non-influenza related VRTI episodes per year in the United States alone. Some respiratory virus illnesses result in death, primarily among young infants and the elderly.

7.2.1.2. Advantages of testing

It is often difficult to distinguish between the different possible causes of respiratory infections. Many viruses have similar symptoms and their precise diagnostic requires microbiological laboratory testing. Tests allowing the detection and identification of the most important viruses within one assay would have a positive impact on patient management (315). Such assays could also test for uncommon respiratory viruses such as SARS or avian influenza (316). Dual respiratory virus infections involving Human Respiratory Syncytial Virus (HRSV) have been shown to reduce INF-γ response, which is associated with a more severe illness (317).

Rapid testing of VRTI has been shown to have many clinical and financial benefits. Woo and collaborators observed that rapid testing for VRTI could lead to a significant reduction in the length of hospital stays in the case of influenza or parainfluenza virus infections (318). They also observed a decrease in antibiotics use when influenza virus, parainfluenza virus and adenovirus were tested. Barenfanger and colleagues estimated that an average $5,716 per patient could be saved by appropriately treating the patients for VRTI after a rapid diagnostic of respiratory viruses, mostly because of a decrease in the length of mean hospital stays (319). Those studies suggest that rapid, in-clinic testing in case of VRTI is advisable, providing additional financial advantages.

Unnecessary use of antibiotics can be linked to parents often expecting clinicians to prescribe their children antibiotics. In a study carried out between 1996 and 2000, it was found that 45% of consultations for VRTI lead to the prescription of antibiotics (320). With the rise of antibiotic-resistant bacteria, it is of utmost importance to limit the misuse of antibiotics (321). The rapid molecular diagnostics of VRTI would give clinicians a new tool to decrease the prescription of antibiotics and increase the appropriate use of the antivirals available for many virus species (322).

7.2.1.3. Current tests available

The gold standard for respiratory virus detection is cell culture. However, this technique requires between 5 to 10 days before the results are available to the clinician (323). In the last fifteen years, many other techniques have been devised to detect respiratory viruses. Those techniques include antigen detection, PCR, and microarray, among others. Antigen detection allows for a rapid detection of any respiratory virus. However, this technique has a limited sensitivity and is not appropriate for multiple virus detection. Over 10 antigen detection tests are available commercially for influenza A detection (323). Multiplex PCR tests for the detection of respiratory viruses have been described in the literature (324-327), but with the limitation of requiring gel electrophoresis for amplicon visualization. Real-time PCR is limited in its multiplex capabilities. Also in use are the more technologically-complex DNA hybridization-based techniques. Such techniques include reverse-line blot (328), semiconductor based hybridization (329), microfluidic flow-thru microarray (330), and tilling microarray (331). All these techniques have the disadvantage of either requiring

highly skilled personnel, of taking many hours of laboratory work, or of being based upon a recent technology that has not yet been commercialized. There is an unmet need to automate a well controlled detection assay for respiratory viruses that is easy to use and provides results in a timely manner.

7.2.1.4. Respiratory viruses

Common cold symptoms can be caused by many viruses. Most of them are RNA viruses, except for the adenoviruses, which are DNA viruses. We selected for inclusion in our respiratory virus assay viruses that produce common cold-like symptoms. These viruses include enteroviruses, influenzaviruses, coronaviruses, rhinoviruses, parainfluenzaviruses, metapneumoviruses, and respiratory syncitial viruses. Most of these viruses lead to different peaks in the level of infection among children throughout the year, but some are endemic (332). However, the peak time of each virus may depend on geographical location. Adults and, particularly, chronic obstructive pulmonary disease patients are also at risk of contracting respiratory viruses. Influenza virus infection is one of the most important threats related to respiratory viruses. Pandemic influenza caused million of deaths in the past hundred years and epidemic influenza causes the death of many young children and elderly people each year. The viruses to be detected in our assay are enumerated in Table 7.1, along with results from published prevalence studies.

7.2.2. Principle of the assay

The Infiniti respiratory virus assay has eight main steps, five of which are performed on the AutoGenomics Infiniti system. Only three steps need to be done by laboratory technicians: sample extraction, reverse transcription, and multiplex PCR.

7.2.2.1. Complete description

A nasopharyngeal aspirate is obtained from the patient and the viral RNA is extracted using the Qiamp Viral mini (cat # 52906) kit from Qiagen (Missisauga, Ontario). A reverse transcription step (using Superscript II protocol from Invitrogen, Canada) is conducted upon the extracted RNA and a small volume of the product is added to the multiplex PCR reagents in a 24-well PCR plate. The multiplex PCR for the respiratory virus assay contains 37 different oligonucleotides, some of which are degenerated. The 40 cycle PCR is done at

an annealing temperature of 55 °C in a conventional thermocycler. The PCR plate is then transferred into the Infiniti System, where all the remaining steps of the assay are performed.

The primer extension reagents are added to each well and the thermal cycling for primer extension is done by the Infiniti System. During the design stage, a unique oligonucleotide tag sequence was added to each detection primer. The detection primers are composed of two parts: a unique detection tag sequence and a target specific primer sequence. During primer extension, the target-specific sequence hybridizes to the amplicons and generates a linear amplification. This primer amplification step has three purposes:

1. Incorporation of fluorescently-labeled nucleotides;
2. Specific amplification of target sequences;
3. Increase in sensitivity.

After primer extension, hybridization solution is added to each well and the mix is added to the chip for hybridization. The tags in 5' of the detection primers will hybridize to the specific complementary probe spotted on the biochip's surface. Chips will then be washed, dried, and scanned using a confocal scanner integrated in the Infiniti apparatus. The results are automatically normalized and a diagnostic is suggested to the user.

7.2.3. Assay design

The quality of a molecular diagnostic assay can be estimated using many criteria. During the development of the respiratory virus assay, we developed the test according to the following criteria.

We wanted the respiratory virus diagnostic assay to follow many guidelines that assess the quality of a molecular diagnostic test. To be useful in clinics, a diagnostic test must be easy to implement and to conduct in a standard clinical setting. The Infiniti system allows for simplified in-clinic procedures, by automating all steps following PCR. Some steps, such as RNA extraction and reverse transcription, are not included in the current test, as they must be performed by laboratory personnel. However, we wanted to keep the test to the fewest

steps possible, so we optimized the assay as a one-tube multiplex PCR, in order to reduce the need for manipulation and pipetting. Also, having one PCR tube reduced the danger for cross-contamination, tube mislabeling or other undesirable handling errors. To achieve this goal, the multiplex PCR contains as few PCR primers as possible for each virus to be detected.

Three important qualities of molecular diagnostic testing ─ sensitivity, specificity, and ubiquity ─ were also strictly followed and evaluated. Sensitivity is the capability of a test to detect as few viruses as possible in order to eliminate false negatives caused by low viral load. Specificity is the ability to distinguish the different viruses amongst each other and without false positives. Ubiquity is the capacity of a test to detect all possible strains of a targeted species without generating a false negative. Qualities such as specificity and ubiquity can be evaluated theoretically during the analysis of known virus sequences. However, they can only truly be assessed by testing many samples and by conducting statistical analyses. Similar approaches need to be implemented to test other qualities such as sensitivity, robustness, and reliability.

7.2.3.1. Bioinformatic tools

Many bioinformatic tools were used in the design of the respiratory virus assay. Sequence alignment was done using Clustal W (333) and visualized using MEGA version 3.1 (215). Phylogenetic and molecular evolutionary analyses were also conducted using MEGA version 3.1. FastPCR was used to estimate some primer properties and to manipulate sequences (http://primerdigital.com/fastpcr.html). The RNAsoft tool Pairfold was used for primer dimer analysis (334). Sequence comparison was done using blast and blast2seq (195). Some sequence analysis tools were created in-house using the Perl language to perform simple repetitive bioinformatic tasks.

7.2.3.2. General sequence analysis

In the first steps of sequence analysis, we downloaded the reference sequences for each virus' genome. In order to determine if it was possible to use common PCR primers for groups of viruses, we aligned the reference genomes of all viruses and grouped the viruses with the most similarities using Clustal W. Using this procedure, we were able to determine

that rhinoviruses and enteroviruses could have common PCR primers. Metapneumovirus and respiratory syncitial virus could also be analyzed together, but further analysis suggested that we use different primers for both species. These analyses also allowed us to group viruses within species to achieve the lowest possible quantity of PCR primers required for amplification. Thus, two groups of virus types were determined for adenoviruses, parainfluenzaviruses and coronaviruses. Following this analysis, alignment of reference sequences was carried out, with the goal of finding a common region, suitable for primer design for each virus.

7.2.3.3. Sequences used for design

Initially, we made first attempts at PCR primer design based solely on reference virus genomes found on Genbank (335). However, when we compared the reference sequences to other sequences available on the NCBI database, we noticed that some reference sequences had many polymorphisms when compared to related sequences found in the NCBI database. For those viruses, there was more homogeneity between non-reference sequences than between reference sequences and non-reference sequences. Thus, we concluded that better sequence analysis would be done using reference sequences along with all the other sequences available in the database for a given virus.

For viruses with high sequence variability, such as type A influenza, for which more than 1,500 sequences for each virus segment are available on the FLU database (336), a random sample of 100 sequences was aligned using Clustal W. This alignment was analyzed to identify regions that were suitable for primer design. We aimed at regions that would minimize the number of primers necessary, while allowing us to detect as many influenza variants as possible. The same process was used with all viruses and virus types targeted by this assay. For many viruses, the number of available sequences was not sufficiently high to allow a randomized selection of sequences used in the design, so all available sequences were used instead. With better epidemiological studies, we may be able to determine the prevalence of certain types that will assist in improving the coverage in a clinical setting.

7.2.3.4. Specificity

The nature of the respiratory virus assay warranted the design of low specificity PCR primers and high specificity detection primers in order to maximize detection. As discussed earlier, we wished to amplify the viruses by using as few PCR primers as possible. Towards this end, we needed to increase the number of sequences amplified by the PCR primers. This leads to less specific detection primers. However, what is lost in specificity during multiplex PCR is not problematic if the detection primers are properly designed with appropriate stringency. To achieve an adequate level of specificity, the sequences targeted by the detection primers should be as different as possible between the virus types amplified by the same PCR primers. However, the most important parameter to insure specificity is the 3' nucleotide of the detection primer. Since the DNA polymerase requires that the 3' nucleotide of the detection primer be hybridized to the target DNA in order to begin DNA synthesis, a mismatch at this base would not allow primer extension. Thus, a careful choice of a detection primer's 3' nucleotides allows the design of very specific detection primers. All primers were blasted on the NCBI database to insure their specificity.

7.2.3.5. Ubiquity

As discussed in section 7.2.3.4, for many of the viruses included in the respiratory virus assay, all related sequences found on the NCBI database were used for both PCR and detection primer design. Thus, their theoretical ubiquity was insured at the primer design step. For primers for which more than 100 sequences were available, we selected a random 100 samples to align and to use in primer design. After primer design, the sequences of the primers were blasted on all the sequences available for their target virus to insure a near-perfect theoretical ubiquity. Still, we must keep in mind that the theoretical ubiquity of primers may not reflect the ubiquity observed with real specimens. We plan to investigate, through DNA sequencing, any sample found to be positive for a virus using another method. We would then use these results to refine our design process, in order to increase the ubiquity of the PCR primers and that of the detection primers.

7.2.3.6. Primer design

For each virus, PCR primers and detection primers were designed to carefully optimize its thermodynamic properties while keeping the primers specific and ubiquitous. One degenerate nucleotide was allowed for each primer. We observed that two or more degenerate bases decreased PCR efficiency. Target amplicon size was around 200 nucleotides and the melting temperature for both types of primers was between 58 °C and 60 °C. Secondary structure of delta G (a measure of free energy; a high delta G suggests the increased possibility of secondary structure, an undesirable attribute for any primer) higher than -6.0 were rejected, especially if the 3' end of the primers were involved. Primer dimers were checked using tools described in section 7.2.3.2. Primers were systematically rejected if their 3' end hybridized to any other primer with a delta G stronger than -6.0. To insure proper labeling of the primer extension products, detection primers were oriented so that many labeled nucleotides could be incorporated during primer extension.

7.2.3.7. Content of the assay

The respiratory virus assay's multiplex PCR is composed of 37 primers. The primer extension mix contains 32 detection primers, each with a different 5' tag. A summary of the number of primers used to detect each virus type is shown in Table 7.2.

7.2.4. Validation of the assay

Careful design of PCR primers and detection primers is of utmost importance, but the quality of a test can only be verified by performing the test on actual samples. The validation stage is comprised of three steps:

1. Simplex PCR validation;
2. Multiplex PCR validation;
3. Primer extension validation.

During the validation stage of the assay, some PCR primers or detection primers may be replaced by new designs. It is important to carefully track all primer sets. A primer set is a

group of primers that are used together to amplify and detect one or more virus types. A primer set may contain more than two primers.

7.2.4.1. Simplex PCR validation

Simplex PCR validation consists of testing all PCR primer sets separately to insure that all viruses can be detected using their respective PCR primers. Since the real test for the PCR primers is their behavior in multiplex PCR, it is not necessary to optimize the PCR conditions and primer concentrations at this step. However, if some primer sets yield no or low amplification, it may be useful to redesign the troublesome primers. Sometimes, a small modification of primer sequence can lead to great improvement. In the case of the respiratory virus assay, all viruses yielded a significant band of appropriate length in gel electrophoresis.

7.2.4.3. Multiplex PCR validation

Multiplex PCR validation requires the creation of a multiplex primer mix that includes all PCR primer sets. This multiplex mix is then used to amplify specimens from each virus type, in order to insure the proper amplification properties of each primer set. At this step, it may be important to qualify each primer set according to the amplification yield. Primer sets that give good amplification can be kept as such, but primer sets that give no or low amplification require attention. Optimization of two parameters is useful at this step. First, the annealing temperature of the thermocycling program can be either increased or decreased, or a gradient PCR can be added to the 6 to 12 first PCR cycles, starting from a higher temperature and decreasing in increment at each step. Second, the concentration of primer sets yielding lower results can be increased by a factor of two or higher, if necessary. If none of those parameters allow a primer set to yield significant amplification, it may be necessary to modify the primer set, or to design a new primer set. However, before concluding that a primer set does not work, it may be advisable to try the multiplex on a different specimen, to insure that the low or absent amplification is actually due to the primer set and not due to the low quality of a specimen, to low DNA concentration or to a variant DNA sequence at the targeted region. Thus, the multiplex PCR validation may also include a test for ubiquity. This test is conducted by amplifying many strains of all virus types, to insure that all are detected. This step will also be needed for primer extension

testing, but doing this experiment at the stage of multiplex PCR evaluation gives insight on the assay's ubiquity.

7.2.4.4. Primer extension validation

Primer extension validation requires that one or more specimens of each virus type be amplified by multiplex PCR and then submitted to primer extension. Primer extension is done using a mix of all detection primers at the same concentration. However, before initiating primer extension, it may be useful to perform an exonuclease and alkaline phosphatase step on the amplicons in order to destroy the remaining free nucleotides and to inactivate the remaining PCR primers. After the primer extension step, the products are hybridized to the Infiniti chips, which are processed using the same protocol as the automated assay. The chips are then read by the Infiniti system and the results are analyzed.

For this first step of primer extension validation, we want to identify four types of signals associated to each virus:

- Significant true positive signals;
- Non-significant true positive signals;
- False negative signals;
- False positive signals.

The first case, where the amplified virus gives a significant positive signal on its targeting probe, is the expected result, and conditions relating to such results should not be modified at this stage.

The second case, where the amplified virus gives a non-significant signal on its targeting probe, suggests that this probe set requires optimization. A first step in the optimization of primer sets giving this type of result is to increase the concentration of primers targeting this virus type in the mix of detection primers. It is also important to compare the signal obtained on the chip to the PCR amplification yield of this virus type. Optimization of the PCR step may be needed for such a primer set.

The third case, where the amplified virus gives no signal on its target probe, is problematic. In this case, the first step is to try the same troubleshooting technique as described above for a non-significant true-positive signal. However, it may be important to reanalyze the sequence of the detection primers to insure that their sequences are correct. Sequencing of the amplicon may be warranted and redesign of the primers may be necessary.

The fourth case, where a false-positive signal is observed on a probe for which the target virus type was not added before PCR, is a problem that will most certainly require a redesign of the detection primer. The most frequent cause of false-positives is primer dimers between detection primers, especially dimers that involve the 3' nucleotides of the detection primer. Thus, the first step in troubleshooting this type of result is to verify the presence of primer dimers between the false-positive detection primer and the other detection primers. If primer dimers involving the 3' extremity of the detection primer are found, the primers will need to be redesigned. The false-positive detection primer can also be compared to the different amplicons to insure that no false-positive result is caused by cross-hybridization.

7.2.5. Conclusion

The respiratory virus assay was designed following the guidelines described in section 7.2.3 and it was validated following the guidelines in section 7.2.4. Upon further validation and optimization of the respiratory virus assay, it will be used as an epidemiological tool in a prospective study on a cohort of over 200 children between 1 and 3 years of age that are seen at our children's hospital in Quebec City, Canada. During this study, we will compare the results obtained with the Infiniti respiratory virus assay with results obtained with standard laboratory methods and with a real-time PCR assay that we are currently developing. This study will allow us to estimate the usefulness of this assay in a clinical setting. Also, we aim to use this assay to conduct a similar study for 3 consecutive years, beginning in the fall of 2006, on children that are seen at our hospital and at a pediatric clinic in Quebec City. The results will help us to understand the epidemiology of the viruses tested and to gain insight in their treatment and control. As respiratory viruses become more and more of a worldwide concern, such a tool will have an increased usefulness for diagnostic and treatment.

7.2.6. Tables

Table 7.1. Viruses to be detected using the respiratory virus assay, including results from published prevalence studies.

Families	**Virus species**	**Types**	**Prevalence studies**		
			n = 200 (327)	**Children < 4 yo n = 536** (337)	**n = 446** (338)
Adenoviridae	Adenovirus	A, B, C, D, E, F	12,9%	16,6%	2,3%
Coronaviridae	Coronavirus	NL63, 229E, HKU1, OC43, SARS	NA	NA	3,4%
Orthomyxoviridae	Influenza	A, B	22,8%	14,6%	8,8%
Paramyxoviridae	Parainfluenza	1, 2, 3, 4	6,3%	28,3%	3,2%
Paramyxoviridae	HRSV	A, B	37,5%	29,3%	43,6%
Paramyxoviridae	HMPV	A, B	NA	NA	4,4%
Picornaviridae	Rhinovirus	A, B	NA	59,6%	31,8%
Picornaviridae	Enterovirus	A, B, C, D	10,6%	NA	2,1%

Table 7.2. Summary of the primer sets included in the respiratory virus assay.

Primer sets		Amplicon length	PCR primers	Detection primers
Virus	**Type**	**(nucleotides)**	**(nb primers)**	**(nb primers)**
Adenovirus	A, B and C	174	4	4
Adenovirus	D, E and F	174	3	3
Coronavirus	SARS	168	2	1
Coronavirus	HKU1 and OC43	123	2	2
Coronavirus	NL63 and 229E	149	2	2
Enterovirus	A, B, C and D	197	3	4 + 1*
Rhinovirus	A and B			2 + 1*
Influenza	A	201	3	2
Influenza	B	201	3	1
HMPV	A and B	148	3	2
HRSV	A and B	245	3	2
HPIV	1	272	2	1
HPIV	3	243	2	1
HPIV	2, 4A and 4B	275	5	3
Total	*30*		*37*	*0*

* Detection primer targetting both rhinovirus and enterovirus.

8. Le diagnostic moléculaire automatisé des virus respiratoires

Cet article a été publié dans le *Journal of Clinical Microbiology,* en 2009 (313).

8.1. Le résumé de l'article

8.1.1. Le résumé en français

Les infections respiratoires d'origine virales sont une préoccupation majeure en santé et elles représentent la première cause de consultation et d'hospitalisation pour les jeunes enfants. Nous avons développé et comparé deux tests qui permettent la détection de 23 virus respiratoires qui infectent souvent les enfants.

La première méthode consistait en une série de tests par qRT-PCR de type Taqman dans une plaque de 96 puits. La seconde méthode consistait en une PCR multiplexe suivie d'une PCR asymétrique et d'une hybridation sur biopuce dans un système de diagnostic moléculaire automatisé, l'analyseur INFINITI. Nos deux tests peuvent détecter les adénovirus A, B, C et E, les coronavirus HKU1, 229E, OC43 et NL63, les entérovirus A, B, C et D, les rhinovirus de génotypes A et B, les virus influenza de types A et B, les métapneumovirus humains (hMPV) de types A et B, les virus respiratoires syncytiaux (HRSV) A et B, et les virus para-influenza de types 1, 2 et 3. Ces tests ont été utilisés pour identifier les virus dans 221 aspirations nasopharyngées provenant d'enfants hospitalisés en 2002 pour des infections des voies respiratoires.

Les virus respiratoires ont été détectés par au moins une des deux méthodes dans 81,4 % des 221 spécimens : 10,0 % étaient positifs pour HRSV-A, 38,0 % pour HRSV-B, 13,1 % pour la grippe A, 8,6 % pour un coronavirus, 13,1 % pour un rhinovirus ou un entérovirus, 7,2 % pour un adénovirus, 4,1 % pour le hMPV et 1,5 % pour le virus para-influenza. Des infections virales multiples ont été trouvées dans 13,1 % des échantillons. Les deux méthodes donnent des résultats concordants dans 94,1 % des échantillons.

Cette étude a permis d'évaluer l'étiologie, en milieu hospitalier, des virus respiratoires infectant les enfants hospitalisés pour une infection des voies respiratoires et aidera les interventions de santé publique.

8.1.2. Abstract

Respiratory virus infections are a major health concern and represent the primary cause of testing, consultation and hospitalization for young children. We developed and compared two assays that allow the detection of up to 23 different respiratory viruses that frequently infect children.

The first method consisted of single Taqman qRT-PCR assays in a 96-well plate format. The second consisted in a multiplex PCR followed by primer extension and microarray hybridization in an integrated molecular diagnostic device, the INFINITI analyzer. Both of our assays can detect adenoviruses groups A, B, C and E, coronaviruses HKU1, 229E, NL63 and OC43, enteroviruses A, B, C and D, rhinoviruses of genotypes A and B, influenzaviruses A and B, human metapneumoviruses (HMPV) A and B, human respiratory syncytial viruses (HRSV) A and B and parainfluenza viruses types 1, 2 and 3. These tests were used to identify viruses in 221 nasopharyngeal aspirates obtained from children hospitalized for respiratory tract infections.

Respiratory viruses were detected with at least one of the two methods in 81.4% of the 221 collected specimens: 10.0% were positive for HRSV-A, 38.0% for HRSV-B, 13.1% for influenza A, 8.6% for any coronaviruses, 13.1% for rhinoviruses or enteroviruses, 7.2% for adenoviruses, 4.1% for HMPV, and 1.5% for parainfluenza viruses. Multiple viral infections were found in 13.1% of the specimens. The two methods yielded concordant results in 94.1% of specimens.

These tests allowed a thorough etiological assessment of respiratory viruses infecting children in hospital settings and would assist public health interventions.

8.2. L'article

Comparison of automated microarray detection with real-time PCR assays for the diagnosis of respiratory viruses in children

Running title: Microarray diagnosis of respiratory viruses

Frédéric Raymond[1], Julie Carbonneau[1], Nancy Boucher[1], Lynda Robitaille[1], Sébastien Boivert[1], Whei-Kuo Wu[2], Gaston De Serres[3], Guy Boivin[1] and Jacques Corbeil[1]

[1] Infectious Disease Research Center of the CHUQ-CHUL and Laval University, Canada.
[2] AutoGenomics Inc., Carlsbad, CA.
[3] Institut national de santé publique du Québec, Canada.

Correspondence: guy.boivin@crchul.ulaval.ca or jacques.corbeil@crchul.ulaval.ca

8.2.1. Introduction

Respiratory tract infections are an important cause of hospitalization in children. Most of these infections are caused by RNA viruses that produce influenza-like symptoms of variable severity (339). Because of cost and technical limitations, virological testing is currently done sporadically and for a limited number of viruses at the clinician's request. The availability of a molecular diagnostic test that allows the detection of all respiratory-related viruses would permit better management of patients and possibly limit unnecessary use of antibiotics (318, 319, 323).

The most frequent virus detected in young children suffering from respiratory tract infections is the human respiratory syncytial virus (HRSV) (340, 341). HRSV is the causal agent in up to 70% of bronchiolitis episodes in infants and young children (342). Other well-known clinically relevant respiratory viruses include influenzavirus, rhinovirus, enterovirus, coronavirus, parainfluenza viruses and adenoviruses. Recently described respiratory pathogens include human metapneumovirus (HMPV) (343-346), coronaviruses SARS, HKU1 and NL63 (347), and bocaviruses (348). When using conventional diagnostic methods, multiple virus infections are observed in 5% of respiratory tract infections (349), whereas co-detection rates of 11 to 20% have been observed when using molecular methods (317, 325, 349).

Proper viral diagnosis has been shown to reduce the length of hospital stay (318, 319). The classic diagnostic methods for the detection of respiratory viruses consist of virus growth on cell culture and direct immunofluorescence assays (350). Although very specific, these methods lack sensitivity, are burdensome, require skilled personnel and can take a few days, if not weeks, before generating results in the case of cell culture. Solid phase immunoassays are often inexpensive and rapid, but they are limited to the detection of a single virus species, and have reduced sensitivity and specificity compared to cell culture (323, 350). In addition, the development of immunological tests is limited for some viruses with many subtypes, for example adenoviruses, enteroviruses and rhinoviruses.

Polymerase chain reaction (PCR) has been used to amplify and detect many respiratory viruses (351). Conventional PCR or real-time PCR has the potential for high sensitivity and

specificity compared to previous methods (351-354). PCR was initially limited by the number of species that could be detected and identified in a single test, often requiring multiple parallel reactions (325, 355). In the last years, numerous tests have been developed using single-tube multiplex PCR to detect many viruses in one assay (326, 327, 356). Single-tube or parallel multiplex PCR assays can be coupled to hybridization using nylon membrane DNA arrays (328), conventional microarrays (331), flow-thru DNA chips (330), semiconductor-based DNA microchips (329) or microspheres (357, 358). Several respiratory virus panels (RVP) using the Luminex technology have been commercialized, such as xTAG RVP from Luminex (308, 309), Multicode PLx RVP from Eragen (358) and Resplex II from QIAgen (312). So far, these tests have not been fully automated, which limits their use in most clinical laboratories (359).

In order to identify the etiology of respiratory tract infections, we developed real-time PCR assays and a microarray assay detection system, allowing the diagnostic of 18 and 23 different respiratory virus types, respectively. The first method consists of qRT-PCR TaqMan assays adapted to the 96-well plate format, each plate allowing the testing of 4 specimens, along with a series of positive and negative controls. The real-time PCR assays have been optimized to reduce hands-on time. The second method consists of a multiplex PCR test followed by primer extension and microarray hybridization. The microarray assay is automated with the INFINITI analyzer manufactured by AutoGenomics inc. (Carlsbad, CA). The INFINITI analyzer was 510K cleared by the FDA for several pharmacogenomic assays. After validation of both assays using laboratory strains of targeted viruses, we compared the performance of the two assays using specimens collected from children ≤ 3 years of age hospitalized for an acute respiratory tract infection (ARTI).

8.2.2. Methods

8.2.2.1. Specimen collection and preparation

This study was accepted by the ethics committee of the Centre Hospitalier Universitaire de Quebec. We tested nasopharyngeal aspirate (NPA) specimens from 221 children ≤ 3 years old who were hospitalized between November 2001 and April 2002 for an ARTI of ≤ 7 days as previously described (343, 360). NPAs (one per patient) were aliquoted and

stored at -80 °C. Clinical information and clinical laboratory results were prospectively collected after informed consent was obtained. Before nucleic acid extraction, NPAs were thawed on ice and 0.5 µl of Hepatitis C, genotype 1a Armored RNA (Ambion Diagnostics, CA) was added to 200 µl of each NPA as an internal control. Nucleic acid extraction was performed with the QiAmp Viral RNA mini kit (QIAGEN, Mississauga, Ontario) using the protocol suggested by the manufacturer, except for the final elution volume, which was 40 µl. Reverse transcription was done using the Superscript II Reverse Transcriptase kit (Invitrogen, Carlsbad, CA).The reaction mixture was composed of 1 µl of 50 ng/µl random primers (Amersham, Piscataway, NJ), 1 µl of 10 µM dNTPs and 10 µl of extracted RNA. The mixture was incubated at 65 °C for 5 min, then put on ice. The following reagents were then added to the solution: 4 µl of 5X first strand Buffer (Invitrogen), 2 µl of 0.1 M DTT (Invitrogen) and 1 µl of 40 U/µl RNAsin (Promega, Madison, WI). The solution was incubated at room temperature for two min, then 200 units of Superscript II (Invitrogen) were added. The solution was incubated at room temperature for 10 min, then at 42 °C for 50 min and finally at 70 °C for 15 min. The cDNA was kept at -20 °C. The reverse-transcripted samples were then tested using both the microarray assay and the qRT-PCR assays (Figure 8.1).

8.2.2.2. Clone generation and sensitivity studies

Laboratory strains or clinical specimens were used to generate the clones for the sensitivity analyses. Amplicons generated using qRT-PCR primers or multiplex PCR primers were cloned in the pCR II (Invitrogen) or pGEM-T easy (Promega) vectors. Insert sequence was verified by DNA sequencing. Prior to sensitivity studies, amplicons were digested with EcoRV (pCR II) or Nde1 (pGEM-T easy) for 120 to 180 min at 37 °C. They were then purified using QIAgen PCR purification kits and quantified with the NanoDrop (Thermo Scientific, Wilmington, DE). Serial dilutions were performed in water to obtain a range of concentrations between 100000 and 0.1 copies/µl.

Both the cycle thresholds for the real-time PCR assays and the signal thresholds for the microarray assay were determined by analyzing the results obtained with laboratory strains and results of the sensitivity study. Then, thresholds were adjusted according to the

background signal of several negative specimens and to the signals obtained with several laboratory strains and clinical specimens.

8.2.2.3. Primers and probes

PCR primers for the real-time PCR assays and for the multiplex PCR targeted the same gene, except for HMPV and adenovirus primers. However, even when targeting the same gene, both assays had distinct primer sets. All primers for qRT-PCR were obtained from Invitrogen Canada. TaqMan probes with MGB quenchers were obtained from Applied Biosystems (Streetsville, Ontario). Sequences of primers and probes used for the qRT-PCR assays are shown in Table 8.1. The multiplex PCR primer mix contains all PCR primers at final concentrations ranging between 50 and 200 nM, depending on the targeted virus. Primers used for primer extension were composed of a proprietary tag sequence followed by a virus specific detection sequence. The primer concentration in the primer extension mix ranged from 250 to 500 nM, depending on the targeted virus. Multiplex PCR primer mix and primer extension mix were obtained from AutoGenomics inc.

8.2.2.4. qRT-PCR assays

All reactions were performed in a 96-well plate using the TaqMan Universal PCR Master Mix (Applied Biosystems) in an ABI 7500 apparatus (Applied Biosystems). PCR primers were used at a 200 nM concentration and TaqMan probes were used at a total probe concentration per well of 250 nM . Each 96-well PCR plate allows for the testing of four specimens, positive controls consisting of cloned amplicons of specific viruses and negative controls consisting of water for each virus. For each specimen tested, 14 wells of the plate were used. One µl of specimen cDNA or plasmid control was added to each well of the plate. The PCR program consisted in the following steps: 2 min at 50 °C, 10 min at 95 °C, followed by 50 cycles of 15 s at 95 °C, 15 s at 55 °C and 40 s at 60 °C. A specimen was considered positive for a virus if its cycle threshold (CT) was lower than a predefined value (Table 8.2).

8.2.2.5. INFINITI microarray assay

First, reverse-transcribed samples were amplified using a highly multiplexed PCR. Then, amplicons were cleaned by enzymatic reactions and were then subjected to primer

extension within the INFINITI analyzer. The assay relies on a proprietary tag system in which amplicons are tagged at the primer extension step. Amplicons are also labelled with fluorescent nucleotides at the primer extension step. Hybridization of the tags to the anti-tags immobilized on the microarray allows specific identification of targets. Each anti-tag hybridizes to three replicates on the microarray. Microarray washing and scanning was done within the INFINITI analyzer without human intervention. Summary of the microarray assay protocol is shown in Figure 8.1.

Multiplex PCR was performed in a T1plus thermocycler (Biometra, Montreal Biotech, Montreal). The multiplex PCR primer mix was composed of 46 primers at concentrations ranging from 50 nM to 200 nM. The amplification solution was composed of 10X buffer, 0.2 µM dNTPs, 1.5 mM $MgCl_2$, multiplex PCR primer mix, 0.5 units of Platinum Taq DNA polymerase (Invitrogen Canada), and 2.5 µl of cDNA, in a final volume of 20 µl. The PCR program consisted in the following steps: 60 s at 94 °C followed by 39 cycles of 30 s at 94 °C, 30 s at 55 °C and 60 s at 72 °C. The reaction was then incubated at 72 °C for 3 min. Then, 3 units of Shrimp Alkaline Phosphatase (Clontech, Mountain View, CA), 7.5 units of exonuclease (Clontech) and 0.25 µl of 50X Titanium DNA polymerase (Clontech) were added to the solution, which was incubated at 37 °C for 50 min and at 94 °C for 20 min. This step allows for the degradation of remaining dNTPs and PCR primers that were not used in the multiplex PCR. The subsequent steps were automated by the INFINITI analyzer (AutoGenomics inc.). The primer extension solution, comprising 34 tagged detection primers (AutoGenomics inc.) was then added to the solution. Primer extension reaction consisted in the following steps: 60 s at 94 °C followed by 39 cycles of 15 s at 94 °C and 15 s at 50 °C. Primer extension was done in the presence of Cy5-dCTP. Following the primer extension reaction, 80 µl of hybridization solution (AutoGenomics inc.) was added to each reaction. The total volume of 120 µl was then hybridized to a DNA microarray (AutoGenomics inc.) for 90 min at 42 °C at high humidity. The tags on the extension primers hybridize to corresponding probes on the microarray. After hybridization, each chip was washed 5 times with 300 µl of 1X SSC. Chips were dried and scanned using a confocal scanner. A specimen was considered positive for a virus if the ratio between the signal for a virus and the background signal was over a defined threshold (Table 8.2) after background correction.

8.2.3. Results

8.2.3.1. Sensitivities of the real-time PCR assays and of the microarray assay

The capacity of both methods to detect each targeted virus was initially validated using laboratory isolates and clinical specimens. Technical specificity was assessed using laboratory strains and clinical specimens validated using culture or sequencing (Table 8.3). No false positives were observed and all described specimens were positive by both assays. Then, sensitivity assays were conducted using cloned amplicons (Table 8.2). Sensitivities of each single TaqMan qRT-PCR assay ranged from 5 to 250 copies of target DNA per reaction, depending on the targeted virus. The sensitivity of the INFINITI microarray assay ranged from 10 copies of target DNA for HRSV B and HMPV A to 2500 copies for parainfluenza viruses type 2 and 3 and coronavirus 229E, and 5000 copies for parainfluenzavirus type 1. In most of the cases, the qRT-PCR assay was more sensitive than the microarray assay. However, similar sensitivities between the two assays were obtained for HRSV B and an increased sensitivity was obtained with the microarray assay for HMPV A.

8.2.3.2. Retrospective study results

The study included 221 specimens collected during the 2001-2002 winter season from children ≤ 3 years old hospitalized for ARTI. All specimens were positive by both methods for the Armored RNA internal control. Specimens were considered positive for one virus if they were positive by either one of both methods. Of the 221 specimens, 81.4% of the specimens were positive for at least one virus. Furthermore, 68.3% of the specimens were positive for one virus, 12.2% were positive for two viruses, and 0.9% was positive for three viruses. Among co-infections, 41.4% involved adenoviruses ($p < 0.001$, Fisher Exact test) and 34.5% involved picornaviruses ($p < 0.05$, Fisher Exact test).

Table 8.4 shows the sample positivity for each virus as detected by the two methods. We considered a specimen positive for a virus if it was positive with either method. The most frequently detected virus was HRSV, including 38% type B and 10% type A. Influenza A and picornaviruses (rhinoviruses or enteroviruses) were both detected in 13.1% of

specimens. Adenoviruses, coronaviruses, HMPV and parainfluenzaviruses were detected in 7.2%, 8.6%, 4.1% and 1.5% of specimens, respectively.

For some viruses, the type was also identified. Due to methodological design, the identification of virus types for adenoviruses and picornaviruses was only possible with the microarray assay. Of the 9 specimens positive by the microarray for adenoviruses, 3 were adenovirus type B, 5 were adenovirus type C, and 1 was positive for both type B and C. Of the 29 specimens positive for respiratory picornaviruses, 16 were positive for rhinovirus type A, one was positive for enterovirus type B and 11 were untyped picornaviruses. Respiratory syncytial virus types were identified by both methods for all HRSV-positive specimens which included 22 type A and 84 type B. Human metapneumovirus types were also identified by both methods for all positive specimens including 5 type A (including one microarray false negative), and 4 type B. Coronaviruses HKU1, NL63 and OC43 were identified by qRT-PCR in 9, 7 and 3 specimens, respectively.

8.2.3.3. Comparison of real-time PCR assays with the microarray assay

Overall, 79.6% of the 221 specimens were positive for at least one virus with both techniques, 18.5% were negative for all viruses by both methods and 1,8% (4/221) were positive for at least one virus by qRT-PCR only (Table 8.5). No viruses were detected with the microarray method only. The results with both methods were compared for each specimen and a concordance in the diagnosis was observed for 94.1% of the specimens. When specific diagnosis for each virus was considered, a perfect concordance between the two methods was observed for HRSV type A and B, parainfluenza viruses, HMPV type B and coronavirus OC43. Of the 5.9% (13/221) specimens with discordant results (all positive with qRT-PCR only), there were 7 adenoviruses, 2 coronaviruses NL63, 1 coronavirus HKU1, 1 HMPV A, 1 influenza A and 1 picornavirus (Table 8.6). In most cases, no signal was observed with the microarray assay for discordant specimens. However, coronavirus HKU1, influenza A and picornavirus discordant specimens gave equivocal ($0 < \text{ratio} < \text{threshold}$) results with the microarray assay and had high cycle thresholds of 38.0, 39.9 and 38.9, respectively, with the qRT-PCR assays. Discordant specimens were tested by DNA sequencing as shown in Table 8.6. Seven discordant specimens were confirmed by DNA sequencing, while we were unable to confirm

5 specimens positive for adenovirus with the qRT-PCR assay. One specimen, positive for coronavirus NL63 with the qRT-PCR assay and giving equivocal coronavirus 229E signal with the microarray assay, was shown to be positive for 229E by DNA sequencing, suggesting possible specificity issues with the qRT-PCR assay. We observed no false positive with the microarray assay when compared to the qRT-PCR assay and to DNA sequencing.

From a technical viewpoint, the microarray assay required 77 events of manual pipetting for 24 specimens, while qRT-PCR required at least 288 manual pipetting events for only four specimens and appropriate controls. Figure 8.1 shows a flowchart comparing the steps and timelines for the real-time PCR assays and for the microarray system. Excluding RNA extraction and reverse transcription, the automated microarray assay required at most 60 min of setup time for up to 24 samples, while the real-time PCR assay required the same time for 96-well plate preparation to test only four specimens.

8.2.4. Discussion

The comparison of the microarray assay automated by the INFINITI analyzer with the qRT-PCR assays showed that both techniques were useful to detect and identify a panel of respiratory viruses in clinical specimens, either present as single agents or as part of a co-infection. Overall, the single qRT-PCR assays were associated with better analytical sensitivity than the multiplex PCR followed by microarray detection. However, results suggest that the 46-primer multiplex PCR assay should, in more than 94% of cases, give results similar to those obtained with the qRT-PCR assay using nasopharygeal aspirates from children.

Because the sensitivity of the qRT-PCR assay is usually higher than that of the microarray assay, discordant specimens could generally be explained by differences in detection limits for the two assays (Tables 8.2 and 8.6). This seems particularly true for adenoviruses, for which the real-time PCR assay could detect as few as 10 copies of the target sequence, while the microarray assay detected only 250 copies of adenovirus C and 1000 copies of adenovirus B. Moreover, the mean cycle threshold of the adenovirus discordant samples (35.7) was higher than that of the adenovirus concordant samples (31.9) ($p = 0.08$, Student

t-test). In all 7 discordant adenovirus cases, no significant signal (< 0.5) was observed on the microarray assay. Only 2 of the 7 discordant adenovirus specimens were confirmed by DNA sequencing. Thus, it is unclear if these specimens are false positive of the qRT-PCR assay or if the viral load is too low to allow DNA sequencing. Differences in primer sequences may explain some discrepancies, either because of sequence variations or different PCR efficiencies. A low sensitivity for adenoviruses detection was also described in other multiplex RVP assays (308, 277).

It is of note that 69.4% (9/13) of discordant specimens had more than one virus detected by qRT-PCR ($p < 0.001$, Fisher Exact test). However, this analysis could be biased by the high rate of discordant adenoviruses that were part of co-infections (6/7). The rate of multiple infection was not different between discordant and concordant adenovirus positive specimens ($p = 0.57$, Fisher Exact test). Still, multiple viruses within a specimen could potentially reduce the sensitivity of the microarray assay for some viruses, but further studies would be required to confirm this hypothesis.

The current version of the qRT-PCR 96-well plate assay, although optimised to reduce pipetting steps, is labor intensive, time consuming and has low throughput, allowing the testing of only four clinical samples per 96-well plate. On the contrary, the microarray assay, when automated using the INFINITI analyzer, requires fewer human interventions and allows the testing of up to 24 samples per run. Due to its automation, this assay is also potentially less susceptible to manipulation errors and to cross-contamination than plate-based qRT-PCR tests. Notably, up to 5h of hands-on-time can be saved by using the automated microarray assay. The reduction in hands-on time with the microarray assay could be a financial advantage of this technique. Samples tested using the qRT-PCR assay could have results available on the same day while samples tested using the microarray assay will have results available on the next day. Although results on the same day would be the ideal scenario, ease of implementation and a higher throughput may still be important factors in choosing a molecular diagnostic assay, especially in high volume laboratories.

As new highly multiplexed molecular diagnostic devices are developed, it will be important to compare these techniques in order to establish their relative sensitivities, specificities and

ease of implementation in a clinical setting. Comparison of the automated microarray assay with commercially available respiratory virus panels, such as xTAG RVP from Luminex (308, 309), Multicode PLx RVP from Eragen (358) and Resplex II from QIAgen (312), is an essential step in assessing the quality and clinical usefulness of the next generation of respiratory virus diagnostic methods. The viruses included in each panel may vary from product to product, according to assay design and approval by regulatory institutions. Also, threshold levels should be validated in each laboratory when implementing a new multiplex molecular assay, because they may vary according to sample preparation methods and other internal issues. The ease of implementation of these techniques in a clinical setting is a critical factor in the selection of a diagnostic assay (359). Currently, because of its automation on the INFINITI analyzer, the microarray assay described herein is the most adaptable system for clinical laboratories. Moreover, the ease of use of this assay could still be improved by performing one-step reverse transcription, and by reducing the time dedicated to or completely removing the nucleotides degradation step. These improvements would further reduce the pipetting events required to test 24 specimens to only three steps per specimen. On the other hand, as for other multiplex assays, the workload of a clinical laboratory is an important criterion when selecting a diagnostic test for respiratory viruses.

While the precise identification of all viruses is not a high priority for clinicians at the moment, the multiplex assays should become increasingly helpful for epidemiological studies, to assess clinical outcome according to virus type or multiple infections, to understand the role of emerging viruses, and to limit the use of antibiotics. Future therapeutic modalities for many respiratory viruses should also increase the usefulness of these assays, in particular for immunocompromised patients and in urgent or intensive care settings. As respiratory viruses are a greater concern worldwide, such a tool will have increased usefulness for diagnosis and adequate use of antibiotics and antiviral agents.

8.2.5. Acknowledgements

FR holds a Ph.D. scholarship from the Canadian Institutes of Health Research. GB is the holder of the Canada Research Chair in Emerging Viruses and Antiviral Resistance. JC is the holder of the Canada Research Chair in Medical Genomics and is supported by the Canadian Institute of Health Research and the Canadian Foundation for Innovation. GDS

holds a clinical researcher scholarship from the Fond de la Recherche en Santé du Québec. The authors thank AutoGenomics inc. for providing reagents necessary to conduct this study.

8.2.6. Tables

Table 8.1. Sequence for primers and probes for the qRT-PCR assay.

Well	Virus	PCR primer sequence	Type	TaqMan sequence
1	Adenovirus	CCCTCACADATTTGCATTTCCCA GGTAATTTATGGACCCACTGGCTG ATTTATGGCCCCACCGGATG ATTTACGGTCCCACYGGGTG ATTTACGGGCCCACCGGC	A, B, C	6FAM-CAGATGGGGGRATCATGT-MGBNFQ 6FAM-CGAGGGTGGAATCATGT-MGBNFQ
2	Adenovirus	AGGATGCATGCGRGGGGA GCMTTCCCTTCCAAGTTGCAT	E	6FAM-AARTTCCCCAAGTGCAC-MGBNFQ
3	Armored RNA	CACTCCCCTGTGAGGAACTACTG AGGCTGCACGACACTCATACTAAC		6FAM-TTCACGCAGAAAGC-MGBNFQ
4	Coronavirus	TTGAAGGCTCAGGAAGGTCTGCT TGCTTAGTKACTTGCTGAGGTTTAG	HKU1 OC43	6FAM-TAAAACAAGATTAGCGATCTC-MGBNFQ VIC-CAGAACAAGACTAGCAATT-MGBNFQ
5	Coronavirus	TAGTCTTAYACACAATGGTARGCCAGTG TGGCTCTTCCATTGTTGGCKCG	229E NL63	6FAM-AATGCGATCTTTGATTACTCCA-MGBNFQ VIC-TATGCGATCTTTAAGTACTCCA-MGBNFQ
6	Enterovirus Rhinovirus	AGCCTGCGTGGCKGCC GAAACACGGACACCCAAAGTAGT TGGCTGCGYTGGCGG	Enterovirus A,B,C,D Rhinovirus A,B	VIC-ATTAGCCGCATTCAGG-MGBNFQ 6FAM-GTTAGCCRCATTCAGG-MGBNFQ
7	Human metapneumovirus	GGCTCCATGCAAATATGAAGTG CATCAGCTCTATCAGTGTTCCTTAAAA	A	6FAM-ACAATTACTGGAGTTGGC-MGBNFQ
8	Human metapneumovirus	GGCTCCATGCAAATATGAAGTG CATCAGCCTTATCWGTGTTTCTTAAAA	B	6FAM-ACAATTACTGGAGTTGGC-MGBNFQ

Well	Virus	PCR primer sequence	Type	TaqMan sequence
9	Human Respiratory syncytial virus	CCATTATGCCTAGACCTGCTGCA CCAGCAGCATTGCCTAATACTACACT GCAGCATTGCCTAATACTACACTGGA GAATGCCTATGGTKCAGGGCAAGT	A B	VIC-TGCTTCTCCACCCAAT-MGBNFQ 6FAM-AGCTTCTCCTCCCAACT-MGBNFQ
10	Influenza	GGCGCTGCAGTCAAAGGART GGTGCTGCAGTCAAAGGART CCCGRCTYTCTCTCACTTGATC	A	6FAM-CAGAATGATCAAACGGG-MGBNFQ 6FAM-CAGAATGGTCAAACGGG-MGBNFQ 6FAM-AGGATGATCAAACGTGG-MGBNFQ
11	Influenza	GGTGTTGCAATCAAAGGAGGTG GGTGTTGCGATCAAAGGAGGTG TGGRTTYCTACTTCCGATCACTTGATC	B	6FAM-CCATTCGATTTATAGGAAGAG-MGBNFQ
12	Parainfluenza	ACAGGAATTGGCTCAGATATGYG GACTTCCCTATATCTGCACATCCTTGAGTG	1	6FAM-ACCATGCAGACGGC-MGBNFQ
13	Parainfluenza	GCTCTTGCAGCATTTTCTGGGGA GCTCCCTGCTGTTTCCTTGC	2	6FAM-CCAGAAATTAAAAGCTCTC-MGBNFQ
14	Parainfluenza	ACAGATGTATATCAACTGTGTTCRACTCC TTGGATGTTCAAGACCTCCATAYCCG	3	6FAM-TGATGAAAGATCAGATTATG-MGBNFQ

Table 8.2. Sensitivities of the qRT-PCR assay and of the microarray assay for each virus.

Virus	Gene	qRT-PCR assays		Microarray assay	
		Threshold[a]	Sensitivity (copy number)	Threshold[b]	Sensitivity (copy number)
Adenovirus	IVA2/L4100kd[c]	40	10	1.5	1000 (B) / 250 (C)
Coronavirus 229E	Nucleocapsid	45	50	3	2500
Coronavirus HKU1	Nucleocapsid	40	10	1.5	500
Coronavirus NL63	Nucleocapsid	40	100	1.5	500
Coronavirus OC43	Nucleocapsid	45	50	2	250
Enterovirus	5' UTR	40	5	1/1,5[d]	250
Rhinovirus	5' UTR	40	250	1/1,5[d]	1000
Influenza A	Nucleocapsid	40	50	1.5	100
Influenza B	Nucleocapsid	42	10	1.5	100
HMPV A	Matrix/F[c]	40	50	1	10
HMPV B	Matrix/F[c]	40	50	3	250
HRSV A	Nucleocapsid	45	250	1	50
HRSV B	Nucleocapsid	45	10	1	10
PIV-1	Hemaglutinin-neuraminidase	43	50	2	5000
PIV-2	Fusion	40	5	2	2500
PIV-3	Hemaglutinin-neuraminidase	44	50	2	2500

[a] Maximum cycle threshold for positivity.

[b] Minimun ratio for positivity.

[c] Target gene different between qRT-PCR/microarray assay.

[d] Threshold for picornaviridae detection/Threshold for species specific detection.

Table 8.3. Specimens tested in the analytical specificity study.

Virus	Number of specimens
Adenovirus B	1
Adenovirus C	1
Adenovirus E	1
Coronavirus 229E	1
Coronavirus NL63	1
Coronavirus OC43	1
Enterovirus B	2
Human metapneumovirus	1
Human respiratory syncytial virus A	1
Human respiratory syncytial virus B	5
Influenza A	3
Influenza B	4
Parainfluenza 1	4
Parainfluenza 2	4
Parainfluenza 3	4
Parainfluenza 4	5
Rhinovirus A	2

Table 8.4. Sample positivity by qRT-PCR and microarray for each virus (n = 221).

Virus	qRT-PCR	Microarray
Adenovirus A	16 (7.2%)[a]	0 (0.0%)
Adenovirus B		4 (1.8%)
Adenovirus C		6 (2.7%)
Adenovirus E		0 (0.0%)
Coronavirus 229E	0 (0.0%)	0 (0.0%)
Coronavirus HKU1	9 (4.1%)	8 (3.6%)
Coronavirus NL63	7 (3.2%)	5 (2.3%)
Coronavirus OC43	3 (1.4%)	3 (1.4%)
Influenza A	29 (13.1%)	28 (12.7%)
Influenza B	0 (0.0%)	0 (0.0%)
Human metapneumovirus A	5 (2.3%)	4 (1.8%)
Human metapneumovirus B	4 (1.8%)	4 (1.8%)
Human parainfluenzavirus 1	1 (0.5%)	1 (0.5%)
Human parainfluenzavirus 2	1 (0.5%)	1 (0.5%)
Human parainfluenzavirus 3	1 (0.5%)	1 (0.5%)
Human respiratory syncytial virus A	22 (10.0%)	22 (10.0%)
Human respiratory syncytial virus B	84 (38.0%)	84 (38.0%)
Picornavirus (Rhinovirus or Enterovirus)	29 (13.1%)[a]	28 (12.7%)[b]
Rhinovirus A	NA[a]	16 (7.2%)
Rhinovirus B	NA[a]	0 (0.0%)
Enterovirus A	NA[a]	0 (0.0%)
Enterovirus B	NA[a]	1 (0.5%)
Enterovirus C	NA[a]	0 (0.0%)
Enterovirus D	NA[a]	0 (0.0%)

[a] Genotyping was not done with the qRT-PCR method.

[b] Independent probe targeting all picornaviruses (enterovirus and rhinovirus). A specimen can be positive for this probe without being positive for a type-specific enterovirus or rhinovirus probe.

Table 8.5. Comparison of qRT-PCR and microarray results for 221 specimens.

Virus	qRT-PCR/microarray				Sensitivity of microarray	Specificity of microarray
	Pos/Pos	Pos/Neg	Neg/Pos	Neg/Neg		
Adenovirus[a]	9	7	0	205	0.563	1.000
Coronavirus 229E	0	0	0	221	NA	1.000
Coronavirus HKU1	8	1	0	212	0.889	1.000
Coronavirus NL63	5	2	0	214	0.714	1.000
Coronavirus OC43	3	0	0	218	1.000	1.000
Enterovirus/Rhinovirus	28	1	0	192	0.966	1.000
Influenza A	28	1	0	192	0.966	1.000
Influenza B	0	0	0	221	NA	1.000
HMPV A	4	1	0	216	0.800	1.000
HMPV B	4	0	0	217	1.000	1.000
HRSV A	22	0	0	199	1.000	1.000
HRSV B	84	0	0	137	1.000	1.000
PIV-1	1	0	0	220	1.000	1.000
PIV-2	1	0	0	220	1.000	1.000
PIV-3	1	0	0	220	1.000	1.000
Any virus	176	4	0	41	0.978	1.000

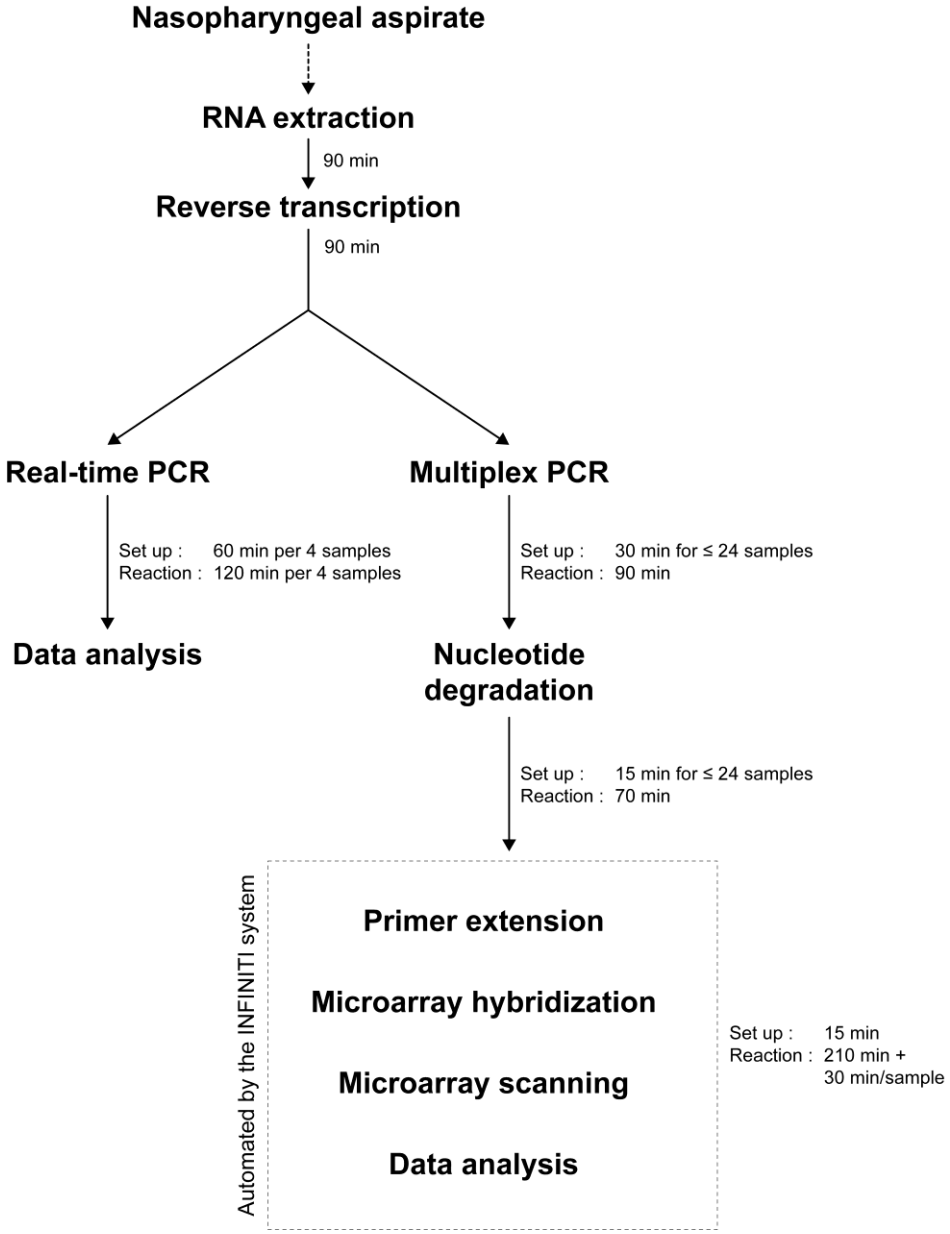
[a] Viruses for which results were significantly different between qRT-PCR and microarray ($p < 0.05$, McNemar).

Table 8.6. Discordant specimens between the qRT-PCR and the microarray assay with signals, internal controls and validation results.

	qRT-PCR assay			*Microarray assay*			
ID	**Detected viruses (discordant in bold)**	**Signal of discordant (Ct)**	**IC (Ct)**	**Detected viruses**	**Signal of discordant (ratio)**	**IC (ratio)**	**Validation sequencing**
032	**HKU1**, HRSV-B	38.04	35.03	HRSV-B	0.92	3.51	Coronavirus HKU1
034	**Influenza A**	39.88	32.09	Negative	0.71	10.82	Influenza A
049	**Adenovirus**, Influenza A, HRSV-A	35.54	32.28	Influenza A, HRSV-A	0.00	8.08	Adenovirus C
069	**NL63**, Influenza A	39.27	30.80	Influenza A	2.44 (229E)	9.00	Coronavirus 229E
075	**Adenovirus**, HRSV-B	37.97	31.16	HRSV-B	0.00	11.33	Negative
081	**Adenovirus**	32.49	33.34	Negative	0.00	2.56	Negative
092	**Adenovirus**, Influenza A	35.35	32.19	Influenza-A	0.00	2.49	Negative
121	**HMPV-A**	38.44	34.54	Negative	0.00	2.88	HMPV A
156	**NL63**, HRSV-B	35.70	34.56	HRSV-B	0.00	11.56	Coronavirus NL63
173	**Adenovirus**, Rhinovirus	34.66	33.04	Rhinovirus A	0.00	8.53	Adenovirus C
182	**Adenovirus**, HRSV-B	38.90	32.48	HRSV-B	0.31	10.34	Negative
186	**Adenovirus**, Rhinovirus	35.31	32.44	Rhinovirus A	0.33	8.86	Negative
218	**Rhinovirus/Enterovirus**	38.90	37.68	Negative	0.78	9.71	Rhinovirus A

8.2.7. Figure

Figure 8.1. Flowchart comparing the protocols for the real-time PCR assay and for the microarray assay. RNA extraction and reverse transcription steps are common to both methods. Real-time PCR assay has only one setup step compared to the microarray assay, which has three. However, the time required to perform the real-time PCR assay multiplied for each four samples to test. The time required for the automated microarray assay is multiplied for each 24 samples. Overall, the qRT-PCR assay requires 60 min of setup time and a total of 120 min of reaction time for 4 specimens, while the microarray assay requires 60 minutes of setup time and 17 hours of reaction time for 24 specimens.



9. Le diagnostic moléculaire de la grippe A (H1N1)

9.1. Le résumé de l'article

9.1.1. Le résumé en français

Le test diagnostique présenté au chapitre 8 a été amélioré afin de permettre la détection de l'influenza A (H1N1) d'origine porcine responsable de la pandémie de 2009. La spécificité de ce test pour la détection de l'influenza A (H1N1) a été validée et la sensibilité technique du test a été estimée à 500 PFU/ml dans 95,0 % des cas. Une étude réalisée d'avril à juin 2009 incluait 154 échantillons obtenus de patients soupçonnés d'infection par l'influenza pandémique. Au total, 24,0 % des échantillons étaient positifs pour l'influenza A (H1N1) et 13,5 % étaient positifs pour un autre virus respiratoire.

9.1.2. Abstract

Swine-origin influenza A/H1N1 virus (S-OIV) and 23 other respiratory viruses were detected using the AutoGenomics RVP Plus assay in 154 samples obtained from patients with suspicion of S-OIV infection during Spring 2009. Overall, 24.0% of samples were positive for S-OIV and 13.5% were positive for another virus without co-infections.

9.2. L'article

Diagnosis of swine-origin influenza A/H1N1 virus (S-OIV) and other respiratory viruses in patients with suspected S-OIV infections using multiplex PCR and automated DNA microarray

Running title: Diagnosis of S-OIV and other respiratory viruses

Frédéric Raymond[1], Marie-Ève Hamelin[1], Julie Carbonneau[1], Nancy Boucher[1], Lynda Robitaille[1], Jacques Corbeil[1] and Guy Boivin[1]

[1] Infectious Disease Research Center of the CHUQ-CHUL and Laval University, Québec City, Québec, Canada.

Correspondence: guy.boivin@crchul.ulaval.ca

9.2.1. Article

Past experience with the SARS coronavirus in 2003 and the A/H5N1 avian influenza virus since 2004 allowed a fast and effective response to the emergence of a new swine-origin A/H1N1 influenza virus (S-OIV) first detected during April 2009 in Mexico (361-363). New cases were subsequently diagnosed in the U.S.A., Canada and then throughout the world. Signature sequences for this reassortant swine virus were rapidly deposited into Genbank, allowing the scientific community to track the evolution and spread of the virus (364). As of June 29, over 70,000 human cases were reported in over 110 countries. The median age of the first 47 cases was 16 years in a U.S. cohort (363). Relatively few cases have been diagnosed in elderly subjects apparently due to the presence of cross-reactive neutralizing antibodies (365). There is a real concern that the virus may undergo reassortment or mutational events leading to increased disease severity.

In order to improve the diagnosis of S-OIV and to provide additional information on the role of mixed viral infections, we updated the recently developed AutoGenomics RVP assay (Carlsbad, CA). The original RVP assay allows for the detection of 23 respiratory viral species and genotypes (313). The new version of the assay, named RVP Plus, was tested on samples prospectively collected from patients with suspicion of S-OIV infection, thus documenting the epidemiology of the initial S-OIV cases in Quebec City. The latter was initially based on travel to an epidemic area or possible contact with confirmed cases and then later based on the presence of typical influenza-like illnesses at a time when epidemic influenza viruses were infrequently detected. All samples were collected according to the ethical procedures of the Centre Hospitalier Universitaire de Québec. Nucleic acids were extracted from 200 µl of nasopharyngeal aspirates (NPA) using the Qiagen viral RNA mini kit (Qiagen, Mississauga, Ontario, Canada) as reported (313) and then RNA was reverse transcribed using the Omniscript Reverse Transcriptase kit (Qiagen). The composition of the RVP Plus assay is similar to that of the regular RVP assay, except that it contains two new PCR primers and one new detection primer targeting the S-OIV nucleoprotein at the same position as the RVP assay, and two new PCR primers, one new detection primer and one probe targeting specifically the S-OIV hemaglutinin gene. All influenza A strains, including S-OIV, are detected by targeting the nucleoprotein

(NP) gene, while specific S-OIV detection is done by targeting the S-OIV hemaglutinin (HA) gene. Primer extension and microarray hybridization steps were performed with the automated Infiniti analyzer (AutoGenomics). To evaluate the performance of the RVP Plus assay for the detection of S-OIV, all samples were tested by conventional PCR assay for the Influenza A matrix gene (366) followed by a specific conventional PCR assay for the S-OIV HA gene performed by the Quebec Public Health Laboratory (367). Partial sequence of all S-OIV positive samples hemaglutinin gene was sequenced bidirectionally to confirm the diagnosis.

The analytical sensitivity of the RVP Plus assay for S-OIV detection was determined using two independent isolates of S-OIV, which were re-grown in vitro. Both cultures were tittered using plaque assays. The re-grown isolates were first diluted to a concentration of 10^6 PFU/ml, which was later diluted using serial 1:10 dilutions to concentrations of 10^5 and 10^4. Dilution series were performed in triplicate. For each dilution series of each isolate, viral RNA was extracted using the QIAgen viral mini kit as previously described (313). Further dilutions were performed on each 10^4 PFU/ml extracted viral RNA to obtain concentrations exquivalent to 1000, 500, 100 and 10 PFU/ml. Samples were then reverse transcribed as previously described (313). The samples were then submitted to the RVP Plus assay. The limit of detection was defined as the dilution point positive for each replicate for both the generic flu probe and the S-OIV specific probe. Twenty replicates of this dilution were performed using the same protocol as for the dilution series, starting with the 10^4 PFU/ml viral RNA sample. The limit of detection of the AutoGenomics RVP Plus assay for S-OIV detection was determined to be 500 PFU/ml in 95% of cases.

Twenty-two clinical specimens previously screened positive for respiratory viruses other than S-OIV were tested using the RVP Plus assay and produced similar results as the original RVP assay, with no false positive or false negative results observed. Genomic DNA from twelve bacteria species frequently found in the respiratory tract did not show any cross reactivity with the RVP Plus assay. The assay was also tested against a panel of representative influenza A and B strains, with positive results for the influenza probe and negative results for the S-OIV probe. List of these specimens is shown in Table 9.1.

A total of 154 clinical NPA specimens collected between April and June 2009 were included in the study, tested with the conventional gold standard PCR and tested by the RVP Plus assay. Of the 37 specimens that were positive for S-OIV as determined by the conventional assays, 33 were also found to be positive with the RVP Plus assay (Table 9.2). The four false negative specimens by the RVP Plus assay contained low copy numbers with quantitative PCR threshold (Ct) values higher than 39.5, as measured by a real-time PCR assay (368). In contrast, the mean Ct value of concordant specimens was 32.0 ($p < 0.05$).

Overall, one specimen was positive for influenza B whereas none was positive for epidemic influenza A viruses. A total of 21 specimens (13.5%) were found to be positive for respiratory viruses other than influenza (Table 9.3). No mixed viral infections were observed whereas 100 (64.9%) specimens were negative for all tested viruses. The negative samples with flu-like symptoms could be due to inadequate sampling time relative to onset of symptoms, bacterial infections or viral agents not targeted by the RVP Plus assay such as PIV-4, bocaviruses, parechoviruses.

Our study reports the validation of a new automated microarray system for detection of S-OIV (sensitivity: 0.89, specificity: 1.00 compared to single conventional PCR assay) and other respiratory viruses during the initial phase of the S-OIV outbreak in Canada. Of note, we did not detect S-OIV in combination with other respiratory viruses in patients sampled during the months of April to June 2009. However, since other respiratory viruses could mimic the clinical presentation associated with S-OIV, it is of prime interest to test samples for a large panel of respiratory viruses. In future influenza seasons, particular emphasis should be taken for patients who will test positive for both epidemic influenza A and the new S-OIV strain since this could lead to reassortment events and increased virulence. Rapid detection using a multiplexed assay that allows the identification of epidemiologically relevant strains of various respiratory viruses is an essential tool for the management of flu-like illnesses in a period not exclusively dominated by circulation of S-OIV. The S-OIV test could be completed in < 16 h and allows for simultaneous testing of 24 samples. It is thus particularly suitable for reference and hospital diagnostic laboratories.

9.2.2. Acknowledgements

FR holds a Ph.D. scholarship from the Canadian Institutes of Health Research. GB is the holder of the Canada Research Chair in Emerging Viruses and Antiviral Resistance and is a Canadian team leader on pandemic influenza supported by the Canadian Institutes of Health Research. JC is the holder of the Canada Research Chair in Medical Genomics and is supported by the Canadian Institute of Health Research and the Canadian Foundation for Innovation. The authors declare competing interests. The authors thank AutoGenomics inc. for providing reagents necessary to conduct this study.

9.2.3. Tables

Table 9.1. Cross-reactivity and inclusivity study results for Influenza A and S-OIV detection in clinical samples of various respiratory viruses and reference strains of influenza viruses or bacteria.

Sample type	Bacteria	Influenza A	S-OIV	Armored RNA (Internal Control)
Non-influenza respiratory viruses (clinical samples)	Adenovirus C (GE10029)	-	-	+
	Adenovirus C (GE10032)	-	-	+
	Coronavirus HKU1 (GE066)	-	-	+
	Coronavirus HKU1 (GE10061)	-	-	+
	Enterovirus A (GE10034)	-	-	+
	Influenza B (GE004)	-	-	+
	HMPV-A (GE072)	-	-	+
	HMPV-B (GE011)	-	-	+
	HRSV-A (GE078)	-	-	+
	HRSV-A (GE049)	-	-	+
	HRSV-A (GE018)	-	-	+
	HRSV-A and HMPV-V (GE075)	-	-	+
	HRSV-B (GE069)	-	-	+
	HRSV-B (GE093)	-	-	+
	HRSV-B (GE099)	-	-	+
	HRSV-B (GE002)	-	-	+
	Negative (GE10100)	-	-	+
	Parainfluenzavirus 1 (P30)	-	-	+
	Parainfluenzavirus 3 (P4)	-	-	+
	Parainfluenzavirus 3 (GE10057)	-	-	+
	Parainfluenzavirus 4 (GE021)	-	-	+
	Rhinovirus or Enterovirus (GE059)	-	-	+
Influenza reference strains	A/WSN/33 (H1N1)	+	-	+
	A/New Caledonia/20/1999 (H1N1-like, vaccine)	+	-	+
	A/Beijing/262/95 (H1N1)	+	-	+
	A/Wisconsin/67/2005 (H3N2-like, vaccine)	+	-	+
	A/Hong Kong/8/68 (H3N2)	+	-	+
	A/Wuham/7244/98	+	-	+
	A/Fujian/2857/03	+	-	+
	A/Panama/5502/98	+	-	+
	A/Sydney/5/97 (H3N2)	+	-	+
	B/Malaysia/2506/2004	-	-	+
	B/Lee/40	-	-	+
	B/Taiwan/2/62	-	-	+
	B/Florida/2006	-	-	+
	B/Harbin/07/94	-	-	+

Sample type	Bacteria	Influenza A	S-OIV	Armored RNA (Internal Control)
Bacteria	*Streptococcus pneumoniae* (ATCC 27336)	-	-	+
	Bordetella pertusis (ATCC 9797)	-	-	+
	Streptococcus salivarius (ATCC 7073)	-	-	+
	Streptococcus pyogenes (ATCC 12384)	-	-	+
	Neisseria meningitidis (ATCC 13077)	-	-	+
	Haemophilus influenzae (ATCC 9006)	-	-	+
	Corynebacter jeikeium (CCRI 15978)	-	-	+
	Moraxella catarrhalis (ATCC 25238)	-	-	+
	Staphylococcus aureus (ATCC 29213)	-	-	+
	Pseudomonas aeruginosa (ATCC 35554)	-	-	+
	Escherichia coli (ATCC 43886)	-	-	+
	Staphylococcus epidermidis (ATCC 35984)	-	-	+

Table 9.2. Contingency table for S-OIV detection using the AutoGenomics RVP Plus assay compared to results obtained from the reference laboratory.

		Reference laboratory[b] (n = 154)	
		Positive	Negative
RVP Plus assay[a]	Positive	33	0
	Negative	4	117

[a] Sensitivity, 33 of 37 (89.2%); specificity; 117 of 117 (100%); positive predictive value, 33 of 33 (100%); negative predictive value, 117 of 121 (96.7%).

[b] Conventional RT-PCR assay for the influenza A matrix gene followed by conventional RT-PCR assay for the S-OIV HA gene.

Table 9.3. Specimens positive for different viruses including S-OIV using the AutoGenomics RVP Plus assay.

Virus	Samples (n = 154)	Percent of samples
Adenovirus A	0	0.0%
Adenovirus B	2	1.3%
Adenovirus C	0	0.0%
Adenovirus E	5	3.2%
Coronavirus 229E	0	0.0%
Coronavirus HKU1	0	0.0%
Coronavirus NL63	0	0.0%
Coronavirus OC43	1	0.6%
Seasonal influenza A	0	0.0%
Swine-origin A/H1N1 influenza A	33	21.4%
Seasonal influenza B	1	0.6%
Human metapneumovirus A	5	3.2%
Human metapneumovirus B	0	0.0%
Human parainfluenzavirus 1	1	0.6%
Human parainfluenzavirus 2	0	0.0%
Human parainfluenzavirus 3	3	1.9%
Human respiratory syncytial virus A	1	0.6%
Human respiratory syncytial virus B	0	0.0%
Untypable picornaviridae	0	0.0%
Rhinovirus A	2	1.3%
Rhinovirus B	0	0.0%
Enterovirus A	0	0.0%
Enterovirus B	0	0.0%
Enterovirus C	0	0.0%
Enterovirus D	0	0.0%
No virus detected	100	64.9%

10. La discussion

En 1999, l'auteur de science-fiction Neal Stephenson a publié un essai, intitulé *In the beginning was the command line,* qui traite de l'histoire de l'informatique et des systèmes d'exploitation. Adaptant ce pastiche biblique à la génomique, nous pourrions le transformer en *In the beginning was the sequence* parce que, sans la capacité de séquencer les génomes, la génomique n'existerait pas, le diagnostic moléculaire non plus. C'est dans cette optique que je vais conclure cette thèse, en soulignant et en discutant les réalisations et les découvertes faites au cours de mon doctorat, et en les reliant par le fil conducteur de la bio-informatique et de la génomique.

De prime abord, mes travaux sur le parasite *Leishmania* (chapitres 4 et 5) et sur le diagnostic des virus respiratoires (chapitres 7, 8 et 9) semblent avoir peu de choses en commun. Cependant, dans les deux cas, le point central de ces études était le matériel génétique de ces organismes. Au fond, toutes deux ont débuté par l'analyse d'interminables séquences de nucléotides. Il s'agissait de donner un sens à une masse informe de séquences afin d'en extraire de l'information biologique pertinente ou de les transformer en un outil diagnostique. Pour tous ces projets, la conception de méthodes bio-informatiques à usage ponctuel ou récurrent a nécessité l'usage des mêmes outils, notamment les langages de programmation Perl et R, et l'usage de la ligne de commande Linux comme un outil de traitement de données et d'automatisation de tâches. À plusieurs moments, il a aussi été nécessaire d'utiliser et d'évaluer des logiciels créés par la communauté scientifique afin de sélectionner ceux qui étaient les mieux adaptés aux problèmes à résoudre. Peu importe l'organisme étudié, l'analyse de la séquence de leur génome reste similaire.

10.1. La gestion des données de transcriptomique

Quand la première génération de la biopuce *Leishmania* a été créée, il a été jugé important de mettre sur pied un système de gestion des données qui faciliterait la conservation et l'analyse des données générées par ces méthodes. Des outils de gestion de données de biopuces étaient disponibles dans la communauté, les plus connus étant la première génération de BASE (189) et MADAM (369), créé par le TIGR Institute.

Cependant, ces deux outils présentaient des limites qui les rendaient peu appropriés pour les projets prévus avec la biopuce *Leishmania.* En effet, la première version de BASE était lourde d'utilisation et il était difficile d'extraire les données qu'on y introduisait. Pour ce qui est de MADAM, le contenu du logiciel aurait été approprié pour les projets prévus, mais le programme ne permettait pas d'accéder aux données à partir de plusieurs postes de travail. Des outils commerciaux ont été évalués, mais leurs coûts étaient prohibitifs. Finalement, nous avons testé un logiciel Web créé à la Faculté de foresterie et de géomatique de l'Université Laval (190). Ce logiciel avait les caractéristiques recherchées pour les projets de biopuce *Leishmania,* mais il avait été conçu pour un projet particulier à la foresterie et il était difficile de l'adapter à l'étude du parasite *Leishmania.* Ainsi, nous nous sommes inspirés des travaux de Hugo Bérubé (190) pour concevoir notre propre système de gestion et d'analyse de biopuces que nous avons nommé LMMAP, pour *Leishmania microarray management and analysis platform*.

De manière générale, le système devait répondre à trois besoins :

1. Être utilisable à partir de plusieurs postes de travail;
2. Conserver les données expérimentales à long terme;
3. Analyser automatiquement les résultats bruts afin de générer des données publiables sans intervention humaine.

Le premier besoin a été comblé en créant une interface Web permettant l'interaction avec une base de données. Le second besoin correspond à la portion cahier de laboratoire électronique (LIMS) de LMMAP et le troisième besoin, à la portion d'analyse de LMMAP.

Plusieurs choix de conception ont été faits pour créer la portion LIMS de LMMAP. Le système a été créé en se basant sur les concepts « Facteur », « Condition », « Expérience » et « Projet » :

- Un facteur correspond à un paramètre expérimental, par exemple la température de culture, l'espèce étudiée, la présence ou l'absence d'un agent antiparasitaire, etc.;

- Une condition peut inclure un ou plusieurs facteurs qui définissent chacun des échantillons testés, par exemple une souche de *L. major* résistante à l'antimoine pentavalent;

- Une expérience inclut deux ou plusieurs conditions que l'on compare pour poser une question biologique, par exemple une expérience qui comparerait une souche de *L. major* résistante à l'antimoine avec une souche sensible;

- Un projet peut inclure plusieurs expériences qui sont liées entre elles par des échantillons biologiques, par exemple si un mutant résistant était comparé à un autre mutant résistant dans une nouvelle expérience.

Ensemble, ces éléments permettent de décrire une expérience dont les grandes lignes pourront ensuite être comprises par un utilisateur autre que son créateur. À eux seuls, ces renseignements permettent de générer l'information particulière qui sera nécessaire à l'analyse statistique de l'expérience de biopuce. En périphérie de ces paramètres de base, de l'information supplémentaire peut être ajoutée à la base de données pour mieux décrire les expériences. La dernière étape de la portion LIMS de LMMAP favorise l'archivage des données brutes de l'expérience, soit les fichiers d'images issus de la numérisation de la biopuce et les résultats numériques extraits de ces images.

L'analyse des biopuces *Leishmania* est faite avec le logiciel d'analyse statistique R et les librairies BioConductor (160). Un avantage de ce logiciel est son intégration facile à une interface Web. Ainsi, une fois le protocole d'analyse conçu, il a été incorporé à LMMAP. La plupart des étapes de normalisation de biopuces étaient réalisables à l'aide de la librairie LIMMA du logiciel R (161). Pour chacune des générations de la biopuce *Leishmania,* un protocole de normalisation des données a été conçu. Le principal défi de conception de ces protocoles était l'association entre les sondes et les gènes des espèces, qui pouvait être étudiée avec les biopuces *Leishmania.* Cela a été réalisé en comparant les sondes avec le génome des différentes espèces de *Leishmania* et en associant chaque sonde à une catégorie représentant son identité nucléotidique avec un génome particulier. Ces groupes de sondes étaient ensuite associés à une pondération qui modulait l'impact de chaque sonde sur la normalisation et l'analyse statistique des résultats.

En utilisation depuis cinq ans, LMMAP a permis de conserver les expériences d'une manière simple et efficace qui permet de consulter aisément les résultats bruts et analysés d'expériences de biopuces. Ainsi, la ligne centrale de LMMAP était parfaitement adaptée aux projets qui ont été réalisés. Jusqu'ici, LMMAP a contribué à la publication de quatre articles scientifiques (voir l'annexe 1), et à la préparation d'autres études en cours d'analyse ou de rédaction, ou soumises à des revues scientifiques.

10.2. La transcriptomique : limites et réussites

L'étude du transcriptome de *Leishmania* en fonction de son stade de développement occupe une place importante dans la recherche sur le parasite *Leishmania.* Cependant, malgré de nombreuses études sur le sujet, les données n'ont pas été unifiées afin de profiter d'une analyse bonifiée par leur intégration. Cette tâche est cependant difficile à accomplir parce les laboratoires utilisent des plateformes différentes. Souvent, les résultats issus d'expériences indépendantes, que ce soit dans un même laboratoire ou dans des laboratoires distincts, produisent des résultats qui semblent, de prime abord, fort différents.

D'un point de vue statistique, la variabilité expérimentale variera selon la méthode, selon l'utilisateur et selon de nombreux autres facteurs propres à l'expérience (370). Ainsi, la puissance expérimentale sera différente entre deux expériences, ce qui modifiera la liste de gènes modulés. Si des sondes de séquences et de longueurs différentes sont utilisées dans deux expériences, elles peuvent entraîner des différences entre les ratios obtenus, ce qui pourrait rendre un gène significatif sur une plateforme et non sur une autre. Aussi, la méthode de correction de la p-value et le seuil de détection utilisés pour considérer un gène significativement modulé font aussi varier les résultats finaux des expériences tels que publiés.

Plus important encore, la méthodologie expérimentale peut influencer l'expression des gènes du parasite. Les méthodes de différenciation de *Leishmania,* les temps d'incubation auxquels les échantillons sont prélevés, les milieux de culture, les traitements effectués sur les cellules avant l'extraction de l'ARN et les souches étudiées sont des facteurs pouvant influencer les résultats obtenus. Comme la différenciation de *Leishmania* est un processus dynamique qui inclut plusieurs étapes dont les processus sont régulés dans le temps, il est

possible que les gènes exprimés à un temps précis soient décalés entre une espèce et une autre, ou entre des méthodes de différenciation ou de culture différentes. De plus, comme la population de *Leishmania* étudiée n'est pas nécessairement synchronisée dans son développement, cela peut aplatir le signal des gènes qui n'ont qu'une expression transitoire très brève. L'utilisation d'un protocole de série temporelle rapprochée, étalé sur la longueur de la culture de parasites, pourrait permettre une meilleure compréhension de la transcriptomique de *Leishmania* au cours de son développement. De plus, la reproduction des expériences à l'échelle génomique reste un défi important. C'est pourquoi une partie des résultats d'expériences de biopuces sont validés par une autre méthode, comme le qRT-PCR.

Les études de la différenciation du parasite *Leishmania* décrites dans la section 4.8 exemplifient bien l'influence de la méthodologie sur l'expression des gènes de *Leishmania.* Notamment, Rochette et ses collaborateurs ont comparé, dans un article publié en 2009 (47), deux cultures de *L. major,* l'une axénique et l'autre issue de lésions. Dans cette étude, seulement 12 % des gènes modulés étaient similaires entre les deux méthodes de différenciation des parasites. De telles observations, faites sur des biopuces traitées par la même personne, dans le même laboratoire avec les mêmes méthodes d'analyse, mettent en exergue la difficulté de comparer entre elles des expériences de biopuces pour étudier le cycle de développement de *Leishmania.* Cela sera encore plus vrai si l'on veut comparer des expériences issues de laboratoires différents.

Au bout du compte, les gènes communs entre les expériences comparées seront des cibles de confiance pour l'étude de la différenciation du parasite, puisqu'ils résistent à la variabilité expérimentale et au biais engendré par les méthodes de culture. Par contre, les gènes dont l'association à un stade de développement n'est pas constante d'une étude à une autre ne doivent pas nécessairement être oubliés. Afin de tirer profit de ces études d'expression génique selon les stades de développement de *Leishmania,* une compilation des différentes études publiées a été effectuée et utilisée pour approfondir l'analyse du génome de *L. tarentolae* décrit dans le chapitre 5. Ainsi, appliquer les données issues d'études d'expression génique à d'autres contextes, comme la génomique comparative, peut offrir un nouvel angle pour l'étude du parasite *Leishmania.*

En plus de permettre l'étude de l'expression des gènes selon le stade de développement du parasite, les biopuces *Leishmania* ont aussi permis la formulation d'hypothèses concernant l'adaptation au stress du parasite. L'étude de l'expression des gènes de *Leishmania* dans des mutants résistants à des antiparasitaires, comme le méthotrexate, a permis d'observer que *Leishmania* avait tendance à amplifier des groupes de gènes contigus lorsqu'ils amplifiaient un gène particulier pour résister aux molécules qui leur seraient autrement fatales. Ainsi, à la fois dans les études de Ubeda et ses collaborateurs (196) et de Leprohon et ses collaborateurs (197), des études de transcriptomiques, permettant de comparer des mutants résistants à un médicament avec des souches sensibles, ont permis de cibler ces régions. À la suite de ces études, les biopuces *Leishmania* ont été hybridées avec de l'ADN génomique afin de vérifier si l'amplification des régions était au niveau de l'ARN messager seulement ou au niveau de l'ADN génomique. Il a été découvert que les régions amplifiées l'étaient au niveau génomique.

Cela coïncide avec ce qui est connu du parasite *Leishmania,* qui amplifie parfois, sous forme d'amplicons extrachromosomiques circulaires ou linéaires, des portions de chromosomes afin de résister à des antiparasitaires (34, 38, 371-377). La formation d'amplicons a aussi été observée en réponse à des tentatives d'inactivation génique (en anglais, *knockout*) (378). Les deux études de biopuces (196, 197) ont permis de cibler ces régions et, à l'aide de la séquence des génomes de *L. infantum* et *L. major,* de déterminer que les amplifications géniques requéraient la présence de séquences répétées au début et à la fin des régions amplifiées. Les répétitions qui étaient dans le même sens sur le chromosome, aussi appelées « répétitions directes », permettraient les amplifications circulaires alors que les répétitions inversées, aussi appelées « répétitions indirectes », permettraient les amplifications linéaires (196). Ces amplicons sont formés par des recombinaisons homologues de ces séquences répétées. Des études approfondissant ces hypothèses en étudiant le contenu en répétitions des génomes de *L. infantum* et de *L. major* sont en cours. Des approches bio-informatiques sont utilisées pour identifier toutes les paires de répétitions possibles dans ces génomes et des approches biologiques sont utilisées pour valider ces hypothèses.

10.3. Le génome de *Leishmania tarentolae*

Jusqu'ici, les études décrites dans cette thèse de doctorat utilisaient des séquences provenant de bases de données publiques pour réaliser les analyses et mieux comprendre le parasite *Leishmania*. Le chapitre 5 de cette thèse décrit le séquençage du génome de *Leishmania tarentolae,* une espèce découverte chez le lézard et qui n'infecte pas l'humain.

L'étude d'un génome nouvellement séquencé se fait en plusieurs étapes :

1. Assemblage du génome;
2. Annotation du génome;
3. Analyse du contenu du génome et génomique comparative.

Au début du projet, les logiciels disponibles pour l'assemblage des génomes séquencés par la technologie 454 étaient limités. Ainsi, nous avons généré plusieurs versions insatisfaisantes du génome de *L. tarentolae* en utilisant CABOG (138), EULER (139) et Newbler version 1 (140). Ce n'est que lorsque la version 2 de Newbler a été disponible que l'assemblage obtenu a été satisfaisant, c'est-à-dire avec un nombre de contigs moins élevés et des erreurs d'assemblage moins fréquentes.

Une fois que cette version a été disponible, nous y avons intégré des données de séquençage non appariées générées avec la technologie Solexa, de Illumina, par le Netherlands Cancer Institute. Bien que la longueur limitée de ces fragments (59 nucléotides) et l'absence d'appariement de séquences limitaient leur utilité, ces données ont permis d'améliorer notre assemblage du génome de *L. tarentolae.* Elles ont aussi permis la validation des séquences obtenues par la technologie 454 lorsque nous avons décelé des erreurs potentielles dans la séquence, en particulier si ces dernières entraînaient des changements de cadre de lecture.

Par contre, l'assemblage en une seule étape de fragments courts et de fragments longs reste un défi bio-informatique qui n'a pas encore été tout à fait résolu (142). Pour cette raison, dans le cas du génome de *L. tarentolae,* les fragments de la technologie 454 et les fragments issus de la technologie Solexa ont été assemblés indépendamment pour être

ensuite combinés à l'aide du logiciel Minimus2 (210). Finalement, afin de transformer l'assemblage en séquences de chromosomes, nous avons ordonné les contigs et les échafaudages obtenus pendant les étapes précédentes en se basant sur le génome dont la séquence était la plus proche de *L. tarentolae : Leishmania major.* Les séquences ainsi obtenues ont ensuite pu être annotées et analysées.

Certaines des limites techniques que nous avons observées avec cette étude seraient minimisées ou absentes si l'on séquençait ce génome avec les technologies offertes en 2011. Premièrement, les fragments séquencés seraient plus longs (> 500 nucléotides au lieu d'environ 250 nucléotides). Deuxièmement, toutes les expériences produiraient des fragments appariés, idéalement avec plusieurs distances séparant les fragments afin de faciliter l'assemblage et la formation d'échafaudages dans les régions répétées. Finalement, l'utilisation des dernières générations de séquenceurs de Illumina limiterait les erreurs de séquençage comme les *carryforwards* ou les erreurs dans les homopolymères tout en produisant des fragments appariés plus grands que 150 nucléotides.

La deuxième étape du processus est l'annotation du génome, qui consiste notamment à trouver la position des gènes dans la séquence et à leur assigner des fonctions potentielles. Dans le cas de *L. tarentolae,* comme d'autres génomes d'espèces de *Leishmania* et de *Trypanosoma* ont déjà été séquencés, la plupart des gènes de *L. tarentolae* ont pu être identifiés par comparaison avec ces autres espèces. Cela a été réalisé en utilisant le logiciel BLAST pour comparer le génome de *L. tarentolae* avec les gènes prédits de *L. infantum,* de *L. major,* de *L. braziliensis* et des trypanosomes, ainsi que pour comparer les cadres de lecture ouverts de plus de 100 acides aminés trouvés dans le génome de *L. tarentolae* avec la séquence des protéines prédites pour les espèces susmentionnées. La combinaison des approches nucléotidiques et protéiques a permis d'obtenir de l'information sur les gènes potentiellement présents dans le génome de *L. tarentolae,* même dans le cas où il s'agirait de pseudogènes potentiels, qu'ils soient réels ou causés par des erreurs de séquençage.

Finalement, afin de détecter des gènes qui seraient uniques à *L. tarentolae* et non annotés chez les autres espèces, nous avons utilisé un logiciel de détection de gènes de type *ab initio,* qui détecte les gènes sans les comparer avec les gènes des autres espèces. Afin d'obtenir de meilleurs résultats, le logiciel Augustus a été adapté à la détection de gènes

chez *Leishmania* en créant un modèle basé sur les gènes des autres espèces de *Leishmania* connues. Même si *Leishmania* est un eucaryote, ses gènes ne contiennent pas d'introns, ce qui ne permet pas d'utiliser les logiciels conçus pour les eucaryotes. C'est pour cette raison que nous avons collaboré avec les créateurs d'Augustus pour qu'ils adaptent leur logiciel, conçu pour les procaryotes, à la détection de gènes chez le parasite *Leishmania.* Les résultats de toutes ces analyses ont finalement été combinés afin de déterminer la position des gènes de *L. tarentolae.* Dans le futur, cette annotation sera combinée avec une nouvelle annotation qui sera réalisée avec le même protocole qui avait été utilisé pour annoter les génomes des autres espèces de *Leishmania* (74).

Une fois que les gènes ont été identifiés dans le génome, il s'est agi d'en analyser le contenu afin de mieux comprendre la biologie des parasites *Leishmania.* Ce qui nous intéressait particulièrement était d'observer les différences entre le génome de *L. tarentolae* et celui des espèces pathogènes pour l'humain. Pour ce faire, nous avons comparé les gènes entre les espèces afin de définir ceux qui étaient partagés et ceux qui étaient uniques à une espèce ou à une autre. Cette étape peut sembler simple, car à première vue, il s'agit simplement de comparer les gènes afin de déterminer lesquels sont partagés et lesquels ne le sont pas.

Le problème avec ce genre d'analyse est la variation entre le pourcentage de similitudes entre les espèces pour leurs différents gènes. Ainsi, lorsqu'on groupe les gènes afin de déterminer lesquels sont partagés, on doit définir un seuil à partir duquel des gènes sont considérés comme partagés entre les espèces, et en dessous duquel les gènes seront considérés comme distincts. Le processus est rendu encore plus complexe lorsque les gènes et les protéines potentielles que l'on compare ont des insertions ou des délétions majeures dans leurs séquences, des longueurs de gènes variables, des domaines répétés ou de faible complexité, ou des erreurs d'annotation, d'assemblage ou de séquençage.

Ainsi, comme première approche, les gènes de *L. tarentolae* ont été comparés à ceux des autres espèces par BLAST nucléotidique et protéique. Les résultats ont ensuite été triés et filtrés selon le pourcentage d'identité entre les gènes et selon le pourcentage de la longueur des deux gènes qui était similaire. Après une optimisation délicate, les résultats obtenus étaient solides, même si le seuil choisi restait arbitraire (un seuil de positivité étant toujours

arbitraire). Le problème avec cette méthode est le fait qu'elle pourrait difficilement être reproduite par un bio-informaticien externe, puisque la plupart des outils utilisés ont été faits maison. De plus, les résultats obtenus quant au compte de gènes divergents entre les espèces ont différé de ceux publiés précédemment par la communauté *Leishmania.* À mon avis, ces différences relèvent surtout de la méthode d'analyse et du seuil de similarité utilisés. Cependant, il a été difficile de justifier ces résultats, puisque l'approche n'était pas standardisée.

Afin d'utiliser une approche plus acceptée par la communauté, les protéines potentielles de *L. tarentolae* ont été comparées à celles des autres espèces en utilisant le logiciel OrthoMCL (216, 217). Cet outil sépare les gènes en groupes d'orthologues en analysant la comparaison par BLASTP (comparaison des séquences d'acides aminés) entre les protéines potentielles. Il groupe ensuite les gènes en se basant sur un modèle statistique de classes latentes (en anglais, *latent class analysis*). Cette méthode étant déjà implantée dans la principale base de données pour les trypanosomes (tritrypdb.org), il devenait plus facile de justifier les résultats à la communauté *Leishmania.* Malgré tout, les chiffres décrivant les gènes divergents entre les espèces pathogènes de *Leishmania* obtenus par cette méthode sont différents de ceux déjà publiés.

Une particularité de cette approche est le fait qu'elle nécessite l'utilisation du terme « groupes d'orthologues divergents entre les espèces » plutôt que du terme « gènes divergents ». Cette méthode entraîne parfois la séparation de gènes relativement proches en groupes d'orthologues différents, ce qui entraîne des gènes qui ont la même description à avoir des distributions différentes entre les espèces. Habituellement, ces différences correspondent à des sous-groupes de gènes de même famille. Cependant, la génomique comparative des espèces de *Leishmania* basée sur leurs groupes d'orthologues a un avantage important pour l'étude de ces parasites, puisqu'elle facilite l'analyse des familles de gènes qui sont en copies multiples, ce qui est fréquent chez le parasite *Leishmania.* Ainsi, avec les résultats de l'analyse OrthoMCL, il est aisé de déterminer les groupes d'orthologues différentiellement distribués entre les espèces ainsi que ceux qui ont un nombre de copies divergent.

Par contre, arrivés à cette étape de l'analyse, il nous a été nécessaire de valider certains résultats afin d'éviter des biais causés par des erreurs issues de la technologie de séquençage utilisée. En effet, cette analyse définissait certains gènes comme absents de *L. tarentolae* alors qu'ils étaient présents. Une analyse en profondeur des étapes du processus a permis d'observer que 10 gènes bel et bien présents chez *L. tarentolae* ne se trouvaient pas dans l'assemblage final du génome, par exemple la tubuline alpha, qui avait déjà été caractérisée chez cette espèce (379). Il semblerait que les logiciels d'assemblage utilisés n'aient pas été en mesure d'assembler ces gènes, ce qui entraîne leur absence dans les analyses subséquentes de génomique comparative. Pourtant, lorsqu'on a recherché la séquence de ces gènes dans les fragments bruts, il a été possible d'assembler chacun de ces 10 gènes en utilisant les fragments triés pour leur similarité avec ces gènes.

Un autre problème observé est le fait que certains gènes présents en plus d'une copie chez *L. tarentolae* se trouvaient en une seule copie dans l'assemblage, puisque deux copies identiques étaient considérées comme une seule par l'assembleur. Pour cette raison, le nombre moyen de fragments séquencés par nucléotide par gène a été calculé pour tous les gènes de *L. tarentolae* afin de les comparer au nombre de gènes de chaque groupe d'orthologues tel que défini par l'annotation du génome. De cette manière, il a été possible d'éliminer des erreurs potentielles qui auraient indiqué que des gènes étaient en nombre de copies variables entre *L. tarentolae* et les autres espèces alors qu'en réalité, ils étaient en nombre de copies similaires. Ainsi, l'analyse des fragments séquencés a permis dans plusieurs cas d'éviter des biais d'analyse causés par l'assemblage ou par l'annotation du génome.

10.4. Les particularités du génome de *Leishmania tarentolae*

Le cycle de vie et les détails de l'infection du lézard par *L. tarentolae* sont peu étudiés. Même si elles sont peu nombreuses, des études en milieu naturel de lézards infectés par *L. tarentolae* ont été réalisées. Deux d'entre elles ont observé des *Leishmania (sauroleishmania)* sous forme amastigote (15, 380), alors qu'une autre n'a observé aucune forme amastigote chez le lézard (18). En 2005, Breton et ses collaborateurs ont montré que *L. tarentolae* pouvait prendre une forme amastigote dans des macrophages humains en culture (19).

D'un autre côté, la réponse immunitaire du lézard à l'infection par *Leishmania* reste nébuleuse, même si certaines caractéristiques du système immunitaire du lézard suggèrent des pistes de réflexion. Premièrement, le lézard est un animal à sang froid et il ne possède pas la capacité d'augmenter sa température corporelle de façon biologique (381). Ainsi, pour faire de la fièvre, le lézard doit se déplacer à un endroit plus chaud afin d'augmenter sa température corporelle. Deuxièmement, l'immunité acquise des reptiles est diminuée, voire perdue, pendant l'hiver, ce qui pourrait modifier leur sensibilité aux infections pendant cette période (381). Chez les espèces de *Leishmania* pathogènes pour l'humain, il a été montré que des changements de température et de pH du milieu entraînaient le passage du stade promastigote au stade amastigote (382).

Le séquençage, l'annotation et la comparaison du génome de *L. tarentolae* à celui des espèces pathogènes de *Leishmania* permettent d'observer une forte synténie entre cette espèce et les espèces de *Leishmania* pathogènes pour l'humain. En tout, plus de 90 % des gènes sont partagés entre les 4 espèces comparées dans cette étude : *L. braziliensis, L. infantum, L. major* et *L. tarentolae.* De plus, 87 % des groupes de gènes orthologues avaient un nombre de copies identique entre toutes les espèces de *Leishmania.* Quant à *L. tarentolae,* il avait un nombre de copies de gènes égal avec au moins une espèce pathogène dans 94,5 % des cas, ce qui laisse environ 5 % des groupes d'orthologues totaux qui pourraient nous informer sur la biologie de *L. tarentolae.*

Les gènes absents ou en plus faible nombre de copies chez *L. tarentolae* incluent des membres de quelques familles bien connues, comme des amastines ou des gènes impliqués dans la modification des lipophosphoglycanes (LPG). Dans le cas des amastines, les gènes de cette famille absents de *L. tarentolae* semblent faire partie d'une sous-famille spécifique aux *Leishmania,* les amastines delta (234). Ces gènes sont habituellement surexprimés au stade amastigote du parasite (55). De leur côté, les LPG sont impliqués dans les interactions avec le vecteur de la leishmaniose (65, 383, 309), la mouche des sables, ainsi que dans la virulence du parasite (51, 247). Cependant, il a été observé que les LPG de *L. tarentolae* étaient plus courts que ceux des autres espèces (246). Cela pourrait avoir un impact sur l'infectivité du parasite et sur ses interactions avec son vecteur. Cependant, *L. tarentolae* est trouvé dans une grande proportion de mouches des sables (4). De plus, des souches de

L. mexicana, dont les gènes codant pour les LPG étaient inactivés, restaient virulentes (384).

D'autres gènes, comme ceux codant pour une subtilisine, sont absents de *L. tarentolae.* Dans le génome de *L. major,* deux gènes sont annotés comme des subtilisines, un type de protéase. L'une des deux est présente chez *L. tarentolae* alors que l'autre est absente. Celle qui est partagée par les deux espèces est impliquée dans le processus de maturation de la machinerie de maturation du trypanothione et, par conséquent, dans la virulence du parasite (227). Par contre, aucune information sur la subtilisine absente de *L. tarentolae* n'a été trouvée dans la littérature. Dans le même esprit, *L. tarentolae* ne semble pas posséder tous les types d'adaptines, des protéines impliquées dans le transport intermembranaire des protéines, présents chez les *Leishmania* pathogènes (385). Notamment, l'adaptine mu est absente de *L. tarentolae.* La délétion de l'adaptine mu chez *L. mexicana* rendait les parasites incapables de survivre dans le macrophage ou dans la souris (266). La même observation a été faite chez *L. mexicana* pour l'adaptine alpha, mais cette dernière est présente chez *L. tarentolae.*

Malgré la présence de certains gènes dont la fonction a été étudiée, la fonction de nombreux gènes absents de *L. tarentolae* reste obscure. Qui plus est, environ les deux tiers de ces gènes ne sont associés à aucune fonction connue, ce qui rend leur analyse difficile. Afin d'approfondir l'analyse de cette liste de gènes, les groupes d'orthologues ont été associés à un stade de développement (promastigote ou amastigote) basé sur leur expression à ces stades dans 13 expériences de transcriptomique ou de protéomique publiées (47, 104, 105, 107, 109-111, 113, 198, 238-241). Cette analyse a montré que 52 des 66 groupes d'orthologues absents de *L. tarentolae* et associés à un stade de développement du parasite étaient principalement retrouvés dans le stade amastigote. Cela suggère que *L. tarentolae* requiert moins de gènes pour survivre dans le stade amastigote que les espèces pathogènes pour l'humain. Cela a peut-être un lien avec le lézard, dont l'infection à certaines périodes de l'année est peut-être plus facile pour le parasite, étant donné la perte d'immunité des lézards pendant l'hiver et leur incapacité à augmenter biologiquement leur température corporelle (381).

De son côté, *L. tarentolae* a plusieurs gènes qui lui sont particuliers ou qu'il possède en un plus grand nombre de copies que les autres espèces de *Leishmania* étudiées. En particulier, *L. tarentolae* possède un nombre élevé de métalloprotéases GP63, aussi appelées « leishmanolysines », et de protéines antigènes de surface de promastigote (*promastigote surface antigen protein* ou PSA31C). Dans les deux cas, le nombre de copies de ces gènes est extrêmement élevé, ce qui a été validé par hybridation Southern (voir la figure 5.5). Cependant, ces gènes en nombre de copies élevé sont difficiles à assembler et les séquences obtenues sont fragmentées, ce qui rend ardues les analyses phylogénétiques poussées. Par contre, une analyse des fragments utilisés pour construire chacun de ces gènes a montré que certaines régions de ces gènes étaient absentes ou en nombre de copies plus limité.

Dans le cas de GP63, la presque totalité du gène semble présente en copies élevées. Cependant, 3 régions d'environ 100 nucléotides semblent absentes ou en nombre de copies réduit chez *L. tarentolae* (voir la figure 5.5A). Ces régions ne correspondent à aucun domaine particulier. Ce n'est pas le cas de PSA31C, qui semble ne contenir que la portion *leucine rich repeat* du gène. Ainsi, il semble manquer la moitié C-terminale de ce gène à *L. tarentolae.* Le parasite ne semble pas non plus contenir le peptide signal et le domaine récepteur de facteur de croissance (en anglais, *growth factor receptor domain*). Les implications biologiques de ces différences restent inconnues. Dans tous les cas, néanmoins, les séquences de GP63 et de PSA31C ont une variabilité qui suggère une forte diversification de ces deux gènes chez *L. tarentolae.*

Globalement, la particularité de *L. tarentolae* serait l'absence de nombreux gènes associés au stade amastigote du parasite et à sa virulence. Bien que ces différences aident à mieux comprendre ce parasite *Sauroleishmania,* le cycle de vie de ce parasite de lézards devra être étudié dans le laboratoire pour qu'on puisse réellement comprendre les implications de ce qui a été observé dans son génome. Ces informations seront d'une importance capitale pour les chercheurs travaillant avec *L. tarentolae.* Concrètement, la séquence du génome de *L. tarentolae* permettra de créer une troisième génération de biopuce *Leishmania,* qui sera utile pour mieux comprendre le développement de ce parasite, la régulation de ses gènes et ses modes de résistance aux antiparasitaires.

Inverser le point de vue sur les données de génomique comparative entre les *Leishmania (sauroleishmania)* et les espèces pathogènes pour l'humain procure de l'information dont l'impact pourrait être encore plus intéressant pour la communauté *Leishmania.* Les gènes de *L. major* et de *L. infantum* qui ont été mis en exergue par cette étude sont des sujets d'étude potentiels qui pourraient nous en apprendre beaucoup sur la virulence du parasite et sur les gènes qui sont essentiels à sa survie dans son hôte humain. Alors que ces résultats confirment l'importance de certains gènes pour l'infection de l'homme par *Leishmania,* ils suggèrent que plusieurs gènes dont la fonction n'est pas connue pourraient être essentiels à l'adaptation de *Leishmania (leishmania)* et de *Leishmania (vianna)* à son hôte humain. Cette étude propose donc plusieurs pistes pour mieux comprendre le processus d'infection du parasite *Leishmania* et ses approches pour survivre à la réponse immunitaire humaine.

10.5. Le diagnostic des infections respiratoires virales

Débuté en 2006, ce projet avait pour objectif de créer un test diagnostique qui permettrait la détection simultanée des virus respiratoires les plus importants. Ainsi, le mandat de départ de ce projet était la création d'un test pour détecter les influenzas A, B et H5N1, les virus respiratoires syncytiaux humains A et B, les métapneumovirus humains A et B, les coronavirus HKU1, OC43, 229E, NL63 et SARS, les adénovirus de A à F, les rhinovirus A et B, les entérovirus de A à D, et les para-influenzavirus de 1 à 4. Malgré ce mandat initial, la composition du test a été modifiée au cours de sa conception. Par exemple, certains analytes ont été retirés du test pour des raisons épidémiologiques (les adénovirus D et F sont peu impliqués dans les infections respiratoires) ou commerciales (la validation de la détection de l'influenza A (H5N1) et du coronavirus SARS rendait difficile la commercialisation d'un test incluant ces virus). La technologie proposée pour ce test était le système INFINITI de AutoGenomics, un appareil automatisé qui effectue du diagnostic moléculaire par hybridation sur biopuce. Conjointement à la création de ce test diagnostique, un second test a été créé, ce dernier utilisant la PCR en temps réel avec des sondes TaqMan. Les deux tests ont été créés en parallèle, mais ils n'utilisent pas nécessairement les mêmes cibles.

Les bases de données de séquences publiques ont permis de créer les alignements de séquences nécessaires à la création de tests de diagnostic moléculaire pour chacune des

cibles testées. Pour plusieurs virus, comme l'influenza, les rhinovirus ou les entérovirus, de nombreuses séquences étaient disponibles, ce qui a permis de créer un test plus ubiquitaire quant à la détection de toutes les souches circulant en clinique. Pour d'autres virus, par exemple le para-influenza de type 4 qui a récemment été découvert, seulement quelques séquences étaient disponibles dans les bases de données, ce qui a limité l'évaluation de l'ubiquité des amorces et des sondes ciblant cet organisme. Par contre, sur le terrain, nous n'avons pas observé de problème majeur d'ubiquité.

La comparaison des deux méthodes de diagnostic, soit la PCR en temps réel de type TaqMan et le test hautement multiplexé RVP sur le INFINITI, suggère que la PCR en temps réel est un peu plus sensible que la PCR multiplexée avec détection sur biopuce. En effet, pour la plupart des virus, la limite de détection des tests TaqMan est inférieure à celle du test multiplexé. Malgré tout, il y a quelques cas pour lesquels le test multiplexé est plus sensible que la PCR en temps réel : HMPV-A et HRSV-A. Lorsque la performance des deux tests a été comparée sur des échantillons cliniques, la sensibilité du test multiplexé était de 97,8 % lorsqu'il était comparé au test TaqMan. En tout, 13 spécimens sur 221 ont montré un virus positif par la méthode TaqMan, alors que ce même virus était négatif par le test multiplexé. Neuf de ces spécimens étaient des infections multiples et sept avaient l'adénovirus comme virus discordant. De plus, les signaux des résultats de PCR en temps réel pour les virus discordants étaient proches du seuil de détection du test, ce qui suggère que ces échantillons contenaient un faible titre viral.

Malgré sa sensibilité légèrement plus faible, le test multiplexé RVP avait un avantage notable sur le test de PCR en temps réel, car il nécessitait beaucoup moins de travail de laboratoire afin de tester plus d'échantillons pour plus de virus. En effet, la PCR en temps réel est faite en plaques et chaque plaque permet de tester six échantillons. De son côté, le test multiplexé sur biopuce permet de tester 24 échantillons à la fois. Comme la plupart des manipulations sont automatisées dans le Infiniti, ce test ne requiert que 77 pipettages pour tester 24 échantillons, alors que le test de PCR en temps réel, même s'il a été optimisé pour réduire le nombre de manipulations, requiert 288 pipettages pour tester seulement 6 échantillons. Le temps de travail manuel pour tester un nombre égal d'échantillons est donc moins élevé pour le test automatisé que pour le test de PCR en temps réel. Le test

RVP permet aussi d'identifier les sous-types des rhinovirus, des entérovirus et des adénovirus, ce que le test de PCR en temps réel ne permet pas.

L'étude clinique effectuée pour valider ces deux tests a généré des résultats intéressants d'un point de vue épidémiologique. Les 221 échantillons de cette étude ont été prélevés à l'automne 2001 et à l'hiver 2002 sur des enfants de moins de trois ans hospitalisés pour des infections respiratoires (343, 360). En tout, 81,4 % de ces spécimens étaient positifs pour au moins un virus respiratoire, le plus fréquent étant le virus respiratoire syncytial (48,0 %). Les deux virus les plus fréquents étaient l'influenza A (13,1 %) et le rhinovirus ou l'entérovirus (13,1 %). Ce portrait indique que les virus respiratoires représentent une forte cause d'hospitalisation des jeunes enfants.

Évidemment, il n'est pas à exclure que certains de ces enfants aient des surinfections bactériennes en plus des infections respiratoires. En effet, il a été observé que les infections par le HRSV, le HMPV ou l'influenza pouvaient prédisposer aux pneumonies à *Streptococcus pneumoniae* (386). Cependant, la plupart de ces pathogènes viraux, en particulier le HRSV et l'influenza, sont connus pour être des causes d'hospitalisation des jeunes enfants (387). Ainsi, plusieurs échantillons (13,1 %) étaient positifs pour plus d'un virus, dont 41,4 % étaient positifs pour un adénovirus et un autre virus, et 34,5 % pour un picornavirus (entérovirus ou rhinovirus) et un autre virus. Une étude récente a montré que les infections multiples qui incluaient un HRSV avaient tendance à être plus sévères que les infections simples à HRSV (388). Dans notre étude, le HRSV était présent dans 57,9 % des infections multiples.

En 2008, le test multiplexé a été mis en marché par la compagnie AutoGenomics sous le nom de AutoGenomics RVP Assay. Cette version commerciale du test a été utilisée au cours des saisons grippales 2006-2007, 2007-2008, 2008-2009 et 2009-2010 pour effectuer des études épidémiologiques sur des enfants de moins de trois ans souffrant d'infections respiratoires qui étaient hospitalisés au Centre hospitalier de l'Université Laval (CHUL) ou qui consultaient un médecin à la Clinique pédiatrique de Sainte-Foy. Au total, plus de 1000 échantillons ont été testés au cours de ces 4 études prospectives (données non incluses dans cette thèse). Ainsi, le test a été utilisé pour réaliser des études épidémiologiques d'envergure qui sont en cours d'analyse par l'Institut national de santé publique du Québec.

10.6. L'amélioration d'un test diagnostique

Au printemps 2009, quelques jours après l'apparition de la grippe A (H1N1) d'origine porcine, des séquences nucléotidiques de cette nouvelle souche d'influenza A ont été déposées dans les bases de données publiques (364). Dès ce moment, il a été possible d'évaluer le test RVP afin de vérifier s'il était en mesure de détecter la nouvelle souche de grippe, ce qui n'était pas le cas.

Comme le test avait été créé en se basant principalement sur les souches de grippes retrouvées chez l'humain, les oligonucléotides du test avaient des polymorphismes avec la nouvelle souche de grippe, dont la séquence divergeait des souches en circulation. Ainsi, les séquences de plusieurs souches de grippe A (H1N1) pandémique ont été comparées et les amorces et sondes du test RVP ont été modifiées afin de détecter à la fois cette nouvelle souche et les souches de grippe saisonnière qu'elles détectaient déjà. La nouvelle version du test porte le nom RVP Plus. En plus de différencier l'influenza A de l'influenza B, le test RVP Plus identifie spécifiquement la grippe A (H1N1) pandémique. Ainsi, pour un échantillon de grippe A (H1N1) pandémique, la sonde détectant l'influenza A et la sonde détectant l'influenza A (H1N1) pandémique seront positives. Pour un échantillon de grippe A saisonnière, seulement la sonde détectant l'influenza A sera positive.

Le test RVP Plus a été validé avec 22 échantillons cliniques de virus autres que l'influenza A, incluant une influenza B et un échantillon négatif, avec 9 souches de référence d'influenza A saisonnière et 5 souches d'influenza B, ainsi qu'avec 12 souches de référence d'espèces bactériennes. Aucun de ces échantillons ne donnait de signal faux positif avec le test RVP Plus. Les limites de détection pour l'influenza A saisonnière et l'influenza A (H1N1) pandémique ont été évaluées à 500 UFP/ml, dans 95 % des cas.

Afin de valider la sensibilité et la spécificité cliniques du test RVP Plus, ce dernier a été comparé avec le protocole de détection de la grippe A (H1N1) pandémique utilisé par l'Agence de santé publique du Canada (367). Une sensibilité clinique de 89,2 % et une spécificité clinique de 100,0 % ont été obtenues en comparant le test RVP Plus d'AutoGenomics au protocole de référence lorsque 154 échantillons suspectés d'être positifs pour la grippe A (H1N1) pandémique ont été testés avec les deux méthodes.

La différence de sensibilité entre les deux méthodes est probablement attribuable à la limite de détection plus sensible du test de référence, qui détecte jusqu'à 20 copies du virus, alors que le test RVP Plus détecte de 250 à 500 copies, au minimum. Cela est corroboré par le signal obtenu en PCR en temps réel de référence pour les échantillons discordants ($Ct > 39,5$). Le signal était significativement plus élevé que pour les échantillons concordants (Ct moyen de 32,0, $p < 0,05$). Les résultats de cette étude ont aussi permis d'observer la proportion de ces échantillons réellement positifs pour la grippe A (H1N1) pandémique et de ceux qui étaient positifs pour un autre virus respiratoire. Au total, le test RVP Plus a trouvé 21,4 % des échantillons positifs pour l'influenza A (H1N1) pandémique, aucun pour l'influenza A saisonnière et 13,3 % pour un autre virus respiratoire.

Cette amélioration du test RVP avait été motivée par une situation de crise qui requérait une modification du test pour répondre à un besoin immédiat. Bien que le besoin soit moins criant, il sera nécessaire, au cours des prochaines années, de réévaluer en totalité le contenu du test et de le comparer aux séquences des différents virus déposés dans des bases de données publiques. De même, il sera important de vérifier si de nouveaux virus respiratoires apparaissent en clinique afin de les ajouter au menu du test RVP Plus comme nous l'avons fait avec la grippe A (H1N1) pandémique.

Cependant, la modification à un test commercial entraîne des coûts élevés pour les fabricants. Néanmoins, pour assurer l'utilité à long terme d'un test diagnostique ciblant des maladies infectieuses dont les souches prédominantes en clinique sont sujettes à modifications, il est nécessaire que le test évolue avec la réalité clinique. La version courante du test RVP Plus permet la détection de 25 virus respiratoires, dont les influenzas A saisonnière et A (H1N1) pandémique, l'influenza B, les virus respiratoires syncytiaux humains A et B, les métapneumovirus humains A et B, les coronavirus HKU1, OC43, 229E et NL63, les adénovirus A, B, C et E, les rhinovirus A et B, les entérovirus de A à D, et les para-influenzavirus de 1 à 4.

Le diagnostic moléculaire hautement multiplexé des infections respiratoires reste un domaine en expansion qui fait peu à peu sa place dans les laboratoires cliniques. Dans cette optique, la comparaison des différents tests sur le marché sera nécessaire afin d'évaluer la

qualité réelle des tests. Comparer ces tests sur une large banque de spécimens cliniques permettrait de mieux évaluer leurs sensibilités et spécificités respectives.

En plus des propriétés scientifiques et cliniques de ces tests, leurs avantages commerciaux auront aussi une influence importante sur leur implantation en clinique. Le système d'hybridation automatisé INFINITI a l'avantage d'avoir été conçu pour l'usage en laboratoires cliniques. Comme ce test automatise plusieurs étapes du processus, dont une partie du pipettage et la totalité de l'hybridation sur biopuce et de la génération des résultats, il sera plus facile de l'implanter qu'un test qui requiert plus de temps et de manipulations. Par contre, le test RVP Plus n'a pas encore été approuvé aux États-Unis par la Food and Drugs Administration (FDA), alors que le xTag de Luminex l'a été (309). L'approbation d'un test diagnostique par les instances gouvernementales aura aussi un rôle clé dans l'implantation de ces systèmes.

11. Conclusion

Comme le montre cette thèse, la disponibilité de la séquence des génomes est essentielle à de nombreux domaines de l'étude des maladies infectieuses. La bio-informatique est l'outil de base qui permet d'analyser ces informations pour mieux comprendre les microorganismes. Les travaux présentés ici s'inscrivent dans les deux extrêmes de la biologie, c'est-à-dire la génération de nouvelles données de séquences qui seront utiles à la communauté scientifique pour étudier le parasite *Leishmania* et l'utilisation de données de séquençage publiées par d'autres groupes afin de créer un test diagnostique dont l'utilisation aura un impact direct sur les malades et sur la santé publique.

L'étude de microorganismes variés, soit les parasites et les virus, avec des méthodes similaires basées sur l'ADN, démontre que la biologie peut être envisagée du point de vue de l'ADN plutôt que selon le microorganisme étudié. En effet, lorsque la biologie est ramenée à son plus petit dénominateur commun, l'ADN génomique, le microorganisme à l'étude devient dans certains cas un outil pour formuler les contraintes de l'analyse plutôt que le centre de notre attention. L'approche bio-informatique présente à sa base une question biologique qui sera ancrée dans notre compréhension de l'organisme à l'étude, mais il est possible que les étapes subséquentes de l'analyse bio-informatique ne nécessitent pas d'ingérence supplémentaire de la microbiologie classique, qui ne deviendra souvent importante qu'au moment d'analyser les résultats. C'est pour cette raison que, du moment que l'analyse est basée sur des informations de séquences et qu'elle consiste en des analyses bio-informatiques, les sources de l'information biologique sur lesquelles cette analyse est basée peuvent être diverses. De fait, les études sur le parasite *Leishmania* et sur les virus respiratoires présentées dans cette thèse découlent d'une même philosophie bio-informatique.

Au final, les travaux sur le parasite *Leishmania* présentés dans cette thèse apporteront des outils à la communauté scientifique pour mieux comprendre ce parasite. Les études de transcriptomique présentées ont à la fois suggéré des hypothèses quant au développement de ce parasite, à sa résistance aux antiviraux et à sa réponse aux stress, tant de sujets qui seront encore approfondis au cours des prochaines années par les groupes de recherches étudiant le parasite *Leishmania.*

En particulier, les études d'hybridation comparatives de génomes qui utilisent la biopuce *Leishmania* ont permis d'élaborer une hypothèse permettant de prédire les amplifications géniques du parasite en réponse au stress, en nous basant sur des méthodes d'analyse bio-informatique. Quoique cette hypothèse ne soit pas explorée dans cette thèse, elle y trouve les éléments qui ont permis de formuler une question bio-informatique permettant de mieux connaître la biologie de *Leishmania* par l'analyse de sa séquence d'un point de vue différent de ce qui avait été fait auparavant (Raymond, Ubeda et coll., en préparation).

De même, les résultats d'études de transcriptomique sur la différenciation du parasite de promastigote en amastigote, en plus de fournir une information biologique favorisant la compréhension de la différenciation du parasite, ont permis d'étudier les différences entre le contenu en gènes de différentes espèces de *Leishmania* en leur associant un stade de développement. Ces données, utilisées pour analyser la séquence du génome de *L. tarentolae,* ont permis de renforcer l'hypothèse que les parasites *Sauroleishmania* sont moins bien équipés pour survivre au stade amastigote. Le séquençage du génome de *L. tarentolae* a permis la comparaison entre cette espèce et les espèces pathogènes. Elle constituera un outil précieux dans l'étude des *Leishmania.*

La séquence du génome des microorganismes est aussi une information essentielle à la création de tests de diagnostic moléculaire. Cela est démontré par la création d'un test diagnostique pour la détection des infections respiratoires les plus communes. Après sa création, ce test a été validé avec des souches de référence et des échantillons cliniques. Les résultats obtenus ont montré que les tests de diagnostic hautement multiplexés génèrent des résultats épidémiologiques importants qui peuvent constituer des atouts majeurs en santé publique. La modification du test pour la pandémie de grippe A (H1N1) de 2009 a montré l'importance de séquencer les génomes des nouveaux pathogènes trouvés en clinique afin d'adapter les tests diagnostiques aux nouvelles réalités. Afin d'évaluer ce test, il devra être comparé aux autres tests de même génération, à la fois sur leur adaptabilité à un laboratoire de virologie clinique et sur leur performance pour la détection de virus dans des échantillons cliniques.

Cette thèse est donc une affirmation que la bio-informatique est un outil qui peut servir à toutes les étapes d'un projet de génomique, qu'il soit fondamental ou appliqué. Bien que la

séquence des génomes ne soit pas la seule information utile pour l'étude des maladies infectieuses, elle constitue malgré tout un aspect essentiel pour la compréhension de tous les organismes vivants. La disponibilité des données de séquençage et leur augmentation constante sont un atout pour les chercheurs de tous les domaines et, lorsqu'on est en mesure de poser les bonnes questions, on peut arriver à des réponses qui permettent d'avancer les connaissances dans des domaines qui auraient été inaccessibles sans la démocratisation du séquençage des génomes.

Bibliographie

1. OMS (2006) Lutte contre la leishmaniose. *Rapport du secrétariat*, 1-7.
2. Zink, A.R., Spigelman, M., Schraut, B., Greenblatt, C.L., Nerlich, A.G. and Donoghue, H.D. (2006) Leishmaniasis in ancient Egypt and Upper nubia. *Emerging infectious diseases*, **12**, 1616-7.
3. Alvar, J., Aparicio, P., Aseffa, A., Den Boer, M., Cañavate, C., Dedet, J.-P., Gradoni, L., Ter Horst, R., López-Vélez, R. and Moreno, J. (2008) The relationship between leishmaniasis and AIDS: The second 10 years. *Clinical microbiology reviews*, **21**, 334-59, 10.1128/CMR.00061-07.
4. Coleman, R.E., Hochberg, L.P., Swanson, K.I., Lee, J.S., McAvin, J.C., Moulton, J.K., Eddington, D.O., Groebner, J.L., OGuinn, M.L. and Putnam, J.L. (2009) Impact of phlebotomine sand flies on U.S. military operations at Tallil Air Base, Iraq: 4. Detection and identification of *Leishmania* parasites in sand flies. *Journal of medical entomology*, **46**, 649-63.
5. Capelli, G., Baldelli, R., Ferroglio, E., Genchi, C., Gradoni, L., Gramiccia, M., Maroli, M., Mortarino, M., Pietrobelli, M., Rossi, L., et al. (2004) Monitoring of canine leishmaniasis in northern Italy: An update from a scientific network. *Parassitologia*, **46**, 193-7.
6. Bañuls, A.-L., Hide, M. and Prugnolle, F. (2007) *Leishmania* and the leishmaniases: A parasite genetic update and advances in taxonomy, epidemiology and pathogenicity in humans. *Advances in Parasitology*, **64**, 1-109, 10.1016/S0065-308X(06)64001-3.
7. Schönian, G., Mauricio, I., Gramiccia, M., Cañavate, C., Boelaert, M. and Dujardin, J.-C. (2008) Leishmaniases in the Mediterranean in the era of molecular epidemiology. *Trends in parasitology*, **24**, 135-42, 10.1016/j.pt.2007.12.006.
8. Aspöck, H., Gerersdorfer, T., Formayer, H. and Walochnik, J. (2008) Sandflies and sandfly-borne infections of humans in Central Europe in the light of climate change. *Wiener klinische Wochenschrift*, **120**, 24-9, 10.1007/s00508-008-1072-8.
9. Wilson, M.E., Jeronimo, S.M.B. and Pearson, R.D. (2005) Immunopathogenesis of infection with the visceralizing *Leishmania* species. *Microbial pathogenesis*, **38**, 147-60, 10.1016/j.micpath.2004.11.002.
10. Awasthi, A., Mathur, R.K. and Saha, B. (2004) Immune response to *Leishmania* infection. *The Indian journal of medical research*, **119**, 238-58.
11. Smith, D.F., Peacock, C.S. and Cruz, A.K. (2007) Comparative genomics: From genotype to disease phenotype in the leishmaniases. *International journal for parasitology*, **37**, 1173-86, 10.1016/j.ijpara.2007.05.015.
12. Chang, K.-P. (2001) Leishmaniases. *Encyclopedia of Life Sciences*, 10.1038/npg.els.0003824.
13. A Bates, P. (2001) *Leishmania*. In Encyclopedia of Life Sciences. *Encyclopedia of Life Sciences*, 10.1038/npg.els.0004265.
14. Davies, C.R., Reithinger, R., Campbell-Lendrum, D., Feliciangeli, D., Borges, R. and Rodriguez, N. (2000) The epidemiology and control of leishmaniasis in Andean countries. *Cad Saude Publica*, **16**, 925-50.
15. Elwasila, M. (1988) *Leishmania tarentolae* Wenyon, 1921 from the gecko *Tarentola annularis* in the Sudan. *Parasitology research*, **74**, 591-92.

16. Simpson, L. and Holz, G. (1988) The status of *Leishmania tarentolae/Trypanosoma platydactyli*. *Parasitol Today*, **4**, 115-8.
17. Momen, H. and Cupolillo, E. (2000) Speculations on the origin and evolution of the genus Leishmania. *Memorias do Instituto Oswaldo Cruz*, **95**, 583-88.
18. Belova, E. (1971) Reptiles and their importance in the epidemiology of leishmaniasis. *Bulletin of the World Health Organization*, **44**, 553-60.
19. Breton, M., Tremblay, M.J., Ouellette, M. and Papadopoulou, B. (2005) Live nonpathogenic parasitic vector as a candidate vaccine against visceral leishmaniasis. *Infection and immunity*, **73**, 6372-82, 10.1128/IAI.73.10.6372-6382.2005.
20. Breton, M., Zhao, C., Ouellette, M., Tremblay, M.J. and Papadopoulou, B. (2007) A recombinant non-pathogenic *Leishmania* vaccine expressing human immunodeficiency virus 1 (HIV-1) Gag elicits cell-mediated immunity in mice and decreases HIV-1 replication in human tonsillar tissue following exposure to HIV-1 infection. *Journal of general virology*, **88**, 217-25, 10.1099/vir.0.81995-0.
21. Matthews, K.R. (2005) The developmental cell biology of *Trypanosoma brucei*. *Journal of cell science*, **118**, 283-90, 10.1242/jcs.01649.
22. Stuart, K., Brun, R., Croft, S., Fairlamb, A., Gürtler, R.E., McKerrow, J., Reed, S. and Tarleton, R. (2008) Kinetoplastids: related protozoan pathogens, different diseases. *Journal of clinical investigations*, **118**, 1301-10, 10.1172/JCI33945.
23. El-Sayed, N.M., Myler, P.J., Blandin, G., Berriman, M., Crabtree, J., Aggarwal, G., Caler, E., Renauld, H., Worthey, E.A., Hertz-Fowler, C., et al. (2005) Comparative genomics of trypanosomatid parasitic protozoa. *Science*, **309**, 404-9, 10.1126/science.1112181.
24. Gonzalez, U., Pinart, M., Rengifo-Pardo, M. nica, Macaya, A., Alvar, J. and Tweed, J.A. (2009) Interventions for American cutaneous and mucocutaneous leishmaniasis. *Cochrane Database Systematic Reviews*, CD004834, 10.1002/14651858.CD004834.pub2.
25. Kedzierski, L., Sakthianandeswaren, A., Curtis, J.M., Andrews, P.C., Junk, P.C. and Kedzierska, K. (2009) Leishmaniasis: current treatment and prospects for new drugs and vaccines. *Current medicinal chemistry*, **16**, 599-614.
26. Sundar, S., More, D.K., Singh, M.K., Singh, V.P., Sharma, S., Makharia, A., Kumar, P.C. and Murray, H.W. (2000) Failure of pentavalent antimony in visceral leishmaniasis in India: report from the center of the Indian epidemic. *Clinical infectious diseases,* **31**, 1104-7, 10.1086/318121.
27. Berman, J.J. (2008) Treatment of leishmaniasis with miltefosine: 2008 status. *Expert Opinion on Drug Metabolism and Toxicology*, **4**, 1209-16, 10.1517/17425255.4.9.1209.
28. Ouellette, M. (2001) Biochemical and molecular mechanisms of drug resistance in parasites. *Tropical medicine and international health*, **6**, 874-82.
29. Zhou, Y., Bhattacharjee, H. and Mukhopadhyay, R. (2006) Bifunctional role of the leishmanial antimonate reductase LmACR2 as a protein tyrosine phosphatase. *Molecular and biochemical parasitology*, **148**, 161-8, 10.1016/j.molbiopara.2006.03.009.
30. Cunningham, M.L., Zvelebil, M.J. and Fairlamb, A.H. (1994) Mechanism of inhibition of trypanothione reductase and glutathione reductase by trivalent organic arsenicals. *European journal of biochemistry / FEBS*, **221**, 285-95.

31. Oza, S.L., Wyllie, S. and Fairlamb, A.H. (2006) Mapping the functional synthetase domain of trypanothione synthetase from *Leishmania major*. *Molecular and biochemical parasitology*, **149**, 117-20, 10.1016/j.molbiopara.2006.05.001.
32. Wyllie, S., Cunningham, M.L. and Fairlamb, A.H. (2004) Dual action of antimonial drugs on thiol redox metabolism in the human pathogen *Leishmania donovani*. *The Journal of biological chemistry*, **279**, 39925-32, 10.1074/jbc.M405635200.
33. Sundar, S. and Goyal, N., with Ashutosh (2007) Molecular mechanisms of antimony resistance in *Leishmania*. *Journal medical microbiology*, **56**, 143-153, 10.1099/jmm.0.46841-0.
34. Grondin, K., Haimeur, A., Mukhopadhyay, R., Rosen, B.P. and Ouellette, M. (1997) Co-amplification of the gamma-glutamylcysteine synthetase gene gsh1 and of the ABC transporter gene pgpA in arsenite-resistant *Leishmania tarentolae*. *The EMBO journal*, **16**, 3057-65, 10.1093/emboj/16.11.3057.
35. Haimeur, A., Guimond, C., Pilote, S., Mukhopadhyay, R., Rosen, B.P., Poulin, R. and Ouellette, M. (1999) Elevated levels of polyamines and trypanothione resulting from overexpression of the ornithine decarboxylase gene in arsenite-resistant *Leishmania*. *Molecular microbiology*, **34**, 726-35.
36. Légaré, D., Papadopoulou, B., Roy, G., Mukhopadhyay, R., Haimeur, A., Dey, S., Grondin, K., Brochu, C., Rosen, B.P. and Ouellette, M. (1997) Efflux systems and increased trypanothione levels in arsenite-resistant *Leishmania*. *Experimental parasitology*, **87**, 275-82, 10.1006/expr.1997.4222.
37. Croft, S.L., Sundar, S. and Fairlamb, A.H. (2006) Drug resistance in Leishmaniasis. *Society*, **19**, 111-26, 10.1128/CMR.19.1.111.
38. Singh, A.K., Papadopoulou, B. and Ouellette, M. (2001) Gene amplification in amphotericin B-resistant *Leishmania tarentolae*. *Experimental parasitology*, **99**, 141-7, 10.1006/expr.2001.4663.
39. Yardley, V., Croft, S.L., De Doncker, S., Dujardin, J.-C., Koirala, S., Rijal, S., Miranda, C., Llanos-Cuentas, A. and Chappuis, F. (2005) The sensitivity of clinical isolates of *Leishmania* from Peru and Nepal to miltefosine. *The American journal of tropical medicine and hygiene*, **73**, 272-5.
40. Pérez-Victoria, F.J., Sánchez-Cañete, M.P., Seifert, K., Croft, S.L., Sundar, S., Castanys, S. and Gamarro, F. (2006) Mechanisms of experimental resistance of *Leishmania* to miltefosine: Implications for clinical use. *Drug resistance updates: Reviews and commentaries in antimicrobial and anticancer chemotherapy*, **9**, 26-39, 10.1016/j.drup.2006.04.001.
41. Castanys-Muñoz, E., Pérez-Victoria, J.M., Gamarro, F. and Castanys, S. (2008) Characterization of an ABCG-like transporter from the protozoan parasite *Leishmania* with a role in drug resistance and transbilayer lipid movement. *Antimicrobial agents and chemotherapy*, **52**, 3573-9, 10.1128/AAC.00587-08.
42. Bray, P.G., Barrett, M.P., Ward, S.A. and Koning, H.P. de (2003) Pentamidine uptake and resistance in pathogenic protozoa: Past, present and future. *Trends in parasitology*, **19**, 232-9.
43. Coelho, A.C., Messier, N., Ouellette, M. and Cotrim, P.C. (2007) Role of the ABC transporter PRP1 (ABCC7) in pentamidine resistance in *Leishmania* amastigotes. *Antimicrobial agents and chemotherapy*, **51**, 3030-2, 10.1128/AAC.00404-07.

44. Ouellette, M., Drummelsmith, J. and Papadopoulou, B. (2004) Leishmaniasis: Drugs in the clinic, resistance and new developments. *Drug resistance updates: Reviews and commentaries in antimicrobial and anticancer chemotherapy*, **7**, 257-66, 10.1016/j.drup.2004.07.002.
45. Ouellette, M., Drummelsmith, J., El-Fadili, A., Kündig, C., Richard, D. and Roy, G. (2002) Pterin transport and metabolism in *Leishmania* and related trypanosomatid parasites. *International journal for parasitology*, **32**, 385-98.
46. Ouameur, A.A., Girard, I. and Ouellette, M. (2008) Functional analysis and complex gene rearrangements of the folate/biopterin transporter (FBT) gene family in the protozoan parasite. *Molecular & Biochemical Parasitology*, **162**, 155-64, 10.1016/j.molbiopara.2008.08.007.
47. Rochette, A., Raymond, F., Corbeil, J., Ouellette, M. and Papadopoulou, B. (2009) Whole-genome comparative RNA expression profiling of axenic and intracellular amastigote forms of *Leishmania infantum*. *Molecular & biochemical parasitology*, **165**, 32-47, 10.1016/j.molbiopara.2008.12.012.
48. Rosenzweig, D., Smith, D., Opperdoes, F., Stern, S., Olafson, R.W. and Zilberstein, D. (2008) Retooling *Leishmania* metabolism: From sand fly gut to human macrophage. *The FASEB journal,* **22**, 590-602, 10.1096/fj.07-9254com.
49. Singh, V. and Singh, D.D. (2008) *Leishmania major:* Genome analysis for identification of putative adhesin-like and other surface proteins. *Experimental Parasitology*, **118**, 139-45, 10.1016/j.exppara.2007.07.006.
50. Lodge, R. and Descoteaux, A. (2005) Modulation of phagolysosome biogenesis by the lipophosphoglycan of *Leishmania*. *Clinical Immunology*, **114**, 256-65, 10.1016/j.clim.2004.07.018.
51. Turco, S.J., Späth, G.F. and Beverley, S.M. (2001) Is lipophosphoglycan a virulence factor? A surprising diversity between *Leishmania* species. *Trends in parasitology*, **17**, 223-6.
52. Yao, C., Donelson, J.E. and Wilson, M.E. (2003) The major surface protease (MSP or GP63) of *Leishmania* sp. Biosynthesis, regulation of expression, and function. *Molecular & biochemical parasitology*, **132**, 1-16.
53. Purdy, J.E., Donelson, J.E. and Wilson, M.E. (2005) Regulation of genes encoding the major surface protease of *Leishmania chagasi* via mRNA stability. *Molecular and biochemical parasitology*, **142**, 88-97, 10.1016/j.molbiopara.2005.03.010.
54. Gomez, M.A., Contreras, I., Hallé, M., Tremblay, M.L., McMaster, R.W. and Olivier, M. (2009) *Leishmania* GP63 alters host signaling through cleavage-activated protein tyrosine phosphatases. *Science signaling*, **2**, ra58, 10.1126/scisignal.2000213.
55. Rochette, A., McNicoll, F., Girard, J., Breton, M., Leblanc, E., Bergeron, M.G. and Papadopoulou, B. (2005) Characterization and developmental gene regulation of a large gene family encoding amastin surface proteins in *Leishmania* spp. *Molecular and biochemical parasitology*, **140**, 205-20, 10.1016/j.molbiopara.2005.01.006.
56. Zhang, W.-W., Mendez, S., Ghosh, A., Myler, P., Ivens, A., Clos, J., Sacks, D.L. and Matlashewski, G. (2003) Comparison of the A2 gene locus in *Leishmania donovani* and Leishmania major and its control over cutaneous infection. *The Journal of biological chemistry*, **278**, 35508-15, 10.1074/jbc.M305030200.

57. Zhang, W.W. and Matlashewski, G. (2001) Characterization of the A2-A2rel gene cluster in *Leishmania donovani:* Involvement of A2 in visceralization during infection. *Molecular microbiology*, **39**, 935-48.
58. Garin, Y.J.F., Meneceur, P., Lorenzo, F., Pratlong, F., Dedet, J.-P. and Derouin, F. (2005) A2 gene of Old World cutaneous *Leishmania* is a single highly conserved functional gene. *BMC infectious diseases*, **5**, 18, 10.1186/1471-2334-5-18.
59. Zhang, W.-W., Peacock, C.S. and Matlashewski, G. (2008) A genomic-based approach combining in vivo selection in mice to identify a novel virulence gene in *Leishmania*. *PLoS Negleglected tropical diseases*, **2**, 10.1371/journal.pntd.0000248.
60. Gregory, D.J. and Olivier, M. (2005) Subversion of host cell signalling by the protozoan parasite *Leishmania*. *Parasitology*, **130 Suppl**, S27-35, 10.1017/S0031182005008139.
61. Olivier, M., Gregory, D.J. and Forget, G. (2005) Subversion mechanisms by which *Leishmania* parasites can escape the host immune response: A signaling point of view. *Clinical microbiology reviews*, **18**, 293-305, 10.1128/CMR.18.2.293-305.2005.
62. Chaussabel, D., Semnani, R.T., McDowell, M.A., Sacks, D., Sher, A. and Nutman, T.B. (2003) Unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites. *Blood*, **102**, 672-81, 10.1182/blood-2002-10-3232.
63. Chang, K.-P. and McGwire, B.S. (2002) Molecular determinants and regulation of *Leishmania* virulence. *Kinetoplastid biology and disease*, **1**, 1.
64. Lo, S.K., Bovis, L., Matura, R., Zhu, B., He, S., Lum, H., Turco, S.J. and Ho, J.L. (1998) *Leishmania* lipophosphoglycan reduces monocyte transendothelial migration: Modulation of cell adhesion molecules, intercellular junctional proteins, and chemoattractants. *Journal of immunology*, **160**, 1857-65.
65. Sacks, D.L. (2001) *Leishmania*-sand fly interactions controlling species-specific vector competence. *Cellular microbiology*, **3**, 189-96.
66. Teixeira, M.J., Teixeira, C.R., Andrade, B.B., Barral-Netto, M. and Barral, A. (2006) Chemokines in host-parasite interactions in leishmaniasis. *Trends in parasitology*, **22**, 32-40, 10.1016/j.pt.2005.11.010.
67. Ato, M., Stäger, S., Engwerda, C.R. and Kaye, P.M. (2002) Defective CCR7 expression on dendritic cells contributes to the development of visceral leishmaniasis. *Nature immunology*, **3**, 1185-91, 10.1038/ni861.
68. Arnoldi, J. and Moll, H. (1998) Langerhans cell migration in murine cutaneous leishmaniasis: Regulation by tumor necrosis factor alpha, interleukin-1 beta, and macrophage inflammatory protein-1 alpha. *Developmental immunology*, **6**, 3-11.
69. Zandbergen, G. van, Klinger, M., Mueller, A., Dannenberg, S., Gebert, A., Solbach, W. and Laskay, T. (2004) Cutting edge: Neutrophil granulocyte serves as a vector for *Leishmania* entry into macrophages. *Journal of immunology*, **173**, 6521-5.
70. Steigerwald, M. and Moll, H. (2005) *Leishmania major* modulates chemokine and chemokine receptor expression by dendritic cells and affects their migratory capacity. *Infection and immunity*, **73**, 2564-7, 10.1128/IAI.73.4.2564-2567.2005.
71. Ritter, U., Moll, H., Laskay, T., Bröcker, E., Velazco, O., Becker, I. and Gillitzer, R. (1996) Differential expression of chemokines in patients with localized and diffuse cutaneous American leishmaniasis. *The Journal of infectious diseases*, **173**, 699-709.

72. Hertz-Fowler, C., Peacock, C.S., Wood, V., Aslett, M., Kerhornou, A., Mooney, P., Tivey, A., Berriman, M., Hall, N., Rutherford, K., et al. (2004) GeneDB: A resource for prokaryotic and eukaryotic organisms. *Nucleic acids research*, **32**, D339-43, 10.1093/nar/gkh007.
73. Aslett, M., Aurrecoechea, C., Berriman, M., Brestelli, J., Brunk, B.P., Carrington, M., Depledge, D.P., Fischer, S., Gajria, B., Gao, X., et al. (2009) TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Research* 10.1093/nar/gkp851.
74. Ivens, A.C., Peacock, C.S., Worthey, E.A., Murphy, L., Aggarwal, G., Berriman, M., Sisk, E., Rajandream, M.-A., Adlem, E., Aert, R., et al. (2005) The genome of the kinetoplastid parasite, *Leishmania major*. *Science*, **309**, 436-42, 10.1126/science.1112680.
75. Peacock, C.S., Seeger, K., Harris, D., Murphy, L., Ruiz, J.C., Quail, M.A., Peters, N., Adlem, E., Tivey, A., Aslett, M., et al. (2007) Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. *Nature Genetics*, **39**, 839-47, 10.1038/ng2053.
76. Weatherly, D.B., Boehlke, C. and Tarleton, R.L. (2009) Chromosome level assembly of the hybrid *Trypanosoma cruzi* genome. *BMC Genomics*, **10**, 10.1186/1471-2164-10-255.
77. El-Sayed, N.M., Myler, P.J., Bartholomeu, D.C., Nilsson, D., Aggarwal, G., Tran, A.-N., Ghedin, E., Worthey, E.A., Delcher, A.L., Blandin, G., et al. (2005) The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science*, **309**, 409-15, 10.1126/science.1112631.
78. Kissinger, J.C. (2006) A tale of three genomes: the kinetoplastids have arrived. *Trends Parasitology*, **22**, 240-3, 10.1016/j.pt.2006.04.002.
79. Loftus, B., Anderson, I., Davies, R., Alsmark, U.C.M., Samuelson, J., Amedeo, P., Roncaglia, P., Berriman, M., Hirt, R.P., Mann, B.J., et al. (2005) The genome of the protist parasite *Entamoeba histolytica*. *Nature*, **433**, 865-8, 10.1038/nature03291.
80. Eichinger, L., Pachebat, J.A., Glöckner, G., Rajandream, M.-A., Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q., et al. (2005) The genome of the social amoeba *Dictyostelium discoideum*. *Nature*, **435**, 43-57, 10.1038/nature03481.
81. Chanda, I., Pan, A., Saha, S.K. and Dutta, C. (2007) Comparative codon and amino acid composition analysis of Tritryps-conspicuous features of *Leishmania major*. *FEBS Letters*, **581**, 5751-8, 10.1016/j.febslet.2007.11.041.
82. Vincendeau, P. and Bouteille, B. (2006) Immunology and immunopathology of African trypanosomiasis. *Anais da Academia Brasileira de Ciências*, **78**, 645-65.
83. Taylor, J.E. and Rudenko, G. (2006) Switching trypanosome coats: What´s in the wardrobe? *Trends in Genetics*, **22**, 614-20, 10.1016/j.tig.2006.08.003.
84. Sung, P., Krejci, L., Van Komen, S. and Sehorn, M.G. (2003) Rad51 recombinase and recombination mediators. *The Journal of biological chemistry*, **278**, 42729-32, 10.1074/jbc.R300027200.
85. Smith, M., Bringaud, F. and Papadopoulou, B. (2009) Organization and evolution of two SIDER retroposon subfamilies and their impact on the *Leishmania* genome. *BMC Genomics*, **10**, 10.1186/1471-2164-10-240.

86. Bringaud, F., Müller, M., Cerqueira, G.C., Smith, M., Rochette, A., El-Sayed, N.M.A., Papadopoulou, B. and Ghedin, E. (2007) Members of a large retroposon family are determinants of post-transcriptional gene expression in *Leishmania*. *PLoS Pathogens*, **3**, 1291-307, 10.1371/journal.ppat.0030136.
87. Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, **419**, 498-511, 10.1038/nature01097.
88. Heng, H.H.Q. (2009) The genome-centric concept: Resynthesis of evolutionary theory. *Bioessays*, **31**, 512-25, 10.1002/bies.200800182.
89. Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W., Iyer, V.N., et al. (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, **450**, 203-218, 10.1038/nature06341.
90. Carroll, S.B. (2008) Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell*, **134**, 25-36, 10.1016/j.cell.2008.06.030.
91. Akopyants, N.S., Kimblin, N., Secundino, N., Patrick, R., Peters, N., Lawyer, P., Dobson, D.E., Beverley, S.M. and Sacks, D.L. (2009) Demonstration of genetic exchange during cyclical development of *Leishmania* in the sand fly vector. *Science*, **324**, 265-8, 10.1126/science.1169464.
92. Myler, P.J., Sisk, E., McDonagh, P.D., Martinez-Calvillo, S., Schnaufer, A., Sunkin, S.M., Yan, S., Madhubala, R., Ivens, A. and Stuart, K. (2000) Genomic organization and gene function in *Leishmania*. *Biochemical Society transactions*, **28**, 527-31.
93. Ginger, M.L. (2005) Trypanosomatid biology and euglenozoan evolution: New insights and shifting paradigms revealed through genome sequencing. *Protist*, **156**, 377-92.
94. Martínez-Calvillo, S., Nguyen, D., Stuart, K. and Myler, P.J. (2004) Transcription initiation and termination on *Leishmania major* chromosome 3. *Eukaryotic Cell*, **3**, 506-17.
95. Campbell, D.A., Thomas, S. and Sturm, N.R. (2003) Transcription in kinetoplastid protozoa: Why be normal? *Microbes and infection*, **5**, 1231-40.
96. Pedrosa, A.L., Ruiz, J.C., Tosi, L.R. and Cruz, A.K. (2001) Characterisation of three chromosomal ends of *Leishmania major* reveals transcriptional activity across arrays of reiterated and unique sequences. *Molecular and biochemical parasitology*, **114**, 71-80.
97. Thomas, S., Green, A., Sturm, N.R., Campbell, D.A. and Myler, P.J. (2009) Histone acetylations mark origins of polycistronic transcription in *Leishmania major*. *BMC Genomics*, **10**, 10.1186/1471-2164-10-152.
98. Clayton, C.E. (2002) Life without transcriptional control? From fly to man and back again. *EMBO Journal*, **21**, 1881-8.
99. Zeiner, G.M., Sturm, N.R. and Campbell, D.A. (2003) The *Leishmania tarentolae* spliced leader contains determinants for association with polysomes. *Journal of biological chemisty*, **278**, 38269-75, 10.1074/jbc.M304295200.
100. Ouellette, M., Olivier, M., Sato, S. and Papadopoulou, B. (2003) [Studies on the parasite *Leishmania* in the post-genomic era]. *Médecine sciences : M/S*, **19**, 900-9.

101. McNicoll, F., Müller, M., Cloutier, S., Boilard, N., Rochette, A., Dubé, M. and Papadopoulou, B. (2005) Distinct 3'-untranslated region elements regulate stage-specific mRNA accumulation and translation in Leishmania. *The Journal of biological chemistry*, **280**, 35238-46, 10.1074/jbc.M507511200.
102. Horn, D. (2008) Codon usage suggests that translational selection has a major impact on protein expression in trypanosomatids. *BMC Genomics*, **9**, 10.1186/1471-2164-9-2.
103. Belli, S.I. (2000) Chromatin remodelling during the life cycle of trypanosomatids. *International journal for parasitology*, **30**, 679-87.
104. Saxena, A., Worthey, E.A., Yan, S., Leland, A., Stuart, K.D. and Myler, P.J. (2003) Evaluation of differential gene expression in *Leishmania major* Friedlin procyclics and metacyclics using DNA microarray analysis. *Molecular & Biochemical Parasitology*, **129**, 103-14.
105. Saxena, A., Lahav, T., Holland, N., Aggarwal, G., Anupama, A., Huang, Y., Volpin, H., Myler, P.J. and Zilberstein, D. (2007) Analysis of the *Leishmania donovani* transcriptome reveals an ordered progression of transient and permanent changes in gene expression during differentiation. *Molecular & Biochemical Parasitology*, **152**, 53-65, 10.1016/j.molbiopara.2006.11.011.
106. Salotra, P., Duncan, R.C., Singh, R., Raju, B.V.S., Sreenivas, G. and Nakhasi, H.L. (2006) Upregulation of surface proteins in *Leishmania donovani* isolated from patients of post kala-azar dermal leishmaniasis. *Microbes and infection*, **8**, 637-44, 10.1016/j.micinf.2005.08.018.
107. Holzer, T.R., McMaster, W.R. and Forney, J.D. (2006) Expression profiling by whole-genome interspecies microarray hybridization reveals differential gene expression in procyclic promastigotes, lesion-derived amastigotes, and axenic amastigotes in *Leishmania mexicana*. *Molecular & Biochemical Parasitology*, **146**, 198-218, 10.1016/j.molbiopara.2005.12.009.
108. Guimond, C., Trudel, N., Brochu, C., Marquis, N., El Fadili, A., Peytavi, R., Briand, G., Richard, D., Messier, N., Papadopoulou, B., et al. (2003) Modulation of gene expression in *Leishmania* drug resistant mutants as determined by targeted DNA microarrays. *Nucleic acids research*, **31**, 5886-96.
109. Leifso, K., Cohen-Freue, G., Dogra, N., Murray, A. and McMaster, W.R. (2007) Genomic and proteomic expression analysis of *Leishmania* promastigote and amastigote life stages: The *Leishmania* genome is constitutively expressed. *Molecular & Biochemical Parasitology*, **152**, 35-46, 10.1016/j.molbiopara.2006.11.009.
110. McNicoll, F., Drummelsmith, J., Müller, M., Madore, E., Boilard, N., Ouellette, M. and Papadopoulou, B. (2006) A combined proteomic and transcriptomic approach to the study of stage differentiation in *Leishmania infantum*. *Proteomics*, **6**, 3567-81, 10.1002/pmic.200500853.
111. Alcolea, P.J., Alonso, A., Sánchez-Gorostiaga, A., Moreno-Paz, M., Gómez, M.J., Ramos, I., Parro, V. and Larraga, V. (2009) Genome-wide analysis reveals increased levels of transcripts related with infectivity in peanut lectin non-agglutinated promastigotes of *Leishmania infantum*. *Genomics*, **93**, 551-64, 10.1016/j.ygeno.2009.01.007.

112. Depledge, D.P., Evans, K.J., Ivens, A.C., Aziz, N., Maroof, A., Kaye, P.M., Smith, D.F. (2009) Comparative expression profiling of *Leishmania*: Modulation in gene expression between species and in different host genetic backgrounds. *PLoS neglected tropical diseases*, **3**, 10.1371/journal.pntd.0000476.
113. Alcolea, P.J., Alonso, A., Gómez, M.J., Moreno, I., Domínguez, M., Parro, V. and Larraga, V. (2010) Transcriptomics throughout the life cycle of *Leishmania infantum:* High down-regulation rate in the amastigote stage. *International journal for parasitology*, **40**, 1497-516, 10.1016/j.ijpara.2010.05.013.
114. Holley, R.W., Apgar, J., Everett, G.A., Madison, J.T., Marquisee, M., Merrill, S.H., Penswick, J.R. and Zamir, A. (1965) Structure of a ribonucleic acid. *Science*, **147**, 1462-5.
115. Holley, R.W., Everett, G.A., Madison, J.T. and Zamir, A. (1965) Nucleotide sequence in the yeast alanine transfer ribonucleic acid. *The Journal of biological chemistry*, **240**, 2122-8.
116. Wu, R. and Taylor, E. (1971) Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA. *Journal of molecular biology*, **57**, 491-511.
117. Wu, R. (1970) Nucleotide sequence analysis of DNA. I. Partial sequence of the cohesive ends of bacteriophage lambda and 186 DNA. *Journal of molecular biology*, **51**, 501-21.
118. Sanger, F. and Coulson, A.R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, **94**, 441-8.
119. Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 560-4.
120. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5463-7.
121. Hutchison, C.A. (2007) DNA sequencing: Bench to bedside and beyond. *Nucleic acids research*, **35**, 6227-37, 10.1093/nar/gkm688.
122. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921, 10.1038/35057062.
123. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001) The sequence of the human genome. *Science (New York, N.Y.)*, **291**, 1304-51, 10.1126/science.1058040.
124. Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.-F., Dougherty, B.A., Merrick, J.M., et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496-512.
125. Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F. and Petersen, G.B. (1982) Nucleotide sequence of bacteriophage lambda DNA. *Journal of molecular biology*, **162**, 729-73.

126. Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520-62, 10.1038/nature01262.
127. Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nature Biotechnology*, **26**, 1135-45, 10.1038/nbt1486.
128. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-ju, Chen, Z., et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376-81, 10.1038/nature03959.
129. Bentley, D.R. (2006) Whole-genome re-sequencing. *Current opinion in genetics & development*, **16**, 545-52, 10.1016/j.gde.2006.10.009.
130. Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., et al. (2009) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, **327**, 78-81, 10.1126/science.1181498.
131. Wicker, T., Schlagenhauf, E., Graner, A., Close, T.J., Keller, B. and Stein, N. (2006) 454 sequencing put to the test using the complex genome of barley. *BMC Genomics*, **7**, 275.
132. Goldberg, S.M.D., Johnson, J., Busam, D., Feldblyum, T., Ferriera, S., Friedman, R., Halpern, A., Khouri, H., Kravitz, S.A., Lauro, F.M., et al. (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 11240-5, 10.1073/pnas.0604351103.
133. Nowrousian, M. (2010) Next-generation sequencing techniques for eukaryotic microorganisms: Sequencing-based solutions to biological problems. *Eukaryotic cell*, **9**, 1300-10, 10.1128/EC.00123-10.
134. Fullwood, M.J., Wei, C.-L., Liu, E.T. and Ruan, Y. (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome research*, **19**, 521-32, 10.1101/gr.074906.107.
135. Chaisson, M.J., Brinza, D. and Pevzner, P. a (2009) De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome research*, **19**, 336-46, 10.1101/gr.079053.108.
136. Huang, X. (1992) A contig assembly program based on sensitive detection of fragment overlaps. *Genomics*, **14**, 18-25.
137. Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H., Remington, K.A., et al. (2000) A whole-genome assembly of *Drosophila*. *Science*, **287**, 2196-204.
138. Miller, J.R., Delcher, A.L., Koren, S., Venter, E., Walenz, B.P., Brownley, A., Johnson, J., Li, K., Mobarry, C. and Sutton, G. (2008) Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, **24**, 2818-24, 10.1093/bioinformatics/btn548.
139. Pevzner, P.A., Tang, H. and Waterman, M.S. (2001) An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 9748-53, 10.1073/pnas.171285098.
140. Rothberg, J.M. and Leamon, J.H. (2008) The development and impact of 454 sequencing. *Nature Biotechnology*, **26**, 1117-24.

141. Zerbino, D.R., Birney, E. and Spring, C. (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 821-29, 10.1101/gr.074492.107.
142. Boisvert, S., Laviolette, F. and Corbeil, J. (2010) Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of Computational Biology*, **17**, 1519-33, 10.1089/cmb.2009.0238.
143. Borodovsky, M., Rudd, K.E. and Koonin, E.V. (1994) Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic acids research*, **22**, 4756-67.
144. Wilson, C.A., Kreychman, J. and Gerstein, M. (2000) Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *Journal of molecular biology*, **297**, 233-49, 10.1006/jmbi.2000.3550.
145. Hegyi, H. and Gerstein, M. (2001) Annotation transfer for genomics: Measuring functional divergence in multi-domain proteins. *Genome research*, **11**, 1632-40, 10.1101/gr. 183801.
146. Staden, R. and McLachlan, A.D. (1982) Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic acids research*, **10**, 141-56.
147. Baxevanis, A.D. (2004) An overview of gene identification: Approaches, strategies, and considerations. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, **Chapter 4**, Unit 4.1, 10.1002/0471250953.bi0401s6.
148. Stanke, M., Schöffmann, O., Morgenstern, B. and Waack, S. (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, **7**, 62, 10.1186/1471-2105-7-62.
149. Delcher, A.L., Harmon, D., Kasif, S., White, O. and Salzberg, S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic acids research*, **27**, 4636-41.
150. Blanco, E. and Abril, J.F. (2009) Computational gene annotation in new genome assemblies using GeneID. *Methods in molecular biology*, **537**, 243-61, 10.1007/978-1-59745-251-9_12.
151. Fontana, P., Cestaro, A., Velasco, R., Formentin, E. and Toppo, S. (2009) Rapid annotation of anonymous sequences from genome projects using semantic similarities and a weighting scheme in gene ontology. *PLoS One*, **4**, 10.1371/journal.pone.0004619.
152. Reese, M.G., Hartzell, G., Harris, N.L., Ohler, U., Abril, J.F. and Lewis, S.E. (2000) Genome annotation assessment in *Drosophila melanogaster*. *Genome research*, **10**, 483-501.
153. Smith, M., Blanchette, M. and Papadopoulou, B. (2008) Improving the prediction of mRNA extremities in the parasitic protozoan *Leishmania*. *BMC Bioinformatics*, **9**, 158, 10.1186/1471-2105-9-158.
154. Tarailo-Graovac, M. and Chen, N. (2002) Current Protocols in Bioinformatics John Wiley & Sons, Inc., Hoboken, NJ, USA.
155. Newberg, L.A. (2009) Error statistics of hidden Markov model and hidden Boltzmann model results. *BMC Bioinformatics*, **10**, 212, 10.1186/1471-2105-10-212.
156. Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Finn, R.D. and Sonnhammer, E.L. (1999) Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic acids research*, **27**, 260-2.

157. Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O. and Davis, R.W. (1996) Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 10614-9.
158. Shalon, D., Smith, S.J. and Brown, P.O. (1996) A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome research*, **6**, 639-45.
159. Miller, M.B. and Tang, Y.-W. (2009) Basic concepts of microarrays and potential applications in clinical microbiology. *Clinical microbiology reviews*, **22**, 611-33, 10.1128/CMR.00019-09.
160. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome biology*, **5**, R80, 10.1186/gb-2004-5-10-r80.
161. Smyth, G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**, Article 3, 10.2202/1544-6115.1027.
162. Gautier, L., Cope, L., Bolstad, B.M. and Irizarry, R.A. (2004) Affy - analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307-15, 10.1093/bioinformatics/btg405.
163. Saeed, A.I., Bhagabati, N.K., Braisted, J.C., Liang, W., Sharov, V., Howe, E.A., Li, J., Thiagarajan, M., White, J.A. and Quackenbush, J. (2006) TM4 microarray software suite. *Methods in Enzymology*, **411**, 134-193, 10.1016/S0076-6879(06)11009-5.
164. Battke, F., Symons, S. and Nieselt, K. (2010) Mayday-integrative analytics for expression data. *BMC bioinformatics*, **11**, 121, 10.1186/1471-2105-11-121.
165. Rehrauer, H., Zoller, S. and Schlapbach, R. (2007) MAGMA: Analysis of two-channel microarrays made easy. *Nucleic acids research*, **35**, W86-90, 10.1093/nar/gkm302.
166. Buness, A., Huber, W., Steiner, K., Sültmann, H. and Poustka, A. (2005) ArrayMagic: Two-colour cDNA microarray quality control and preprocessing. *Bioinformatics*, **21**, 554-6, 10.1093/bioinformatics/bti052.
167. Hokamp, K., Roche, F.M., Acab, M., Rousseau, M.-E., Kuo, B., Goode, D., Aeschliman, D., Bryan, J., Babiuk, L.A., Hancock, R.E.W., et al. (2004) ArrayPipe: A flexible processing pipeline for microarray data. *Nucleic acids research*, **32**, W457-9, 10.1093/nar/gkh446.
168. Silver, J.D., Ritchie, M.E. and Smyth, G.K. (2008) Microarray background correction: Maximum likelihood estimation for the normal-exponential convolution. *Biostatistics*, **10**, 352-63.
169. Ritchie, M.E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A. and Smyth, G.K. (2007) A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, **23**, 2700-7, 10.1093/bioinformatics/btm412.
170. Silver, J.D., Ritchie, M.E. and Smyth, G.K. (2008) Microarray background correction: Maximum likelihood estimation for the normal-exponential convolution. *Biostatistics*, **10**, 352-63, 10.1093/biostatistics/kxn042.
171. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249-64, 10.1093/biostatistics/4.2.249.

172. Millenaar, F.F., Okyere, J., May, S.T., Zanten, M. van, Voesenek, L.A.C.J. and Peeters, A.J.M. (2006) How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results. *BMC bioinformatics*, **7**, 137, 10.1186/1471-2105-7-137.
173. Naef, F. and Magnasco, M.O. (2003) Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays. *Physical review. E, Statistical, nonlinear, and soft matter physics*, **68**, 011906.
174. Sásik, R., Calvo, E. and Corbeil, J. (2002) Statistical analysis of high-density oligonucleotide arrays: A multiplicative noise model. *Bioinformatics*, **18**, 1633-40.
175. Chen, J.J., Wang, S.-J., Tsai, C.-A. and Lin, C.-J. (2007) Selection of differentially expressed genes in microarray data analysis. *The pharmacogenomics journal*, **7**, 212-20, 10.1038/sj.tpj.6500412.
176. Dennis, G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C. and Lempicki, R.A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome biology*, **4**, P3.
177. Shannon, P.T., Reiss, D.J., Bonneau, R. and Baliga, N.S. (2006) The Gaggle: An open-source software system for integrating bioinformatics software and data sources. *BMC bioinformatics*, **7**, 176, 10.1186/1471-2105-7-176.
178. Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.-L. and Ideker, T. (2010) Cytoscape 2.8: New features for data integration and network visualization. *Bioinformatics*, **27**, 431-2, 10.1093/bioinformatics/btq675.
179. Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P. and Morissette, J. (2008) Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics*, **41**, 706-16, 10.1016/j.jbi.2008.03.004.
180. Mockler, T.C., Chan, S., Sundaresan, A., Chen, H., Jacobsen, S.E. and Ecker, J.R. (2005) Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, **85**, 1-15, 10.1016/j.ygeno.2004.10.005.
181. Haquin, S., Oeuillet, E., Pajon, A., Harris, M., Jones, A.T., Tilbeurgh, H. van, Markley, J.L., Zolnai, Z. and Poupon, A. (2008) Data management in structural genomics: An overview. *Methods in molecular biology*, **426**, 49-79, 10.1007/978-1-60327-058-8_4.
182. Barrett, T. and Edgar, R. (2006) Gene Expression Omnibus: Microarray data storage, submission, retrieval, and analysis. *Methods in Enzymology*, **411**, 352-69, 10.1016/S0076-6879(06)11019-8.
183. Brazma, A., Kapushesky, M., Parkinson, H., Sarkans, U. and Shojatalab, M. (2006) Data storage and analysis in ArrayExpress. *Methods in Enzymology*, **411**, 370-86, 10.1016/S0076-6879(06)11020-4.
184. Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G.G., et al. (2003) ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, **31**, 68-71.
185. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., et al. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature genetics*, **29**, 365-71, 10.1038/ng1201-365.

186. Ritchie, M.E., Diyagama, D., Neilson, J., Laar, R. van, Dobrovic, A., Holloway, A. and Smyth, G.K. (2006) Empirical array quality weights in the analysis of microarray data. *BMC bioinformatics*, **7**, 261, 10.1186/1471-2105-7-261.
187. Zahurak, M., Parmigiani, G., Yu, W., Scharpf, R.B., Berman, D., Schaeffer, E., Shabbeer, S. and Cope, L. (2007) Pre-processing Agilent microarray data. *BMC bioinformatics*, **8**, 142, 10.1186/1471-2105-8-142.
188. Edwards, D. (2003) Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics*, **19**, 825-33.
189. Saal, L.H., Troein, C., Vallon-Christersson, J., Gruvberger, S., Borg, A. and Peterson, C. (2002) BioArray Software Environment (BASE):0 a platform for comprehensive management and analysis A of microarray data. *Genome biology*, **3**, SOFTWARE0003.
190. Bérubé, H. (2006) Mise en place d'une chaîne d'analyse et de traitement de biopuces. Mémoire. Université Laval.
191. Vallon-Christersson, J., Nordborg, N., Svensson, M. and Häkkinen, J. (2009) BASE-2nd generation software for microarray data management and analysis. *BMC bioinformatics*, **10**, 330, 10.1186/1471-2105-10-330.
192. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000) Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, **25**, 25-9, 10.1038/75556.
193. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., et al. (2008) KEGG for linking genomes to life and the environment. *Nucleic acids research*, **36**, D480-4, 10.1093/nar/gkm882.
194. Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., Goto, S. and Kanehisa, M. (2008) KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic acids research*, **36**, W423-6, 10.1093/nar/gkn282.
195. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *Journal of molecular biology*, **215**, 403-10, 10.1006/jmbi.1990.9999.
196. Ubeda, J.-M., Légaré, D., Raymond, F., Ouameur, A.A., Boisvert, S., Rigault, P., Corbeil, J., Tremblay, M.J., Olivier, M., Papadopoulou, B., et al. (2008) Modulation of gene expression in drug resistant Leishmania is associated with gene amplification, gene deletion and chromosome aneuploidy. *Genome biology*, **9**, R115, 10.1186/gb-2008-9-7-r115.
197. Leprohon, P., Légaré, D., Raymond, F., Madore, É., Hardiman, G., Corbeil, J. and Ouellette, M. (2009) Gene expression modulation is associated with gene amplification, supernumerary chromosomes and chromosome loss in antimony-resistant *Leishmania infantum*. *Nucleic acids research*, **37**, 1387-99.
198. Rochette, A., Raymond, F., Ubeda, J.-M., Smith, M., Messier, N., Boisvert, S., Rigault, P., Corbeil, J., Ouellette, M. and Papadopoulou, B. (2008) Genome-wide gene expression profiling analysis of *Leishmania major* and *Leishmania infantum* developmental stages reveals substantial differences between the two species. *BMC Genomics*, **9**, 255, 10.1186/1471-2164-9-255.
199. Murray, H.W., Berman, J.D., Davies, C.R. and Saravia, N.G. (2005) Advances in leishmaniasis. *The Lancet*, **366**, 1561-77, 10.1016/S0140-6736(05)67629-5.

200. Noyes, H.A., Chance, M.L., Croan, D.G. and Ellis, J.T. (1998) *Leishmania (sauroleishmania):* A comment on classification. *Parasitology today*, **14**, 167.
201. Wilson, V. and Southern, B. (1979) Lizard *Leishmania*. In Lumsden, W., Evans, D. (eds), *Biology of Kinetoplastida*. Academic press, New York, pp. 242-68.
202. Killick-Kendrick, R., Lainson, R., Rioux, J.-A. and Saf'janova, V.M. (1986) The taxonomy of *Leishmania*-like parasites in reptiles. In *Leishmania. Taxonomie et phylogenèse*. CNRS/INSERM/OMS. Colloque international, pp. 143-48.
203. Taylor, V.M., Muñoz, D.L., Cedeño, D.L., Vélez, I.D., Jones, M.A. and Robledo, S.M. (2010) *Leishmania tarentolae:* Utility as an in vitro model for screening of antileishmanial agents. *Experimental parasitology*, **126**, 471-5, 10.1016/j.exppara.2010.05.016.
204. Ouellette, M., Hettema, E., Wüst, D., Fase-Fowler, F. and Borst, P. (1991) Direct and inverted DNA repeats associated with P-glycoprotein gene amplification in drug resistant *Leishmania*. *The EMBO journal*, **10**, 1009-16.
205. Petrillo-Peixoto, M.L. and Beverley, S.M. (1988) Amplified DNAs in laboratory stocks of *Leishmania tarentolae:* Extrachromosomal circles structurally and functionally similar to the inverted-H-region amplification of methotrexate-resistant *Leishmania major*. *Molecular and cellular biology*, **8**, 5188-99.
206. White, T.C., Fase-Fowler, F., Luenen, H. van, Calafat, J. and Borst, P. (1988) The H circles of *Leishmania tarentolae* are a unique amplifiable system of oligomeric DNAs associated with drug resistance. *The Journal of biological chemistry*, **263**, 16977-83.
207. Simpson, L., Aphasizhev, R., Gao, G. and Kang, X. (2004) Mitochondrial proteins and complexes in *Leishmania* and *Trypanosoma* involved in U-insertion/deletion RNA editing. *RNA*, **10**, 159-70.
208. Basile, G. and Peticca, M. (2009) Recombinant protein expression in *Leishmania tarentolae*. *Molecular biotechnology*, **43**, 273-8, 10.1007/s12033-009-9213-5.
209. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Molecular cloning: A laboratory manual, second edition, Vols. 1, 2 and 3. In *Sambrook J E F Fritsch and T Maniatis Molecular Cloning A Laboratory Manual Second Edition Vols 1 2 and 3* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA. 999 pp.
210. Sommer, D.D., Delcher, A.L., Salzberg, S.L. and Pop, M. (2007) Minimus: A fast, lightweight genome assembler. *BMC Bioinformatics*, **8**, 64, 10.1186/1471-2105-8-64.
211. Stanke, M. and Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19**, 215-25, 10.1093/bioinformatics/btg1080.
212. Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. and Morgenstern, B. (2006) AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic acids research*, **34**, W435-39, 10.1093/nar/gkl200.
213. Krogh, A., Sj lander, K., Brown, M., Mian, I.S. and Haussler, D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *Journal of molecular biology*, **235**, 1501-31, 10.1006/jmbi.1994.1104.
214. Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. and Robles, M. (2005) Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674-6, 10.1093/bioinformatics/bti610.

215. Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular biology and evolution*, **24**, 1596-9, 10.1093/molbev/msm092.
216. Chen, F., Mackey, A.J., Stoeckert, C.J. and Roos, D.S. (2006) OrthoMCL-DB: Querying a comprehensive multi-species collection of ortholog groups. *Nucleic acids research*, **34**, D363-8, 10.1093/nar/gkj123.
217. Li, L., Stoeckert, C.J. and Roos, D.S. (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome research*, **13**, 2178-89, 10.1101/gr.1224503.
218. Huang, X. and Madan, A. (1999) CAP3: A DNA sequence assembly program. *Genome Reearchs*, **9**, 868-877.
219. Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589-95, 10.1093/bioinformatics/btp698.
220. Mousavi, S.A., Malerød, L., Berg, T. and Kjeken, R. (2004) Clathrin-dependent endocytosis. *The Biochemical journal*, **377**, 1-16, 10.1042/BJ20031000.
221. Ohno, H. (2006) Physiological roles of clathrin adaptor AP complexes: Lessons from mutant animals. *Journal of biochemistry*, **139**, 943-8, 10.1093/jb/mvj120.
222. Weise, F., Thilo, L., Engstler, M., Wiese, M., Benzel, I., Kühn, C., Bühring, H.-J. and Overath, P. (2005) Binding affinity and capacity of putative adaptor-mediated sorting of a Type I membrane protein in Leishmania mexicana. *Molecular and biochemical parasitology*, **142**, 203-11, 10.1016/j.molbiopara.2005.04.002.
223. Cantrell, D.A. (2001) Phosphoinositide 3-kinase signalling pathways. *Journal of cell science*, **114**, 1439-45.
224. Caberoy, N.B., Zhou, Y. and Li, W. (2010) Tubby and tubby-like protein 1 are new MerTK ligands for phagocytosis. *The EMBO journal*, **29**, 3898-910, 10.1038/emboj.2010.265.
225. Wallar, B.J. and Alberts, A.S. (2003) The formins: Active scaffolds that remodel the cytoskeleton. *Trends in cell biology*, **13**, 435-46.
226. Welch, M.D. and Mullins, R.D. (2002) Cellular control of actin nucleation. *Annual review of cell and developmental biology*, **18**, 247-88, 10.1146/annurev.cellbio.18.040202.112133.
227. Swenerton, R.K., Knudsen, G.M., Sajid, M., Kelly, B.L. and McKerrow, J.H. (2010) *Leishmania* subtilisin is a maturase for the trypanothione reductase system and contributes to disease pathology. *The Journal of biological chemistry*, **285**, 31120-9, 10.1074/jbc.M110.114462.
228. Wilson, M.A. (2011) The role of cysteine oxidation in DJ-1 function and dysfunction. *Antioxidants & redox signaling*, 10.1089/ars.2010.3481.
229. Jodko, K. and Litwinienko, G. (2010) [Oxidative stress in the neurodegenerative diseases-potential antioxidant activity of catecholamines]. *Postepy biochemii*, **56**, 248-59.
230. Butcher, B.A., Turco, S.J., Hilty, B.A., Pimenta, P.F., Panunzio, M. and Sacks, D.L. (1996) Deficiency in beta1, 3-galactosyltransferase of a *Leishmania major* lipophosphoglycan mutant adversely influences the Leishmania-sand fly interaction. *The Journal of biological chemistry*, **271**, 20573-9.

231. Dobson, D.E., Mengeling, B.J., Cilmi, S., Hickerson, S., Turco, S.J. and Beverley, S.M. (2003) Identification of genes encoding arabinosyltransferases (SCA) mediating developmental modifications of lipophosphoglycan required for sand fly transmission of *Leishmania major*. *The Journal of biological chemistry*, **278**, 28840-8, 10.1074/jbc.M302728200.
232. Williams, D.B. (2006) Beyond lectins: The calnexin/calreticulin chaperone system of the endoplasmic reticulum. *Journal of cell science*, **119**, 615-23, 10.1242/jcs.02856.
233. Wu, Y., El Fakhry, Y., Sereno, D., Tamar, S. and Papadopoulou, B. (2000) A new developmentally regulated gene family in *Leishmania* amastigotes encoding a homolog of amastin surface proteins. *Molecular and biochemical parasitology*, **110**, 345-57.
234. Jackson, A.P. (2010) The evolution of amastin surface glycoproteins in trypanosomatid parasites. *Molecular biology and evolution*, **27**, 33-45, 10.1093/molbev/msp214.
235. Schneider, P., Rosat, J.P., Bouvier, J., Louis, J. and Bordier, C. (1992) *Leishmania major:* Differential regulation of the surface metalloprotease in amastigote and promastigote stages. *Experimental parasitology*, **75**, 196-206.
236. Ilgoutz, S.C. and McConville, M.J. (2001) Function and assembly of the *Leishmania* surface coat. *International journal for parasitology*, **31**, 899-908.
237. Devault, A. and Bañuls, A.-L. (2008) The promastigote surface antigen gene family of the *Leishmania* parasite: differential evolution by positive selection and recombination. *BMC evolutionary biology*, **8**, 292, 10.1186/1471-2148-8-292.
238. Brotherton, M.-C., Racine, G., Foucher, A.L., Drummelsmith, J., Papadopoulou, B. and Ouellette, M. (2010) Analysis of stage-specific expression of basic proteins in *Leishmania infantum*. *Journal of proteome research*, **9**, 3842-53, 10.1021/pr100048m.
239. Depledge, D.P., Evans, K.J., Ivens, A.C., Aziz, N., Maroof, A., Kaye, P.M. and Smith, D.F. (2009) Comparative expression profiling of *Leishmania*: Modulation in gene expression between species and in different host genetic backgrounds. *PLoS neglected tropical diseases*, **3**, e476, 10.1371/journal.pntd.0000476.
240. Paape, D., Barrios-Llerena, M.E., Le Bihan, T., Mackay, L. and Aebischer, T. (2010) Gel free analysis of the proteome of intracellular *Leishmania mexicana*. *Molecular and biochemical parasitology*, **169**, 108-14, 10.1016/j.molbiopara.2009.10.009.
241. Srividya, G., Duncan, R., Sharma, P., Raju, B.V.S., Nakhasi, H.L. and Salotra, P. (2007) Transcriptome analysis during the process of in vitro differentiation of *Leishmania donovani* using genomic microarrays. *Parasitology*, **134**, 1527-39, 10.1017/S003118200700296X.
242. Chain, P.S.G., Grafham, D.V., Fulton, R.S., Fitzgerald, M.G., Hostetler, J., Muzny, D., Ali, J., Birren, B., Bruce, D.C., Buhay, C., et al. (2009) Genomics. Genome project standards in a new era of sequencing. *Science*, **326**, 236-37, 10.1126/science.1180614.
243. Naderer, T., Vince, J.E. and McConville, M.J. (2004) Surface determinants of *Leishmania* parasites and their role in infectivity in the mammalian host. *Current molecular medicine*, **4**, 649-65.
244. Descoteaux, A. and Turco, S.J. (1999) Glycoconjugates in *Leishmania* infectivity. *Biochimica et biophysica acta*, **1455**, 341-52.

245. Yao, C. (2010) Major surface protease of trypanosomatids: One size fits all? *Infection and immunity*, **78**, 22-31, 10.1128/IAI.00776-09.
246. Previato, J.O., Jones, C., Wait, R., Routier, F., Saraiva, E. and Mendonça-Previato, L. (1997) *Leishmania adleri*, a lizard parasite, expresses structurally similar glycoinositolphospholipids to mammalian *Leishmania*. *Glycobiology*, **7**, 687-95.
247. Lodge, R. and Descoteaux, A. (2008) *Leishmania* invasion and phagosome biogenesis. *Subcellular Biochemistry*, **47**, 174-81.
248. Becker, I., Salaiza, N., Aguirre, M., Delgado, J., Carrillo-Carrasco, N., Kobeh, L.G., Ruiz, A., Cervantes, R., Torres, A.P., Cabrera, N., et al. (2003) *Leishmania* lipophosphoglycan (LPG) activates NK cells through toll-like receptor-2. *Molecular and biochemical parasitology*, **130**, 65-74.
249. Aebischer, T., Bennett, C.L., Pelizzola, M., Vizzardelli, C., Pavelka, N., Urbano, M., Capozzoli, M., Luchini, A., Ilg, T., Granucci, F., et al. (2005) A critical role for lipophosphoglycan in proinflammatory responses of dendritic cells to *Leishmania mexicana*. *European journal of immunology*, **35**, 476-86, 10.1002/eji.200425674.
250. Späth, G.F., Lye, L.-F., Segawa, H., Sacks, D.L., Turco, S.J. and Beverley, S.M. (2003) Persistence without pathology in phosphoglycan-deficient *Leishmania* major. *Science*, **301**, 1241-3, 10.1126/science.1087499.
251. Späth, G.F., Garraway, L.A., Turco, S.J. and Beverley, S.M. (2003) The role(s) of lipophosphoglycan (LPG) in the establishment of *Leishmania major* infections in mammalian hosts. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9536-41, 10.1073/pnas.1530604100.
252. Ilg, T. (2000) Lipophosphoglycan is not required for infection of macrophages or mice by *Leishmania mexicana*. *The EMBO journal*, **19**, 1953-62, 10.1093/emboj/19.9.1953.
253. Button, L.L. and McMaster, W.R. (1988) Molecular cloning of the major surface antigen of *Leishmania*. *The Journal of experimental medicine*, **167**, 724-9.
254. Joshi, P.B., Kelly, B.L., Kamhawi, S., Sacks, D.L. and McMaster, W.R. (2002) Targeted gene deletion in *Leishmania major* identifies leishmanolysin (GP63) as a virulence factor. *Molecular and biochemical parasitology*, **120**, 33-40.
255. Brittingham, A., Morrison, C.J., McMaster, W.R., McGwire, B.S., Chang, K.P. and Mosser, D.M. (1995) Role of the *Leishmania* surface protease gp63 in complement fixation, cell adhesion, and resistance to complement-mediated lysis. *Journal of immunology*, **155**, 3102-11.
256. McGwire, B.S., Chang, K.-P. and Engman, D.M. (2003) Migration through the extracellular matrix by the parasitic protozoan *Leishmania* is enhanced by surface metalloprotease gp63. *Infection and immunity*, **71**, 1008-10.
257. Hallé, M., Gomez, M.A., Stuible, M., Shimizu, H., McMaster, W.R., Olivier, M. and Tremblay, M.L. (2009) The *Leishmania* surface protease GP63 cleaves multiple intracellular proteins and actively participates in p38 mitogen-activated protein kinase inactivation. *The Journal of biological chemistry*, **284**, 6893-908, 10.1074/jbc.M805861200.
258. Contreras, I., Gómez, M.A., Nguyen, O., Shio, M.T., McMaster, R.W. and Olivier, M. (2010) *Leishmania*-induced inactivation of the macrophage transcription factor AP-1 is mediated by the parasite metalloprotease GP63. *PLoS pathogens*, **6**, e1001148, 10.1371/journal.ppat.1001148.

259. Gregory, D.J., Godbout, M., Contreras, I., Forget, G. and Olivier, M. (2008) A novel form of NF-kappaB is induced by *Leishmania* infection: Involvement in macrophage gene expression. *European journal of immunology*, **38**, 1071-81, 10.1002/eji.200737586.
260. Azizi, H., Hassani, K., Taslimi, Y., Najafabadi, H.S., Papadopoulou, B. and Rafati, S. (2009) Searching for virulence factors in the non-pathogenic parasite to humans *Leishmania tarentolae*. *Parasitology*, 1-13, 10.1017/S0031182009005873.
261. Campbell, D.A., Kurath, U. and Fleischmann, J. (1992) Identification of a gp63 surface glycoprotein in *Leishmania tarentolae*. *FEMS microbiology letters*, **75**, 89-92.
262. Beetham, J.K., Donelson, J.E. and Dahlin, R.R. (2003) Surface glycoprotein PSA (GP46) expression during short- and long-term culture of *Leishmania chagasi*. *Molecular and biochemical parasitology*, **131**, 109-17.
263. Teixeira, S.M., Russell, D.G., Kirchhoff, L.V. and Donelson, J.E. (1994) A differentially expressed gene family encoding "amastin, " a surface protein of *Trypanosoma cruzi* amastigotes. *The Journal of biological chemistry*, **269**, 20509-16.
264. Boucrot, E., Saffarian, S., Zhang, R. and Kirchhausen, T. (2010) Roles of AP-2 in clathrin-mediated endocytosis. *PloS one*, **5**, e10597, 10.1371/journal.pone.0010597.
265. Vince, J.E., Tull, D.L., Spurck, T., Derby, M.C., McFadden, G.I., Gleeson, P.A., Gokool, S. and McConville, M.J. (2008) *Leishmania* adaptor protein-1 subunits are required for normal lysosome traffic, flagellum biogenesis, lipid homeostasis, and adaptation to temperatures encountered in the mammalian host. *Eukaryotic cell*, **7**, 1256-67, 10.1128/EC.00090-08.
266. Gokool, S. (2003) Sigma 1- and mu 1-Adaptin homologues of *Leishmania mexicana* are required for parasite survival in the infected host. *The Journal of biological chemistry*, **278**, 29400-9, 10.1074/jbc.M304572200.
267. Besteiro, S., Tonn, D., Tetley, L., Coombs, G.H. and Mottram, J.C. (2008) The AP3 adaptor is involved in the transport of membrane proteins to acidocalcisomes of *Leishmania*. *Journal of cell science*, **121**, 561-70, 10.1242/jcs.022574.
268. Tomsig, J.L. and Creutz C.E. (2002) Copines: a ubiquitous family of Ca(2+)-dependent phospholipid-binding proteins. *Cellular and molecular life sciences,* **59**, 1467-77.
269. Pannekoek W.J., Kooistra M.R., Zwartkruis F.J. and Bos J.L. (2009) Cell-cell junction formation: the role of Rap1 and Rap1 guanine nucleotide exchange factors. *Biochimica et biophysica acta,* **1788,** 790-6.
270. McKerrow J.H., Pino-Heiss S., Lindquist R. and Werb Z. (1985) Purification and characterization of an elastinolytic proteinase secreted by cercariae of *Schistosoma mansoni. The Journal of biological chemistry*, **260,** 3703-7.
271. Button L.L., Reiner N.E. and McMaster W.R. (1991) Modification of GP63 genes from diverse species of *Leishmania* for expression of recombinant protein at high levels in *Escherichia coli*. *Molecular and biochemical parasitology*, **44,** 213-24.
272. Firat-Karalar, E.N. and Welch, M.D. (2010) New mechanisms and functions of actin nucleation. *Current opinion in cell biology*, 10.1016/j.ceb.2010.10.007.
273. Krauth-Siegel, R.L. and Comini, M.A. (2008) Redox control in trypanosomatids, parasitic protozoa with trypanothione-based thiol metabolism. *Biochimica et biophysica acta*, **1780**, 1236-48, 10.1016/j.bbagen.2008.03.006.

274. Alkan, C., Sajjadian, S. and Eichler, E.E. (2010) Limitations of next-generation genome sequence assembly. *Nature methods*, **8**, 61-5, 10.1038/nmeth.1527.
275. Utokaparch, S., Marchant, D., Gosselink, J.V., McDonough, J.E., Thomas, E.E., Hogg, J.C. and Hegele, R.G. (2010) The relationship between respiratory viral loads and diagnosis in children presenting to a pediatric hospital emergency department. *The Pediatric infectious disease journal*, **30**, e18-23, 10.1097/INF.0b013e3181ff2fac.
276. Hsieh, Y.-J., Chin, H., Chiu, N.-C. and Huang, F.-Y. (2010) Hospitalized pediatric parainfluenza virus infections in a medical center. *Journal of Microbiology, Immunology and Infection*, **43**, 360-65, 10.1016/S1684-1182(10)60057-6.
277. Mahony, J.B. (2008) Detection of respiratory viruses by molecular methods. *Clinical microbiology reviews*, **21**, 716-47, 10.1128/CMR.00037-07.
278. Watanabe, A., Carraro, E., Kamikawa, J., Leal, E., Granato, C. and Bellei, N. (2010) Rhinovirus species and their clinical presentation among different risk groups of non-hospitalized patients. *Journal of medical virology*, **82**, 2110-5, 10.1002/jmv.21914.
279. Nichols, W.G., Peck Campbell, A.J. and Boeckh, M. (2008) Respiratory viruses other than influenza virus: Impact and therapeutic advances. *Clinical microbiology reviews*, **21**, 274-90, 10.1128/CMR.00045-07.
280. Sung, J.Y., Lee, H.J., Eun, B.W., Kim, S.H., Lee, S.Y., Lee, J.Y., Park, K.U. and Choi, E.H. (2010) Role of human coronavirus NL63 in hospitalized children with croup. *The Pediatric infectious disease journal*, **29**, 822-6, 10.1097/INF.0b013e3181e7c18d.
281. Fiore, A.E., Bridges, C.B. and Cox, N.J. (2009) Seasonal influenza vaccines. *Current topics in microbiology and immunology*, **333**, 43-82, 10.1007/978-3-540-92165-3_3.
282. Keech, M. and Beardsworth, P. (2008) The impact of influenza on working days lost: A review of the literature. *PharmacoEconomics*, **26**, 911-24.
283. Listed, N. authors (2000) Easing the burden: The challenge of managing influenza. *The American journal of managed care*, **6**, S276-81.
284. Molinari, N.-A.M., Ortega-Sanchez, I.R., Messonnier, M.L., Thompson, W.W., Wortley, P.M., Weintraub, E. and Bridges, C.B. (2007) The annual impact of seasonal influenza in the US: Measuring disease burden and costs. *Vaccine*, **25**, 5086-96, 10.1016/j.vaccine.2007.03.046.
285. Welliver, T.P., Garofalo, R.P., Hosakote, Y., Hintz, K.H., Avendano, L., Sanchez, K., Velozo, L., Jafri, H., Chavez-Bueno, S., Ogra, P.L., et al. (2007) Severe human lower respiratory tract illness caused by respiratory syncytial virus and influenza virus is characterized by the absence of pulmonary cytotoxic lymphocyte responses. *The Journal of infectious diseases*, **195**, 1126-36, 10.1086/512615.
286. Paramore, L.C., Ciuryla, V., Ciesla, G. and Liu, L. (2004) Economic impact of respiratory syncytial virus-related illness in the US: An analysis of national databases. *PharmacoEconomics*, **22**, 275-84.
287. Nuolivirta, K., Koponen, P., He, Q., Halkosalo, A., Korppi, M., Vesikari, T. and Helminen, M. (2010) *Bordetella pertussis* infection is common in nonvaccinated infants admitted for bronchiolitis. *The Pediatric infectious disease journal*, **29**, 1013-5.

288. File, T.M. and Marrie, T.J. (2010) Burden of community-acquired pneumonia in North American adults. *Postgraduate medicine*, **122**, 130-41, 10.3810/pgm.2010.03.2130.
289. Hoyo, I., Linares, L., Cervera, C., Almela, M., Marcos, M.A., Sanclemente, G., Cofán, F., Ricart, M.J. and Moreno, A. (2010) Epidemiology of pneumonia in kidney transplantation. *Transplantation proceedings*, **42**, 2938-40, 10.1016/j.transproceed.2010.07.082.
290. Lynch, J.P. and Zhanel, G.G. (2010) *Streptococcus pneumoniae:* Epidemiology and risk factors, evolution of antimicrobial resistance, and impact of vaccines. *Current opinion in pulmonary medicine*, **16**, 217-25, 10.1097/MCP.0b013e3283385653.
291. Jones, R.N., Jacobs, M.R. and Sader, H.S. (2010) Evolving trends in *Streptococcus pneumoniae* resistance: Implications for therapy of community-acquired bacterial pneumonia. *International journal of antimicrobial agents*, **36**, 197-204, 10.1016/j.ijantimicag.2010.04.013.
292. McDonough, E.A., Barrozo, C.P., Russell, K.L. and Metzgar, D. (2005) A multiplex PCR for detection of *Mycoplasma pneumoniae, Chlamydophila pneumoniae, Legionella pneumophila,* and *Bordetella pertussis* in clinical specimens. *Molecular and cellular probes*, **19**, 314-22, 10.1016/j.mcp.2005.05.002.
293. Jennings, L.C., Anderson, T.P., Beynon, K.A., Chua, A., Laing, R.T.R., Werno, A.M., Young, S.A., Chambers, S.T. and Murdoch, D.R. (2008) Incidence and characteristics of viral community-acquired pneumonia in adults. *Thorax*, **63**, 42-8, 10.1136/thx.2006.075077.
294. Aebi, T., Weisser, M., Bucher, E., Hirsch, H.H., Marsch, S. and Siegemund, M. (2010) Co-infection of influenza B and streptococci causing severe pneumonia and septic shock in healthy women. *BMC infectious diseases*, **10**, 308, 10.1186/1471-2334-10-308.
295. Talbot, H.K. and Falsey, A.R. (2010) The diagnosis of viral respiratory disease in older adults. *Clinical infectious diseases*, **50**, 747-51, 10.1086/650486.
296. Byington, C.L., Castillo, H., Gerber, K., Daly, J.A., Brimley, L.A., Adams, S., Christenson, J.C. and Pavia, A.T. (2002) The effect of rapid respiratory viral diagnostic testing on antibiotic use in a children´s hospital. *Archives of pediatrics & adolescent medicine*, **156**, 1230-4.
297. Goldmann, D.A., Weinstein, R.A., Wenzel, R.P., Tablan, O.C., Duma, R.J., Gaynes, R.P., Schlosser, J. and Martone, W.J. (1996) Strategies to prevent and control the emergence and spread of antimicrobial-resistant microorganisms in hospitals. A challenge to hospital leadership. *JAMA : The journal of the American Medical Association*, **275**, 234-40.
298. Tenover, F.C. (2010) Potential impact of rapid diagnostic tests on improving antimicrobial use. *Annals of the New York Academy of Sciences*, **1213**, 70-80, 10.1111/j.1749-6632.2010.05827.x.
299. Loeffelholz, M. and Chonmaitree, T. (2010) Advances in diagnosis of respiratory virus infections. *International journal of microbiology*, **2010**, 126049, 10.1155/2010/126049.
300. Ginocchio, C.C., Zhang, F., Manji, R., Arora, S., Bornfreund, M., Falk, L., Lotlikar, M., Kowerska, M., Becker, G., Korologos, D., et al. (2009) Evaluation of multiple test methods for the detection of the novel 2009 influenza A (H1N1) during the New York City outbreak. *Journal of clinical virology*, **45**, 191-5, 10.1016/j.jcv.2009.06.005.

301. Takahashi, H., Otsuka, Y. and Patterson, B.K. (2010) Diagnostic tests for influenza and other respiratory viruses: Determining performance specifications based on clinical setting. *Journal of infection and chemotherapy*, **16**, 155-61, 10.1007/s10156-010-0035-y.
302. Chkhaidze, I., Manjavidze, N. and Nemsadze, K. (2006) Serodiagnosis of acute respiratory infections in children in Georgia. *Indian journal of pediatrics*, **73**, 569-72.
303. Hindiyeh, M., Hillyard, D.R. and Carroll, K.C. (2001) Evaluation of the Prodesse Hexaplex multiplex PCR assay for direct detection of seven respiratory viruses in clinical specimens. *American journal of clinical pathology*, **116**, 218-24, 10.1309/F1R7-XD6T-RN09-1U6L.
304. Lin, B., Wang, Z., Vora, G.J., Thornton, J.A., Schnur, J.M., Thach, D.C., Blaney, K.M., Ligler, A.G., Malanoski, A.P., Santiago, J., et al. (2006) Broad-spectrum respiratory tract pathogen identification using resequencing DNA microarrays. *Genome research*, **16**, 527-35, 10.1101/gr.4337206.
305. Quan, P.-L., Palacios, G., Jabado, O.J., Conlan, S., Hirschberg, D.L., Pozo, F., Jack, P.J.M., Cisterna, D., Renwick, N., Hui, J., et al. (2007) Detection of respiratory viruses and subtype identification of influenza A viruses by GreeneChipResp oligonucleotide microarray. *Journal of clinical microbiology*, **45**, 2359-64, 10.1128/JCM.00737-07.
306. Cannon, G.A., Carr, M.J., Yandle, Z., Schaffer, K., Kidney, R., Hosny, G., Doyle, A., Ryan, J., Gunson, R., Collins, T., et al. (2010) A low density oligonucleotide microarray for the detection of viral and atypical bacterial respiratory pathogens. *Journal of virological methods*, **163**, 17-24, 10.1016/j.jviromet.2009.07.005.
307. Fox, J.D. (2007) Nucleic acid amplification tests for detection of respiratory viruses. *Journal of Clinical Virology*, **1**, 15-23.
308. Krunic, N., Yager, T., Himsworth, D., Merante, F., Yaghoubian, S. and Janeczko, R. (2007) xTAG™ RVP assay: Analytical and clinical performance. *Journal of Clinical Virology*, **40**, S39-46, 10.1016/S1386-6532(07)70009-4.
309. Merante, F., Yaghoubian, S. and Janeczko, R. (2007) Principles of the xTAG™ respiratory viral panel assay (RVP Assay). *Journal of Clinical Virology*, **40**, S31-5, 10.1016/S1386-6532(07)70007-0.
310. Balada-Llasat, J.-M., Larue, H., Kelly, C., Rigali, L. and Pancholi, P. (2010) Evaluation of commercial ResPlex II v2.0, MultiCode(®)-PLx, and xTAG(®) respiratory viral panels for the diagnosis of respiratory viral infections in adults. *Journal of clinical virology*, **50**, 42-5, 10.1016/j.jcv.2010.09.022.
311. Marshall, D.J., Reisdorf, E., Harms, G., Beaty, E., Moser, M.J., Lee, W.-M., Gern, J.E., Nolte, F.S., Shult, P. and Prudent, J.R. (2007) Evaluation of a multiplexed PCR assay for detection of respiratory viral pathogens in a public health laboratory setting. *Journal of clinical microbiology*, **45**, 3875-82, 10.1128/JCM.00838-07.
312. Li, H., McCormac, M.A., Estes, R.W., Sefers, S.E., Dare, R.K., Chappell, J.D., Erdman, D.D., Wright, P.F. and Tang, Y.-W. (2007) Simultaneous detection and high-throughput identification of a panel of RNA viruses causing respiratory tract infections. *Journal of clinical microbiology*, **45**, 2105-9, 10.1128/JCM.00210-07.

313. Raymond, F., Carbonneau, J., Boucher, N., Robitaille, L., Boisvert, S., Wu, W.-K., De Serres, G., Boivin, G. and Corbeil, J. (2009) Comparison of automated microarray detection with real-time PCR assays for detection of respiratory viruses in specimens obtained from children. *Journal of clinical microbiology*, **47**, 743-50, 10.1128/JCM.01297-08.
314. Fendrick, A.M., Monto, A.S., Nightengale, B. and Sarnes, M. (2003) The economic burden of non-influenza-related viral respiratory tract infection in the United States. *Archives of internal medicine*, **163**, 487-94.
315. Wallace, L.A., Collins, T.C., Douglas, J.D.M., McIntyre, S., Millar, J. and Carman, W.F. (2004) Virological surveillance of influenza-like illness in the community using PCR and serology. *Journal of clinical virology*, **31**, 40-5, 10.1016/j.jcv.2003.12.003.
316. Speers, D.J. (2006) Clinical applications of molecular biology for infectious diseases. *The Clinical biochemist. Reviews / Australian Association of Clinical Biochemists*, **27**, 39-51.
317. Aberle, J.H., Aberle, S.W., Pracher, E., Hutter, H.-P., Kundi, M. and Popow-Kraupp, T. (2005) Single versus dual respiratory virus infections in hospitalized infants: Impact on clinical course of disease and interferon-gamma response. *The Pediatric infectious disease journal*, **24**, 605-10.
318. Woo, P.C., Chiu, S.S., Seto, W.H. and Peiris, M. (1997) Cost-effectiveness of rapid diagnosis of viral respiratory tract infections in pediatric patients. *Journal of clinical microbiology*, **35**, 1579-81.
319. Barenfanger, J., Drake, C., Leon, N., Mueller, T. and Troutt, T. (2000) Clinical and financial benefits of rapid detection of respiratory viruses: An outcomes study. *Journal of clinical microbiology*, **38**, 2824-8.
320. Nyquist, A.C., Gonzales, R., Steiner, J.F. and Sande, M.A. (1998) Antibiotic prescribing for children with colds, upper respiratory tract infections, and bronchitis. *JAMA : The journal of the American Medical Association*, **279**, 875-7.
321. Mainous, A.G., Hueston, W.J. and Clark, J.R. (1996) Antibiotics and upper respiratory infection: Do some folks think there is a cure for the common cold. *The Journal of family practice*, **42**, 357-61.
322. Abed, Y. and Boivin, G. (2006) Treatment of respiratory virus infections. *Antiviral research*, **70**, 1-16, 10.1016/j.antiviral.2006.01.006.
323. Vega, R. (2005) Rapid viral testing in the evaluation of the febrile infant and child. *Current opinion in pediatrics*, **17**, 363-7.
324. Osiowy, C. (1998) Direct detection of respiratory syncytial virus, parainfluenza virus, and adenovirus in clinical respiratory specimens by a multiplex reverse transcription-PCR assay. *Journal of clinical microbiology*, **36**, 3149-54.
325. Bellau-Pujol, S., Vabret, A., Legrand, L., Dina, J., Gouarin, S., Petitjean-Lecherbonnier, J., Pozzetto, B., Ginevra, C. and Freymuth, F. (2005) Development of three multiplex RT-PCR assays for the detection of 12 respiratory RNA viruses. *Journal of virological methods*, **126**, 53-63, 10.1016/j.jviromet.2005.01.020.
326. Gruteke, P., Glas, A.S., Dierdorp, M., Vreede, W.B., Pilon, J.-W. and Bruisten, S.M. (2004) Practical implementation of a multiplex PCR for acute respiratory tract infections in children. *Journal of clinical microbiology*, **42**, 5596-603, 10.1128/JCM.42.12.5596-5603.2004.

327. Gröndahl, B., Puppe, W., Hoppe, A., Kühne, I., Weigl, J.A. and Schmitt, H.J. (1999) Rapid identification of nine microorganisms causing acute respiratory tract infections by single-tube multiplex reverse transcription-PCR: Feasibility study. *Journal of clinical microbiology*, **37**, 1-7.
328. Coiras, M.T., López-Huertas, M.R., López-Campos, G., Aguilar, J.C. and Pérez-Breña, P. (2005) Oligonucleotide array for simultaneous detection of respiratory viruses using a reverse-line blot hybridization assay. *Journal of medical virology*, **76**, 256-64, 10.1002/jmv.20350.
329. Lodes, M.J., Suciu, D., Elliott, M., Stover, A.G., Ross, M., Caraballo, M., Dix, K., Crye, J., Webby, R.J., Lyon, W.J., et al. (2006) Use of semiconductor-based oligonucleotide microarrays for influenza a virus subtype identification and sequencing. *Journal of clinical microbiology*, **44**, 1209-18, 10.1128/JCM.44.4.1209-1218.2006.
330. Kessler, N., Ferraris, O., Palmer, K., Marsh, W. and Steel, A. (2004) Use of the DNA flow-thru chip, a three-dimensional biochip, for typing and subtyping of influenza viruses. *Journal of clinical microbiology*, **42**, 2173-85.
331. Wang, D., Coscoy, L., Zylberberg, M., Avila, P.C., Boushey, H.A., Ganem, D. and DeRisi, J.L. (2002) Microarray-based detection and genotyping of viral pathogens. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 15687-92, 10.1073/pnas.242579699.
332. Mäkelä, M.J., Puhakka, T., Ruuskanen, O., Leinonen, M., Saikku, P., Kimpimäki, M., Blomqvist, S., Hyypiä, T. and Arstila, P. (1998) Viruses and bacteria in the etiology of the common cold. *Journal of clinical microbiology*, **36**, 539-42.
333. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947-8, 10.1093/bioinformatics/btm404.
334. Andronescu, M., Aguirre-Hernández, R., Condon, A. and Hoos, H.H. (2003) RNAsoft: A suite of RNA secondary structure prediction and design software tools. *Nucleic acids research*, **31**, 3416-22.
335. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W. (2010) GenBank. *Nucleic acids research*, **39**, D32-7, 10.1093/nar/gkq1079.
336. Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J. and Lipman, D. (2008) The influenza virus resource at the National Center for Biotechnology Information. *Journal of virology*, **82**, 596-601, 10.1128/JVI.02005-07.
337. Monto, A.S. (2002) Epidemiology of viral respiratory infections. *The American journal of medicine*, **112 Suppl**, S4-12.
338. Freymuth, F., Vabret, A., Cuvillon-Nimal, D., Simon, S., Dina, J., Legrand, L., Gouarin, S., Petitjean, J., Eckart, P. and Brouard, J. (2006) Comparison of multiplex PCR assays and conventional techniques for the diagnostic of respiratory virus infections in children admitted to hospital with an acute respiratory illness. *Journal of medical virology*, **78**, 1498-504, 10.1002/jmv.20725.
339. Kelly, H. and Birch, C. (2004) The causes and diagnosis of influenza-like illness. *Australian Family Physician*, **33**, 305-9.

340. Fleming, D.M., Pannell, R.S., Elliot, A.J. and Cross, K.W. (2005) Respiratory illness associated with influenza and respiratory syncytial virus infection. *Archives of disease in childhood*, **90**, 741-6, 10.1136/adc.2004.063461.
341. Greenough, A. (2002) Respiratory syncytial virus infection: Clinical features, management, and prophylaxis. *Current opinion in pulmonary medicine*, **8**, 214-7.
342. Smyth, R. and Openshaw, P. (2006) Bronchiolitis. *The Lancet*, **368**, 312-22, 10.1016/S0140-6736(06)69077-6.
343. Boivin, G., De Serres, G., Côté, S., Gilca, R., Abed, Y., Rochette, L., Bergeron, M.G. and Déry, P. (2003) Human metapneumovirus infections in hospitalized children. *Emerging infectious diseases*, **9**, 634-40.
344. Deffrasnes, C., Hamelin, M.-E. and Boivin, G. (2007) Human metapneumovirus. *Seminars in respiratory and critical care medicine*, **28**, 213-21, 10.1055/s-2007-976493.
345. Hamelin, M.-E., Abed, Y. and Boivin, G. (2004) Human metapneumovirus: A new player among respiratory viruses. *Clinical infectious diseases*, **38**, 983-90, 10.1086/382536.
346. Hamelin, M.-E. and Boivin, G. (2005) Human metapneumovirus: A ubiquitous and long-standing respiratory pathogen. *The Pediatric infectious disease journal*, **24**, S203-7.
347. Fouchier, R.A., Rimmelzwaan, G.F., Kuiken, T. and Osterhaus, A.D. (2005) Newer respiratory virus infections: Human metapneumovirus, avian influenza virus, and human coronaviruses. *Current opinion in infectious diseases*, **18**, 141-6.
348. Kahn, J.S. (2007) Newly discovered respiratory viruses: Significance and implications. *Current opinion in pharmacology*, **7**, 478-83, 10.1016/j.coph.2007.07.004.
349. Drews, A.L., Atmar, R.L., Glezen, W.P., Baxter, B.D., Piedra, P.A. and Greenberg, S.B. (1997) Dual respiratory virus infections. *Clinical infectious diseases*, **25**, 1421-9.
350. Storch, G.A. (2000) Diagnostic virology. *Clinical infectious diseases*, **31**, 739-51, 10.1086/314015.
351. Weinberg, G.A., Erdman, D.D., Edwards, K.M., Hall, C.B., Walker, F.J., Griffin, M.R. and Schwartz, B. (2004) Superiority of reverse-transcription polymerase chain reaction to conventional viral culture in the diagnosis of acute respiratory tract infections in children. *The Journal of infectious diseases*, **189**, 706-10, 10.1086/381456.
352. Kuypers, J., Wright, N., Ferrenberg, J., Huang, M.-L., Cent, A., Corey, L. and Morrow, R. (2006) Comparison of real-time PCR assays with fluorescent-antibody assays for diagnosis of respiratory virus infections in children. *Journal of clinical microbiology*, **44**, 2382-8, 10.1128/JCM.00216-06.
353. Kuroiwa, Y., Nagai, K., Okita, L., Ukae, S., Mori, T., Hotsubo, T. and Tsutsumi, H. (2004) Comparison of an immunochromatography test with multiplex reverse transcription-PCR for rapid diagnosis of respiratory syncytial virus infections. *Journal of clinical microbiology*, **42**, 4812-4, 10.1128/JCM.42.10.4812-4814.2004.
354. Vernet, G. (2004) Molecular diagnostics in virology. *Journal of clinical virology*, **31**, 239-47, 10.1016/j.jcv.2004.06.003.

355. Perkins, S.M., Webb, D.L., Torrance, S.A., El Saleeby, C., Harrison, L.M., Aitken, J.A., Patel, A. and DeVincenzo, J.P. (2005) Comparison of a real-time reverse transcriptase PCR assay and a culture technique for quantitative assessment of viral load in children naturally infected with respiratory syncytial virus. *Journal of clinical microbiology*, **43**, 2356-62, 10.1128/JCM.43.5.2356-2362.2005.
356. Pol, A.C. van de, Loon, A.M. van, Wolfs, T.F.W., Jansen, N.J.G., Nijhuis, M., Breteler, E.K., Schuurman, R. and Rossen, J.W.A. (2007) Increased detection of respiratory syncytial virus, influenza viruses, parainfluenza viruses, and adenoviruses with real-time PCR in samples from patients with respiratory symptoms. *Journal of clinical microbiology*, **45**, 2260-2, 10.1128/JCM.00848-07.
357. Mahony, J., Chong, S., Merante, F., Yaghoubian, S., Sinha, T., Lisle, C. and Janeczko, R. (2007) Development of a respiratory virus panel test for detection of twenty human respiratory viruses by use of multiplex PCR and a fluid microbead-based assay. *Journal of clinical microbiology*, **45**, 2965-70, 10.1128/JCM.02436-06.
358. Nolte, F.S., Marshall, D.J., Rasberry, C., Schievelbein, S., Banks, G.G., Storch, G.A., Arens, M.Q., Buller, R.S. and Prudent, J.R. (2007) MultiCode-PLx system for multiplexed detection of seventeen respiratory viruses. *Journal of clinical microbiology*, **45**, 2779-86, 10.1128/JCM.00669-07.
359. Holland, C.A. and Kiechle, F.L. (2005) Point-of-care molecular diagnostic systems — past, present and future. *Current Opinion in Microbiology*, **5**, 504-9, 10.1016/j.mib.2005.08.001.
360. Boivin, G., Baz, M., Côté, S., Gilca, R., Deffrasnes, C., Leblanc, E., Bergeron, M.G., Déry, P. and De Serres, G. (2005) Infections by human coronavirus-NL in hospitalized children. *The Pediatric infectious disease journal*, **24**, 1045-8.
361. Smith, G.J.D., Vijaykrishna, D., Bahl, J., Lycett, S.J., Worobey, M., Pybus, O.G., Ma, S.K., Cheung, C.L., Raghwani, J., Bhatt, S., et al. (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*, **459**, 1122-5, 10.1038/nature08182.
362. Center for disease control (2009) Outbreak of swine-origin influenza A (H1N1) virus infection - Mexico, March-April 2009. *Morbidity and mortality weekly report*, **58**, 467-70.
363. Center for disease control (2009) Update: infections with a swine-origin influenza A (H1N1) virus - United States and other countries, April 28, 2009. *Morbidity and mortality weekly report*, **58**, 431-3.
364. Dawood, F.S., Jain, S., Finelli, L., Shaw, M.W., Lindstrom, S., Garten, R.J., Gubareva, L.V., Xu, X., Bridges, C.B. and Uyeki, T.M. (2009) Emergence of a novel swine-origin influenza A (H1N1) virus in humans. *The New England journal of medicine*, **360**, 2605-15, 10.1056/NEJMoa0903810.
365. Center for disease control (2009) Serum cross-reactive antibody response to a novel influenza A (H1N1) virus after vaccination with seasonal influenza vaccine. *Morbidity and mortality weekly report*, **58**, 521-4.
366. Fouchier, R.A., Bestebroer, T.M., Herfst, S., Van Der Kemp, L., Rimmelzwaan, G.F. and Osterhaus, A.D. (2000) Detection of influenza A viruses from different species by PCR amplification of conserved sequences in the matrix gene. *Journal of clinical microbiology*, **38**, 4096-101.

367. LeBlanc, J.J., Li, Y., Bastien, N., Forward, K.R., Davidson, R.J. and Hatchette, T.F. (2009) Switching gears for an influenza pandemic: Validation of a duplex reverse transcriptase PCR assay for simultaneous detection and confirmatory identification of pandemic (H1N1) 2009 influenza virus. *Journal of clinical microbiology*, **47**, 3805-13, 10.1128/JCM.01344-09.
368. Boivin, G., Côté, S., Déry, P., De Serres, G. and Bergeron, M.G. (2004) Multiplex real-time PCR assay for detection of influenza and human respiratory syncytial viruses. *Journal of clinical microbiology*, **42**, 45-51.
369. Quackenbush, J. (2002) Microarray data normalization and transformation. *Nature genetics*, **32**, 496-501, 10.1038/ng1032.
370. Bammler, T., Beyer, R.P., Bhattacharya, S., Boorman, G.A., Boyles, A., Bradford, B.U., Bumgarner, R.E., Bushel, P.R., Chaturvedi, K., Choi, D., et al. (2005) Standardizing global gene expression analysis between laboratories and across platforms. *Nature methods*, **2**, 351-6, 10.1038/nmeth754.
371. Beverley, S.M. (1991) Gene amplification in *Leishmania*. *Annu Rev Microbiol*, **45**, 417-444, 10.1146/annurev.mi.45.100191.002221.
372. Haimeur, A. and Ouellette, M. (1998) Gene amplification in *Leishmania tarentolae* selected for resistance to sodium stibogluconate. *Antimicrobial agents and chemotherapy*, **42**, 1689-94.
373. Olmo, A., Arrebola, R., Bernier, V., González-Pacanowska, D. and Ruiz-Pérez, L.M. (1995) Co-existence of circular and multiple linear amplicons in methotrexate-resistant *Leishmania*. *Nucleic acids research*, **23**, 2856-64.
374. Grondin, K., Roy, G. and Ouellette, M. (1996) Formation of extrachromosomal circular amplicons with direct or inverted duplications in drug-resistant *Leishmania tarentolae*. *Molecular and cellular biology*, **16**, 3587-95.
375. Beverley, S.M. and Coburn, C.M. (1990) Recurrent de novo appearance of small linear DNAs in *Leishmania major* and relationship to extra-chromosomal DNAs in other species. *Molecular and Biochemical Parasitology*, **42**, 133-41.
376. Papadopoulou, B., Roy, G. and Ouellette, M. (1992) A novel antifolate resistance gene on the amplified H circle of *Leishmania*. *EMBO Journal*, **11**, 3601-8.
377. Grondin, K., Kündig, C., Roy, G. and Ouellette, M. (1998) Linear amplicons as precursors of amplified circles in methotrexate-resistant *Leishmania tarentolae*. *Nucleic acids research*, **26**, 3372-8.
378. Riet, B., Dumas, C., Papadopoulou, B., Luenen, H.G.A.M.V. and Borst, P. (2005) Formation of linear inverted repeat amplicons following targeting of an essential gene in Leishmania. *Cloning*, **33**, 1699-709, 10.1093/nar/gki304.
379. Yakovich, A.J., Ragone, F.L., Alfonzo, J.D., Sackett, D.L. and Werbovetz, K.A. (2006) Leishmania tarentolae: Purification and characterization of tubulin and its suitability for antileishmanial drug screening. *Experimental parasitology*, **114**, 289-96, 10.1016/j.exppara.2006.04.008.
380. Kazemi, B., Tahvildar-Bideroni, G., Hashemi Feshareki, S. and Javadian, E. (2004) Isolation a Lizard *Leishmania* promastigote from its Natural Host in Iran. *Journal of Biological Sciences*, **4**, 620-3.
381. Zimmerman, L.M., Vogel, L.A. and Bowden, R.M. (2010) Understanding the vertebrate immune system: Insights from the reptilian perspective. *The Journal of experimental biology*, **213**, 661-71, 10.1242/jeb.038315.

382. Zilberstein, D. and Shapira, M. (1994) The role of pH and temperature in the development of *Leishmania* parasites. *Annual review of microbiology*, **48**, 449-70, 10.1146/annurev.mi.48.100194.002313.
383. Dobson, D.E., Kamhawi, S., Lawyer, P., Turco, S.J., Beverley, S.M. and Sacks, D.L. (2010) *Leishmania major* survival in selective *Phlebotomus papatasi* sand fly vector requires a specific SCG-encoded lipophosphoglycan galactosylation pattern. *PLoS pathogens*, **6**, e1001185, 10.1371/journal.ppat.1001185.
384. Myskova, J., Svobodova, M., Beverley, S.M. and Volf, P. (2007) A lipophosphoglycan-independent development of *Leishmania* in permissive sand flies. *Microbes and infection / Institut Pasteur*, **9**, 317-24, 10.1016/j.micinf.2006.12.010.
385. Denny, P.W., Morgan, G.W., Field, M.C. and Smith, D.F. (2005) *Leishmania major:* Clathrin and adaptin complexes of an intra-cellular parasite. *Experimental parasitology*, **109**, 33-7, 10.1016/j.exppara.2004.10.007.
386. Ampofo, K., Bender, J., Sheng, X., Korgenski, K., Daly, J., Pavia, A.T. and Byington, C.L. (2008) Seasonal invasive pneumococcal disease in children: Role of preceding respiratory viral infection. *Pediatrics*, **122**, 229-37, 10.1542/peds.2007-3192.
387. Bourgeois, F.T., Valim, C., McAdam, A.J. and Mandl, K.D. (2009) Relative impact of influenza and respiratory syncytial virus in young children. *Pediatrics*, **124**, e1072-80, 10.1542/peds.2008-3074.
388. Franz, A., Adams, O., Willems, R., Bonzel, L., Neuhausen, N., Schweizer-Krantz, S., Ruggeberg, J.U., Willers, R., Henrich, B., Schroten, H., et al. (2010) Correlation of viral load of respiratory pathogens and co-infections with disease severity in children hospitalized for lower respiratory tract infection. *Journal of clinical virology*, **48**, 239-45, 10.1016/j.jcv.2010.05.007.

Annexe 1. Liste des études utilisant la biopuce *Leishmania*

Les articles suivants ont utilisé la biopuce *Leishmania* de première génération :

- Leprohon P., D. Légaré, **F. Raymond**, E. Madore, G. Hardiman, J. Corbeil et M. Ouellette. (2009) « Gene expression modulation is associated with gene amplification, supernumerary chromosomes and chromosome loss in antimony-resistant *Leishmania infantum* ». *Nucleic Acids Research.* 37:1387-99.

- Rochette A., **F. Raymond**, J. Corbeil, M. Ouellette et B. Papadopoulou. (2009) « Whole-genome comparative RNA expression profiling of axenic and intracellular amastigote forms of *Leishmania infantum* ». *Molecular and Biochemical Parasitology.* 165:32-47.

- Rochette A., **F. Raymond**, J. M. Ubeda, M. Smith, N. Messier, S. Boisvert, P. Rigault, J. Corbeil, M. Ouellette et B. Papadopoulou. 2008. « Genome-wide gene expression profiling analysis of *Leishmania major* and *Leishmania infantum* developmental stages reveals substantial differences between the two species ». *BMC Genomics*. 9:255.

- Ubeda J. M., D. Légaré, **F. Raymond**, A. A. Ouameur, S. Boisvert, P. Rigault, J. Corbeil, M. J. Tremblay, M. Olivier, B. Papadopoulou et M. Ouellette. 2008. « Modulation of gene expression in drug resistant *Leishmania* is associated with gene amplification, gene deletion and chromosome aneuploidy ». *Genome Biology.* 9:R115.

Les articles suivants ont utilisé la biopuce *Leishmania* de deuxième génération :

- Do Monte Neto R. L., A. C. Coelho, **F. Raymond**, M. N. Melo, F. J. G. Frézard, M. Ouellette. Gene expression profilinf and molecular characterization of antimony resistance in *Leishmania amazonensis. PLoS Neglected Tropical Diseases.* **5**:e1167. doi:10.1371/journal.pntd.0001167.

- **Raymond F.,** S. Boisvert, G. Roy, J.-F. Ritt, D. Légaré, M. Stanke, M. Olivier, M. J. Tremblay, B. Papadopoulou, M. Ouellette, J. Corbeil. Genome analysis of the lizard parasite *Leishmania tarentolae* reveals expansion of insect-specific and loss of mammalian-specific genes in comparison to human pathogenic *Leishmania* species. En préparation.

Annexe 2. Programme Perl pour la création de fichiers SpotType

```
#!/usr/bin/perl

use strict;
use warnings;

### Script par Frédéric Raymond


### Input : Probe list ; Gene list ; blastfile -m 8
### Output : fichier spottype pour l'organisme blaste pour utiliser dans R

### Read probe list - uses list file (one id per line) not fasta file.

my %probelist;

open(PROBELIST, shift)|| die ("Could not read probe list file - Use list of ID not fasta");
while(<PROBELIST>){
	chomp $_;
	my @a;
	$a[0] = "Empty";
	$probelist{$_} = \@a;
}
close(PROBELIST);

my %genelist;
open(GENELIST, shift)|| die ("Could not read Genelist file - should be
id\tbegin\t\end\tDescription");
while(<GENELIST>){
	chomp $_;
	my @genedescription = split(/\t/, $_);
	$genelist{$genedescription[0]}=
	"$genedescription[0]\t$genedescription[1]\t$genedescription[2]\t\"$genedescription
[3]\"";
}
close(GENELIST);

# restrict the number of digits after the decimal point
sub restrict_num_decimal_digits {
my $num=shift;
my $digs_to_cut=shift;
if ($num=~/\d+\.(\d){$digs_to_cut,}/) {
	$num=sprintf("%.".($digs_to_cut-1)."f", $num);
}
```

```
return $num;
}

open(BLAST, shift) || die ("Could not open blast formated file");
while (<BLAST>){
	chomp $_;
	my $call;
	my @blast = split(/\t/, $_);
	my $score = &restrict_num_decimal_digits(($blast[2]*$blast[3]/100),1);
	if($score==60){
		$call = "Perfect";
	}
	if(($score>=58)&&($score<=59)){
		$call = "1-2";
	}
	if(($score>=53)&&($score<=57)){
		$call = "3-7";
	}
	if(($score>=1)&&($score<=52)){
		$call = "badhit";
	}
	my $toprint = "$_\t$score\t$call";
	if($probelist{$blast[0]} eq "Empty"){
		my @a;
		$a[0] = $toprint;
		$probelist{$blast[0]} = \@a;
	} else {
		push @{$probelist{$blast[0]}}, "$toprint";
	}
}

print "SpotType\tGeneName\tColor\tscore\tBestHit\tBegin\tEnd\tID\tPerfect\t1-2\t3-7\n";
print "ControlAgilent\t*\tblack\t0\tNone\t0\t0\t\"Control\"\t0\t0\t0\n";

foreach my $probe(keys %probelist){
	my $current = 0;
	my $finalcall = "NoMatch";
	my $genehit = "None\t0\t0\t\"None\"";
	my $perfect = "None";
	my $undeux = "None";
	my $troissept = "None";
	foreach my $entry(@{$probelist{$probe}}){
		if($entry ne "Empty"){
			my @currententry = split(/\t/, $entry);
			if($currententry[12]>$current){
				$finalcall = $currententry[13];
				$current = $currententry[12];
```

```
                        if(exists($genelist{$currententry[1]})){
                            $genehit = $genelist{$currententry[1]};
                        }
                    }
                }
                if($entry ne "Empty"){
                    my @currententry = split(/\t/, $entry);
                    if($currententry[13] eq "Perfect"){
                        if ($perfect =~ /$currententry[1]/){} else {
                            if ($perfect eq "None"){
                                $perfect = "$currententry[1]";
                            } else {
                                $perfect = "$perfect;$currententry[1]";
                            }
                        }
                    }
                    if($currententry[13] eq "1-2"){
                        if ($undeux =~ /$currententry[1]/){} else {
                            if ($undeux eq "None"){
                                $undeux= "$currententry[1]";
                             } else {
                                $undeux= "$undeux;$currententry[1]";
                            }
                        }
                    }
                    if($currententry[13] eq "3-7"){
                        if ($troissept =~ /$currententry[1]/){} else {
                            if($troissept eq "None"){
                                $troissept = "$currententry[1]";
                            } else {
                                $troissept = "$troissept;$currententry[1]";
                            }
                        }
                    }
                }
            }
            print
"$finalcall\t$probe\tblack\t$current\t$genehit\t$perfect\t$undeux\t$troissept\n";
}
```

Annexe 3. Protocole d'analyse R de la biopuce *Leishmania* de première génération

Programme par Frédéric Raymond
frederic.raymond (arobas) crchul.ulaval.ca
Centre de recherche en infectiologie
2008-05-21

 # Chargement des librairies R utilisées pour l'analyse

```
library(limma)
library(marray)
library(convert)
```

 # Sélection du répertoire de travail

```
setwd("C:/Répertoire de Travail")
```

Description des paramètres utilisés lors de l'analyse

 # Biopuce de référence
```
parameters.reference="A"
```

 # Logiciel de quantification des images
```
parameters.read.maimages.source="genepix"
```

 # Liste des types de sondes décrites dans le fichier spottype
```
parameters.modifyWeights.w_keys = c("excluded", "BLANK", "Common", "Common-1to2", "LinJ-3to7", "LinJ-none", "LmaJ-3to7", "LmaJ-none")
```

 # Pondération de chacun des types de sondes
```
parameters.modifyWeights.w_values = c(0, 0, 1, 1, 0, 0, 1, 1)
```

 # Multiplicateur de pondération à appliquer aux points marqués comme mauvais par le logiciel de quantification des biopuces
```
parameters.read.maimages.wt.fun=wtflags(0.0)
```

 # Nom du fichier targets
```
parameters.readTargets.targets_file="targets.txt"
```

 # Nom du fichier GAL
```
parameters.readGAL.gal_file="control-corrected_060210_corbeil_gal.gal"
```

 # Nom du fichier spottypes
```
parameters.readSpotTypes.spottypes_file="spottypes.txt"
```

```
# Méthode de correction du bruit de fond
parameters.backgroundCorrect.method="normexp"

# Méthode de normalisation interne de chaque biopuce et paramètres
parameters.normalizeWithinArrays.method="printtiploess"
parameters.normalizeWithinArrays.span=0.3
parameters.normalizeWithinArrays.iterations=4

# Méthode de normalisation globale de l'expérience
parameters.normalizeBetweenArrays.method="scale"

# Code d'identification de l'expérience
parameters.normalization.unique_name="M123"

# Paramètres de création de fichier de résultats
parameters.quote=FALSE
parameters.dec="."

## Paramètres du test statistique
parameters.decideTests.method="global"
parameters.decideTests.ajust.method="fdr"
parameters.decideTests.p=0.05
parameters.decideTests.cutoff <- c(0.75)

# Fonctions leish_get_gpr_columns utilisées avec la permission de Sebastien Boisvert
# Début de leish_get_gpr_columns

leish_debugger <-local( function(message, verbosity = FALSE) {
if ( verbosity == TRUE) {
print (message)
}
})

leish_get_columns_name <- local(function (first_file) {
leish_debugger("----> leish_get_columns_name")
conn=file(first_file,"r")
content <- readLines ( conn ,n=2)
line_with_amount_of_header_lines = content[2]
columns=strsplit(line_with_amount_of_header_lines,split="\t",
fixed=TRUE)
nb_lines_in_header =as.numeric(columns[[1]][1])
readLines(conn, n= nb_lines_in_header)
line_with_columns = readLines (conn,n=1)
while (length( grep ("Block", line_with_columns)) == 0) {
line_with_columns = readLines (conn,n=1)
}
close(conn)
```

```
columns <- strsplit(line_with_columns, split="\t")
columns_to_process <- columns [[1]]
columns_to_return <- c()
for (a_column in columns_to_process)    {
a_clean_column = gsub("\"", "", a_column)
columns_to_return <- cbind(c(a_clean_column), columns_to_return)
}
columns_to_return
})

leish_get_columns_using_a_pattern <- local(function ( columns_name, pattern ) {
leish_debugger ("------> leish_get_columns_using_a_pattern")
columns <- c()
leish_debugger("BEGIN;")
leish_debugger(columns_name)
leish_debugger("COMMIT;")
for ( column_name in columns_name ) {
leish_debugger(paste("COLUMN : ", column_name))
leish_debugger (paste("Pattern : ", pattern))
if (length( grep ( pattern, column_name)) > 0) {
columns <- cbind (columns, c(column_name))
}
}
columns
})

leish_get_mean_columns <- local( function ( columns_name ) {
leish_debugger ("--------> leish_get_mean_columns")
leish_get_columns_using_a_pattern (columns_name, "^F.+Mean$")
})
leish_get_median_columns <-local( function ( columns_name ) {
leish_debugger ("-----------> leish_get_median_columns")
leish_get_columns_using_a_pattern (columns_name, "^B.+Median$")
})
leish_get_wave_lengths <- local(function (mean_columns) {
leish_debugger ("-----------> leish_get_wave_lengths")
wave_lengths <- c()
leish_debugger (paste("nb_means : ", length(mean_columns)))
for (mean_column in mean_columns ){
wave_length <- gsub ( "[^0-9]","",mean_column)
leish_debugger ("wave length  : ")
leish_debugger (wave_length)
wave_lengths <- cbind (wave_lengths, c(wave_length))
}
leish_debugger ("wave_lengths : ")
leish_debugger (wave_lengths)
wave_lengths
```

```
})
leish_get_columns_list <- local(function (mean_columns, median_columns,
red_wave_length, green_wave_length){
leish_debugger("----------------> leish_get_columns_list")
red_mean <- NULL # "F647 Mean"
green_mean <- NULL # "F555 Mean"
red_median <- NULL # "B647 Median"
green_median <- NULL # "B555 Median"
leish_debugger (paste("red pattern : ", red_wave_length))
leish_debugger (paste("green pattern : " , green_wave_length))
for( mean_column in mean_columns){
if (length(grep (red_wave_length, mean_column))>0){
red_mean = mean_column
} else if (length(grep (green_wave_length, mean_column))>0) {
green_mean = mean_column
}
}
for (median_column in median_columns) {
if (length(grep (red_wave_length, median_column))>0) {
red_median = median_column
} else if (length(grep ( green_wave_length, median_column))>0) {
green_median = median_column
}
}
list(R= red_mean, G= green_mean,
Rb=red_median, Gb= green_median)
})

leish_get_gpr_columns<- local( function(targets){
leish_debugger("---------> leish_get_gpr_columns")
first_file <- targets$FileName[1]
columns_name <- leish_get_columns_name(first_file)
leish_debugger ("column names ")
leish_debugger (columns_name)
mean_columns <- leish_get_mean_columns (columns_name)
median_columns <- leish_get_median_columns (columns_name)
leish_debugger("mean columns : ")
leish_debugger (mean_columns)
wave_lengths <- leish_get_wave_lengths (mean_columns)
leish_debugger("wave_lengths : " )
leish_debugger ( wave_lengths)
red_wave_length <- NULL
green_wave_length <- NULL
if (wave_lengths[1] < wave_lengths[2]){
red_wave_length <- wave_lengths[2]
green_wave_length <- wave_lengths[1]
}else if (wave_lengths[2] < wave_lengths[1]){
```

```
red_wave_length <- wave_lengths[1]
green_wave_length <- wave_lengths[2]
}
leish_get_columns_list (mean_columns, median_columns, red_wave_length,
green_wave_length)
})

# Fin de leish_get_gpr_columns

# Lecture des fichiers de quantification de biopuces
targets=readTargets(parameters.readTargets.targets_file)
RG<-NULL
gpr_columns = leish_get_gpr_columns (targets)
RG <- read.maimages(targets, columns=gpr_columns, annotation =
c("Block","Row","Column","ID","Name"), source=parameters.read.maimages.source,
wt.fun=parameters.read.maimages.wt.fun)

# Lecture du fichier GAL
RG$genes<-readGAL(parameters.readGAL.gal_file)
RG$printer<-getLayout(RG$genes)

#Lecture du fichier spottypes et calcul de la pondération initiale de chaque sonde
spottypes<-readSpotTypes(parameters.readSpotTypes.spottypes_file)
RG$genes$Status<-controlStatus(spottypes,RG)
w<-modifyWeights(RG$weights, RG$genes$Status,
parameters.modifyWeights.w_keys, parameters.modifyWeights.w_values)
p_weights=w

# Correction du bruit de fond
RGb<-backgroundCorrect(RG,method=parameters.backgroundCorrect.method,
offset=50)

# Normalisation interne de chaque biopuce
MA<-normalizeWithinArrays(RGb,
method=parameters.normalizeWithinArrays.method,weights=p_weights, span=
parameters.normalizeWithinArrays.span,
iterations=parameters.normalizeWithinArrays.iterations)

# Normalisation globale de l'expérience
MA.between<-normalizeBetweenArrays(MA,
method=parameters.normalizeBetweenArrays.method)

# Création du modèle de design expérimental
reference = parameters.reference #condition de reference.
design <- modelMatrix(targets, ref = reference)

# Calcul de la pondération de chaque biopuce et modification de la pondération des sondes
```

```
array.wts <- arrayWeights(MA, design, weights=NULL)
arraymatrix.wts <- matrix(rep(array.wts, each=dim(MA.between)[1]),
dim(MA.between)[1], dim(MA.between)[2])
w.array=arraymatrix.wts*w
weightArray <- cbind(RG$targets, array.wts)

# Calculs préalables au test statistique (analyse des duplicats et modélisation linéaire
bayésienne)
corfit <- duplicateCorrelation(MA.between, design, ndups =2, spacing =200)
fit <- lmFit(MA.between, design, correlation=corfit$consensus, ndups =2, spacing =200,
weights = w.array)
fitbayes <- eBayes(fit)

# Test statistique
results <- decideTests(fitbayes,method=parameters.decideTests.method,adjust.method =
parameters.decideTests.ajust.method, p.value=parameters.decideTests.p,
lfc=parameters.decideTests.cutoff)

# Génération de diagrammes de Venn
par(mfrow=c(3,1))
vennDiagram(results, main = "Venn diagram of modulated genes")
vennDiagram(results, include ="up", main = "Venn diagram of upregulated genes")
vennDiagram(results, include ="down", main = "Venn diagram of downregulated genes")


# Sauvegarde d'un fichier de résultats
write.fit(fitbayes, results=results, file="results.txt", digits=12,
adjust=parameters.decideTests.ajust.method, sep="\t")
```

Annexe 4. Protocole d'analyse R de la biopuce *Leishmania* de seconde génération

```
### Microarray analysis script for R
### Created by Frederic Raymond
### Centre de recherche en infectiologie de l'Université Laval
### frederic.raymond (arobas) crchul.ulaval.ca
### May 20, 2009

# Chargement des librairies R utilisées pour l'analyse
library(limma)
library(convert)

# Sélection du répertoire de travail
setwd("C:/Répertoire de Travail")

# Biopuce de référence
parameters.reference="A"

# Logiciel de quantification des images
parameters.read.maimages.source="agilent"

# Liste des types de sondes décrites dans le fichier spottype
parameters.modifyWeights.w_keys = c("Perfect", "1-2", "3-7", "badhit", "NoMatch",
"ControlAgilent")

# Pondération de chacun des types de sondes
parameters.modifyWeights.w_values = c(1, 1, 0.5, 0, 0, 0)

# Multiplicateur de pondération à appliquer aux points marqués comme mauvais par le
logiciel de quantification des biopuces
parameters.read.maimages.wt.fun=wtflags(0.0)

# Nom du fichier targets
parameters.readTargets.targets_file="targets.txt"

# Nom du fichier GAL
parameters.readGAL.gal_file="control-corrected_060210_corbeil_gal.gal"

# Nom du fichier spottypes
parameters.readSpotTypes.spottypes_file="LmjFv52-withnames.spottypes"

# Méthode de correction du bruit de fond
parameters.backgroundCorrect.method="edwards"

# Méthode de normalisation interne de chaque biopuce et paramètres
```

```
parameters.normalizeWithinArrays.method="loess"
parameters.normalizeWithinArrays.span=0.3
parameters.normalizeWithinArrays.iterations=4

# Méthode de normalisation globale de l'expérience
parameters.normalizeBetweenArrays.method="Aquantile"

# Code d'identification de l'expérience
parameters.normalization.unique_name="ID"

## Paramètres du test statistique
parameters.decideTests.method="global"
parameters.decideTests.ajust.method="fdr"
parameters.decideTests.p=0.05
parameters.decideTests.cutoff <- c(0.75)

# Lecture des fichiers de quantification de biopuces
targets <- readTargets(file=parameters.readTargets.targets_file)
RG <- read.maimages(targets, source=parameters.read.maimages.source,
other.columns=c("gIsBGNonUnifOL",      "gIsBGPopnOL",      "gIsFeatNonUnifOL",
        "gIsFeatPopnOL",    "gIsFound",    "gIsInNegCtrlRange",
        "gIsLowPMTScaledUp",      "gIsPosAndSignif",    "gIsSaturated",
        "gIsUsedInMD",      "gIsWellAboveBG",   "rIsBGNonUnifOL",  "rIsBGPopnOL",
        "rIsFeatNonUnifOL", "rIsFeatPopnOL",    "rIsFound",    "rIsInNegCtrlRange",
        "rIsLowPMTScaledUp",      "rIsPosAndSignif",    "rIsSaturated",
        "rIsUsedInMD",      "rIsWellAboveBG"))

#Lecture du fichier spottypes et calcul de la pondération initiale de chaque sonde
spottypes<-readSpotTypes(parameters.readSpotTypes.spottypes_file)
RG$genes$Status<-controlStatus(spottypes,RG)
RG$weights <- matrix(data = c(1), nrow=length(RG$R[,1]), ncol=length(RG$R[1,]))
w<-modifyWeights(RG$weights, RG$genes$Status, parameters.modifyWeights.w_keys,
parameters.modifyWeights.w_values)
colnames(w)=colnames(RG$G)

# Correction du bruit de fond
RGb <- backgroundCorrect(RG, method = parameters.backgroundCorrect.method)

# Normalisation interne de chaque biopuce
MA <- normalizeWithinArrays(RGb, method = parameters.normalizeWithinArrays.method,
span = parameters.normalizeWithinArrays.span , iterations =
parameters.normalizeWithinArrays.iterations, weights = w)

# Normalisation globale de l'expérience
MA.between <- normalizeBetweenArrays(MA, method =
parameters.normalizeBetweenArrays.method)
```

```
# Calcul de la pondération de chaque biopuce et modification de la pondération des sondes
array.wts <- arrayWeights(MA, weights=NULL)

# Création du modèle de design expérimental
targets2 <- targetsA2C(targets)
u <- unique(targets2$Target)
f <- factor(targets2$Target, levels=u)
design <- model.matrix(~0+f)
colnames(design) <- u

# Calculs préalables au test statistique (analyse des duplicats et modélisation linéaire
bayésienne)
corfit <- intraspotCorrelation(MA.between, design)
design2 <- modelMatrix(targets, ref = parameters.reference)
fit <- lmFit(MA.between, design2, correlation=corfit$consensus, weights = w*array.wts)
fitbayes <- eBayes(fit)

# Test statistique
results <- decideTests(fitbayes,method=parameters.decideTests.method,adjust.method =
parameters.decideTests.ajust.method, p.value=parameters.decideTests.p,
lfc=parameters.decideTests.cutoff)

# Génération de diagrammes de Venn
par(mfrow=c(3,1))
vennDiagram(results, main = "Venn diagram of modulated genes")
vennDiagram(results, include ="up", main = "Venn diagram of upregulated genes")
vennDiagram(results, include ="down", main = "Venn diagram of downregulated genes")

# Sauvegarde d'un fichier de résultats
fitbayes$genes$GeneName <- spottypes$BestHit[pmatch(fitbayes$genes$GeneName,
spottypes$GeneName)]
fitbayes$genes$SystematicName <- spottypes$BestHit[pmatch(fitbayes$genes$ProbeName,
spottypes$GeneName)]
fitbayes$genes$Description <- spottypes$ID[pmatch(fitbayes$genes$ProbeName,
spottypes$GeneName)]
write.fit(fitbayes, results=results, file="results.txt", digits=12,
adjust=parameters.decideTests.ajust.method, sep="\t")
```