CARACTÉRISATION INTÉGRATIVE ET DÉVELOPPEMENT D'OUTILS MOLÉCULAIRES CHEZ LA BACTÉRIE *MESOPLASMA FLORUM*

par

Dominick Matteau

Thèse présentée au Département de biologie en vue
de l'obtention du grade de docteur ès sciences (Ph.D.)

FACULTÉ DES SCIENCES
UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, 19 mai 2020

Le 19 mai 2020

Le jury a accepté la thèse de Monsieur Dominick Matteau dans sa version finale.

Membres du jury

Professeur Sébastien Rodrigue
Directeur de recherche
Département de biologie, Université de Sherbrooke

Professeur Pierre-Étienne Jacques
Codirecteur de recherche
Département de biologie, Université de Sherbrooke

Docteur Pascal Sirand-Pugnet
Évaluateur externe
UMR Biologie du Fruit et Pathologie, INRAE, Université de Bordeaux

Professeur Jean-Philippe Côté
Évaluateur interne
Département de biologie, Université de Sherbrooke

Professeur Nicolas Gévry
Président-rapporteur
Département de biologie, Université de Sherbrooke

Simplicity is the ultimate sophistication.

-Leonardo da Vinci

Despite decades of scientific progress, we still are missing many -perhaps most- of the critical design principles and details necessary to engineer a cell.

-Thomas F. Knight

SOMMAIRE

L'émergence de la biologie synthétique marque l'entrée dans une nouvelle ère où il sera possible de modifier et reprogrammer des génomes entiers afin de répondre à des besoins spécifiques. Ce domaine de recherche est par conséquent appelé à jouer un rôle de premier plan dans le développement de nouvelles technologies visant à s'attaquer à certains des plus grands défis du 21e siècle tels que la multirésistance aux antibiotiques, la production d'énergies renouvelables et le traitement de maladies comme le cancer ou le diabète. Notre habileté actuelle à programmer des comportements cellulaires prévisibles est cependant très limitée, principalement parce que les organismes modèles couramment utilisés possèdent une complexité qui dépasse nos capacités d'analyse et que les règles fondamentales qui gouvernent le fonctionnement global des cellules demeurent encore mal comprises.

En raison de leurs génomes remarquablement petits, les bactéries appartenant à la classe des Mollicutes représentent des candidats particulièrement intéressants afin de décortiquer le fonctionnement intégral de cellules via les approches intégratives de la biologie des systèmes et de la génomique synthétique. La majorité de ces microorganismes sont toutefois caractérisés par un style de vie parasitaire, des capacités métaboliques réduites et une croissance relativement lente nécessitant l'utilisation de milieux de culture complexes. Conjointement au manque d'outils génétiques efficaces, ces caractéristiques restreignent considérablement leur manipulation en laboratoire. Certains Mollicutes se démarquent néanmoins en tant qu'organismes modèles pour l'avancement de la biologie synthétique et de la biologie des systèmes. C'est le cas pour *Mesoplasma florum*, une bactérie étroitement apparentée aux mycoplasmes du groupe de *Mycoplasma mycoides* (*mycoides cluster*). Contrairement à la plupart des mycoplasmes, *M. florum* ne possède aucun pouvoir pathogène connu et croît rapidement en conditions de laboratoire. De plus, *M. florum* possède un génome comprenant seulement 793 224 paires de bases et 685 séquences codantes pour des protéines, ce qui positionne cette bactérie parmi les organismes à réplication autonome les plus simples connus à ce jour.

Malgré ces avantages considérables, seulement quelques études avaient jusqu'à tout récemment spécifiquement exploré la biologie de *M. florum*, et ce même si sa découverte remonte à près de 40 ans. Ainsi, lors du commencement de mon doctorat, plusieurs aspects importants concernant ce microorganisme demeuraient toujours à définir. Par exemple, pratiquement aucune donnée quantitative sur la physiologie de cette bactérie était à ce moment-là disponible dans la littérature, et aucune étude sur l'expression de ses gènes n'avait encore été entreprise. De plus, très peu voire même aucun outil moléculaire n'était disponible afin de modifier le génome de *M. florum*, ce qui constituait une limitation technique importante à l'étude de la biologie de cet organisme, en plus de restreindre son utilisation en tant que châssis cellulaire pour l'ingénierie microbienne et le développement d'applications biotechnologiques.

Face à cette problématique, j'ai tout d'abord développé un système de culture en continu flexible et peu dispendieux permettant de faire croître *M. florum* dans des conditions contrôlées, stables et hautement reproductibles. Cet appareil offre plusieurs modes de fonctionnement pour accommoder les différents besoins rencontrés en laboratoire, et nous avons rendu les détails de sa conception entièrement disponibles pour l'ensemble de la communauté scientifique. En diminuant les fluctuations physiologiques des cellules, ce système de culture permet de réduire les variations expérimentales lors de l'étude de *M. florum*, et ainsi de générer des données plus facilement interprétables et comparables entre expériences.

J'ai ensuite développé les tout premiers plasmides spécifiquement conçus pour se répliquer chez *M. florum.* Basés sur l'origine de réplication du chromosome, ces plasmides ont permis de tester la fonctionnalité de différents marqueurs de sélection aux antibiotiques, en plus de mettre au point différentes méthodes de transformation pour cette bactérie. Grâce à leur tendance naturelle à recombiner avec le chromosome, ces plasmides ont d'ailleurs servi de fondement à la technique développée par notre laboratoire afin de cloner le génome complet de *M. florum* dans la levure. Cette souche de levure peut maintenant servir de plateforme afin de modifier efficacement le génome de *M. florum* et ensuite le transplanter dans une cellule réceptrice.

Finalement, j'ai procédé à la caractérisation approfondie de cette bactérie quasi minimale en combinant différentes méthodes expérimentales et approches intégratives. Cette caractérisation intégrative comprend la mesure de plusieurs aspects physiques et physiologiques propres à *M. florum*, incluant son temps de doublement, diamètre cellulaire, masse cellulaire sèche, ainsi que la définition des fractions macromoléculaires de celle-ci. J'ai également réalisé les premières analyses du transcriptome et du protéome de ce microorganisme afin de définir les unités transcriptionnelles, estimer les abondances moléculaires absolues de chacun des transcrits et protéines exprimées, de même qu'évaluer l'importance globale des fonctions cellulaires prédites. En plus d'augmenter nos connaissances fondamentales sur différents aspects de la biologie de *M. florum*, ces efforts de caractérisation serviront de fondation pour le développement d'un modèle à l'échelle du génome décrivant le métabolisme de cette bactérie.

L'ensemble de ces efforts visent à acquérir les connaissances et les outils moléculaires nécessaires afin de transformer *M. florum* en une plateforme simplifiée, hautement caractérisée et spécialement conçue pour explorer les règles gouvernant l'organisation et la plasticité des génomes, ainsi que les mécanismes cellulaires à la base du fonctionnement des cellules. Une telle plateforme a le potentiel de transformer la biologie synthétique en une discipline logique, prévisible et reproductible, rendant ainsi possible le prototypage rationnel et efficace de génomes dans le but de produire des souches bactériennes capables d'accomplir des tâches bien précises.

Mots-clés : Mollicutes, *Mesoplasma florum*, biologie des systèmes, biologie synthétique, génomique synthétique, outils moléculaires, caractérisation intégrative.

REMERCIEMENTS

Je souhaiterais tout d'abord remercier très sincèrement mon directeur de recherche, le Pr Sébastien Rodrigue, pour avoir cru en mon potentiel et m'avoir offert l'opportunité de travailler au sein de son laboratoire. Sébastien a su reconnaître et mettre à profit mes qualités afin que je puisse pleinement m'accomplir en tant que scientifique. Je ne saurais d'ailleurs exprimer à quel point je suis reconnaissant pour toute la confiance et la liberté qu'il m'a accordée au fil des années passées à ses côtés, de même que pour l'encadrement irréprochable qu'il a su m'offrir malgré ses multiples tâches et responsabilités. Par son regard passionné, son dévouement, son enthousiasme, sa patience, son audace, sa sagesse et son calme légendaire, Sébastien représente pour moi une véritable source d'inspiration.

Je tiens également à exprimer ma gratitude envers mon codirecteur, le Pr Pierre-Étienne Jacques, ainsi que mes conseillers, les Prs Vincent Burrus et Nicolas Gévry, pour le soutien et les précieux conseils qu'ils m'ont fournis tout au long du périple que représente un doctorat. Leur savoir et leur expérience ont indéniablement contribué à éclaircir plusieurs aspects relatifs à mon projet de recherche, en plus de m'avoir aidé à faire face à l'adversité lorsque rencontrée. J'aimerais aussi remercier Pascal Sirand-Pugnet et Jean-Philippe Côté pour avoir accepté d'évaluer ma thèse, ainsi que l'ensemble des membres des laboratoires Rodrigue et Jacques pour l'entraide inestimable dont ils ont fait preuve. Ces derniers ont contribué à créer un climat de travail exceptionnellement agréable et stimulant. Je souhaite remercier plus particulièrement Vincent, Marie-Eve, Jean-Christophe, Frédéric et Joëlle pour leurs importantes contributions à mon projet de recherche. Ce fut un réel plaisir de côtoyer ces personnes lors de mes études graduées.

Je désire finalement témoigner ma plus chaleureuse reconnaissance envers ma famille ; mon père Marcel et ma mère Manon, qui m'ont supporté et encouragé depuis le tout début, ainsi que mes frères, Gabriel et Francis, qui ont toujours cru en moi. Sans eux, je ne serais pas l'homme que je suis aujourd'hui. Un merci tout spécial à ma conjointe Mélissa et mes deux enfants, Laurier et Jules. Merci d'embellir ma vie au quotidien, de partager mes joies comme mes peines, de m'accepter comme je suis. Vous êtes mes amours, mes rayons de soleil.

TABLE DES MATIÈRES

LISTE DES ABRÉVIATIONS

3’ -UTR	région 3’ non-traduite
5’ -RACE	5’-*rapid amplification of cDNA ends*
5’ -UTR	région 5’ non-traduite
ADN	acide désoxyribonucléique
ADNc	ADN complémentaire
ADNg	ADN génomique
ARN	acide ribonucléique
ARNm	ARN messager
ARNnc	ARN non-codant
ARNr	ARN ribosomal
ChIP	immunoprécipitation de la chromatine
ChIP-seq	immunoprécipitation de la chromatine couplée au séquençage
CMI	concentration minimale inhibitrice
COBRA	analyse de reconstruction basée sur les contraintes
DNase-seq	séquençage des sites hypersensibles à la DNase I
dTTP	désoxythymidine triphosphate
dUTP	désoxyuridine triphosphate
ESI	ionisation par électronébuliseur
FAIRE-seq	*formaldehyde-assisted identification of regulatory elements*
FPKM	nombre de fragments par kb de transcrit pour un million de lectures alignées

GAM	*growth-associated maintenance*
GEM	modèle à l'échelle du génome
HPLC	chromatographie en phase liquide à haute performance
IHGSC	*International Human Genome Sequencing Consortium*
kb	kilobase
LC-MS/MS	MS/MS couplé à la séparation par chromatographie en phase liquide
MALDI	désorption-ionisation laser assistée par matrice
MS/MS	spectrométrie de masse en tandem
NCBI	*National Center for Biotechnology Information*
NGAM	*non-growth-associated maintenance*
NGS	séquençage nouvelle génération
NHGRI	*National Human Genome Research Institute*
NSAF	*normalized spectrum abundance factor*
oriC	origine de réplication du chromosome
pb	paires de bases
PCA	*polymerase chain assembly*
PCR	réaction en chaîne par polymérase
PEG	polyéthylène glycol
RNA-seq	séquençage de l'ARN
RPKM	nombre de lectures par kb de transcrit pour un million de lectures alignées
Rend-seq	*end-enriched RNA-sequencing*

RT-qPCR	transcription inverse couplée à une réaction en chaîne par polymérase quantitative
SDS	dodécylsulfate de sodium
TAR	*transformation-associated recombination*
TSS	site d'initiation de la transcription
UT	unité transcriptionnelle
UFC	unités formatrices de colonies
VCCD	*Versatile Continuous Culture Device*

LISTE DES TABLEAUX

LISTE DES FIGURES

CHAPITRE 1
INTRODUCTION GÉNÉRALE

1.1 L'avènement de l'ère post-génomique

C'est en 1990 que la communauté scientifique débuta le séquençage du génome humain en utilisant l'approche Sanger (Sanger et al., 1977), un projet d'envergure colossale estimé à 3 milliards de dollars et regroupant 20 centres de séquençage à travers le monde (*International Human Genome Sequencing Consortium;* IHGSC) (Watson and Jordan, 1989). Après plus d'une décennie d'efforts, c'est finalement en février 2001 que l'IHGSC et *Celera Genomics*, une compagnie privée partageant le même objectif, rapportèrent conjointement la toute première séquence brute du génome de l'*Homo sapiens* comprenant 2,91 milliards de paires de bases (pb) (Lander et al., 2001; Venter et al., 2001). Devant les coûts faramineux reliés au séquençage de larges génomes comme celui de l'humain, le *National Human Genome Research Institute* (NHGRI) lança quelques années plus tard un programme visant à réduire significativement les coûts reliés au séquençage des génomes, avec comme cible le séquençage du génome humain pour 1000 $ d'ici dix ans (Schloss, 2008). Ce programme stimula alors le développement de nouvelles technologies de séquençage d'acides nucléiques afin de remplacer la méthode Sanger qui demeure, malgré les progrès réalisés au courant des décennies précédentes (automatisation, marquage fluorescent, etc.), une méthode à relativement faible débit. On assista alors à la naissance des technologies de séquençage qu'on appelle communément séquençage à haut débit ou nouvelle génération (NGS), celles-ci offrant un débit beaucoup plus grand que celui de la méthode Sanger, et ce, à un prix considérablement plus bas (van Dijk et al., 2014). Cette opportunité sans précédent de pouvoir déterminer rapidement la séquence génomique de pratiquement n'importe quel organisme amena ainsi la démocratisation du NGS; on observa une diminution dramatique des coûts associés au séquençage de l'acide désoxyribonucléique (ADN), passant de près de 1000 $ par million de pb en 2004 à environ 0,1 $ en 2011 (Spencer and Fernandes, 2010; Wetterstrand, 2019). Il s'en suivit une véritable explosion du nombre de séquences d'ADN accessibles publiquement ; au début de l'année 2000, on comptait près de

huit milliards de pb associées aux génomes complets déposés dans les principales bases de données publiques, tandis que dix ans plus tard, soit en 2010, ce nombre se situait à plus ou moins 280 milliards (Spencer and Fernandes, 2010). La séquence du génome de la plupart des organismes modèles était maintenant disponible et offrait de nouvelles opportunités quant à la compréhension des règles fondamentales de la vie. L'ère post-génomique avait débuté.

1.2 La génomique fonctionnelle

Suite à la publication de la séquence génomique complète de la plupart des organismes modèles, il devint possible, en principe, d'identifier la fonction de tous les gènes participant aux différents processus cellulaires d'un organisme donné. Puisque le nombre de gènes contenus dans les génomes dépasse généralement plusieurs milliers (à titre d'exemple, la bactérie modèle *Escherichia coli* MG1655 compte environ 4500 gènes [Blattner et al., 1997]), l'atteinte de cet objectif était et demeure encore étroitement reliée à notre capacité à développer des technologies capables d'interroger efficacement les différents niveaux de complexité des cellules (génome, transcriptome, protéome, métabolome, etc.), une sous-discipline de la biologie moléculaire communément appelée la génomique fonctionnelle (Bunnik and Le Roch, 2013; Gasperskaja and Kučinskas, 2017).

La génomique fonctionnelle se concentre sur l'analyse des différents processus cellulaires dynamiques impliqués dans l'expression des gènes et dans le contrôle de l'activité de leur produit, c'est-à-dire majoritairement les protéines, afin de déterminer la fonction de ceux-ci. En plus des approches ciblées comme la délétion génique et la complémentation en *trans*, cette discipline implique l'utilisation de technologies à l'échelle du génome conçues pour analyser l'ensemble des acides ribonucléiques (ARN) transcrits (transcriptome), des protéines traduites (protéome), des processus régulant l'expression de gènes (régulome), ainsi que des interactions protéine-protéine (interactome) de la cellule, sans pour autant être restreinte à cette liste (Bunnik and Le Roch, 2013; Gasperskaja and Kučinskas, 2017). En raison de l'aspect global et très souvent quantitatif des techniques employées en génomique fonctionnelle, l'analyse des

données générées requiert généralement l'utilisation d'algorithmes et d'outils bio-informatiques spécialisés, une approche également utilisée en biologie des systèmes (voir section 1.3) (Bunnik and Le Roch, 2013; Kohl et al., 2010).

Au départ, les analyses à haut débit effectuées en génomique fonctionnelle reposaient principalement sur des méthodologies basées sur les puces à ADN (*microarrays*), celles-ci présentant toutefois de nombreuses limitations quant à l'étude à grande échelle des processus cellulaires, spécialement pour les génomes de taille élevée (Bumgarner, 2013; van Vliet, 2010). Ces technologies furent rapidement remplacées par des approches beaucoup plus fiables et sensibles suite à l'apparition du NGS et des technologies modernes de spectrométrie de masse. Au fil des ans, les méthodes basées sur le NGS et la spectrométrie de masse se sont grandement raffinées et diversifiées ; il existe de nos jours une variété très impressionnante de méthodes permettant de quantifier efficacement et avec précision les différents constituants cellulaires et l'ensemble de leurs interactions. Parmi les techniques les plus couramment utilisées en génomique fonctionnelle, on compte très certainement le séquençage de l'ARN (RNA-seq), l'immunoprécipitation de la chromatine couplée au séquençage (ChIP-seq), ainsi que le séquençage des protéines par spectrométrie de masse (Bunnik and Le Roch, 2013; Gasperskaja and Kučinskas, 2017). Bien évidemment, ces méthodes à l'échelle du génome ont subi plusieurs améliorations et adaptations depuis leur description initiale dans la littérature et ne cessent de se perfectionner. Celles-ci profitent désormais des plus récents progrès technologiques afin d'explorer une variété de processus biologiques toujours croissante, et ce, avec une précision encore plus grande. En raison de la place importante que ces techniques occupent dans le contexte de la génomique fonctionnelle et dans mon projet de recherche, celles-ci seront explorées plus en détail dans les prochaines sections.

1.2.1 Le RNA-seq

Chez les eucaryotes comme les procaryotes, la transcription constitue l'étape à la base de l'expression des gènes. Par conséquent, ce processus cellulaire représente également un point

de régulation clé afin d'assurer une adaptation rapide des cellules à leur environnement lorsque requis. En modulant le niveau de transcription de gènes cibles, les différents mécanismes de régulation permettent de contrôler efficacement les abondances intracellulaires des transcrits d'ARN associés, ce qui résulte en la modification de la composition en protéines des cellules.

En raison de la place centrale que la transcription occupe dans le fonctionnement global des cellules, il n'est point surprenant que l'étude de ce processus ait toujours été d'un très grand intérêt pour les biologistes, et ce bien avant le séquençage du génome humain. Les techniques de biologie moléculaire utilisées pour y parvenir ont toutefois bien changé avec les années. Originalement, ces techniques reposaient principalement sur l'hybridation d'oligonucléotides à des séquences d'acides nucléiques cibles (Croucher and Thomson, 2010; van Vliet, 2010). Parmi les plus couramment utilisées, on comptait l'immunobuvardage de type *Northern* (Alwine et al., 1977) ainsi que la transcription inverse couplée à une réaction en chaîne par polymérase quantitative (RT-qPCR) (Bustin, 2000). Bien que l'immunobuvardage de type *Northern* et le RT-qPCR soient encore utilisés aujourd'hui, leur faible débit a rapidement incité la communauté scientifique à développer des méthodologies alternatives capables d'analyser un très grand nombre de transcrits simultanément, voire même l'ensemble du transcriptome. C'est au milieu des années 1990 que les premières technologies d'étude des transcrits à l'échelle du génome firent leur apparition (Schena et al., 1995). Celles-ci se basaient sur l'utilisation de puces à ADN pouvant contenir jusqu'à plusieurs milliers de sondes d'ADN distinctes, ce qui permettait de potentiellement mesurer le niveau d'expression de la totalité des gènes de l'organisme étudié (Bumgarner, 2013; van Vliet, 2010). Les molécules d'ARN contenues dans les échantillons à analyser étaient premièrement converties en ADNc, marquées, puis hybridées sur la matrice de sondes contenue dans la puce. D'autres techniques telles que le *serial analysis of gene expression* (*SAGE*) firent également leur apparition au milieu des années 1990 (Velculescu et al., 1995). Toutefois, le débit supérieur des puces à ADN ainsi que l'absence d'étapes de clonage contribuèrent à en faire la méthode par excellence pour l'étude du transcriptome.

Même si les puces à ADN représentaient à l'époque une révolution dans l'étude de l'expression des gènes, celles-ci comportaient de nombreux désavantages. En effet, cette technologie possédait généralement un fort niveau de bruit de fond et une sensibilité variable, était restreinte à l'analyse de transcrits prédéfinis, n'informait pas sur la structure des gènes et reposait sur des méthodes de normalisation complexes afin de comparer les niveaux d'expression entre expériences (Bumgarner, 2013; van Vliet, 2010). De plus, la résolution et la couverture des génomes de grandes tailles pouvaient s'avérer plutôt problématiques en raison du nombre élevé de régions codantes à couvrir. Pour l'ensemble de ces raisons, ces technologies ont rapidement perdu en popularité lorsque les premières méthodes basées sur le RNA-seq, décrit plus bas, firent leur apparition (Bainbridge et al., 2006; Nagalakshmi et al., 2008). Le NGS étant au cœur de la technique du RNA-seq, cette dernière possédait une sensibilité et une résolution beaucoup plus grandes que les techniques employant les puces à ADN, permettant ainsi de mesurer le niveau d'expression de l'ensemble des transcrits de la cellule. De plus, puisque le RNA-seq reposait sur l'alignement *in silico* des lectures obtenues suite au séquençage sur la séquence d'un génome d'intérêt, cette technologie n'était pas restreinte à la détection de régions génomiques sélectionnées, ce qui constituait un atout majeur par rapport aux puces à ADN. De ce fait, le RNA-seq permettait non seulement de quantifier l'expression des gènes, mais possédait aussi le potentiel d'informer sur l'organisation de ceux-ci et sur la nature des régions 5' et 3' non-traduites (5' -UTR ; 3' -UTR), en plus de rendre possible la découverte de nouveaux transcrits tels que les petits ARN non-codants (ARNnc) (Creecy and Conway, 2015).

La procédure du RNA-seq peut être subdivisée en trois grandes étapes : la préparation de la librairie de séquençage, le séquençage à proprement dit et l'analyse des lectures obtenues. Certains détails méthodologiques entourant la préparation de la librairie peuvent varier considérablement selon la plateforme de NGS utilisée (*Ion Torrent*, *454*, *SOLiD*, *Solexa/Illumina*). Toutefois, en raison de son haut débit de séquençage, son faible coût par nucléotide séquencé et son faible taux d'erreur, le séquençage de type *Illumina* domine présentement le marché des technologies NGS, et représente la principale technologie utilisée pour la préparation des librairies RNA-seq (van Dijk et al., 2014; Liu et al., 2012a).

Originalement, les techniques utilisées afin de fabriquer les librairies RNA-seq ne permettaient pas de préserver l'identité du brin transcrit, ce qui amenait une importante perte d'information quant à la structure des gènes (Borodina et al., 2011; Croucher and Thomson, 2010). Peu de temps après, plusieurs groupes de recherche développèrent des approches palliant ce défaut : on parlait alors de librairies RNA-seq directionnelles. Deux stratégies légèrement différentes sont généralement employées pour construire des librairies de RNA-seq directionnelles visant à couvrir l'entièreté des transcrits (Levin et al., 2010). La première repose sur la ligation séquentielle d'adaptateurs directement sur les extrémités 5' et 3' des molécules d'ARN simple brin, suivie d'une étape de synthèse d'ADNc par transcription inverse, tandis que la seconde utilise le marquage du brin non transcrit via l'incorporation de désoxyuridine triphosphate (dUTP) lors de la synthèse de l'ADNc. Alors que ces deux stratégies possèdent leurs forces et faiblesses respectives, la ligation d'adaptateurs directement sur l'ARN est de moins en moins utilisée pour le RNA-seq de type conventionnel, notamment en raison de la susceptibilité de l'ARN à la dégradation et à la simplicité des protocoles utilisant le marquage au dUTP. De nombreuses trousses commerciales sont d'ailleurs maintenant disponibles afin de faciliter la préparation des librairies RNA-seq utilisant le dUTP. En contrepartie, la méthodologie utilisant la ligation sur l'ARN est encore exploitée dans plusieurs protocoles dérivés du RNA-seq classique pour lesquels l'approche dUTP s'avère moins pratique ou difficilement utilisable, comme par exemple pour le séquençage de petits ARNnc ou pour l'enrichissement des débuts et fins des transcrits (Carraro et al., 2014; Hafner et al., 2008; Lalanne et al., 2018; Matteau and Rodrigue, 2015a; Radmila et al., 2017; Shinhara et al., 2011). Ceci dit, il existe aujourd'hui une multitude de protocoles basés sur le NGS et conçus pour interroger différents aspects relatifs à la transcription (Radmila et al., 2017), mais la présente section se concentrera plutôt sur le RNA-seq standard visant à séquencer et couvrir la totalité des ARN messagers (ARNm) de la cellule.

Alors que certaines étapes peuvent différer selon les besoins spécifiques à l'utilisateur ainsi que le type de cellules et la trousse utilisée, la stratégie générale de RNA-seq dUTP demeure relativement semblable entre protocoles. Celle-ci peut être résumée en quelques étapes (Figure 1.1). Cette méthodologie consiste tout d'abord à purifier l'ensemble des ARN de

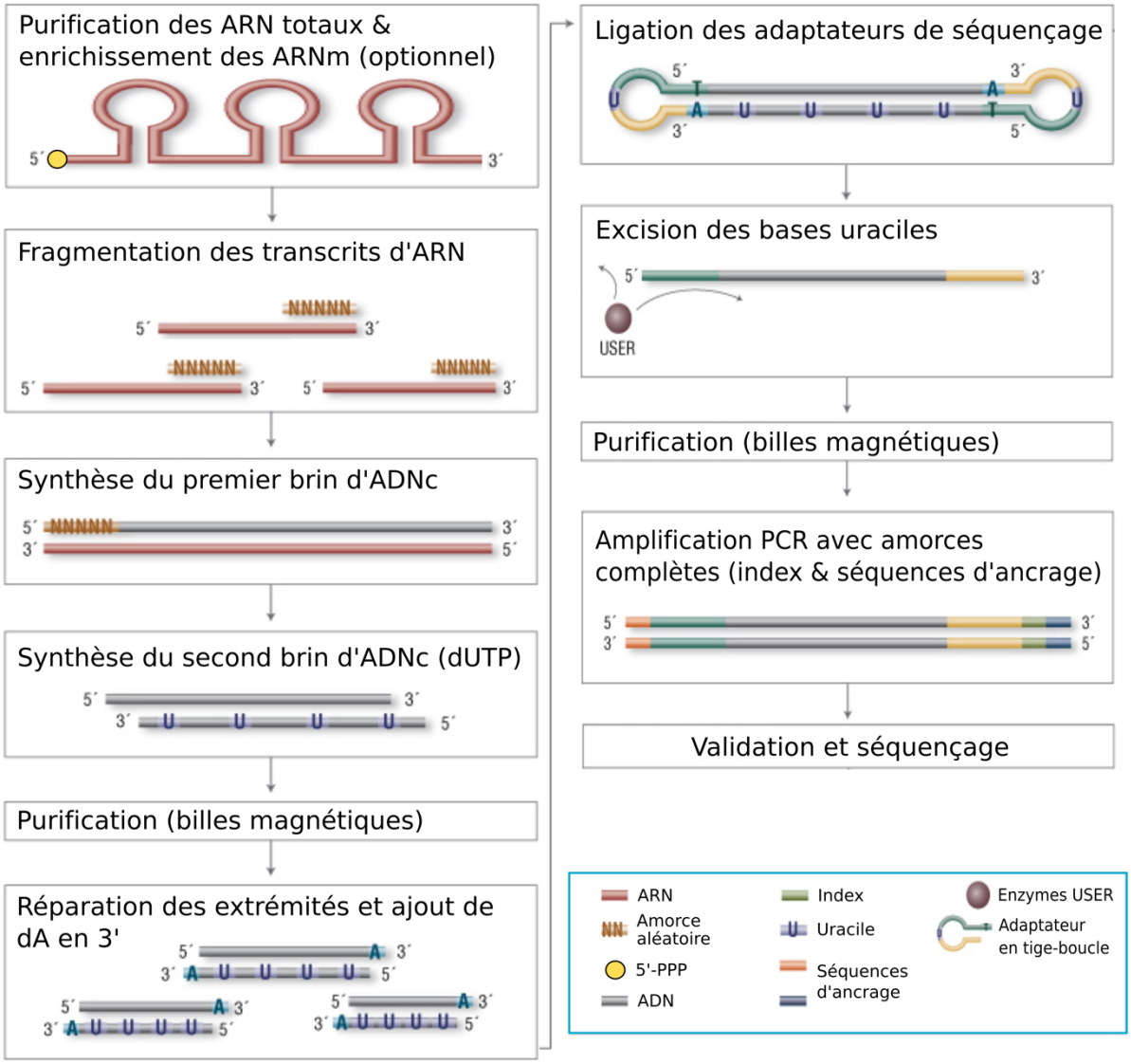

Figure 1.1. Exemple représentatif d'un protocole de préparation de librairies RNA-seq de type dUTP. Certains détails ont été omis pour des raisons de simplicité. Figure adaptée du manuel de la trousse *NEBNext Ultra II Directional RNA Library Prep Kit for Illumina* de la compagnie *New England BioLabs* (www.neb.com).

la cellule, puis à fragmenter ces derniers en plus courtes molécules, généralement grâce à une hydrolyse dépendante de l'action combinée de la chaleur et d'ions de magnésium (Mg^{+2}). Il s'ensuit alors la synthèse du premier brin d'ADNc via l'utilisation des amorces aléatoires et de la réverse transcriptase, puis de la synthèse du second brin en incorporant des dUTP au lieu des désoxythymidines triphosphate (dTTP) classiquement utilisés dans la synthèse de l'ADN. Les

extrémités des molécules d'ADN double brins sont ensuite réparées pour devenir franches et une adénine est ajoutée en 3' de chacun des brins afin de permettre la ligation d'adaptateurs de séquençage. Un mélange d'enzymes (*USER enzymes*) reconnaissant spécifiquement les uraciles est alors ajouté, ce qui permet de dégrader le second brin synthétisé et d'ainsi procéder à un séquençage directionnel des transcrits d'ARN. Les librairies sont finalement amplifiées par réaction en chaîne par polymérase (PCR), leur qualité et concentration est vérifiée à l'aide d'appareils spécialisés, puis le séquençage à haut débit des molécules peut avoir lieu. Il est à noter que la plupart des étapes de purification conventionnelles (précipitation à l'éthanol, purification sur colonne de silice) requises entre les réactions enzymatiques ont maintenant été remplacées par des purifications sur billes magnétiques, ce qui permet de diminuer les pertes de matériel entre les étapes, faciliter l'automatisation ainsi que simplifier et augmenter l'efficacité générale de la procédure. Il est donc possible de construire des banques de RNA-seq avec aussi peu que quelques ng d'ARN de départ – contrairement aux quelques µg nécessaires dans les protocoles originaux – ce qui a même rendu possible le développement d'approches appliquées aux cellules uniques (*single-cell RNA-seq*) (Ziegenhain et al., 2017). Il est également important de mentionner qu'une étape de déplétion des ARN ribosomaux (ARNr) ou d'enrichissement des ARNm est généralement effectuée au début ou à la toute fin du protocole de préparation des librairies de RNA-seq, sans quoi plus de 90 % des lectures s'avèreraient être associées aux transcrits d'ARNr (Levin et al., 2010). En effet, chez les eucaryotes comme les procaryotes, les ARNm représentent seulement environ 5 % de tous les ARN exprimés par la cellule (Bionumbers, 2015; Westermann et al., 2012). Chez les eucaryotes, l'étape d'enrichissement des ARNm est couramment effectuée grâce à l'utilisation de billes magnétiques ou de cellulose comportant des oligonucléotides de thymidines à leur surface, ce qui permet de lier la queue d'adénines située en 3' des ARNm matures et ainsi les isoler des autres ARN de la cellule (Radmila et al., 2017). Alternativement, des amorces contenant un homopolymère de désoxythyimidine peuvent être utilisées en remplacement des amorces aléatoires lors de la synthèse du premier brin d'ADNc. Les procaryotes étant déficients d'une telle structure en 3' de leur ARNm, l'enrichissement de ceux-ci nécessite l'application de stratégies légèrement plus complexes, généralement axées sur l'élimination ou la dégradation des ARNr et autres ARN indésirables. Cela peut s'effectuer via l'utilisation de billes portant des sondes d'ADN capables

de lier spécifiquement les ARNr, ou bien par l'action de nucléases spécialisées telles que la *Duplex specific nuclease* et l'exonucléase XRN-1 (Croucher and Thomson, 2010; Matteau and Rodrigue, 2015a; Radmila et al., 2017; Yi et al., 2011).

Dû à la très grande variété de protocoles de RNA-seq disponibles et à toutes ses variantes possibles, l'analyse des données générées est habituellement réalisée à l'aide de pipelines bio-informatiques développés sur mesure et répondant aux besoins spécifiques de l'expérimentateur. Ces pipelines regroupent généralement plusieurs outils et programmes bio-informatiques spécialisés, connectés ensemble par la préparation de scripts dédiés à cette tâche. Le choix des outils à utiliser dépend de multiples facteurs tels que le type d'organisme étudié (procaryote ou eucaryote), la taille du génome, le nombre d'échantillons à analyser, la puissance de calcul disponible, etc. Malgré la grande diversité de programmes développés pour l'analyse des données de RNA-seq, la plupart de ces analyses s'inscrivent dans une démarche comprenant seulement quelques étapes principales, résumées ici à la Figure 1.2. Lors d'expériences de RNA-seq typiques, la qualité des lectures obtenues est premièrement évaluée, puis les lectures et/ou bases de mauvaise qualité sont retranchées avant de procéder à l'alignement des lectures d'ADN sur le génome d'intérêt. Des programmes tels que *FastQC* (Andrews, 2018) et *Trimmomatic* (Bolger et al., 2014) permettent d'accomplir ce type de tâche. Il existe ensuite plusieurs programmes spécialisés dans l'alignement de courtes séquences d'ADN, possédant tous leurs propres caractéristiques, options et paramètres. Parmi les plus couramment utilisés, on compte *Bowtie* (Langmead and Salzberg, 2012), *BWA* (Li and Durbin, 2009), ainsi que *TopHat* (Kim et al., 2013), ce dernier utilisant *Bowtie* afin de résoudre les jonctions créées par l'épissage alternatif de l'ARN. Suite à l'alignement des lectures sur le génome, il est possible d'évaluer la qualité globale des correspondances et filtrer les alignements de faible qualité avec des outils comme *SAMStat* (Lassmann et al., 2011) et *SAMtools* (Li et al., 2009), mais il peut s'avérer toutefois préférable de conserver l'ensemble des alignements dans certaines circonstances. Dans le cas des eucaryotes, il faut généralement procéder à l'assemblage des transcrits avant de pouvoir quantifier l'expression des gènes en raison de l'épissage alternatif de l'ARN. Cet assemblage peut s'effectuer avec ou sans annotation de référence, et est la plupart

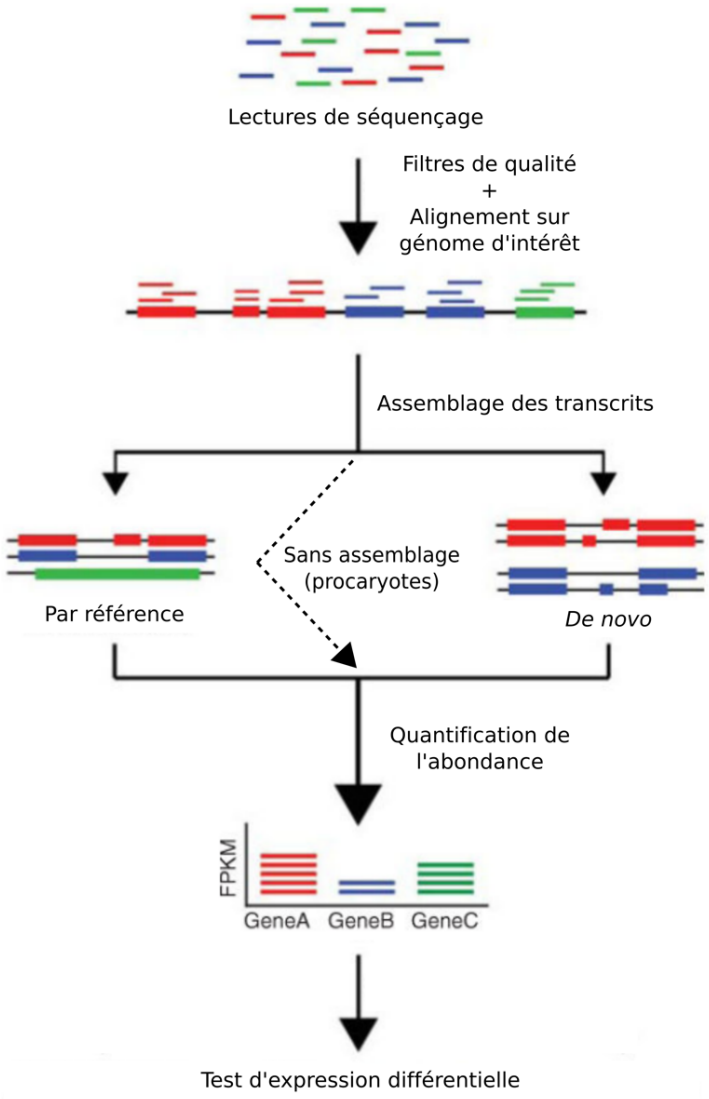

Figure 1.2. Vue d'ensemble des principales étapes impliquées dans l'analyse des données de RNA-seq. Suivant le séquençage des librairies de RNA-seq, les lectures obtenues sont habituellement filtrées pour leur qualité, puis alignées sur la séquence du génome d'intérêt. Il s'ensuit alors d'une étape d'assemblage des transcrits, soit par référence, c'est-à-dire selon une liste de transcrits déjà connus, ou bien seulement à partir des lectures obtenues (*de novo*). L'abondance des transcrits est ensuite évaluée et normalisée selon le nombre de lectures assignées à chacun d'entre eux, puis l'expression différentielle des gènes entre conditions peut être effectuée via différentes approches statistiques. D'autres outils peuvent également être utilisés suivant la quantification de l'expression des gènes afin de répondre à différentes questions, par exemple pour évaluer l'expression spécifique aux allèles ou bien pour l'identification de loci de caractères quantitatifs. Figure adaptée de (Kukurba and Montgomery, 2015).

du temps non requis pour les organismes procaryotes, à moins bien sûr que le but principal de l'expérience consiste à déterminer la structure des unités transcriptionnelles (UT) et des opérons. Dans ce cas, d'autres méthodologies adaptées du RNA-seq sont alors plus appropriées pour déterminer avec exactitude les frontières des transcrits procaryotes, comme les méthodes du 5'-*rapid amplification of cDNA ends* (5' -RACE) (Liu et al., 2018; Matteau and Rodrigue, 2015a) et du *end-enriched RNA-sequencing* (Rend-seq) (Lalanne et al., 2018). La préparation de librairies de séquençage de 3^{e} génération produisant de très longues lectures est également idéale à cet effet (Abdel-Ghany et al., 2016; Bolisetty et al., 2015; Kono and Arakawa, 2019; Wang et al., 2016).

Quoi qu'il en soit, les programmes comme *Cufflinks* (Trapnell et al., 2010, 2012), *Trinity* (Grabherr et al., 2011), *StringTie* (Pertea et al., 2015), *TopHat* (Kim et al., 2013), *STAR* (Dobin et al., 2013) et *rnaSPAdes* (Bushmanova et al., 2019) représentent de bons exemples d'outils spécialisés dans l'assemblage des isoformes d'ARN à partir de données de RNA-seq. Une fois les étapes d'alignement des lectures et d'assemblage des transcrits effectuées, le niveau d'expression des gènes et des isoformes peut être évalué. Certains programmes utilisés lors de l'assemblage des transcrits remplissent également cette fonction, notamment *Cufflinks* (Trapnell et al., 2010) et *StringTie* (Pertea et al., 2015), mais cette tâche est habituellement accomplie par des outils distincts tels que *DESeq2* (Love et al., 2014), *EDGE-pro* (Magoc et al., 2013), *featureCounts* (Liao et al., 2014), *GenomicAlignments* (Lawrence et al., 2013) et *HTSeq-count* (Anders et al., 2015). Ces programmes retournent généralement un compte normalisé exprimé en nombre de lectures par kilobase (kb) de transcrit pour un million de lectures alignées (RPKM). Si le séquençage a été effectué en paires (*paired-end sequencing*), c'est-à-dire qu'une lecture de séquençage a été générée pour chacune des deux extrémités des molécules séquencées, ce compte peut également être exprimé en nombre de fragments par kb de transcrit pour un million de lectures alignées (FPKM) (Babarinde et al., 2019). Le séquençage en paires facilite d'ailleurs l'assemblage des transcrits en raison de la connectivité existante entre les deux lectures obtenues ainsi que la meilleure couverture des molécules séquencées. À ce point, les valeurs de RPKM ou FPKM obtenues constituent une bonne approximation de l'abondance des

transcrits de la cellule, quoique certains biais comme le pourcentage GC, la longueur des gènes et la présence de structures secondaires peuvent affecter la quantification du niveau d'expression des transcrits (Babarinde et al., 2019). Ces biais sont aussi présents dans l'ensemble des méthodes de quantification impliquant une amplification PCR des transcrits, le RT-qPCR n'échappant donc pas à ce phénomène. Pour cette raison, la majorité des études comparent le niveau d'expression d'un ou plusieurs gènes cibles entre différentes conditions ou expériences, ce qui permet d'éliminer les biais attribuables à la séquence puisque celle-ci ne devrait point différer entre les conditions comparées. On parle alors de la mesure du changement du niveau d'expression ou de l'expression relative entre conditions. Ce type d'analyse peut être effectuée grâce à différents programmes tels que *edgeR* (Robinson et al., 2010), *DESeq2* (Love et al., 2014) et *Cuffdiff*, ce dernier étant intégré à la suite d'outils *Cufflinks* (Trapnell et al., 2010, 2012). Ces programmes utilisent des approches statistiques variées afin d'évaluer la significativité de l'augmentation ou de la diminution du niveau d'expression observé pour chacun des gènes à tester, tout en tenant compte de la variation d'expression existante entre réplicats à l'intérieur d'une même condition.

La mesure du changement d'expression entre conditions représente une approche très puissante afin d'identifier la fonction de gènes ou groupes de gènes d'un organisme donné. Malgré le nombre relativement élevé d'étapes incluses dans l'approche RNA-seq, celle-ci représente une méthode éprouvée et mature pour y parvenir. De plus, il existe un bon nombre de trousses commerciales facilitant la préparation des librairies, et le coût relié au NGS est encore aujourd'hui en constante décroissance. L'existence de plateformes web spécialement conçues pour intégrer et faciliter l'utilisation de nombreux programmes – comme le serveur *Galaxy* (Afgan et al., 2018) – permet également aux personnes ne possédant pas de connaissances approfondies en bio-informatique d'être tout de même en mesure de réaliser l'ensemble des analyses typiquement requises dans le contexte du RNA-seq.

1.2.2 Le ChIP-seq

Les protéines liant l'ADN sont à la base du fonctionnement de tous les organismes vivants sur terre, des eucaryotes pluricellulaires que nous sommes à la plus simple des bactéries. Ces protéines jouent un rôle crucial dans la régulation de plusieurs processus cellulaires majeurs tels que la transcription et la traduction, la réplication du matériel génétique ainsi que la réparation de l'ADN. Elles occupent également un rôle essentiel dans la régulation et la propagation de tous les virus et dans la dissémination des éléments génétiques mobiles, comme les éléments intégratifs et conjugatifs (ICEs) (Carraro and Burrus, 2014), démontrant clairement leur rôle ubiquitaire dans le fonctionnement des systèmes biologiques.

Dû à leur importance, l'identification des cibles génomiques reconnues par ces protéines est depuis longtemps le sujet d'un vaste intérêt pour les biologistes. La compréhension approfondie des interactions existant entre les protéines et l'ADN est non seulement critique pour la l'avancement des connaissances fondamentales en biologie, mais les progrès en médecine moderne en sont également tributaires. Avant l'arrivée de technologies à haut débit, ces interactions étaient classiquement étudiées par des approches telles que le retardement sur gel, l'empreinte à la DNAse, l'immunoprécipitation de la chromatine (ChIP) et les essais basés sur le *phage display* et les gènes rapporteurs (Dey et al., 2012). Puisque le ChIP représentait une des seules techniques permettant d'analyser les interactions protéines-ADN dans un contexte *in vivo*, cette approche devint très rapidement la référence au sein de la communauté scientifique, d'autant plus qu'elle est relativement simple à utiliser comparativement à plusieurs autres techniques (Figure 1.3). Brièvement, la première étape de la méthode ChIP typique consiste à traiter les cellules à analyser avec un agent capable de ponter de manière covalente mais réversible les complexes protéines-ADN (Dey et al., 2012; Matteau and Rodrigue, 2015b). Cet agent est habituellement le formaldéhyde, mais l'utilisation d'autres agents chimiques ou la fixation médiée par les rayons ultraviolets sont également possibles. Les cellules fixées sont ensuite lysées par sonication à l'aide d'appareils spécialisés, ce qui fragmente par le fait même l'ADN génomique (ADNg) lié par les protéines en courts fragments mesurant généralement

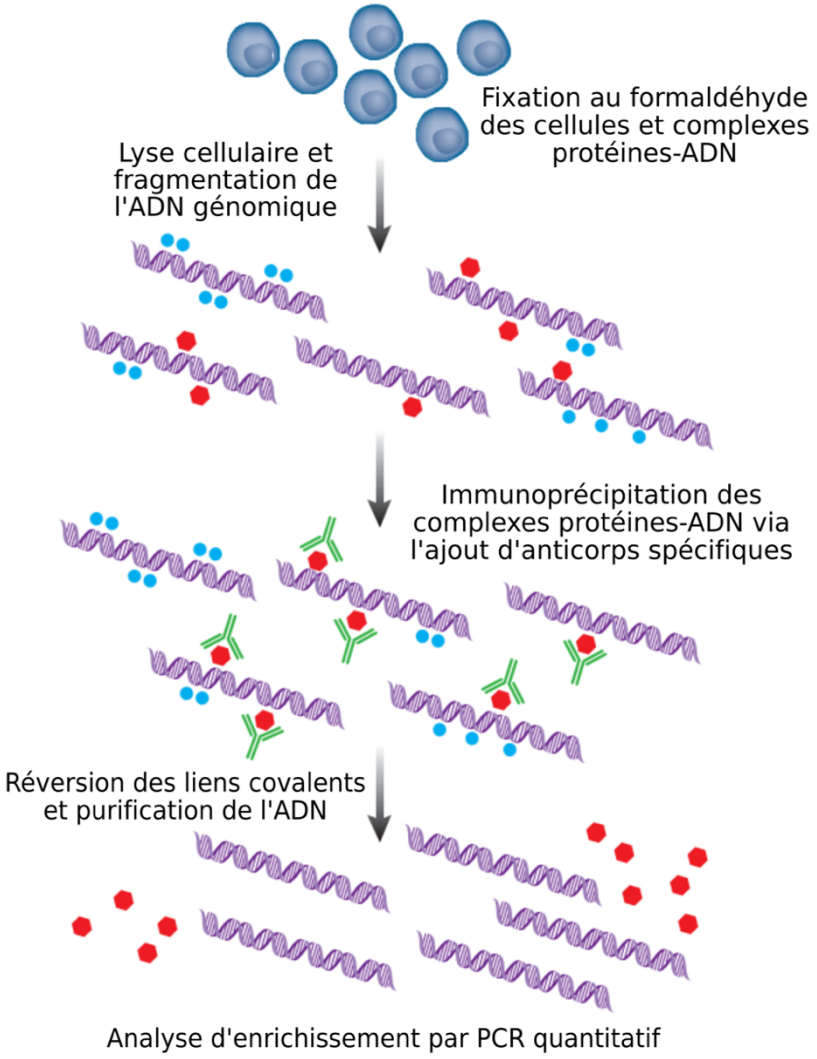

Figure 1.3. Résumé des étapes impliquées dans la méthode ChIP classique. Les cellules à analyser sont tout d'abord fixées au formaldéhyde, puis lysées par sonication, ce qui libère et fragmente l'ADNg lié par les protéines. Les protéines d'intérêt sont ensuite immunoprécipitées à l'aide d'anticorps spécifiques, ce qui enrichit également les fragments d'ADN liés par celles-ci par rapport aux autres constituants de la cellule. Les liens covalents créés entre les protéines et l'ADN par le formaldéhyde sont finalement rompus en conditions dénaturantes, puis l'ADN est purifié avant d'être analysé par PCR quantitative. L'étape de PCR quantitative est remplacée par une étape de séquençage haut débit dans la méthode de ChIP-seq. Figure modifiée de (Mardis, 2007).

entre ~100 et 1000 pb. Les protéines d'intérêt sont alors immunoprécipitées grâce à l'ajout d'anticorps spécifiques couplés à des billes magnétiques ou formées de divers polymères. Il est à noter qu'un anticorps démontrant une excellente avidité en immunobuvardage de type *Western*

n'est pas assurément gage d'une immunoprécipitation efficace en situation de ChIP. En effet, la partie de la protéine reconnue par l'anticorps doit être exposée et accessible en contexte *in vivo*, ce qui n'est malheureusement pas toujours le cas. Si aucun anticorps convenable n'est disponible, l'ajout d'étiquette en partie terminale de la protéine étudiée peut s'avérer une option valable. Une fois l'immunoprécipitation complétée, les liens covalents créés par la fixation au formaldéhyde sont rompus par la chaleur et l'action de détergents ioniques forts comme le dodécylsulfate de sodium (SDS) ou en chauffant les échantillons en condition de très haute salinité. Les protéines présentes sont ensuite hydrolysées grâce à l'action de la protéinase K, puis les fragments d'ADN libérés sont finalement purifiés, généralement à l'aide de colonnes de silice commerciales tolérant les hautes concentrations de SDS et de sels.

En ChIP classique, l'ADN recueilli suite à l'immunoprécipitation est normalement analysé à l'aide d'approches de PCR quantitative utilisant des amorces spécifiques à des régions génomiques d'intérêt. L'enrichissement – ou à l'inverse la déplétion – de ces régions entre conditions (p. ex. cellules traitées vs non-traitées) est synonyme d'une variation au niveau de la quantité de protéines cibles liées à ces sites (Dey et al., 2012). La plus grande faiblesse de la technique ChIP classique réside dans sa capacité limitée à interroger plusieurs sites génomiques parallèlement, technique nécessitant une réaction PCR pour chaque locus génomique à étudier. Ceci a toutefois changé lorsque les approches basées sur les puces à ADN (*ChIP on chip*) et sur le NGS (ChIP-seq) firent leur apparition, celles-ci permettant d'étudier plusieurs milliers de sites génomiques parallèlement (Dey et al., 2012; Kim and Ren, 2006; Mardis, 2007; Park, 2009; Tanita Casci, 2001). Sans grande surprise, le ChIP-seq éclipsa rapidement les méthodes utilisant les puces à ADN en raison de sa couverture entière du génome étudié, sa meilleure sensibilité, ainsi que sa simplicité générale d'exécution. Avec l'approche ChIP-seq, l'étape de PCR quantitative normalement effectuée suite à la purification de l'ADN immunoprécipité est remplacée par la préparation d'une librairie de NGS. Tout comme pour le RNA-seq, c'est la technologie *Illumina* qui est aujourd'hui la plateforme de NGS privilégiée pour la méthodologie ChIP-seq. La préparation de librairies de ChIP-seq est par contre beaucoup plus simple que la préparation de librairies de RNA-seq, principalement parce qu'aucune étape de transcription

inverse n'est nécessaire et que l'ADN à analyser est déjà fragmenté à la taille adéquate en raison de l'étape de lyse par sonication. En bref, il s'agit d'une étape de ligation d'adaptateurs de séquençage suivie d'une amplification PCR avec amorces spécifiques à la technologie de NGS choisie. Cette procédure est donc quasiment identique à celle utilisée pour le séquençage de l'ADNg total à des fins d'assemblage de séquences génomiques. Toutefois, la quantité d'ADN de départ est généralement beaucoup plus faible avec le ChIP-seq en raison des étapes d'immunoprécipitation. Diverses compagnies offrent maintenant des trousses commerciales facilitant la préparation des librairies de NGS à partir d'ADN fragmenté ou non, et certaines d'entre elles ont été optimisées pour supporter des très faibles concentrations, ce qui convient parfaitement aux librairies de ChIP-seq (Sundaram et al., 2016).

Grâce à la puissance du NGS, la technologie de ChIP-seq a été utilisée afin d'étudier une multitude de processus cellulaires importants tels que la régulation de la transcription, la modification post-traductionnelle des histones et la réparation de l'ADN (Dey et al., 2012; Park, 2009). Plusieurs méthodologies similaires au ChIP-seq et exploitant les technologies de NGS ont d'ailleurs été développées au fil des ans pour répondre à des objectifs bien précis. Le *formaldehyde-assisted identification of regulatory elements* (FAIRE-seq) et le séquençage des sites hypersensibles à la DNase I (DNase-seq), par exemple, permettent d'enrichir et séquencer les régions génomiques accessibles et exemptes de nucléosomes, caractéristiques de régions régulatrices transcriptionnellement actives (Bianco et al., 2015; Simon et al., 2012; Song et al., 2011). À mi-chemin entre le ChIP-seq et le RNA-seq, le profilage ribosomique ou Ribo-seq représente pour sa part une méthode extrêmement puissante afin d'identifier et quantifier les ARNm activement traduits en protéines (Brar and Weissman, 2015; Ingolia et al., 2012). Une variante améliorée du ChIP-seq, appelée le ChIP-exo, a aussi été récemment développée dans le but de déterminer la localisation génomique des protéines liant l'ADN avec une précision jusqu'alors inégalée frôlant la paire de base unique (Matteau and Rodrigue, 2015b; Rhee and Pugh, 2011, 2012; Serandour et al., 2013). Originalement exploitée pour identifier précisément les sites chromosomiques reconnus par différents facteurs de transcription de la levure et de l'humain, cette méthodologie tire son épingle du jeu en exploitant le pouvoir processif

d'exonucléases afin de digérer les extrémités des molécules d'ADN liées par les protéines lors de l'immunoprécipitation (Figure 1.4). Le ChIP-exo surpasse ainsi la résolution habituellement observée avec le ChIP-seq, celle-ci n'étant plus limitée par la résolution de fragmentation de l'ADNg mais bien par la zone protégée par la protéine étudiée. L'utilisation d'exonucléases permet également de diminuer considérablement le bruit de fond normalement observé avec l'approche de ChIP-seq classique, ce qui facilite l'analyse des signaux obtenus et rend possible la détection de sites plus faiblement liées par les protéines. En contrepartie, la nature partiellement simple brin des molécules d'ADN obtenues avec le ChIP-exo complexifie substantiellement la préparation des librairies de séquençage, quoique certains efforts ont récemment été investis afin de diminuer le temps et les coûts reliés à cette étape (Rossi et al., 2018). Par conséquent, aucune compagnie n'offre actuellement de trousse commerciale consacrée à la fabrication des librairies de ChIP-exo. Toutefois, quelques protocoles hautement détaillés sont disponibles dans la littérature afin d'y parvenir (Matteau and Rodrigue, 2015b; Rhee and Pugh, 2012).

Puisque les données obtenues suite au séquençage des librairies de RNA-seq et de ChIP-seq/ChIP-exo sont virtuellement les mêmes, c'est-à-dire des courtes séquences d'ADN accompagnées de scores de qualité (généralement contenues dans un fichier de format *FASTQ*), les processus d'analyse qui en découlent partagent de nombreuses similarités (Bailey et al., 2013; Furey, 2012). En effet, tout comme avec le RNA-seq, les étapes initiales consistent habituellement à appliquer un filtre de qualité sur l'ensemble des lectures de séquençage obtenues, suivi de leur alignement sur la séquence du génome étudié (voir la section précédente pour des exemples de programmes recommandés et la Figure 1.2 pour le schéma récapitulatif de l'analyse des données de RNA-seq). La différence majeure entre l'analyse des données de RNA-seq et de ChIP-seq réside principalement dans les étapes suivant l'alignement des lectures sur le génome; au lieu de procéder à la reconstruction des transcrits et au compte des lectures s'alignant sur chacun d'entre eux, la démarche usuelle employée avec le ChIP-seq consiste à identifier les loci génomiques présentant un nombre de lectures statistiquement plus élevé que ce qui attendu simplement par chance, une approche communément appelée le *peak calling*.

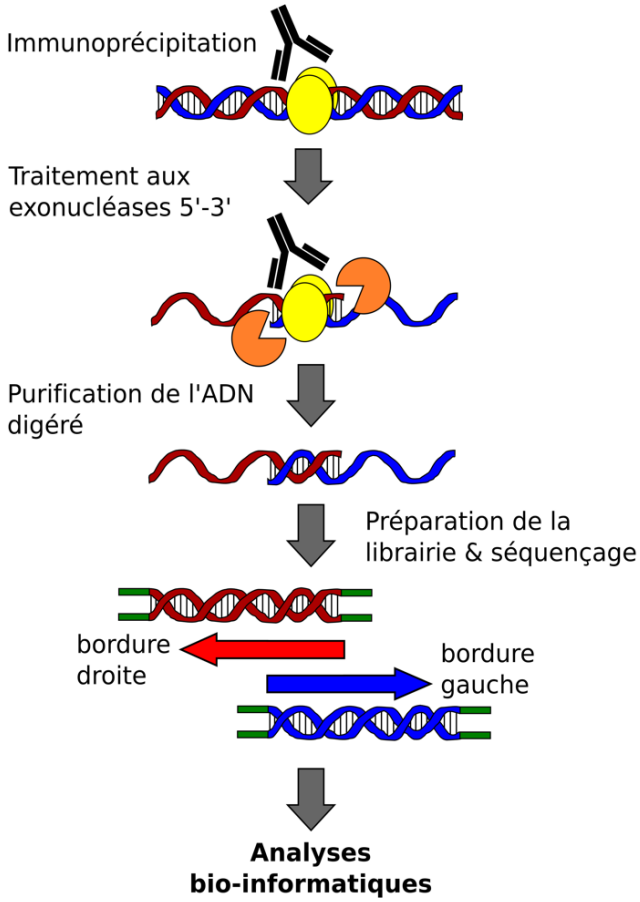

Figure 1.4. Sommaire de l'adaptation ChIP-exo permettant au ChIP-seq de bénéficier d'une résolution accrue. Contrairement au ChIP-seq classique, les complexes protéines-ADN sont soumis à un traitement aux exonucléases à activité 5'-3' lors de l'immunoprécipitation, ce qui crée des molécules d'ADN partiellement simple brin. Seule la région protégée par la protéine d'intérêt conserve sa nature double brin. Les molécules d'ADN hydrolysées sont ensuite purifiées et dénaturées, puis la frontière de digestion de chaque brin est marquée à l'aide d'une ligation avec un adaptateur de séquençage. Les librairies sont finalement amplifiées par PCR, purifiées, puis validées avant d'être séquencées. Puisque chacun des deux brins de l'ADN recueilli est utilisé lors de la préparation de la librairie, chaque site génomique séquencé sera couvert à la fois sur le brin positif et négatif de l'ADN. La portion couverte sur les deux brins correspondra à la partie liée et protégée par la protéine étudiée. Figure adaptée de (Matteau and Rodrigue, 2015b).

Depuis la popularisation du NGS et de ses applications dérivées, plusieurs algorithmes et outils bio-informatiques ont été développés afin d'accomplir ce genre d'analyse d'enrichissement. *MACS* (*model-based analysis of ChIP-seq*) (Feng et al., 2011; Zhang et al., 2008) figure très

certainement parmi les outils les plus utilisés, mais d'autres programmes tels que *GenoGAM* (Stricker et al., 2017), *FindPeaks* (Fejes et al., 2008), *ZINBA* (Rashid et al., 2011), *HOMER* (Benner, 2019) et *QuEST* (Valouev et al., 2008) proposent une variété de paramètres et fonctionnalités qui peuvent s'avérer intéressants selon les besoins spécifiques liés aux analyses à effectuer (Koohy et al., 2014; Wilbanks and Facciotti, 2010). Évidemment, plusieurs variables telles que la taille du génome, la longueur des lectures obtenues, le niveau de bruit de fond et l'envergure des régions protégées peuvent dramatiquement affecter le comportement des outils de *peak calling*. Les paramètres des outils sélectionnés doivent par conséquent être ajustés pour chacun des jeux de données à analyser afin de maximiser la détection des régions génomiques véritablement liées par des protéines. Par exemple, plusieurs de ces outils permettent de comparer les données issues des expériences d'immunoprécipitation à un contrôle non enrichi (*input*) ou sans anticorps afin d'éliminer les sites potentiellement enrichis en raison de biais d'amplification ou d'alignement (faux positifs). Cette comparaison est toutefois généralement superflue dans le cas du ChIP-exo puisque le traitement aux exonucléases amène une perte quasi totale du bruit du fond. D'ailleurs, pour profiter de toute la résolution offerte par le ChIP-exo, il est conseillé de ne conserver que la première paire de bases séquencée de chacune des lectures alignées, ce qui permet d'obtenir des signaux d'enrichissement extrêmement définis et correspondant précisément aux frontières des régions protégées par les protéines immunoprécipitées (Carraro et al., 2014; Matteau and Rodrigue, 2015b; Poulin-Laprade et al., 2015; Rhee and Pugh, 2011, 2012). Ces signaux peuvent ensuite être visualisés à l'aide d'outils de visualisation comme *IGV* (Thorvaldsdóttir et al., 2013) ou le *UCSC genome browser* (Kent et al., 2002), ce dernier étant très prisé en raison de ses multiples fonctionnalités et de sa richesse en données publiquement accessibles.

Une fois la liste des régions génomiques affichant un enrichissement significatif générée, plusieurs types d'analyses sont possibles afin de nous aider à interpréter les résultats obtenus. Premièrement, la comparaison avec d'autres jeux de données génomiques peut s'avérer très informative selon les cas. À titre d'exemple, la colocalisation d'une protéine de fonction inconnue avec des marques épigénétiques typiquement associées à l'initiation de la transcription

permet de formuler l'hypothèse que celle-ci est potentiellement impliquée dans ce processus. Ce type d'association peut s'effectuer de manière bio-informatique avec des outils de traitement de fichiers de coordonnées génomiques tels que *BEDTools* (Quinlan and Hall, 2010) et *BEDOPS* (Neph et al., 2012), mais peut aussi être réalisé de manière manuelle grâce aux outils de visualisation de données génomiques comme le *UCSC genome browser* (Kent et al., 2002). À l'instar du RNA-seq, il est aussi possible de procéder à des analyses d'enrichissement différentiel entre conditions (p. ex. mutant vs type sauvage) via l'utilisation d'outils comme *DESeq2* (Love et al., 2014) et *edgeR* (Robinson et al., 2010). Advenant que des différences d'enrichissement significatives soient constatées et que celles-ci puissent être associées à des gènes de fonction connue, différents outils peuvent être employés afin d'évaluer si certains de ces gènes partagent des similarités au niveau de leur rôle biologique ou leur localisation cellulaire. Cette démarche est aussi fréquemment utilisée afin d'analyser les gènes différentiellement exprimés en RNA-seq, et repose généralement sur l'enrichissement en termes associés aux bases de données publiques comme le *Gene Ontology* (GO) *Consortium* (Ashburner et al., 2000) ou le *KEGG Orthology* (KO) *Database* (Kanehisa et al., 2016). Finalement, les analyses de recherche de motifs sont chose commune dans l'univers du ChIP-seq et du ChIP-exo. Celles-ci permettent de bâtir une séquence consensus à partir des sites génomiques enrichis et ainsi identifier la séquence de reconnaissance de la protéine étudiée si elle n'était pas connue au départ. Énormément d'outils existent afin d'exécuter ce type d'analyse (Hashim et al., 2019). DREME (Bailey, 2011), MEME (Bailey et al., 2009), et HOMER (Benner, 2019) en sont de bons exemples et figurent parmi les plus populaires. De plus, certains de ces programmes permettent de comparer les motifs trouvés avec ceux déjà connus et disponibles dans les bases de données publiques.

Somme toute, le ChIP-seq et le ChIP-exo constituent des techniques extrêmement puissantes afin d'investiguer les sites de liaison de protéines à l'échelle du génome et pour déterminer la fonction des gènes. De ce fait, celles-ci occupent présentement une place très importante dans l'étude des systèmes biologiques et sont appelées à jouer un rôle déterminant dans plusieurs

domaines de recherche émergents, notamment en biologie des systèmes et en génomique synthétique.

1.2.3 Le séquençage des protéines par spectrométrie de masse

Les protéines représentent les principales entités fonctionnelles des cellules. Par leurs fonctions hautement diversifiées, celles-ci occupent un rôle essentiel dans pratiquement tous les processus cellulaires, à quelques exceptions près. Comparativement au séquençage des acides nucléiques, la détermination à haut débit de la séquence peptidique des protéines s'est avérée beaucoup plus difficile, principalement en raison du très grand niveau de diversité et de complexité de ces macromolécules (Altelaar et al., 2013). En effet, en plus de reposer sur un alphabet comprenant 22 acides aminés différents (incluant la sélénocystéine et la pyrrolysine), les protéines sont sujettes à de nombreuses modifications post-traductionnelles, peuvent s'associer sous forme de complexes multimériques, et sont fréquemment exprimées sous différentes isoformes en raison de codons de départ alternatifs ou de l'épissage alternatif des ARNm. Par conséquent, l'étude de la séquence des protéines a pendant longtemps été confinée à l'analyse individuelle de celles-ci (Aslam et al., 2017). Les expériences typiques consistaient généralement à purifier la ou les protéines d'intérêt grâce à diverses techniques de chromatographie, suivie parfois d'une séparation par électrophorèse sur gel de polyacrylamide, pour finalement se terminer par la détermination de la séquence N-terminale des protéines par la méthode de dégradation d'Edman (Edman, 1949). Alors que la dégradation d'Edman représentait à l'époque une avancée majeure dans l'étude des protéines, cette technique ne permettait pas de mesurer l'abondance de la protéine étudiée. L'utilisation de techniques complémentaires comme l'immunobuvardage de type *Western*, la méthode de Bradford ou bien les essais immuno-enzymatiques de type ELISA était alors nécessaire pour y parvenir (Noble and Bailey, 2009). Quoi qu'il en soit, les courtes séquences peptidiques obtenues par la dégradation d'Edman étaient généralement suffisantes pour déterminer l'identité des protéines séquencées, pourvu qu'une de base de données existait pour l'organisme étudié. Cette technique fut d'ailleurs à l'origine des premières études dites protéomiques, c'est-à-dire portant sur l'étude à relativement grande échelle des protéines

(Graves and Haystead, 2002). Combinant la séparation des protéines par électrophorèse bidimensionnelle à la dégradation d'Edman, ces études pionnières étaient néanmoins limitées par l'absence de la séquence génomique complète des organismes couramment étudiés, ce qui rendait l'identification de nouvelles protéines très difficile.

Ce n'est toutefois que vers le fin des années 1990 que le domaine de la protéomique a connu son réel essor, en partie grâce à l'accessibilité de la séquence génomique complète de la plupart des organismes modèles, mais principalement en raison des progrès importants effectués en spectrométrie de masse (Andersen and Mann, 2000; Graves and Haystead, 2002; Griffiths, 2008). En effet, alors que l'analyse de petites molécules organiques était devenue routine en spectrométrie de masse, les protéines et autres macromolécules représentaient à l'époque un défi beaucoup plus important. Cela changea lorsque les techniques de désorption-ionisation laser assistée par matrice (MALDI) ainsi que d'ionisation par électronébuliseur (ESI) furent inventées. Celles-ci permettaient de transférer efficacement les molécules de haut poids moléculaire de la phase liquide à la phase gazeuse injectée et analysée par le spectromètre de masse (Andersen and Mann, 2000; Graves and Haystead, 2002; Griffiths, 2008). Ceci fut également suivi d'améliorations importantes au niveau de la précision, de la rapidité et de la sensibilité des appareils de spectrométrie de masse, qui, jumelées aux nouvelles méthodes de purification et de fractionnement des protéines, ont mené aux approches actuelles d'analyse et d'identification de mélanges complexes de protéines. Bien entendu, ces approches n'auraient pas été possibles sans les avancées réalisées en bio-informatique, en raison de la quantité importante de données à traiter dans ce type d'analyse.

Les approches de spectrométrie de masse en tandem (MS/MS), c'est-à-dire combinant le pouvoir analytique de deux ou plusieurs spectromètres de masse connectés en série, s'avèrent particulièrement puissantes afin d'explorer le protéome des cellules (Delahunty and Yates, 2005; Mann et al., 2001; Sabidó et al., 2012). En plus de rendre possible la quantification ainsi que la détermination de la séquence peptidique de plusieurs centaines, voire de milliers de protéines différentes simultanément, les techniques de MS/MS sont facilement automatisables,

possèdent la capacité d'identifier l'ensemble des modifications post-traductionnelles des protéines, et montrent une sensibilité grandement supérieure à la méthode de dégradation d'Edman (Graves and Haystead, 2002; Han et al., 2008). Pour ces raisons, les méthodes basées sur le MS/MS ont de nos jours essentiellement remplacé la dégradation d'Edman pour le séquençage des protéines et constituent la méthode par excellence afin d'étudier le contenu protéique des cellules. Pour profiter au maximum de la puissance offerte par le MS/MS, les protéines contenues dans les échantillons protéiques à analyser sont habituellement séparées au préalable en utilisant différentes approches chromatographiques ou de migration sur gel. Tout comme avec le RNA-seq, la méthode exacte employée peut différer considérablement selon les besoins de l'expérimentateur et le type d'échantillon à analyser. Les techniques de MS/MS couplées à la séparation par chromatographie en phase liquide (LC-MS/MS) offrent toutefois une capacité de résolution du protéome extrêmement intéressante dans la plupart des cas (Figure 1.5). Avec cette approche, les protéines extraites sont généralement séparées par électrophorèse sur gel de polyacrylamide, puis hydrolysées en peptides de longueur variable grâce à l'action de différentes protéases, la trypsine étant la plus couramment utilisée (Aslam et al., 2017; Delahunty and Yates, 2005). Les peptides sont ensuite séparés par chromatographie en phase liquide en utilisant des colonnes présentant une affinité variable selon la séquence et les propriétés des peptides en question, par exemple selon leur degré d'hydrophobicité, leur taille ou leur potentiel ionique. Il s'ensuit alors une étape d'ionisation des peptides, – les techniques ESI et MALDI étant les deux méthodes d'ionisation les plus répandues – ce qui permet d'injecter les peptides-ions formés (appelés précurseurs) à l'intérieur d'un premier spectromètre de masse pour analyser leur masse et charge respectives. Les peptides analysés sont ensuite fragmentés en peptides-ions encore plus courts, puis analysés une seconde fois dans un autre spectromètre de masse connecté en série. Puisque ces étapes s'effectuent très rapidement, l'acquisition des données de masse et de charge, appelées spectres, doit également se faire dans un laps de temps très court. Le processus d'acquisition des spectres constitue d'ailleurs un point critique dans la détermination des séquences peptidiques analysées par le spectromètre de masse. Cette étape peut s'effectuer de manière indépendante aux données, c'est-à-dire que l'ensemble des ions précurseurs est analysé peu importe leurs caractéristiques, ou bien à l'inverse, selon une approche dépendante de la taille, masse ou charge des peptides

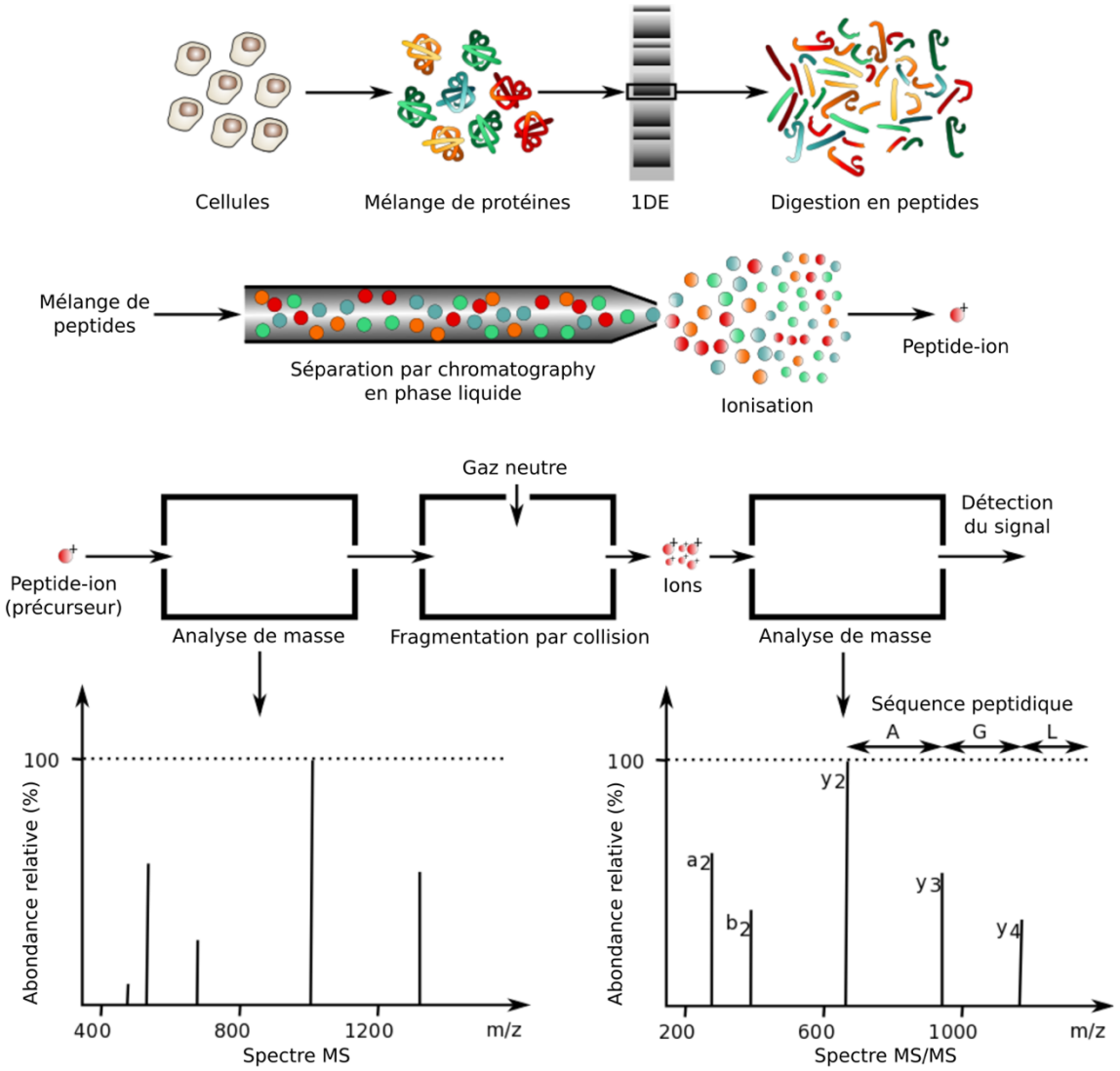

Figure 1.5. Résumé de la technique LC-MS/MS utilisée pour séquencer des extraits protéiques complexes. Suite à l'extraction des protéines, celles-ci sont généralement séparées par électrophorèse sur gel de polyacrylamide, puis hydrolysées en peptides par l'action de protéases comme la trypsine. Les peptides générés sont ensuite fractionnés par chromatographie en phase liquide, ionisés, puis analysés selon leur masse et leur charge en passant sous forme gazeuse dans deux spectromètres de masse successifs. L'analyse de spectres obtenus permet de déduire la séquence des protéines analysées selon la base de données choisie, ainsi que quantifier l'abondance de celles-ci dans l'extrait protéique original. Figure adaptée de (Barillot et al., 2012).

analysés. Ces critères permettent de sélectionner seulement certains types de précurseurs pour la fragmentation subséquente et l'analyse dans le deuxième spectre de masse (Mann et al., 2001; Wolf-Yadlin et al., 2016).Une fois les spectres recueillis, ceux-ci doivent passer par une série

d'étapes d'analyse afin d'en déduire la séquence des peptides associés et ainsi procéder à la détection des protéines analysées. Malheureusement, les fichiers renfermant les spectres sont typiquement encodés dans un format fermé et spécifique à l'appareil utilisé. La compatibilité entre les différents formats de fichiers et les programmes d'analyse de spectres disponibles peut donc constituer une limitation importante dans ce contexte (Kessner et al., 2008; Röst et al., 2016). De plus, plusieurs de ces programmes requièrent une licence d'utilisation et ne prennent en charge qu'une quantité très restreinte de formats de fichiers différents. Heureusement, la tendance au libre accès (*open-source*) des dernières décennies n'ayant pas échappé au domaine de la spectrométrie de masse, plusieurs alternatives de programmes ouverts et gratuits tels que *ProteoWizard* (Kessner et al., 2008), *OpenMS* (Röst et al., 2016) et *PeptideShaker* (Vaudel et al., 2015) sont maintenant disponibles pour analyser les données de séquençage de protéines. De plus, certains de ces programmes sont compatibles avec la quasi-totalité des formats de fichiers générés par les appareils de spectrométrie de masse.

Les analyses typiques de données de séquençage des protéines par MS/MS débutent habituellement par la conversion des formats de fichiers en formats ouverts, étape durant laquelle toute une gamme d'opérations de filtration, de transformation et d'annotation des spectres peuvent être effectuées afin d'éliminer les peptides de mauvaise qualité ou réduire le bruit de fond observé dans les signaux bruts (Deutsch et al., 2008; Veltri, 2008). Bien que ces opérations puissent avoir une répercussion non négligeable sur les résultats finaux, celles-ci demandent une connaissance très approfondie des appareils utilisés et de leurs caractéristiques spécifiques, ce qui est rarement le cas pour le biologiste moléculaire moyen. Dans la plupart des cas, les spectres obtenus sont ensuite comparés à une base de données afin de déterminer la séquence peptidique associée à chacun d'entre eux et inférer l'identité des protéines analysées. Dans le cas d'expériences ciblées, cette base de données peut être de nature relativement simple et ne comporter que la séquence de quelques protéines cibles. Inversement, celle-ci peut contenir l'ensemble des séquences protéiques associées à un génome de référence, voire même la totalité des cadres de lecture ouverts possibles de ce génome (*6-frame database*). Ce type de base de données est particulièrement puissant pour découvrir de nouvelles protéines absentes de

l'annotation classique des génomes (Andrews and Rothnagel, 2014; Mouilleron et al., 2016; Ravikumar et al., 2018; Vanderperre et al., 2013; Yang et al., 2014). Tout comme le processus d'acquisition des spectres, la base de données choisie aura un impact considérable sur la nature des peptides et protéines identifiés. Une base de données plus vaste offrira davantage de possibilités aux algorithmes de recherche utilisés pour la détection des peptides, ce qui causera un nombre total de résultats faux-positifs forcément plus élevé (Deutsch et al., 2008; Veltri, 2008). Cet aspect peut d'ailleurs être exacerbé si les algorithmes utilisés permettent aux peptides de contenir des modifications post-traductionnelles, ce qui augmente encore plus le nombre de possibilités. Il faut donc faire preuve de prudence dans le choix de la base de données à utiliser, et advenant que de nouvelles protéines soient découvertes, la confirmation par des approches complémentaires doit être de mise. Il est d'ailleurs possible de fournir une liste de contaminants fréquemment observés (p. ex. la kératine) aux programmes d'identification des spectres afin d'augmenter la confiance envers les peptides et les protéines trouvés.

En plus de permettre la détection de nouvelles protéines et l'identification de modifications post-traductionnelles, le séquençage des protéines par MS/MS est fréquemment utilisé pour quantifier l'abondance des protéines entre différentes conditions. De façon similaire à la quantification de l'expression des gènes en RNA-seq, l'abondance des protéines peut être calculée selon le nombre de lectures ou spectres associés à chacune d'entre elles, puis normalisée selon la longueur des protéines et de leur nombre total dans l'expérience (McIlwain et al., 2012; Zhu et al., 2010). On parle alors d'un compte normalisé selon le nombre de spectres associés (*normalized spectrum abundance factor* ; NSAF). Il existe également d'autres méthodes afin d'estimer l'abondance des protéines identifiées par MS/MS, certaines se basant plutôt sur l'intensité des spectres associés aux peptides que sur leur nombre (McIlwain et al., 2012; Zhu et al., 2010). Tout comme pour le RNA-seq, ces méthodes de quantification comportent des biais notables sur la mesure des abondances absolues, principalement en raison de différences dans le potentiel d'ionisation des peptides et au niveau de l'efficacité d'extraction des protéines. Malgré cela, les abondances calculées à partir des valeurs de NSAF représentent *grosso modo* les niveaux intracellulaires des protéines séquencées. Ces biais peuvent néanmoins

être contournés en utilisant différents types de standards lors des analyses de MS/MS (Bantscheff et al., 2007; Kito and Ito, 2008). De plus, dans le cas de la mesure du niveau d'expression relatif entre conditions, ces biais n'ont pratiquement plus d'importance, car les protéines sont comparées avec elles-mêmes, abolissant ainsi tout biais relatif à la séquence peptidique.

Alors que la spectrométrie de masse fût jadis un domaine de recherche très fermé, celle-ci se transpose peu à peu vers une discipline ouverte, mature et axée sur la découverte de nouvelles possibilités. Son importance en biologie est maintenant incontestable ; en plus de pouvoir mesurer avec précision l'ensemble du protéome des cellules, la spectrométrie de masse représente également un outil puissant d'analyse des métabolites, lipides, polysaccharides et interactions protéines-protéines (Girolamo et al., 2013; Iacobucci et al., 2018; Smith et al., 2014). Au fur et à mesure que les technologies de MS/MS se perfectionnent et que davantage d'outils d'analyses en libre accès se développent, la spectrométrie de masse démontrera alors son plein potentiel dans l'étude des systèmes biologiques, profitant plus particulièrement aux disciplines hautement intégratives telles que la biologie des systèmes.

1.3 La biologie des systèmes

La biologie des systèmes est un domaine de recherche interdisciplinaire qui vise à comprendre et décrire de manière quantitative les systèmes biologiques à l'aide d'approches holistiques (Breitling, 2010; Hillmer, 2015). Ces systèmes biologiques peuvent être de nature relativement ciblée – l'ensemble des gènes d'un réseau métabolique et leurs interactions, par exemple – ou à l'inverse, extrêmement large – l'ensemble des organismes et des interactions propres à un écosystème. Contrairement aux approches réductionnistes classiques, cette discipline repose généralement sur l'intégration d'une grande quantité de données biologiques via l'utilisation de modèles mathématiques et informatiques. Ces modèles permettent de représenter les interactions entre les différentes variables propres à un système biologique sous forme d'un ensemble organisé d'équations mathématiques (Breitling, 2010; Hillmer, 2015). Cette

représentation mathématique sert de plateforme afin d'étudier le comportement global du système et rapidement explorer l'impact de différents paramètres (la vitesse d'une réaction métabolique par exemple) sur celui-ci. Ainsi, les modèles computationnels constituent des outils particulièrement efficaces afin de générer des hypothèses pouvant servir de point de départ aux expérimentations plus approfondies (Breitling, 2010; Hillmer, 2015).

Même si l'importance d'étudier et de comprendre les systèmes biologiques dans leur ensemble est reconnue depuis le début du 20^{e} siècle, la biologie des systèmes est demeurée pendant longtemps limitée par notre incapacité à valider expérimentalement les modèles mathématiques proposés, faute de technologies adéquates (Marin-Sanguino et al., 2018). Ce n'est qu'autour des années 2000 que cette branche de la biologie a réellement pris son envol grâce à l'augmentation du nombre de séquences de génomes complets disponibles ainsi qu'aux avancées réalisées en bio-informatique et en génomique fonctionnelle (Bunnik and Le Roch, 2013; Gasperskaja and Kučinskas, 2017). Par leur nature globale et quantitative, les données provenant des approches comme le RNA-seq et le MS/MS représentent des substrats idéaux afin de valider les modèles computationnels développés en biologie des systèmes. Aujourd'hui, soit près de 20 ans après l'introduction de la biologie des systèmes dans le vocabulaire courant de la biologie moderne, cette discipline intégrative s'est grandement épanouie et diversifiée. Celle-ci profite notamment des plus récents progrès en bio-informatique et en génomique fonctionnelle afin d'étudier les relations complexes qui existent à l'intérieur des systèmes biologiques et arriver à une compréhension complète et détaillée de ceux-ci. L'approche scientifique de la biologie des systèmes est maintenant employée dans pratiquement tous les domaines touchant de près ou de loin aux sciences biologiques, allant de la compréhension des étapes menant à la cancérogenèse à l'étude des communautés microbiennes, en passant par l'ingénierie métabolique et la découverte de nouvelles cibles pharmaceutiques (Chuang et al., 2010; Gu et al., 2019; Kitano, 2002).

Même si les principes propres à la biologie des systèmes sont maintenant utilisés dans de nombreux domaines de recherche, le but le plus fondamental et ambitieux de cette discipline

demeure cependant la compréhension absolue des principes et contraintes à la base de la vie (Breitling, 2010). Si l'on se base sur la célèbre citation du physicien Richard Feynman : *"what I cannot create, I do not understand"*, la démonstration ultime de l'atteinte de cet objectif résidera dans notre capacité à créer des organismes entièrement nouveaux strictement à partir de nos connaissances sur les systèmes biologiques (Elowitz and Lim, 2010). Ces organismes synthétiques pourraient très bien devenir réalité grâce aux technologies récentes de synthèse et d'assemblage de fragments d'ADN ainsi qu'aux techniques de manipulation des génomes complets, un nouveau domaine de recherche hautement intégratif portant le nom de génomique synthétique (voir section 1.4.2) (Montague et al., 2012; Schindler et al., 2018; van der Sloot and Tyers, 2017).

1.3.1 Les modèles métaboliques à l'échelle du génome

Les approches mathématiques et computationnelles utilisées pour modéliser le comportement des systèmes biologiques sont aussi nombreuses et diverses que ces derniers. Les approches d'analyse de reconstruction basées sur les contraintes (COBRA) représentent toutefois l'une des méthodes les plus reconnues en biologie des systèmes afin de modéliser le fonctionnement des cellules (Becker et al., 2007; Ebrahim et al., 2013; Lloyd et al., 2018). Les méthodes COBRA sont notamment utilisées pour la reconstruction de modèles métaboliques à l'échelle du génome (GEMs) (Durot et al., 2009; Gu et al., 2019; O'Brien et al., 2015; Oberhardt et al., 2009). Les GEMs constituent une manière efficace de regrouper et représenter l'ensemble des connaissances génomiques, biochimiques et métaboliques propres à un organisme donné. De plus, ces modèles sont particulièrement puissants pour intégrer les données de génomique fonctionnelle dans le contexte de la physiologie cellulaire et générer des prédictions à partir de l'information génomique disponible (Bordbar et al., 2014; Ebrahim et al., 2016; Gu et al., 2019; Joyce and Palsson, 2006; Kim et al., 2016).

Brièvement, les GEMs consistent en une structure mathématique ordonnée décrivant le métabolisme des organismes dans laquelle l'ensemble des réactions métaboliques est représenté

sous forme d'une matrice stœchiométrique (Figure 1.6). La relation entre les réactions et les métabolites est représentée sous forme de vecteurs de flux dans cette matrice, et l'ensemble du réseau est assumé comme étant à l'équilibre, c'est-à-dire que le produit des vecteurs avec la matrice est égal à zéro. Puisque plusieurs combinaisons de flux peuvent toutefois remplir cette condition, il existe alors un espace de solutions plutôt qu'une solution unique à cette équation. Si l'on se replace dans le contexte de la cellule, cela reviendrait à dire que celle-ci possède la faculté d'adapter les flux associés aux différentes parties de son réseau métabolique en fonction des conditions environnementales rencontrées (température, nutriments, stress, etc.) en conservant tout de même l'ensemble du réseau dans un état d'équilibre dynamique, un des principes à la base de la robustesse des cellules. Toutefois, puisque les cellules et les réactions métaboliques opèrent sous tout un lot de contraintes physico-chimiques, cinétiques et environnementales, il existe un nombre fini de solutions possibles aux conditions et paramètres définis dans le modèle, dont certaines s'avèrent plus optimales que d'autres pour la cellule. Il est donc possible, avec les GEMs, de prédire l'impact de différentes contraintes sur la physiologie des cellules (p. ex. limitations nutritionnelles, délétions géniques, etc.) et ainsi déterminer les solutions les plus optimales pour maximiser un certain phénotype (le taux de croissance, par exemple). Ce type de prédiction est généralement effectué grâce aux approches de programmation linéaire telles que le *flux balance analysis* (FBA) (Orth, Jeffrey D., Ines Thiele, 2010). Afin de générer des prédictions fiables, les modèles doivent cependant reposer sur des données expérimentales les plus complètes et précises possibles. La prédiction des taux de croissance, par exemple, demande de définir les principales macromolécules de la cellule (ADN, ARN, protéines, lipides, glucides), ainsi que différents composants cellulaires tels que les métabolites, les coenzymes et les ions inorganiques (Feist and Palsson, 2010; Lachance et al., 2019b). Les données générées en génomique fonctionnelle (voir section 1.2) sont particulièrement utiles dans ce contexte puisqu'elles permettent d'interroger un grand nombre d'éléments simultanément et ainsi définir des contraintes s'appliquant à l'ensemble du réseau métabolique, comme les vitesses réactionnelles des enzymes ou le niveau d'expression de protéines clés dans le métabolisme.

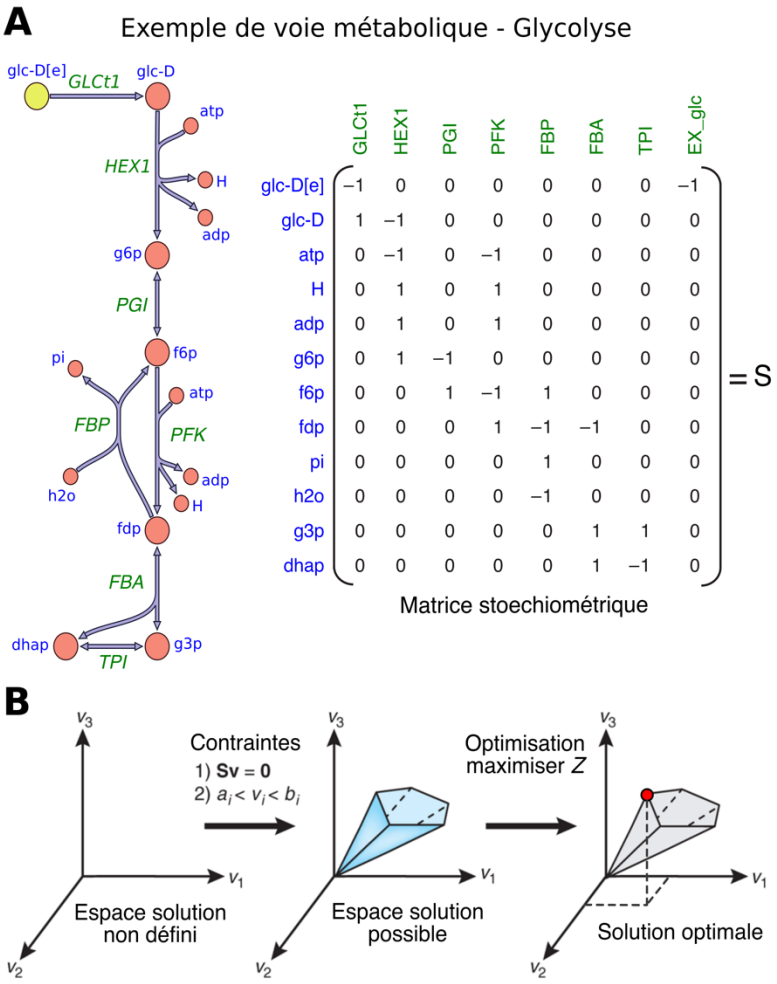

Figure 1.6. Description de l'approche COBRA utilisée dans les GEMs. A) Exemple de représentation mathématique des réseaux métaboliques sous forme d'une matrice stœchiométrique [S]. Chacune des colonnes de cette matrice représente une réaction métabolique, alors que chaque rangée désigne un métabolite. Ces métabolites peuvent soit être consommés (-1), produits (1) ou inutilisés (0) par ces différentes réactions métaboliques, créant ainsi les vecteurs [v] de cette matrice. La voie de la glycolyse est montrée ici en exemple, mais l'ensemble des réseaux cellulaires peuvent être représentés sous cette forme. B) Puisque l'ensemble des réactions incluses dans les modèles doivent forcément obéir aux règles physico-chimiques, l'ensemble du réseau se trouve alors à l'équilibre, c'est-à-dire que le produit des vecteurs de flux [v] avec la matrice [S] est égal à 0. Toutefois, plusieurs combinaisons de flux peuvent résulter en un état d'équilibre du système, ce qui crée un espace de solutions plutôt qu'une solution unique. Puisque les cellules opèrent sous différentes contraintes (physico-chimiques, cinétiques, environnementales, etc.) et pressions sélectives, cet espace peut être restreint aux solutions dites possibles ainsi qu'optimisé pour maximiser un certain phénotype (*Z*), le taux de croissance par exemple. Figure adaptée de (Becker et al., 2007).

Depuis le développement du tout premier GEM en 1999 pour la bactérie *Haemophilus influenzae* RD (Edwards and Palsson, 1999; Fleischmann et al., 1995), ce type de modèle a été reconstruit pour plus de 180 organismes différents, sans compter les modèles reconstruits à l'aide d'algorithmes automatiques (Gu et al., 2019; Karlsen et al., 2018; Machado et al., 2018). Ces modèles sont désormais disponibles pour la plupart des organismes couramment étudiés, incluant *E. coli*, *Saccharomyces cerevisiae* et l'humain (Gu et al., 2019). Les GEMs sont aujourd'hui utilisés dans un large éventail d'applications, notamment pour la création de souches dédiées aux processus industriels, l'étude des communautés microbiennes et la découverte de nouveaux antibiotiques (Gu et al., 2019). Alors que les techniques de génomique fonctionnelle ne cessent de se perfectionner, la quantité et la précision des données biologiques intégrées dans ces modèles vont assurément continuer d'augmenter. Ceci permettra la modélisation de processus cellulaires au-delà du métabolisme et l'utilisation des GEMs dans de nouveaux secteurs d'applications. L'intégration récente des mécanismes responsables de la synthèse de la machinerie d'expression des gènes en est un bon exemple (Lerman et al., 2012; Lloyd et al., 2018; Thiele et al., 2012). Ce type de modèle (*ME-model*) permet d'expliquer jusqu'à 80 % de la masse en protéines des cellules. En intégrant de plus en plus de processus cellulaires et de données biologiques de haute qualité, les GEMs pourraient rapidement devenir des outils très intéressants dans le contexte de la génomique synthétique (Chalkley et al., 2019; Rees et al., 2018). En effet, malgré le potentiel incomparable de ce domaine de recherche émergent, les outils actuellement disponibles afin d'évaluer de manière systématique les configurations génomiques proposées sont pratiquement inexistants. L'utilisation de GEMs pour prédire l'impact de réorganisations génomiques sur la cellule constituerait alors un avantage considérable par rapport aux méthodes actuellement utilisées pour la programmation des génomes.

1.3.2 Les appareils de culture en continu

Afin de générer des prédictions phénotypiques, les techniques de FBA utilisées dans les GEMs doivent assumer que l'ensemble des flux associés aux réactions métaboliques de la cellule se

trouvent à l'équilibre, c'est-à-dire dans un état physiologique appelé *steady-state* (Feist and Palsson, 2010; Orth, Jeffrey D., Ines Thiele, 2010). Alors que les GEMs constituent une façon efficace pour intégrer différents types de données biologiques dans un contexte de physiologie cellulaire, spécialement les données à l'échelle du génome, il est important de produire ces données dans des conditions de culture qui reflètent le mieux possible cet état d'équilibre métabolique. Contrairement aux *batch cultures* conventionnelles qui produisent des variations environnementales complexes et difficiles à modéliser, les appareils de culture en continu représentent un bon moyen pour recréer cet état d'équilibre cellulaire. En effet, ces appareils permettent d'établir des conditions de croissance hautement contrôlées et reproductibles, dans lesquelles la plupart des paramètres physico-chimiques du milieu demeurent stables sur de longues périodes (Bull, 2010; Gresham and Dunham, 2014; Hoskisson and Hobbs, 2005; Winder and Lanthaler, 2011). Ces conditions de croissance particulières contribuent à réduire les fluctuations physiologiques des cellules comme le taux de croissance et ainsi générer des données plus facilement interprétables et comparables entre expériences. Ceci est particulièrement important dans le contexte de la biologie des systèmes, puisque l'identité des gènes essentiels ainsi que la composition intracellulaire en ARN, protéines et métabolites sont grandement influencées par les conditions physiologiques des cellules (Klumpp et al., 2009; Lalanne et al., 2018; Minato et al., 2019).

Il existe aujourd'hui une variété impressionnante d'appareils de culture en continu, possédant tous leurs caractéristiques spécifiques et leurs utilités propres, allant de la culture de cellules uniques à la croissance en chambres pouvant accueillir plusieurs litres de culture (de Crécy et al., 2007; Gopalakrishnan et al., 2019; Guarino et al., 2019; Hoffmann et al., 2017; Lee et al., 2011; Long et al., 2013; Markx et al., 1991; Matteau et al., 2015; McGeachy et al., 2019; Miller et al., 2013; Moffitt et al., 2012; Skelding et al., 2018; Takahashi et al., 2015; Toprak et al., 2013; Zhang et al., 2006). Toutefois, leur principe général d'opération demeure sensiblement pareil d'un système à l'autre, peu importe qu'ils soient destinés à usage commercial, industriel ou académique. Celui-ci consiste essentiellement à ajouter du milieu de culture frais à la chambre de croissance afin de diluer les cellules et les déchets métaboliques au fur et à mesure

qu'ils s'accumulent. La principale différence réside ensuite dans la nature des paramètres utilisés afin de provoquer l'ajout de milieu frais dans le système. Deux grandes catégories d'appareils sont alors distinguables, les chemostats et les turbidostats, quoique d'autres types de systèmes ont également été développés au fil des années. Les chemostats, pour « *chemical environment is static* », maintiennent les cellules dans un état de *steady-state* en diluant les cultures à un taux fixe grâce à l'ajout de milieu limité pour certains nutriments clés, généralement le carbone, l'azote, le phosphore ou le soufre, ce qui contraint les cellules à croître à un rythme bien spécifique (Bull, 2010; Gresham and Dunham, 2014; Hoskisson and Hobbs, 2005; Winder and Lanthaler, 2011). Les turbidostats, de leur côté, procèdent à l'ajout de milieu frais en réponse à une augmentation de la densité optique de la culture, ce qui permet de maintenir les cellules dans un intervalle étroit de concentrations cellulaires. Puisqu'aucune limitation en nutriments n'est appliquée, l'utilisation des turbidostats est généralement privilégiée au détriment des chemostats dans les expériences où les cellules doivent proliférer à leur taux maximal de croissance.

Développés au tout début des années 1950, les premiers appareils de culture en continu furent à l'origine de plusieurs décennies de recherche innovante concernant la physiologie cellulaire (Bull, 2010; Monod, 1950; Novick and Szilard, 1950). L'utilisation de cette méthode de culture s'est ensuite graduellement estompée au profit de la biologie moléculaire, la génomique et autres technologies à l'échelle du génome denses en information. Même si les *batch cultures* demeurent encore aujourd'hui la méthode la plus utilisée grâce à sa simplicité et son faible coût, on assiste présentement à une recrudescence de la culture en continu en raison des multiples avantages qu'elle offre dans divers domaines de recherche émergents comme la biologie des systèmes et la biologie synthétique (Bull, 2010; Gresham and Dunham, 2014; Guarino et al., 2019; Hoskisson and Hobbs, 2005; McGeachy et al., 2019; Takahashi et al., 2015; Winder and Lanthaler, 2011). En plus de faciliter l'interprétation des données à l'échelle du génome, les appareils de culture en continu sont particulièrement utiles pour la réalisation d'expériences d'évolution (accélérées ou non) en laboratoire et possèdent un fort potentiel pour la production de molécules à haute valeur industrielle (de Crécy et al., 2007; Dragosits and Mattanovich, 2013;

Gopalakrishnan et al., 2019; Gresham and Dunham, 2014; Peebo and Neubauer, 2018). De plus, une quantité croissante de systèmes simples, abordables, flexibles et adaptés aux conditions de laboratoire sont désormais décrits dans la littérature, dont certains avec un niveau élevé de détails, ce qui va certainement encourager la communauté scientifique à adopter la méthode de culture en continu pour leurs expériences à venir.

1.4 La biologie synthétique

Les cellules vivantes représentent des systèmes moléculaires uniques possédant un niveau d'organisation et une complexité sans pareil. Cette complexité leur permet d'accomplir l'ensemble des tâches à la base de la vie, notamment de croître, se répliquer et répondre à leur environnement, mais également d'exécuter une variété de fonctions pratiquement infinie. Aussi incroyable qu'il puisse paraître, l'ensemble de ces fonctions se trouvent encodées sous forme d'une ou plusieurs molécules d'ADN dans la cellule, l'équivalent du code source d'un programme que la cellule exécute et transmet à ses descendants. Cette caractéristique fait des cellules des substrats uniques et extrêmement puissants pour l'ingénierie, puisque la simple modification de l'information génétique contenue dans l'ADN est suffisante pour reprogrammer l'ensemble de leurs fonctions. Évidemment, ce genre de travaux ne serait pas possible sans les techniques modernes de biologie moléculaire, mais requiert également une compréhension approfondie du fonctionnement des cellules et l'utilisation de concepts à la base des disciplines d'ingénierie. Le domaine de recherche unifiant ces aspects, appelé la biologie synthétique, vise à transformer la biologie moléculaire en une discipline robuste, prévisible et standardisée dans le but de créer ou modifier des systèmes biologiques ayant de nouvelles propriétés intéressantes (Andrianantoandro et al., 2006; Cheng and Lu, 2012; Endy, 2005; Lokody, 2013). Ces propriétés peuvent servir d'outils afin de mieux comprendre les systèmes biologiques naturels, mais ont également le potentiel d'accomplir diverses tâches utiles à l'humain. En effet, grâce à la biologie synthétique, il est possible d'imaginer la programmation d'organismes synthétiques capables de produire des biocarburants de manière efficace, de synthétiser des antibiotiques de manière moins coûteuse, de détecter et dégrader des produits toxiques ou de traiter certaines

maladies telles que le cancer et le diabète (Alper et al., 2010; Khalil and Collins, 2010). Ce domaine de recherche multidisciplinaire émergent est donc appelé à jouer un rôle prédominant dans le développement de nouvelles technologies visant à s'attaquer à certains des défis les plus importants du 21e siècle.

La biologie synthétique prend racine dans certains travaux réalisés au début des années 2000 décrivant la création de circuits génétiques artificiels organisés de manière à reproduire le comportement d'un interrupteur et d'un oscillateur, deux composantes clés à la base du fonctionnement des circuits développés en génie électrique (Cameron et al., 2014; Elowitz and Leibler, 2000; Gardner et al., 2000). Pour ce faire, ces circuits génétiques utilisaient la combinaison de différents éléments génétiques simples tels que des promoteurs, activateurs, répresseurs et terminateurs, soit l'équivalent des résistors, transistors et condensateurs utilisés en génie électrique. Puisque tout comme en génie électrique, ces pièces sont à la base de la construction des circuits complexes, il s'en est suivi de nombreux efforts visant à développer, caractériser, standardiser, distribuer et assembler une grande variété de composantes génétiques simples utiles pour la conception des circuits génétiques (Alper et al., 2010; Cameron et al., 2014; Khalil and Collins, 2010). Ces pièces caractérisées furent non seulement employées pour la fabrication d'interrupteurs et d'oscillateurs génétiques variés (Danino et al., 2010; Fung et al., 2005; Ham et al., 2008; Isaacs et al., 2003; Stricker et al., 2008; Tigges et al., 2009), mais également pour la création de circuits agissant à titre de portes logiques (Guet et al., 2002; Rinaudo et al., 2007; Win and Smolke, 2008), filtres de signaux (Basu et al., 2005; Hooshangi et al., 2005; Sohka et al., 2009) et biosenseurs génétiques (Bayer and Smolke, 2005; Kobayashi and Kærn, 2004). Certains circuits génétiques légèrement plus complexes ont également été développés à partir de pièces semblables, comme des circuits capables de contrôler la densité cellulaire, de répondre à la lumière, d'envahir des cellules cancéreuses ou de cibler certaines maladies infectieuses (Alper et al., 2010; Cameron et al., 2014; Khalil and Collins, 2010). Même si ces efforts de programmation génétique constituaient des accomplissements fort remarquables dans le domaine, il n'en demeurait pas moins que la majorité des circuits développés présentaient une complexité relativement limitée, n'impliquant souvent qu'un faible nombre

d'acteurs. De plus, la plupart des circuits génétiques artificiels développés en biologie synthétique se concluaient soit par un échec, un rendement sous-optimal ou un comportement imprévu (Kwok, 2010). En fait, notre habileté à programmer des comportements cellulaires prévisibles était et demeure encore très restreinte, principalement parce que la complexité des organismes modèles couramment utilisés en biologie synthétique (*E. coli*, *S. cerevisiae*) surpasse nos capacités d'analyse, de compréhension et de modélisation globale des cellules (Gardner, 2013; Knight, 2005; Lu et al., 2009). En effet, même pour *E. coli*, le GEM le plus complet produit à ce jour (*i*JL1678-ME) ne comprend qu'environ un tiers des gènes prédits chez ce microorganisme (Lloyd et al., 2018). Par conséquent, les règles fondamentales qui gouvernent le fonctionnement global des cellules demeurent mal caractérisées, et la majorité des circuits génétiques artificiels sont développés à l'aide d'une approche laborieuse, coûteuse et reposant souvent sur des décisions arbitraires appuyées que par très peu d'évidences expérimentales (Arkin, 2008; Endy, 2005).

Face à cette problématique, il devenu rapidement évident que des organismes aux génomes simplifiés ou minimaux, sélectionnés pour être beaucoup plus faciles à comprendre dans leur intégralité et mieux adaptés aux outils retrouvés en laboratoire, pourraient constituer des plateformes de prototypage intéressantes pour la programmation de circuits génétiques artificiels (Gibson and Benders, 2008; Lachance et al., 2019a; Morowitz, 1984; Xavier et al., 2014; Zhang et al., 2010). Dans ce contexte, deux types d'approches sont généralement préconisées afin d'obtenir de tels organismes : le *bottom-up* et le *top-down* (Fritz et al., 2010; Jewett and Forster, 2010). Brièvement, le *bottom-up* consiste à synthétiser une cellule minimale en choisissant et en assemblant correctement l'ensemble des molécules biologiques jugées comme essentielles à la réplication cellulaire autonome à l'intérieur d'une bicouche lipidique artificielle. À l'inverse, le *top-down* vise plutôt à identifier et ultimement retirer tous les éléments non essentiels du génome d'un microorganisme simple de sorte à ne conserver que ceux nécessaires au maintien de la vie minimale. Avec suffisamment de connaissances et d'outils, ces châssis cellulaires simplifiés ou minimaux ont le potentiel de transformer la biologie synthétique en une discipline d'ingénierie robuste et prévisible, et ainsi accélérer le

développement d'organismes programmés sur mesure pour accomplir des tâches bien spécifiques.

1.4.1 Les génomes minimaux

Suite à la publication des premiers génomes complets à la fin des années 1990, il devint théoriquement possible de définir l'identité des gènes essentiels à la vie, c'est-à-dire reformuler la proverbiale question « qu'est-ce que la vie? » par « qu'est-ce qu'un ensemble minimal de gènes essentiels? » (Fleischmann et al., 1995; Fraser et al., 1995; Mushegian and Koonin, 1996). En fait, le concept du génome minimal remonte bien avant l'avènement de la génomique, et provient principalement de l'étude de certaines des formes de vie autonomes les plus simples connues à ce jour, les Mollicutes (voir section 1.5) (Glass et al., 2017; Morowitz, 1984). Le génome de *Mycoplasma genitalium*, par exemple, ne contient qu'environ 580 kb et 500 séquences codantes, faisant de cette bactérie un candidat extrêmement intéressant pour étudier les règles fondamentales de la vie (Fraser et al., 1995). Malgré cette simplicité remarquable, il s'est avéré non longtemps après la publication de la séquence génomique complète de *M. genitalium* que celle-ci comprenait une proportion importante (~20 %) de gènes non essentiels à sa croissance en laboratoire (Glass, 2006; Hutchison et al., 1999). Ceci laissait ainsi supposer que des génomes encore plus simples pourraient en fait encoder l'ensemble des fonctions essentielles à la vie cellulaire autonome. Stimulés par l'apparition des technologies NGS et l'émergence de la biologie synthétique, plusieurs projets visant à définir l'identité des gènes essentiels chez différents organismes ont ainsi été mis sur pied au début des années 2000, ayant bien entendu comme objectif final la création de la toute première cellule minimale artificielle capable de réplication autonome en conditions de laboratoire (Mushegian, 1999; Peterson and Fraser, 2001).

Certes, la création des premiers organismes minimaux ne constituerait pas une finalité en soi, mais bien le point de départ pour le développement de châssis cellulaires spécifiquement conçus pour étudier les règles qui définissent les génomes et faciliter les efforts effectués en biologie

synthétique. Effectivement, avec les organismes minimaux, le nombre d'éléments à interroger et modéliser avec les approches de la génomique fonctionnelle et de la biologie des systèmes diminue considérablement, ce qui rend la caractérisation exhaustive de leurs mécanismes cellulaires plus facilement réalisable comparativement aux organismes modèles conventionnels (Esvelt and Wang, 2013; Lachance et al., 2019a; Xavier et al., 2014). Une fois atteinte, cette compréhension approfondie permettra de prévoir avec plus de précision les interactions entre les circuits artificiels à implanter et les mécanismes cellulaires naturellement présents chez l'organisme en question. De plus, puisque seuls les mécanismes essentiels ou importants sont conservés dans les génomes minimaux, les probabilités d'interactions non désirées parfois responsables de la défaillance des circuits seront également significativement réduites. Sélectionnés pour être beaucoup plus faciles à comprendre dans leur intégralité et mieux adaptés aux outils moléculaires retrouvés en laboratoire, ces organismes contribueront à transformer le processus laborieux de programmation des circuits génétiques synthétiques en une approche beaucoup plus rationnelle et efficace, accélérant ainsi le développement d'une multitude d'applications biotechnologiques. Évidemment, cette simplicité génomique se fera au détriment de la robustesse de l'organisme utilisé comme châssis cellulaire, limitant ce dernier à des conditions environnementales hautement spécifiques et contrôlées. Une fois que les règles et les mécanismes cellulaires à la base du fonctionnement des cellules seront bien compris, les niveaux de complexité contribuant à la robustesse des cellules, par exemple, pourront être incorporés puis décortiqués afin de conférer au châssis cellulaire minimal les propriétés et caractéristiques désirées. Idéalement, le châssis cellulaire utilisé pour cette tâche devrait être sécuritaire, facilement programmable, prévisible, modulable, stable génétiquement, mais surtout facilement débogable si des problèmes survenaient au niveau de ses circuits génétiques (Cambray et al., 2011).

Jusqu'à maintenant, plusieurs études ont cherché à déterminer le contenu en gènes essentiels de microorganismes, incluant *S. cerevisiae* (Dow et al., 2002), *E. coli* (Baba et al., 2006; Goodall et al., 2018), *Bacillus subtilis* (Commichau et al., 2013; Kobayashi et al., 2003), *M. genitalium* (Glass, 2006; Hutchison et al., 1999), *Mycoplasma pneumoniae* (Lluch-Senar et al., 2015) et

Mycoplasma mycoides (Hutchison et al., 2016). Différentes approches sont possibles afin d'y parvenir, mais elles ne sont pas nécessairement équivalentes (Juhas et al., 2011). Il faut tout d'abord préciser que le caractère essentiel d'un gène est intimement lié à son contexte génétique et aux conditions environnementales rencontrées par l'organisme. Parmi les techniques les plus utilisées, on compte très certainement les analyses de génomique comparative, la mutagenèse aléatoire par transposons ainsi que les méthodes de délétion génique. Brièvement, la génomique comparative consiste à comparer la séquence génomique d'organismes plus ou moins apparentés afin d'en identifier les gènes conservés (Koonin, 2003). Lorsque ces organismes sont phylogénétiquement très rapprochés, des souches d'une même espèce par exemple, les gènes conservés ont de fortes probabilités d'être essentiels pour la survie de ceux-ci, du moins en conditions très similaires à celles rencontrées dans leur habitat naturel. Ces gènes peuvent toutefois s'avérer facultatifs en conditions de laboratoire, principalement parce que ces conditions sont beaucoup plus stables et contrôlées que celles rencontrées dans l'environnement, et que plusieurs phénomènes tels que la compétition inter-espèces ou la restriction nutritionnelle y sont absents. L'approche de la génomique comparative a d'ailleurs été utilisée afin de comparer les deux premiers génomes bactériens séquencés, celui de *H. influenzae* et *M. genitalium*, mettant ainsi en évidence 256 gènes conservés entre les deux espèces (Mushegian and Koonin, 1996). La mutagenèse aléatoire par transposons, de son côté, utilise la capacité naturelle des transposons de s'intégrer dans le génome et causer par le fait même une interruption dans la séquence nucléotidique du locus d'insertion (Chao et al., 2016). Si le transposon s'intègre à l'intérieur de la séquence codante d'un gène, ce dernier aura de très fortes chances d'être inactivé, excepté certains cas très particuliers où l'expression et la fonction du gène interrompu sont maintenues. Avec cette méthode, l'ensemble du génome est criblé de transposons, à raison d'environ une insertion par cellule, et seulement les cellules ne possédant pas d'insertion dans un gène essentiel survivent. Les sites d'insertion sont ensuite séquencés, généralement à l'aide des approches de NGS, puis les gènes n'affichant aucune ou très peu d'insertions sont jugés comme essentiel pour la croissance. Il faut toutefois être prudent avec la notion d'essentialité, car un gène non essentiel ne veut pas nécessairement dire qu'il n'est pas important. Les gènes impliqués dans la fiabilité de la réplication de l'ADN, par exemple, ne sont pas essentiels à strictement parler, mais leur absence entraînera une hausse du taux de mutations

par génération, ce qui n'est pas réellement souhaitable dans le contexte du développement d'un châssis cellulaire réduit pour la biologie synthétique. De plus, certains gènes peuvent être interrompus individuellement, mais leur inactivation combinée résultera en la mort de la cellule, un phénomène portant le nom de létalité synthétique. Toutefois, ces interactions génétiques ne se limitent pas seulement à des interactions négatives; un gène jugé essentiel peut s'avérer non essentiel lorsqu'interrompu conjointement à un autre gène (viabilité synthétique). Pour ces raisons, les expériences de mutagenèse aléatoire par transposons sont parfois complémentées par différents types d'analyses, comme par exemple les analyses de génomique comparative, ce qui permet d'identifier les gènes à la fois non conservés et non essentiels. Cette catégorie est particulièrement propice à la délétion dans l'établissement d'un génome simplifié ou minimal.

Au lieu de procéder à des analyses de génomique comparative ou à des essais de mutagenèse par transposons, certains groupes de recherche ont tout simplement tenté d'effectuer la délétion individuelle de chacun des gènes prédits chez un organisme. Plusieurs techniques moléculaires peuvent être utilisées afin d'effectuer ce genre de manipulation génétique, mais la plus utilisée demeure de loin celle du *recombineering*. Cette méthode utilise l'opéron λ-Red du phage λ afin de stimuler la recombinaison d'un fragment d'ADN à un locus génomique précis lors du processus de réplication de l'ADNg (Datsenko and Wanner, 2000). La technique du *recombineering* fut notamment employée afin de systématiquement retirer chacun des gènes prédits d'*E. coli*, ce qui a permis de générer une banque impressionnante de près de 4000 mutants individuels (*Keio collection*) et identifier un noyau de 303 gènes essentiels (Baba et al., 2006). Ce nombre représente cependant une sous-estimation du nombre réel de gènes essentiels chez *E. coli*, car tout comme la mutagenèse aléatoire par transposons, la délétion individuelle des gènes ne tient pas compte du phénomène de létalité synthétique. Cette lacune peut toutefois être contournée en accumulant les délétions géniques à l'intérieur d'une seule et même souche, ce qui permet d'obtenir des organismes aux génomes significativement réduits. En utilisant cette stratégie, certains groupes de recherche sont parvenus à créer des souches d'*E. coli* possédant des génomes considérablement réduits, éliminant jusqu'à ~39 % de son génome original (pour une taille finale d'environ 2,8 millions de pb) (Hashimoto et al., 2005; Hirokawa et al., 2013;

Iwadate et al., 2011; Pósfai et al., 2006). Une approche similaire a également permis de générer des souches de *B. subtilis* possédant des génomes jusqu'à ~36 % plus petits que leur génome d'origine, pour une taille finale d'environ 2,7 millions de pb (Morimoto et al., 2008; ReuB et al., 2017). Alors que ces travaux impressionnants représentent des accomplissements majeurs dans le domaine de la biologie synthétique, ces génomes, malgré leur réduction importante, sont encore très loin de la taille du génome de *M. genitalium* (~580 kb) et donc d'un génome soi-disant minimal. De plus, l'approche de délétions cumulatives employée pour générer ces souches peut s'avérer excessivement laborieuse, est difficilement automatisable et chaque ronde de délétion subséquente augmente les risques d'acquisition et d'accumulation de mutations dans le génome.

Les approches basées sur la génomique synthétique offrent toutefois de nouvelles opportunités en ce qui a trait au développement de châssis cellulaires réduits ou minimaux (Montague et al., 2012; Schindler et al., 2018). Celles-ci furent notamment à l'origine de la création de la toute première cellule artificielle dite minimale (JCVI-syn3.0) (Hutchison et al., 2016). Basée sur le génome de *M. mycoides* sous-espèce *capri,* cette bactérie synthétique possède un génome de seulement 531 kb, ce qui en fait l'organisme à réplication autonome le plus simple connu à ce jour.

1.4.2 La génomique synthétique

Depuis la toute première synthèse *in vitro* d'un gène complet (Agarwal et al., 1970) et le développement de la technique du PCR (Saiki et al., 1985), les techniques de synthèse de molécules d'ADN synthétiques ont grandement évolué en termes d'efficacité et de capacité (Hughes and Ellington, 2017; Schindler et al., 2018). Alors qu'au début des années 2000 les oligonucléotides commandés sur mesure dépassaient rarement la centaine de paires de bases, les progrès réalisés au cours des dernières années offrent maintenant la possibilité de commander des fragments d'ADN synthétiques dépassant les 10 kb, et ce, à des coûts relativement abordables (Baker, 2011). De plus, le développement récent de nouvelles méthodes

d'assemblage de fragments d'ADN rend désormais la synthèse et l'assemblage de chromosomes entiers beaucoup plus envisageable qu'auparavant, une discipline émergente portant le nom de génomique synthétique (Baby et al., 2019; Montague et al., 2012; Schindler et al., 2018). Il est même devenu possible de synthétiser des molécules d'ADN comportant des analogues artificiels aux nucléotides naturels, ouvrant ainsi la porte à la création d'un code génétique augmenté pour des applications en biologie synthétique (Georgiadis et al., 2015; Hoshika et al., 2019).

La génomique synthétique tire ses origines au début des années 2000 avec la synthèse de génomes viraux (Baby et al., 2019; Schindler et al., 2018). En raison de leur taille restreinte (Mahmoudabadi and Phillips, 2018), ces génomes représentèrent des matrices idéales pour les premiers efforts de synthèse de génomes complets (Tableau 1.1). La première démonstration de synthèse complète d'un génome viral fut celle du virus de la polio, réalisée en 2002 (Cello et al., 2002). Mesurant 7,5 kb, le génome de ce virus fut synthétisé à l'aide d'une approche d'assemblage hiérarchique qui prit plusieurs mois à accomplir. Celle-ci consistait tout d'abord en la synthèse de fragments d'ADN mesurant environ 400-600 pb à partir de courts oligonucléotides (~70 pb) en utilisant une technique similaire au PCR appelée le *polymerase chain assembly* (PCA) (TerMaat et al., 2009). Les amplicons obtenus étaient ensuite clonés individuellement, validés, puis assemblés en fragments plus longs avant d'être clonés et validés une deuxième fois pour l'assemblage final. Même si cette preuve de concept reposait sur une démarche plutôt laborieuse, celle-ci permit de générer les toutes premières particules virales infectieuses à partir d'un génome entièrement synthétique, un accomplissement qui n'est certainement pas passé inaperçu compte tenu de son énorme potentiel pour la recherche fondamentale et pour le développement d'applications thérapeutiques (Cello et al., 2002). Cette approche d'assemblage hiérarchique fut d'ailleurs nettement améliorée et courant des années qui suivirent et permis de synthétiser plusieurs génomes viraux supplémentaires, dont certains mesurant plus de 100 kb (Tableau 1.1), mais également le tout premier chromosome bactérien synthétique réalisé en 2008 par l'équipe de John Craig Venter (Gibson and Benders, 2008). Pratiquement identique au génome de la bactérie quasi minimale *M. genitalium* (G37), la synthèse de ce chromosome circulaire a nécessité l'assemblage hiérarchique de 101 cassettes

Tableau 1.1. Tableau récapitulatif des projets de chromosomes et génomes synthétiques terminés jusqu'à présent. Adapté de (Baby et al., 2019; Schindler et al., 2018).

Année	Espèce	Taille en kb	% du génome original	Référence
Génomes synthétiques viraux				
2002	Virus de la Polio PV1 (M)	~7,5	-	(Cello et al., 2002)
2003	Bactériophage φX174	5,4	-	(Smith et al., 2003)
2005	Virus de la grippe espagnole de 1918 (H1N1 influenza A)	~13,0	-	(Tumpey et al., 2005)
2007	Rétrovirus endogène humain (HERV-K) – séquence consensus	9,5	-	(Young and Bieniasz, 2007)
2008	Coronavirus associé au syndrome respiratoire aigu sévère (SRAS) de la chauve-souris (Bat-SCoV)	29,7	-	(Becker et al., 2008)
2011	Virus de l'immunodéficience humaine (VIH-1-C)	11,0	-	(Nauwelaers et al., 2011)
2011	Bactériophage G4	5,6	-	(Yang et al., 2011)
2012	Virus de l'hépatite C sous-type 1a	9,6	-	(Munshaw et al., 2012)
2012	Bactériophage S13-like	5,4	-	(Liu et al., 2012b)
2013	*Chikungunya virus* (CHIKV)	11,9	-	(Scholte et al., 2013)
2014	Virus de la mosaïque du tabac	6,4	-	(Cooper, 2014)
2014	*Grapevine Algerian latent virus* (GALV-Nf)	4,7	-	(Lovato et al., 2014)
2015	Virus du syndrome reproducteur et respiratoire du porc (PRRSV)	15,4	-	(Vu et al., 2015)
2017	*Autographa californica nucleopolyhedrovirus* (AcMNPV)	145,3		(Shang et al., 2017)
2018	Virus de la vaccine (*horsepox*)	212,8	-	(Noyce et al., 2018)
Génomes synthétiques bactériens				
2008	*Mycoplasma genitalium* G37	583,0	-	(Gibson and Benders, 2008)
2010	*Mycoplasma mycoides* sous-espèce *capri* GM12 (JCVI-syn1.0)	1077,9	-	(Gibson et al., 2010a)
2016	*Mycoplasma mycoides* JCVI-syn3.0	531,5	49,3%	(Hutchison et al., 2016)
2019	*Caulobacter ethensis-2.0* (*C. crescentus*)	785,7	19,4%	(Venetz et al., 2019)
2019	*Escherichia coli* Syn61	3978,9	85,8%[(a)]	(Fredens et al., 2019)
Génomes synthétiques d'organelles eucaryotes				
2010	ADN mitochondrial de la souris (*Mus musculus*)	16,3	-	(Gibson et al., 2010b)
Génomes synthétiques eucaryotes				
2011	*Saccharomyces cerevisiae* synIXR	91,0	101,9%[(b)]	(Dymond et al., 2011)

Tableau 1.1. Tableau récapitulatif des projets de chromosomes et génomes synthétiques terminés jusqu'à présent (suite). Adapté de (Baby et al., 2019; Schindler et al., 2018).

2012	*Saccharomyces cerevisiae* synIII	272,2	86,2%	(Annaluru et al., 2014)
2017	*Saccharomyces cerevisiae* synII	770,1	94,7%	(Shen et al., 2017)
2017	*Saccharomyces cerevisiae* synV	536,0	92,9%	(Xie et al., 2017)
2017	*Saccharomyces cerevisiae* synVI	242,7	89,9%	(Mitchell et al., 2017)
2017	*Saccharomyces cerevisiae* synX	707,5	94,9%	(Wu et al., 2017)
2017	*Saccharomyces cerevisiae* synXII	999,4	92,7%[(c)]	(Zhang et al., 2017)

[(a)] Par rapport à la souche K-12 MG1655 originale.

[(b)] La taille du chromosome synIXR est légèrement plus grande que la version sauvage en raison de l'insertion de sites de recombinaison loxPsym.

[(c)] Les tailles indiquées ne tiennent pas compte des multiples copies de l'opéron de l'ADN ribosomal du chromosome XII.

mesurant entre 5 et 7 kb chacune, pour un produit final mesurant près de 583 kb. Alors que les étapes initiales d'assemblage furent accomplies dans *E. coli*, les étapes avancées et l'assemblage final ont été réalisés dans la levure *S. cerevisiae* sous forme de plasmides centromériques. Puisque le génome de *M. genitalium* était à ce moment-là le plus petit génome jamais observé dans une bactérie autonome (Fraser et al., 1995), celui-ci représentait le candidat idéal pour un tel effort de synthèse. De plus, les analyses de génomique comparative et les données de mutagenèse par transposons précédemment générées pour *M. genitalium* permettaient déjà d'envisager la création d'un châssis cellulaire minimal à partir de ce génome (Glass, 2006; Hutchison et al., 1999). Ceci n'eut toutefois jamais lieu, de même que la démonstration que ce chromosome circulaire synthétique basé sur *M. genitalium* était viable. En effet, l'isolation du génome complet de *M. genitalium* cloné dans la levure suivie de sa transformation dans une bactérie réceptrice, procédure portant le nom de transplantation de génome, n'aura jamais été réussie (Gibson and Benders, 2008; Labroussaa et al., 2019). Par le biais de sélection, cette procédure complexe permet de recueillir des organismes possédant le génotype et le phénotype conférés par le génome transplanté. Pour l'instant, encore très peu de choses sont connues au sujet de la transplantation de génome, mais il semblerait que la distance phylogénétique entre la bactérie réceptrice et le génome à transplanter jouerait un rôle déterminant dans la réussite de la méthode (Baby et al., 2017; Labroussaa et al., 2016, 2019).

Devant cet échec, l'approche de synthèse utilisée par le groupe de Venter fut reprise quelques années plus tard afin de synthétiser le génome d'un autre mycoplasme, *M. mycoides* sous-espèce *capri* (Gibson et al., 2010a). Même si *M. mycoides capri* possède un génome considérablement plus grand que celui de *M. genitalium* (1085 kb vs 580 kb), celui-ci pouvait être transplanté avec succès dans une cellule réceptrice de *Mycoplasma capricolum* sous-espèce *capricolum* déficiente pour son système de restriction naturel (Lartigue et al., 2007). De plus, *M. mycoides* croit beaucoup plus rapidement que *M. genitalium* en laboratoire, ce qui facilite la tenue de ce type d'expérience. Quoi qu'il en soit, ce projet estimé à près de 40 millions de dollars se conclut avec la création de la première bactérie contrôlée par un génome entièrement synthétique, *M. mycoides* JCVI-syn1.0 (Gibson et al., 2010a; Sleator, 2010). De manière encore plus impressionnante, cette approche de synthèse hiérarchique fut répétée quelques années plus tard par le même groupe de recherche afin de générer la toute première cellule artificielle considérée comme minimale, *M. mycoides* JCVI-sy3.0 (Hutchison et al., 2016). Alors que le génome de la souche JCVI-syn1.0 consistait en une copie pratiquement exacte de celui de *M. mycoides capri*, le génome de JCVI-syn3.0 représentait pour sa part une réduction de près de 50 % de la taille du génome de cet organisme. Avec un génome totalisant seulement 531 kb, JCVI-syn3.0 représente l'organisme à réplication autonome le plus simple connu à ce jour. Ce véritable tour de force fut possible grâce à une adroite combinaison des approches *top-down* et *bottom-up*, principalement via l'accumulation d'une grande quantité de données d'essentialité des gènes obtenues par mutagenèse par transposons, le tout jumelé à plusieurs itérations de conception, construction et test du génome synthétique (Figure 1.7) (Hutchison et al., 2016).

Il est maintenant envisageable de synthétiser et d'assembler des génomes beaucoup plus volumineux que celui de *M. mycoides*. La construction récente d'une souche d'*E. coli* utilisant seulement 57 des 64 codons en est un bon exemple (Fredens et al., 2019), tout comme la création de souches de *S. cerevisiae* possédant des versions synthétiques de chacun des 16 chromosomes (Richardson et al., 2017), un projet d'envergure colossale impliquant la synthèse de plus de 11 millions de paires de bases (Tableau 1.1). Nous entrons donc dans une ère où la synthèse complète de génomes microbiens deviendra plus facile et courante. Cependant, notre capacité à

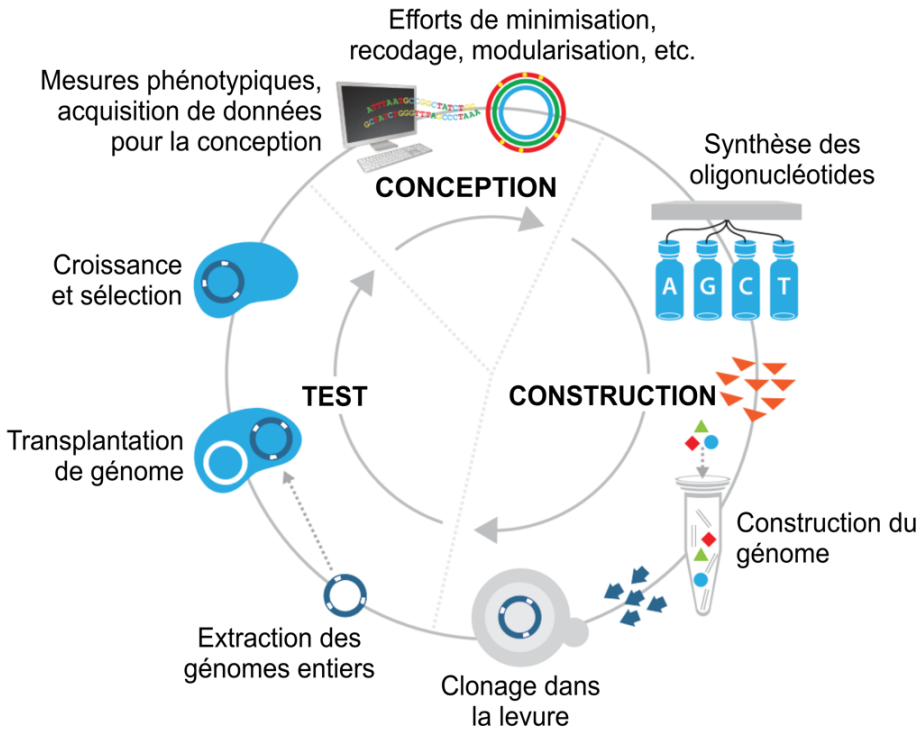

Figure 1.7. Résumé de l'approche de conception, construction et test utilisée afin de créer la bactérie minimale artificielle JCVI-syn3.0. Utilisée ici pour minimiser le génome de *M. mycoides capri*, cette approche peut toutefois être appliquée à des fins différentes (recodage, modularisation, etc.). Débutant avec les données d'essentialité des gènes obtenues chez la souche JCVI-syn1.0, une première étape de conception est effectuée afin de retirer les gènes non essentiels de segments représentant chacun un huitième du génome de *M. mycoides capri*. Chaque variant génomique est construit et assemblé à partir d'oligonucléotides, puis cloné sous forme d'un plasmide centromérique de levure. Le génome cloné est ensuite transplanté dans une cellule réceptrice (*M. capricolum*), puis analysé pour sa viabilité et ses caractéristiques phénotypiques. Cette approche est appliquée en parallèle pour les huit segments du génome de JCVI-syn1.0, puis en combinaison les uns avec les autres. La combinaison de segments présentant le plus petit génome et une croissance satisfaisante est utilisée pour les cycles de réduction suivants, et de nouvelles expériences d'essentialité des gènes sont réalisées à chacun des cycles afin de planifier les rondes suivantes. Figure adaptée de (Hutchison et al., 2016).

concevoir des génomes dont la composition ou l'organisation est inédite demeure extrêmement limitée (Baker, 2011; Knight, 2005; van der Sloot and Tyers, 2017; Wang, 2010). L'utilisation d'organismes simplifiés ou minimaux comme *M. mycoides* JCVI-sy3.0 représente une avenue particulièrement prometteuse dans ce contexte. En plus d'être plus faciles à comprendre dans leur intégralité, ces microorganismes ont une faible taille de génome, ce qui diminue le nombre

de combinaisons et de réorganisations génomiques possibles à tester par des approches de génomique synthétique, ainsi que les coûts associés à leur synthèse. Adroitement jumelés aux approches de la biologie des systèmes, ces organismes minimaux constitueront des outils puissants d'apprentissage afin de mieux comprendre les principes sous-tendant la conception des génomes et la programmation des circuits génétiques synthétiques (Schindler et al., 2018; van der Sloot and Tyers, 2017).

1.5 Les Mollicutes

Les bactéries appartenant à la classe des Mollicutes sont principalement caractérisées par leurs génomes exceptionnellement petits (~580-2200 kb), leur contenu en G-C particulièrement bas, leur absence de paroi cellulaire, ainsi que leur faible taille (~0,2-0,6 μm) (Dybvig and Voelker, 1996; Pettersson and Johansson, 2002; Sirand-Pugnet et al., 2007a). De plus, la plupart de ces organismes utilisent un code génétique alternatif (*Mycoplasma/Spiroplasma genetic code*) dans lequel le codon UGA normalement utilisé pour l'arrêt de la traduction signale plutôt l'incorporation d'un tryptophane à la chaîne peptidique en cours de synthèse (Navas-Castillo et al., 1992). Quoique très simples, ces bactéries ne représentent pas d'anciennes formes de vie primitives. Au contraire, celles-ci auraient évolué de bactéries à Gram positif à bas contenu en G-C via un processus de réduction génomique et de perte massive de gènes (Sirand-Pugnet et al., 2007a). Cette simplification de leur génome a également amené une baisse de leurs capacités métaboliques, avec plusieurs voies métaboliques manquantes ou incomplètes (Dybvig and Voelker, 1996; Pollack et al., 1997). La majorité des espèces de Mollicutes ont notamment adopté un style de vie parasitaire, infectant un bon nombre de plantes et d'animaux, dont l'humain. *M. pneumoniae*, par exemple, est estimé être responsable de plus de 40 % des cas de pneumonies humaines acquises en communauté (Bajantri et al., 2018; Waites and Talkington, 2004). À l'exception de espèces parasitaires obligatoires du genre *Phytoplasma* et des mycoplasmes hemotrophiques, ces microorganismes ne requièrent toutefois pas la présence de cellules hôtes pour croître en laboratoire, mais leur culture nécessite l'utilisation de milieux particulièrement riches pour pallier leurs déficiences métaboliques.

Dû à leur simplicité génomique remarquable, les Mollicutes ont depuis longtemps été proposées comme modèles pour l'étude des principes fondamentaux de la vie (Morowitz, 1984). Jadis étudiées principalement pour leur rôle causatif dans diverses maladies, ces bactéries attirent de plus en plus l'attention depuis quelques années en raison de leur potentiel en tant que châssis cellulaires simplifiés pour la biologie synthétique et la génomique synthétique. Le manque de plasmides naturels utilisables comme vecteurs de clonage et la faible quantité d'outils génétiques disponibles rendent toutefois la modification de leur génome relativement difficile avec les méthodologies conventionnelles (Sirand-Pugnet et al., 2007b, 2007a). Cette limitation peut être contournée grâce aux approches récentes de synthèse et d'assemblage de chromosome complets, comme démontré avec la création des bactéries artificielles JCVI-syn1.0 et JCVI-syn3.0 (Gibson et al., 2010a; Hutchison et al., 2016). Bien que puissantes, ces approches impliquent généralement des dépenses qui peuvent facilement excéder la barre du million de dollars, et ce, seulement pour les frais reliés à la synthèse chimique des chromosomes (Richardson et al., 2017; Sleator, 2010). Une stratégie alternative et moins dispendieuse consiste à isoler le génome natif de la bactérie, puis procéder à son clonage complet dans la levure *S. cerevisiae* (Karas et al., 2013). Initialement décrite pour *M. mycoides capri* (Lartigue et al., 2009), cette méthode a maintenant été utilisée avec plus d'une dizaine de génomes bactériens différents, dont la grande majorité font partie de la classe des Mollicutes (Labroussaa et al., 2019). Maintenus sous forme de plasmides centromériques, ces génomes peuvent ensuite être modifiés efficacement à l'aide des outils moléculaires disponibles dans la levure, par exemple le système d'endonucléase guidé CRISPR/Cas9 (Tsarmpopoulos et al., 2016). Jusqu'à présent, seuls quelques génomes de Mollicutes ont pu être transplantés avec succès de la levure vers une cellule réceptrice (Labroussaa et al., 2016, 2019).

Afin d'insérer les éléments génétiques nécessaires au clonage des génomes dans la levure, c'est-à-dire un centromère, une origine de réplication et un marqueur de sélection, plusieurs stratégies sont possibles. L'utilisation de transposons s'intégrant aléatoirement dans le génome est une façon simple d'y parvenir, et fut d'ailleurs l'approche employée pour cloner le génome naturel de *M. mycoides capri* (Benders et al., 2010; King and Dybvig, 1991; Lartigue et al., 2009). Ceci

peut également être accompli via le développement de plasmides artificiels basés sur l'origine de réplication du chromosome (*oriC*) (voir Chapitre 3). Suite à leur transformation, les plasmides *oriC* peuvent se répliquer grâce à l'interaction entre la protéine DnaA et des séquences spécifiques incluses dans l'*oriC* appelées boîtes DnaA (Lartigue et al., 2003; Messer, 2002). Leur tendance naturelle à recombiner avec le chromosome peut aussi être exploitée afin d'y introduire les éléments de réplication, partition et sélection de la levure. Cette méthode fut récemment démontrée par notre laboratoire avec la bactérie quasi minimale *Mesoplasma florum* (Baby et al., 2017), un Mollicute étroitement apparenté aux mycoplasmes du groupe de *M. mycoides* (*mycoides cluster*) (Figure 1.8). Pour les organismes hautement réfractaires au processus de transformation de courtes molécules d'ADN, comme *Mycoplasma hominis* par exemple, la méthode de clonage par recombinaison (*transformation-associated recombination*; TAR) peut aussi s'avérer une option intéressante (Rideau et al., 2017).

1.5.1 *Mesoplasma florum*

Outre *M. genitalium* et *M. mycoides capri*, d'autres espèces de Mollicutes représentent des modèles d'étude très intéressants pour la biologie des systèmes, la biologie synthétique et les efforts en génomique synthétique. *M. pneumoniae*, par exemple, a été le sujet de vastes études portant sur le métabolisme, le transcriptome et le protéome d'une cellule quasi minimale, ce qui en fait l'un des Mollicutes les mieux caractérisés à ce jour (Chen et al., 2016; Krishnakumar et al., 2010; Kühner et al., 2009; Lloréns-Rico et al., 2015; Lluch-Senar et al., 2015; Miravet-Verde et al., 2019; Wodke et al., 2013, 2015; Yus et al., 2009, 2019). Initialement décrite en 1984 sous le nom d'*Acholeplasma florum* (McCoy et al., 1984), la bactérie *M. florum* possède également plusieurs caractéristiques avantageuses dans ce contexte. Bien que phylogénétiquement très proche des mycoplasmes pathogéniques du groupe *mycoides* (Figure 1.8), *M. florum* ne possède aucun pouvoir pathogène connu, ce qui facilite sa manipulation en laboratoire et sa distribution à travers la communauté scientifique (Gupta et al., 2019; Iriarte et al., 2011; Labroussaa et al., 2016; Sirand-Pugnet et al., 2007a). De plus, celle-ci présente un taux de croissance plus rapide que celui de la plupart des mycoplasmes, et possède

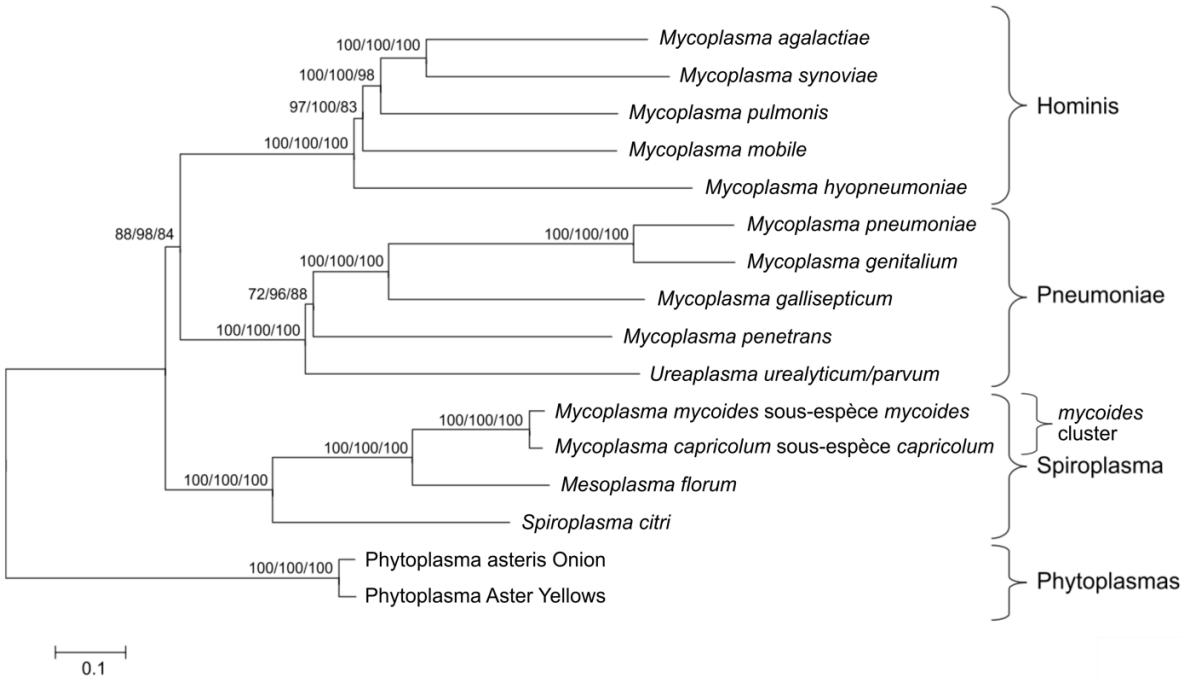

Figure 1.8. Arbre phylogénétique de la classe des Mollicutes. Cet arbre a été généré à partir de l'alignement de 91 protéines orthologues. Le nombre d'itérations confirmant la structure des embranchements représentés est indiqué (évalué à l'aide de trois méthodes différentes). L'échelle illustrée indique le nombre de substitutions par position. Figure modifiée de (Sirand-Pugnet et al., 2007a).

un génome plus petit que celui de *M. mycoides* et *M. capricolum* (Tableau 1.2). Le génome de la souche L1, isolée à partir d'une fleur de citronnier et séquencée en 2004 par une équipe du Broad Institute, totalise seulement 793 224 pb et 685 séquences codantes prédites, ce qui positionne cette bactérie parmi les organismes à réplication autonome les plus simples identifiés jusqu'à maintenant (Baby et al., 2018; McCoy et al., 1984). De plus, son génome ne contient qu'un seul facteur sigma prédit et seulement une dizaine de facteurs de transcription, ce qui indique un réseau de régulation relativement simple. Naturellement, *M. florum* possède aussi un métabolisme grandement simplifié (Pollack et al., 1997). L'absence des gènes dont les produits sont impliqués dans le cycle de l'acide citrique, par exemple, suggère que la production d'énergie de cet organisme repose presque entièrement sur la glycolyse et la fermentation. Par conséquent, sa croissance en laboratoire requiert l'utilisation de milieux de culture extrêmement riches composés notamment de sérum et d'extrait de levure, sans toutefois nécessiter

d'équipement spécialisé ou particulièrement dispendieux (Matteau et al., 2015, 2017). Tout comme la plupart des mycoplasmes, *M. florum* utilise le codon UGA pour l'incorporation du tryptophane lors de la traduction (*Mycoplasma/Spiroplasma genetic code*) (Iriarte et al., 2011; Navas-Castillo et al., 1992; Sirand-Pugnet et al., 2007a). De plus, la très faible fréquence du codon arginine CGG (<10 occurrences pour l'ensemble des séquences codantes prédites) et l'absence du gène codant pour l'ARN de transfert correspondant ($tRNA_{CCG}$) laisse supposer que ce codon n'est peut-être pas utilisé chez ce Mollicute, comme c'est le cas pour *M. mycoides capri*. Ces particularités ont pour effet de prévenir les échanges indésirables de matériel génétique entre *M. florum* et des espèces non apparentées, mais impliquent toutefois des considérations supplémentaires pour l'expression de protéines hétérologues dans ce microorganisme.

Tableau 1.2. Caractéristiques de différentes espèces de Mollicutes d'intérêt.

Espèce	Hôte naturel	Taille du génome (pb)	Nombre de gènes codants	Nombre de gènes essentiels	Numéro d'accession GenBank	Temps de doublement
M. florum L1	Insectes/ plantes	793 224	685	~290-466[a]	AE017263.1	~31-33 min[b]
M. mycoides sous-espèce *mycoides* SC PG1	Ruminant	1 211 703	1017	N.D.[c]	BX203980.2	N.D.
M. mycoides sous-espèce *capri* GM12	Ruminant	1 084 586	841	473[d]	CP001668.1	~60-80 min[d,e]
M. capricolum sous-espèce *capricolum* CK (ATCC 27343)	Ruminant	1 010 023	792	N.D.	CP000123.1	~100 min[e]
M. genitalium G37	Humain	580 076	509	382[f]	L43967.2	~8-16 hrs[g]
M. pneumoniae M129 (ATCC 29342)	Humain	816 394	691	342-435[h]	U00089.2	~8-60 hrs[i]

[a] (Baby et al., 2018).
[b] Voir Chapitre 4.
[c] Non disponible.
[d] (Breuer et al., 2019; Hutchison et al., 2016).
[e] (Lartigue et al., 2007).
[f] (Glass, 2006).
[g] (Karr et al., 2012; Sleator, 2010).
[h] (Lluch-Senar et al., 2015).
[i] (Wodke et al., 2013; Yus et al., 2009).

Pour l'ensemble de ces caractéristiques, *M. florum* apparaît comme un candidat de choix pour le développement d'un châssis cellulaire minimal dédié aux efforts de programmation de

génomes et de circuits génétiques artificiels. Le retrait des gènes déterminés comme dispensables en conditions de laboratoire constituerait un point de départ logique dans ce contexte (Tableau 1.2) (Baby et al., 2018). Même si une première cellule minimale artificielle a été créée (Hutchison et al., 2016), la souche JCVI-syn3.0 ne représente qu'une seule configuration minimale possible, fort probablement sous-optimale pour la cellule. En effet, celle-ci est caractérisée par un temps de doublement considérablement plus élevé que la souche de *M. mycoides capri* de départ (~180 min vs ~60 min), et ce malgré un génome deux fois plus petit à répliquer (Hutchison et al., 2016). Une multitude d'organisations alternatives sont possibles et pourraient être étudiées, autant à partir d'un même génome parental qu'entre chromosomes d'espèces distinctes. En effet, il a été montré récemment que 57 gènes déterminés comme essentiels chez *M. florum* n'ont pas d'homologue chez *M. mycoides* JCVI-syn3.0, ce qui suggère que différentes compositions de génomes minimaux existent probablement, et ce, même pour des espèces étroitement apparentées (Baby et al., 2018).

1.6 Objectifs et hypothèses du projet de recherche

Dans le laboratoire de Sébastien Rodrigue, nous pensons qu'il est possible de développer une plateforme simple, sécuritaire et hautement caractérisée dédiée au prototypage efficace de génomes synthétiques à l'aide d'une démarche intégrative visant à décortiquer le fonctionnement intégral de *M. florum.* Cette démarche repose principalement sur la production et l'analyse d'une grande quantité de données expérimentales obtenues à l'échelle du génome, incluant des données d'expression, d'essentialité et de conservation des gènes, le tout jumelé aux approches de la biologie des systèmes tels que les GEMs. En intégrant l'ensemble de ces données en contexte du métabolisme et de la physiologie cellulaire de *M. florum*, il sera beaucoup plus facile d'évaluer et de prédire avec précision l'impact de réorganisations génomiques sur le comportement global de la cellule. Cela permettra de minimiser le nombre de génomes intéressants à tester avec les approches de la génomique synthétique, et ainsi arriver plus efficacement à des compositions optimales selon les critères choisis. Bien entendu, cette démarche peut être appliquée de façon itérative afin de guider pas à pas vers les modifications

génomiques à effectuer selon le cas. En plus de faciliter le développement de circuits génétiques artificiels, cette démarche vise à mieux comprendre les règles qui définissent les génomes ainsi que les principes à la base de la programmation des organismes vivants.

Bien que *M. florum* présente déjà plusieurs caractéristiques intéressantes à titre de châssis cellulaire simplifié pour la programmation de génomes, la littérature portant spécifiquement sur la biologie de ce microorganisme est relativement restreinte comparée à plusieurs mycoplasmes. Effectivement, une simple recherche sur la base de données du *National Center for Biotechnology Information* (NCBI) est suffisante pour le constater ; en date d'écriture de ces lignes, une recherche en utilisant les termes « *Mesoplasma florum* ou *Acholeplasma florum* » retourne 25 résultats différents, alors que pour « *Mycoplasma pneumoniae* » ce nombre se situe au-delà de 6000. Ceci est en quelque sorte l'envers de la médaille de n'avoir jamais été associé à aucune maladie, contrairement aux mycoplasmes qui sont responsables de diverses maladies engendrant d'importantes pertes économiques (Egwu et al., 1996; Maes et al., 2018; McGowin and Anderson-Smits, 2011; Welte et al., 2012; Yan et al., 2016). Par conséquent, au commencement de mon doctorat, beaucoup d'aspects concernant la biologie de *M. florum* restaient à explorer afin d'arriver à une compréhension approfondie de son fonctionnement global. Par exemple, aucune donné quantitative d'expression des gènes de *M. florum* n'était disponible, et ce, même si son génome était séquencé depuis 2004. De plus, très peu d'évidences expérimentales appuyaient les connaissances sur la physiologie et le métabolisme spécifiques à cette bactérie, celles-ci étant fondées principalement sur des prédictions et des comparaisons avec d'autres espèces de Mollicutes. De manière encore plus critique, pratiquement aucun outil moléculaire ou méthode de transformation n'avait été validé comme étant fonctionnel chez *M. florum*. Ceci constituait une limitation technique très importante pour l'étude de ce microorganisme et pour la modification de son génome en vue du développement d'un châssis cellulaire réduit.

Face à cette problématique, je m'étais fixé trois objectifs principaux à atteindre au terme de mon doctorat :

1) Le développement d'un système de culture en continu simple, flexible et abordable. Cet appareil sera utilisé afin de faire croître *M. florum* dans des conditions de culture stables, contrôlées et reproductibles, et servira entre autres à diminuer les variations expérimentales lors de l'étude de ce microorganisme. Le développement de cet appareil est présenté au Chapitre 2.

2) Le développement des premiers plasmides spécialement conçus pour se répliquer chez *M. florum*. Basés sur l'*oriC*, ces plasmides serviront d'outils afin de valider la fonctionnalité de différents marqueurs de sélection aux antibiotiques, en plus de mettre au point plusieurs méthodes de transformation pour cette bactérie. Ces plasmides pourront ensuite être utilisés comme outils moléculaires afin de modifier le génome de *M. florum*. Cette partie de mon projet de recherche est décrit au Chapitre 3.

3) La caractérisation intégrative de *M. florum*. En combinant différentes approches et méthodes expérimentales, incluant notamment l'intégration de données de 5' -RACE, RNA-seq et MS/MS, cette caractérisation physique, physiologique et moléculaire de *M. florum* vise à obtenir une compréhension plus approfondie des mécanismes cellulaires globaux de cette bactérie. De plus, les données générées pourront servir de base expérimentale pour la reconstruction et la validation d'un GEM pour *M. florum*. Cet effort de caractérisation est présenté au Chapitre 4. Au cours de mon doctorat, cet objectif m'a également permis de développer une expertise en génomique, plus précisément dans la réalisation d'expériences de 5' -RACE, RNA-seq et ChIP-exo. Ceci m'a donné l'opportunité de collaborer avec l'équipe du Pr Vincent Burrus afin d'étudier les mécanismes de régulation du transfert d'éléments génétiques mobiles impliqués dans la dissémination de la résistance aux antibiotiques (Carraro et al., 2014, 2015; Poulin-Laprade et al., 2015). De plus, j'ai été invité à publier des protocoles détaillés concernant les méthodes du 5' -RACE et du ChIP-exo (Matteau and Rodrigue, 2015b, 2015a). L'information relative à ces méthodes, de même que celle relative aux publications découlant des collaborations, est présentée en annexe I.

L'ensemble de ces objectifs vise à acquérir les outils et les connaissances nécessaires afin d'entreprendre le développement d'un châssis cellulaire simplifié, sécuritaire et hautement

caractérisé spécifiquement conçu pour accélérer le développement de la biologie synthétique et pour faciliter les efforts en génomique synthétique.

CHAPITRE 2

A SMALL-VOLUME, LOW-COST, AND VERSATILE CONTINUOUS CULTURE DEVICE

2.1 Présentation de l'article et contributions

Contrairement aux *batch cultures* classiques qui produisent des changements environnementaux complexes et dynamiques, les appareils de culture en continu représentent d'excellents outils afin de maintenir les populations bactériennes dans un état d'équilibre physiologique appelé *steady-state* (Bull, 2010; Hoskisson and Hobbs, 2005; Winder and Lanthaler, 2011). L'atteinte de cet état d'équilibre est particulièrement importante dans le contexte de la biologie des systèmes puisque les conditions physiologiques de la cellule influencent directement (et indirectement) l'identité des gènes essentiels à sa croissance, ainsi que sa composition intracellulaire d'ARN, de protéines et de métabolites (Klumpp et al., 2009; Lalanne et al., 2018; Minato et al., 2019). En diminuant les fluctuations physiologiques des cellules, les appareils en culture en continu permettent de réduire les variations expérimentales et ainsi générer des données plus facilement interprétables et comparables entre expériences.

Grâce à ces avantages considérables, il s'est avéré évident dès le début de mon doctorat que nous voulions utiliser un système de culture en continu dans le contexte d'expériences impliquant *M. florum*. Cependant, relativement peu de modèles simples et abordables étaient à ce moment décrits dans la littérature, et la majorité d'entre eux présentaient un niveau de complexité élevé afin de répondre à des besoins très spécifiques. De plus, seulement quelques publications comportaient suffisamment de détails techniques afin de permettre leur fabrication et utilisation par les tiers (Esvelt et al., 2011; Miller et al., 2013; Takahashi et al., 2015; Toprak et al., 2013). D'autre part, la plupart des appareils disponibles commercialement étaient et sont encore aujourd'hui très coûteux, ne sont généralement pas conçus pour manipuler de faibles volumes, et manquent de flexibilité pour accommoder les besoins variés rencontrés en laboratoire. J'ai donc entrepris, au tout début de mon doctorat, le développement d'un système

de culture en continu à la fois simple, flexible, abordable, et *open-source* afin de répondre à cette problématique. Ce projet, initié par un ancien étudiant du laboratoire (Vincent Baby), a été le fruit d'une collaboration étroite et grandement appréciée entre les départements de biologie et de physique de l'Université de Sherbrooke. Malgré plusieurs défis et embûches rencontrés tout au long de sa conception, nous avons su tirer profit des forces et compétences propres à chaque domaine d'études afin de faire face à l'adversité et proposer un système simple, facile d'utilisation et offrant différents modes de fonctionnement pour répondre à divers besoins expérimentaux. Dans la configuration initialement proposée, le *Versatile Continuous Culture Device* (VCCD) offrait trois chambres de culture pouvant être contrôlées indépendamment, ainsi qu'un système de photodiodes et photorécepteurs permettant le suivi de croissance par turbidité ou activité métabolique (Matteau et al., 2015). De plus, différents modes de rafraîchissement des cultures étaient sélectionnables afin que le système agisse de manière équivalente à un chemostat ou un turbidostat. Nous avons aussi rendu accessible l'ensemble des informations nécessaires à la construction et au fonctionnement du VCCD pour que les utilisateurs potentiels puissent bénéficier du système.

Soumis initialement en mai 2015 à la revue *open-access PLoS ONE*, le manuscrit décrivant le système VCCD a été extrêmement bien accueilli par les évaluateurs. Ces derniers n'exigeant que quelques corrections mineures, sa publication suivit dans les mois suivants et fut très bien accueillie par la communauté scientifique. Cet article compte désormais plus de 20 citations à son actif. Par sa description détaillée et sa conception qui se veut hautement personnalisable, notre système a d'ailleurs inspiré plusieurs laboratoires à développer leur propre instrument de culture en continu et rendre accessibles les informations nécessaires à leur fabrication (Gopalakrishnan et al., 2019; Guarino et al., 2019; Hoffmann et al., 2017; McGeachy et al., 2019; Pilizota and Yang, 2018). Jusqu'à maintenant, notre système a été utilisé dans plusieurs contextes au laboratoire du Pr Sébastien Rodrigue, notamment pour mesurer la stabilité de plasmides développés pour *M. florum* (voir Chapitre 3) et pour quantifier le niveau d'expression des gènes de cette bactérie (voir Chapitre 4). Nous envisageons également utiliser cet appareil

dans plusieurs expériences à venir, par exemple pour le criblage des gènes essentiels à la croissance de *M. florum* en milieu défini.

L'article présenté dans ce chapitre représente l'aboutissement d'un travail collectif impliquant plusieurs personnes. Sous la supervision du Pr Sébastien Rodrigue, j'ai été chargé de concevoir le support physique et le système de culture du VCCD, en plus de réaliser les expériences de croissance microbienne pour valider le système. J'ai aussi assisté Stéphane Pelletier (département de physique) pour le développement du logiciel de contrôle de l'appareil et la conception des modules électroniques de ce dernier. Stéphane Pelletier a été chargé de fabriquer les modules électroniques inclus dans l'appareil. Benoît Couture et Frédéric Francoeur, aussi au département de physique, ont fabriqué les pièces du support physique et ont participé à la conception de celles-ci. Vincent Baby et le Pr Sébastien Rodrigue ont développé les versions prototypes du VCCD et ont contribué à la conception de la version actuelle. Thomas F. Knight et Marie-Eve Pepin ont également participé aux discussions concernant la conception du VCCD. Conjointement avec le Pr Sébastien Rodrigue, j'ai rédigé le manuel d'utilisation et le manuscrit du VCCD. Vincent Baby et Alain Lavigueur ont participé à la révision du manuscrit.

Référence bibliographique : Matteau, D., Baby, V., Pelletier, S., Rodrigue, S. (2015). A Small-Volume, Low-Cost, and Versatile Continuous Culture Device. PLoS ONE *10*, e0133384.

2.2 Title page

A small-volume, low-cost, and versatile continuous culture device

Dominick Matteau[1], Vincent Baby[1], Stéphane Pelletier[2] & Sébastien Rodrigue[1].

1- Département de biologie, Université de Sherbrooke, Sherbrooke, Québec, Canada.
2- Département de physique, Université de Sherbrooke. Sherbrooke, Québec, Canada.

Corresponding author: sebastien.rodrigue@usherbrooke.ca

Running title:
Customizable continuous culture device

2.3 Abstract

2.3.1 Background

Continuous culture devices can be used for various purposes such as establishing reproducible growth conditions or maintaining cell populations under a constant environment for long periods. However, commercially available instrument are expensive, were not designed to handle small volumes in the milliliter range, and can lack the flexibility required for the diverse experimental needs found in several laboratories.

2.3.2 Methodology/Principal Findings

We developed a versatile continuous culture system and provide detailed instructions as well as a graphical user interface software for potential users to assemble and operate their own instrument. Three culture chambers can be controlled simultaneously with the proposed configuration, and all components are readily available from various sources. We demonstrate that our continuous culture device can be used under different modes, and can easily be programmed to behave either as a turbidostat or chemostat. Addition of fresh medium to the culture vessel can be controlled by a real-time feedback loop or simply calibrated to deliver a defined volume. Furthermore, the selected light-emitting diode and photodetector enable the use of phenol red as a pH indicator, which can be used to indirectly monitor the bulk metabolic activity of a cell population rather than the turbidity.

2.3.3 Conclusions/Significance

This affordable and customizable system will constitute a useful tool in many areas of biology such as microbial ecology as well as systems and synthetic biology.

2.4 Introduction

Since their introduction in 1950 [1,2], continuous culture devices have been useful to study the biology of various species and were adopted in many industrial contexts such as brewing and waste treatment [3]. Their principle of operation essentially consists in using fresh medium to dilute cells and waste products as they accumulate in the growth chamber. The two main continuous culture devices are chemostats and turbidostats, but variants and other types of systems have been described [4-13]. Chemostats, for "chemical environment is static", are used to maintain cells in a physiological steady state through dilution of a liquid culture at a specified rate with a medium limited for specific nutrients, thus restraining cell growth to a fixed rate. In contrast, turbidostats maintain cell density by adding fresh medium in response to an increased optical density of a culture. Turbidity is constantly monitored and the culture is refreshed through a feedback control loop to constrain cell concentration within a narrow range. Turbidostats are more appropriate than chemostats for experiments that require cells to proliferate at their maximum growth rate, since no limitation of nutrients is applied.

Continuous culture conditions result in a stable and controllable set of physico-chemical conditions in which growth rate, pH, biomass, as well as the concentrations of dissolved oxygen, proteins, and metabolites reach a dynamic equilibrium and remain approximately constant over an extended period of time [3,9,14,15]. These highly-reproducible parameters contribute to the establishment of a physiological steady state in cell populations, which decreases variability in quantitative experiments such as RNA sequencing and mass spectrometry [14-17]. Systems biology that aims at analyzing and integrating large amounts of data to understand complex interactions between components of biological systems would thus particularly benefit from controlled and highly-reproducible growth parameters provided by continuous culture devices [14,18].

Despite of its multiple advantages, the use of continuous culture, more specifically in the context of systems biology, is not common. This is not surprising since commercially available

continuous culture instruments are expensive, not designed to handle small volumes, rarely customizable, and can lack the flexibility required to accommodate a large diversity of experimental conditions. As a consequence, most studies are thus conducted using batch cultures, a simple, easy to implement, low cost, and highly scalable cultivation procedure that however results in a continuously fluctuating chemical environment [17,19]. Some laboratories have developed continuous culture instruments for their specific needs. However, the design and utilization of these custom instruments is often limited to particular experimental conditions and/or specific types of organisms [3]. In addition, relatively few publications have described these devices, and even fewer provide detailed technical information along with a software to operate a device [7,20-22]. To address this issue, we have conceived and built the Versatile Continuous Culture Device (VCCD), a flexible, open-source, small-volume, and low-cost continuous cultivation system that can easily be programmed to act either as a turbidostat or chemostat. Here, we describe the features and capabilities of the VCCD. We also provide the required information for the construction, operation and modification of this system along with the source code of the associated software. We anticipate that the VCCD will be particularly useful in the fields of microbial ecology, functional genomics, synthetic biology, and any other discipline in which constant and controllable growth conditions are advantageous or crucial.

2.5 Materials and Methods

2.5.1 Strains and growth conditions

E. coli strain BW25113 was obtained from the Coli Genetic Stock Center (CGSC) (strain 7636) and was grown at 30°C in LB broth. *Mesoplasma florum* strain L1 (ATCC 33453) was grown at 34°C in ATCC1161 medium containing 200 U/mL penicillin. *Saccharomyces cerevisiae* strain VL6-48 (ATCC MYA-3666) was grown at 30°C in Yeast Extract-Peptone-Adenine-Dextrose (YPAD) medium with 2% glucose. All strains were preserved at -80°C in their respective growth medium containing 25% (vol/vol) glycerol. Batch cultures were grown from frozen stocks in an orbital shaker incubator and used to inoculate VCCD culture vessels (55 mL

PYREX tubes, see CS-L1 in Table S2.2). VCCD cultures were grown in the same conditions as described for batch cultures in a temperature-controlled chamber.

2.5.2 VCCD hardware

Complete construction, assembly and operation instructions of the VCCD are regrouped in Manual S2.1 (user manual). All frame, culture system, and electronics parts were purchased from various sources and are listed in Tables S2.1-S2.3, respectively. Reference codes for every VCCD components are listed in Tables S2.1-S2.3 and are used for part identification in Figs. S2.1-S2.8. VCCD Frame parts listed in Table S2.1 were designed using SolidWorks education edition 2014 and can be machined as described in Appendix S2.1 using standard academic or commercial fabrication services. Machined Frame parts are also available in 3D CAD file format in File S2.1. Frame and culture system assemblies are depicted in Figs. S2.1 and S2.2, respectively. Electronics schematic diagrams (Figs. S2.5 and S2.6) and Printed circuit boards (PCB) layouts (Figs. S2.7 and S2.8) were designed using the National Instruments (NI) Circuit Design Suite 11.0 software. PCBs were created using double-sided presensitized copper clad boards and a Kinsten KVB-30 UV exposure box according to manufacturer's specifications (see EL-A1, FR-O1 to O3, and FR-P1 to P3 in Table S2.3). Soldering of components through PCBs holes (see Table S2.3, Figs. S2.7 and S2.8) was done by hand using a Weller EC1002 soldering iron and Kester 44 Rosin Sn63Pb37 0.031″ diameter solder wire (Digi-Key KE1102-ND). Transmittance data acquisition is performed by a NI USB-6008 DAQ (see EL-C1 in Table S2.3) and reported to the VCCD software by a 3.0 USB port. Assembly of the electronics box is depicted in Figs S2.3 and S2.4. An example of the complete VCCD assembly is available in 3D CAD file format in File S2.2 (assembly without electronics details).

2.5.3 VCCD software

VCCD software (version 1.0) was designed using the NI LabVIEW Professional Development System 2009 SP1 software. Executable version of the software (VCCD 1.0.exe) and its

dependencies were compiled in an installation package using the LabVIEW 2009 application builder. The installation package and the source code of the software are available at: http://lab-rodrigue.recherche.usherbrooke.ca/VCCD_en/#Software_Download
Current version of the VCCD software was tested on Windows 7 Enterprise Service Pack 1 (64 bits).

2.5.4 VCCD calibration

VCCD calibration procedure is described in Manual S2.1. Briefly, the system was first calibrated by setting the 0% transmittance to correspond to total obscurity. LB broth, water or YPAD medium with 2% glucose were used to set the 100% transmittance value in experiments involving *E. coli*, *M. florum*, and *S. cerevisae* respectively.

2.5.5 Phenol red as a pH indicator for growth measurement

25 ml of ATCC 1161 medium was progressively acidified by adding increasing volumes of HCl 1N. Transmittance at 560nm was measured at 34°C using the VCCD system calibrated with water for 100% transmittance and total obscurity for 0% transmittance. Acquisition rate and averaging parameters were set to 200 msec and 5 data points, respectively. pH was measured using a VWR SB20 SympHony pH meter calibrated at 34°C.

2.5.6 Batch culture monitoring

30 mL of *E. coli*, *M. florum*, and *S. cerevisae* were separately grown in a VCCD culture chamber and 560nm transmittance was monitored using the VCCD system calibrated with the appropriate solutions. Growth of *E. coli* and *M. florum* was monitored with the acquisition rate and averaging parameters respectively set to 1 sec and 5 data points, while 5 sec and 1 data point were used for *S. cerevisae.* For the *E. coli* culture, absorbance at 600nm was also measured periodically by a GE Healthcare Ultrospec 2100 pro UV/Visible Spectrophotometer calibrated

with LB broth. To calculate cell concentrations, 10 µL aliquots were taken at different transmittance values throughout the experiment and then diluted serially with the appropriate culture medium. Dilutions were plated in triplicates on LB, ATCC 1161 with 200 U/mL penicillin or on YPAD 2% glucose plates for *E. coli*, *M. florum*, and *S. cerevisae* respectively. Plates were incubated at the appropriate temperature until colonies were visible. Colonies were counted and colony-forming units (CFU) were calculated according to the corresponding dilution.

2.5.7 Continuous culture experiments

20 mL of *E. coli*, *M. florum*, and *S. cerevisae* were grown in VCCD culture chambers and 560nm transmittance was monitored using the VCCD system calibrated with the appropriate solutions. Growth of *E. coli* and *S. cerevisae* was monitored with the acquisition rate and averaging parameters respectively set to 5 sec and 1 data point, while 0.5 sec and 2 data points were used for *M. florum*. *E. coli* and *S. cerevisae* cultures were refreshed using the Threshold-activated mode with a minimum transmittance set to 50% and a pinch time set to 3 sec for 1 cycle. *M. florum* culture was refreshed using the Real-time feedback loop mode with minimum and a maximum transmittance values set to 11.5% and 12%, respectively. For all experiments, the refresh flow rate was previously adjusted to approximately 1 mL/sec using sterilized water instead of fresh medium. Cell concentrations were calculated as described in the batch culture monitoring section.

2.6 Results

2.6.1 VCCD fabrication and principle of operation

We developed the VCCD, a low-cost and open-source device composed of three independently controlled continuous culture units mounted on a plexiglass acrylic frame (Fig. 2.1A, Figs. S2.1 and S2.2). Each culture unit includes a culture vessel whose turbidity is measured using a

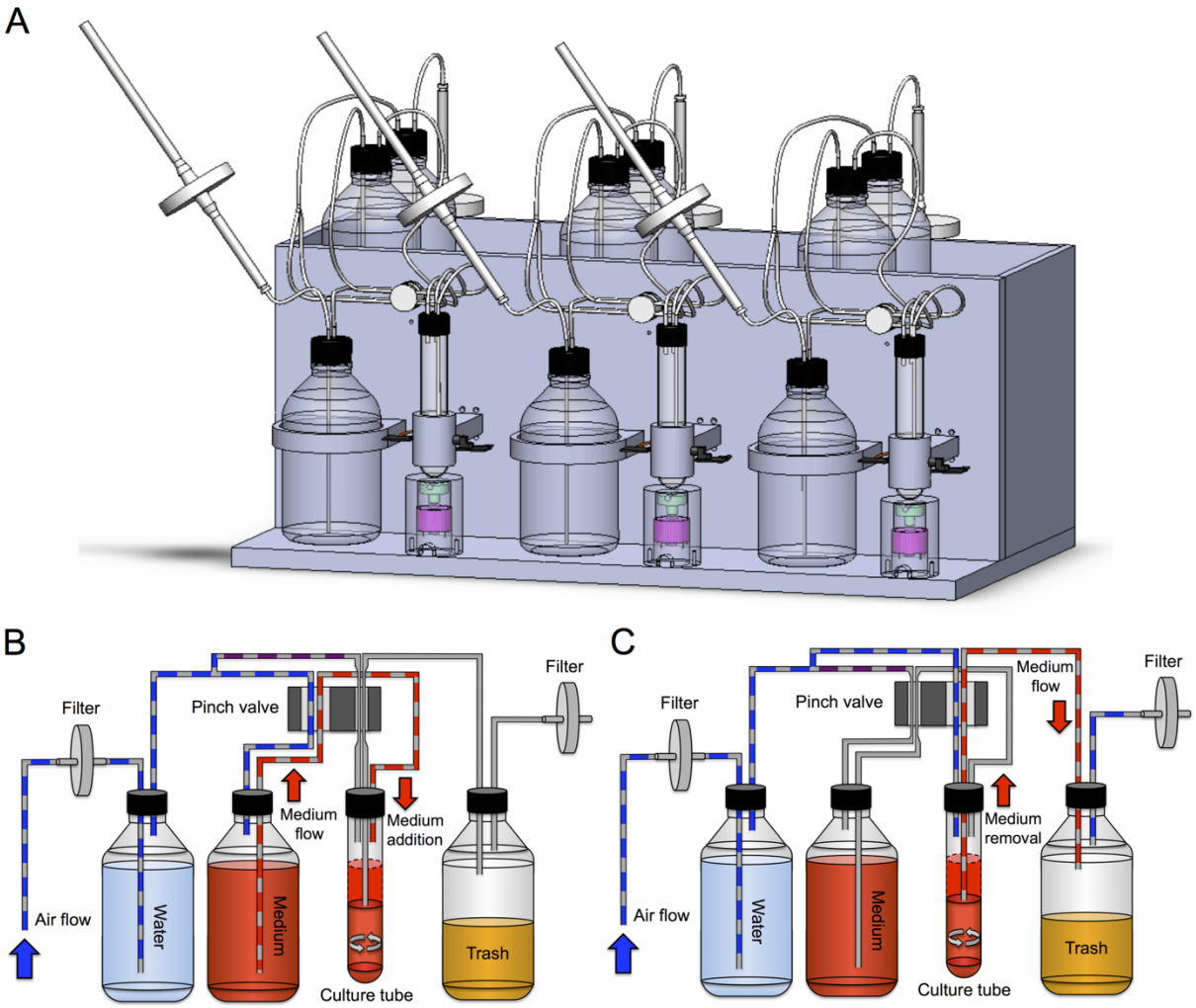

Figure 2.1. Hardware configuration and schematic depiction of culture-refreshing steps. (A) Three-dimensional representation of the versatile continuous cultivation device (VCCD). The system consists of three independently controlled continuous culture units supported by a plexiglass acrylic structure. For each unit, transmittance is measured through the culture chamber by a light-emitting diode coupled to a photo receiver and then reported to a user interface control system. The culture refreshing capability is provided by computer-controlled pinch valves that manage air and liquid flows inside each culture unit. (B) First step of a culture refresh cycle (culture dilution). Upon pinch valve activation, two tubes of the culture unit get pinched, and the air flow is diverted to the medium bottle resulting in the addition of medium into the culture tube. (C) Second step of a culture refresh cycle (excess culture removal). By returning the pinch valve to its original position, two tubes of the culture unit are pinched, which then redirects the air flow to the culture tube and causes the excess of volume to be evacuated into the trash bottle.

conventional 560nm light emitting diode (LED) paired with a photo receiver (PHR). Transmittance at 560nm is monitored in real-time while cultures are gently agitated with adjustable speed magnetic stir bars. To establish continuous cultures, silicone tubing connects each culture vessel to an adjustable air pump that provides a low air pressure sufficient for liquid displacement inside the system (Figs. 2.1A and B, Figs. S2.1 and S2.2). The culture vessel is connected between a bottle containing fresh medium (in which water saturated air is injected to displace the liquid when needed) and a trash bottle. The silicone tubing network passes through a four way computer-controlled pinch valve that, once specified conditions are reached, changes the air flow path resulting in culture dilution (Fig. 2.1B). When culture dilution is completed, the pinch valve returns to its original position and the excess volume is removed and disposed inside the trash bottle (Fig. 2.1C). The entire continuous culture unit is flanked by two 0.2 µm filters, and can be autoclaved for sterilization (see Manual S2.1). With the proposed frame configuration (see Appendix S2.1), the VCCD is most likely to be used in a temperature-controlled room, but could easily be adapted to fit in a common incubator.

Electrical alimentation of LEDs, PHRs, mixing motors and pinch valves are supported by low-cost PCBs that can be easily fabricated and assembled using standard procedures (Table S2.3, Figs. S2.3-S2.8). Transmittance measurement components were designed to use circular connectors so that different sets of LEDs and PHRs could be rapidly exchanged without requiring any specialized tool. Since the frequency of light emission of LEDs and the acquisition rate of PHRs are synchronized, the impact of ambient light on the PHR is nearly abolished, thus eliminating the need to use the VCCD in total obscurity to acquire accurate transmittance signals. Data acquisition is performed by a multifunction NI USB-6008 DAQ (see EL-C1 in Table S2.3) connected by a standard 3.0 USB port to a computer executing the VCCD software, which is used to display transmittance signal graphs and control culture dilution cycles.

In general, culture transmittance declines as the cell population grows and gradually blocks the incident light. This phenomenon is due to an increase in turbidity that is easily visible when following the growth of *E. coli* in LB broth (Fig. 2.2A). However, some microorganisms do not

possess the physiological characteristics required to proportionally increase culture opacity when growing. For instance, very small cells (<500 nm) such as mollicutes (e.g. *Mycoplasma sp* and *Mesoplasma sp*) do not absorb or scatter light efficiently, and as a consequence, are not properly quantified using a turbidity based approach [23,24]. Under these circumstances, our selected LED and PHR allow using phenol red as a growth medium pH indicator to monitor the acidification caused by metabolic activity and proliferation of cells. As pH decreases, the absorbance of phenol red at 560 nm drops (which results in a concomitant increase of the 560nm transmittance), and is visible through a color change of the medium from red to orange (Fig. 2.2B).

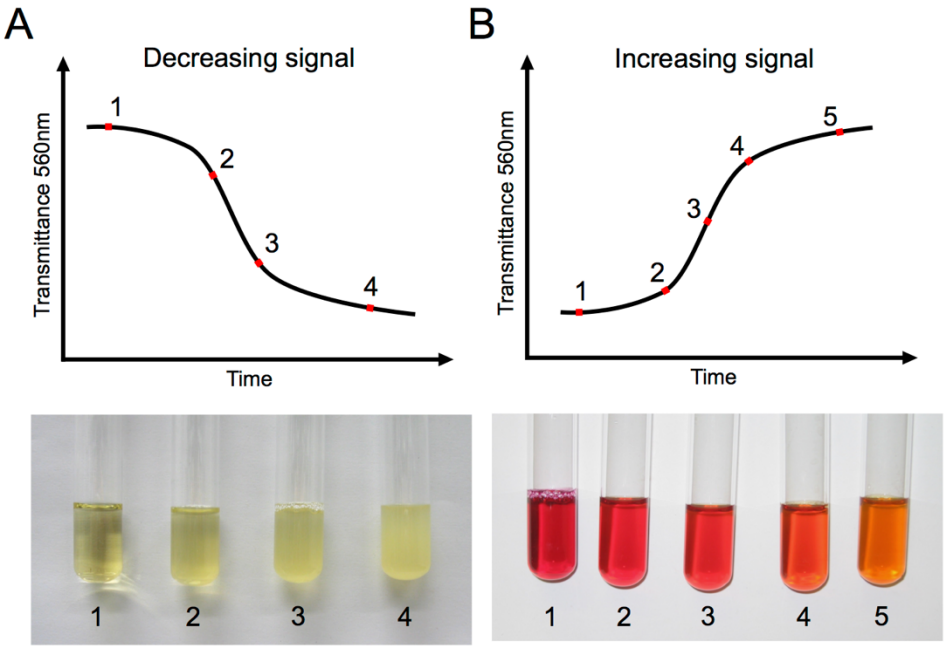

Figure 2.2. Typical transmittance curves associated with bacterial growth. (A) Increasing opacity observed when monitoring transmittance of an *E. coli* culture in LB broth, where the growing cells increasingly block the incident light generated by a light emitting diode. (B) In some cases, growth can alternatively be measured by adding phenol red to the growth medium as a pH indicator. This method is especially employed to follow the growth of *M. florum* in ATCC 1161 medium, where medium acidification caused by metabolic activity is reported by an increase in the 560nm transmittance.

2.6.2 The VCCD can measure the growth of various microorganisms

In order to use the VCCD to keep cell populations at a constant density for long periods, we first verified that the instrument could accurately measure the growth of different model microorganisms. Using the VCCD to monitor the growth of an *E. coli* BW25113 [25] batch culture in LB broth, we observed a decreasing 560nm transmittance signal as cells grew (Figs. S2.9A and S2.10A), and more importantly, we noticed a strong correlation according to the Beer-Lambert law between 560nm transmittance values acquired by our system and 600nm absorbance measured by a conventional spectrophotometer (Fig. 2.3A). This suggests that transmittance quantification by our LED-PHR approach offers performances comparable to commercially available instruments, even for measurements performed under ambient light. We also observed a strong linear correlation between log of cell concentrations and relative 560nm transmittance ranging from approximately 10^7 to 10^8 CFU/mL and 70% to 25%, respectively (Fig. 2.3B). This range of transmittance roughly corresponds to an OD_{600nm} signal between 0.2 and 0.5 (Fig. 2.3A), an optical density interval typically associated with *E. coli* exponential growth phase. A similar pattern was also observed with *S. cerevisiae* growing in YPAD medium with 2% glucose (Fig. 2.3C, Figs. S2.10C and S2.10D), clearly showing that the selected LED and PHR for the VCCD are suitable to follow the growth of commonly studied microorganisms.

We then sought to establish that our system is suitable to monitor the growth of microorganisms whose characteristics impose the use of phenol red as a growth indicator. To do so, we measured the transmittance at 560nm of ATCC 1161 medium employed to grow diverse bacteria belonging to the class of mollicutes, and observed that the color change caused by acidification of the medium was accurately detected by the VCCD. A linear correlation was observed between transmittance at 560nm and pH from ~6 to 7 (Fig. 2.3D). Next, we monitored the growth of a batch culture of *M. florum,* a mollicute closely related to mycoplasmas [26,27]. We observed that the transmittance signal followed a typical pattern for this bacterium (Figs. S2.9B and S2.10E), in which 560nm transmittance initially increases from about 8% to 16% and then decreases due to cell agglomeration. Importantly, medium acidification caused by metabolic

activity, as well as measured transmittance at 560nm, were very well correlated with cell concentrations from approximately 10^9 and 10^{10} CFU/mL (Figs. 2.3E and F), showing that our device is suitable to indirectly monitor the growth of microorganisms using phenol red.

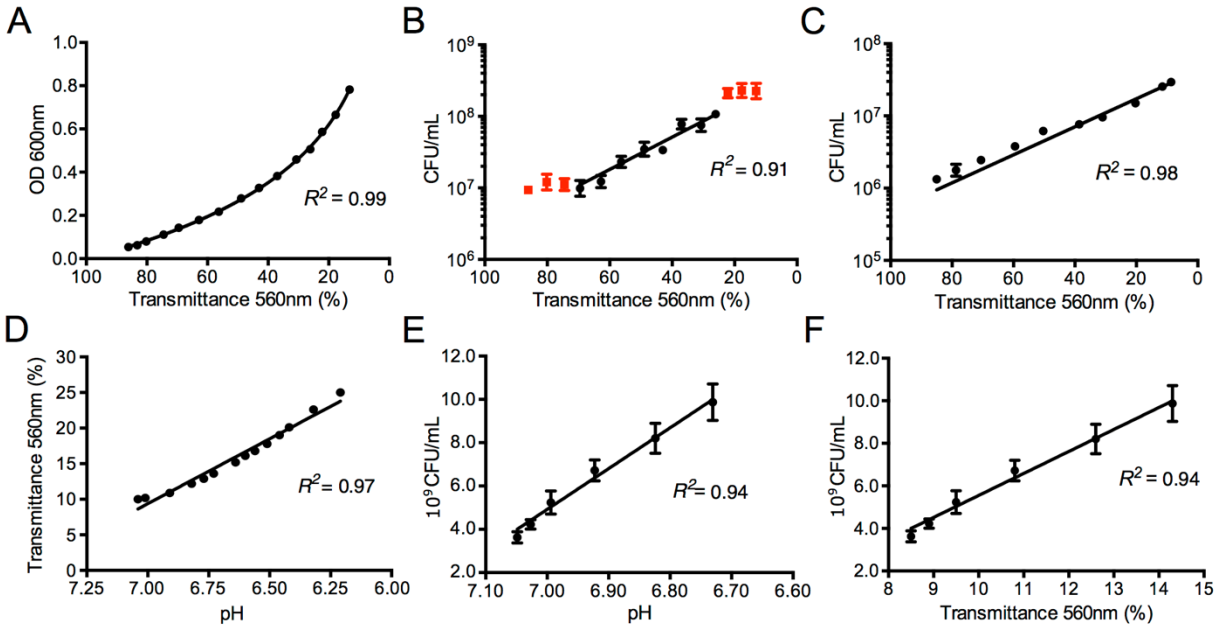

Figure 2.3. Calibration of the Versatile continuous culture device (VCCD) using batch cultures. (A) Comparison of 560nm transmittance measured by the VCCD and 600nm absorbance measured by a conventional spectrophotometer of an *E. coli* culture in LB broth. (B) Relationship between relative 560nm transmittance measured by the VCCD and cell density of an *E. coli* culture grown in LB broth. Red squares are excluded from the correlation determination because they are not part of the exponential growth phase. (C) Relationship between relative 560nm transmittance measured by the VCCD and cell density of *S. cerevisiae* growing in YPAD 2% glucose medium. (D) Relative 560nm transmittance of ATCC 1161 medium through different pH values generally observed during *M. florum* growth. (E) Cell density of *M. florum* growing in ATCC 1161 medium through pH decrease. (F) Relationship between relative 560nm transmittance measured by the VCCD and cell density of an *M. florum* culture grown in ATCC 1161 medium.

2.6.3 Available culture refresh modes

After establishing a strong correlation between cell density and measured 560nm transmittance for different microorganisms, we investigated whether the VCCD could maintain cell populations at a desired concentration for extended periods. To achieve this goal, we first

verified that our continuous culture system configuration could accurately dilute cell population to perform culture refreshing. As expected, volume of liquid added to the culture vessel was well correlated with pinch valve opening time (Fig. S2.11), at a flow rate of ~1 mL/sec in our assay. Because selected culture vessels have a maximal volume capacity of 55 mL, we designed our system to support cultures of about 20 mL to allow the possibility of diluting them in half in a single refresh cycle. If a greater dilution factor is required, multiple dilution cycles could be programmed with the software to avoid any overflow (see below).

In order to establish continuous cultures, we included different options and parameters in the VCCD software to accommodate various experimental parameters. Among those, one major setting that has to be considered before executing a continuous culture experiment is the desired general behavior of the system, which is defined according to the selected culture refresh mode (Fig. 2.4A). For example, the Real-time feedback loop mode makes the VCCD behave like a turbidostat, in which the culture is diluted by fresh medium in response to a specific transmittance threshold until a second threshold is reached. In that mode, a maximum culture refresh time (pinch time) can be set to avoid culture overflow during a refresh cycle, thus forcing the system to execute multiple dilution cycles to obtain the desired transmittance value (see Manual S2.1). If a chemostat behavior is more convenient for a specific experiment, the Time interval mode can be chosen to refresh the culture with a constant dilution rate. In that context, different options must be selected to specify whether the interval timer starts at a specified transmittance value or at a specified time, and if culture refreshing is stopped according to a defined pinch time or a transmittance threshold (Fig. 2.4B).

2.6.4 The VCCD can establish continuous cultures of different microorganisms

To verify if our VCCD modes can effectively maintain cell populations at a desired concentration, we conducted continuous culture experiments on the three microorganisms previously chosen for batch culture monitoring experiments, that is *E. coli*, *M. florum* and *S. cerevisiae*. When performing continuous culture experiments on *E. coli* and *M. florum* using the

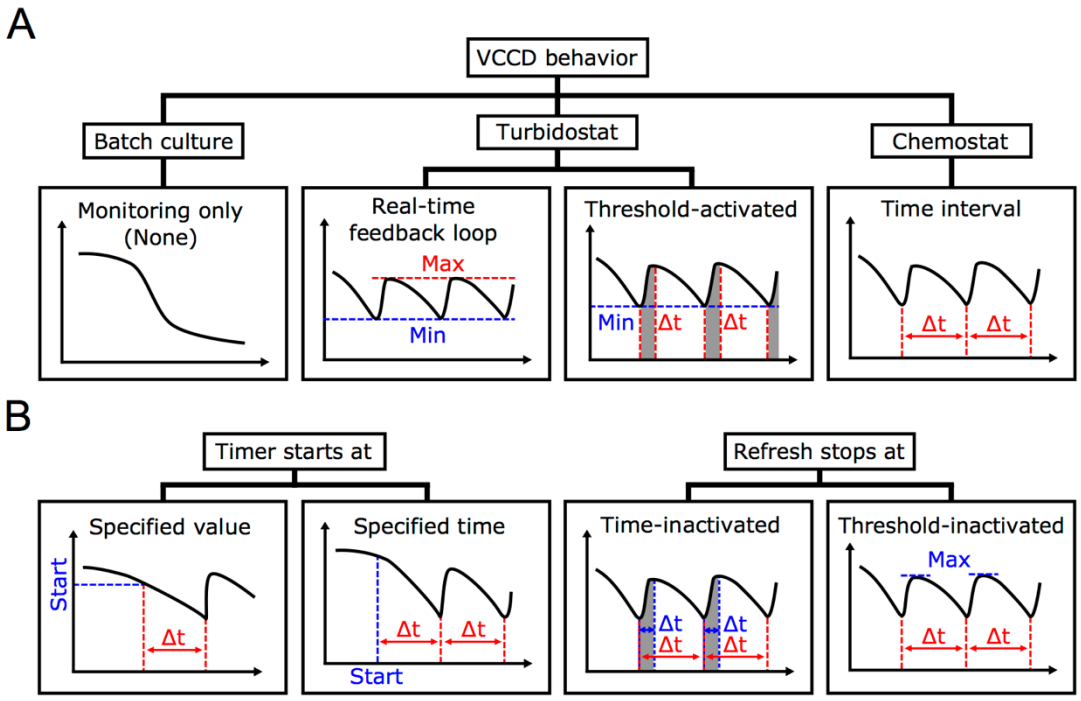

Figure 2.4. Illustration of available continuous culture modes used to maintain cell growth. (A) Under its present software configuration, the versatile continuous cultivation device (VCCD) can behave like a turbidostat or a chemostat, or simply be used to measure the transmittance of a batch culture without performing any culture refresh. In the turbidostat mode, the culture is refreshed when a desired transmittance value is detected until: 1) a second value is reached [Real-time feedback loop] or 2) a specified refresh time has elapsed [Threshold-activated]. Alternatively, the chemostat behavior uses a Time interval mode to refresh the culture with a constant specified dilution rate. (B) In the Time interval mode, additional options are available to choose if the interval timer starts at a specified value or at a specified time, and to decide if refreshes are stopped in a time-dependent manner or using a transmittance threshold.

Real-time feedback loop refresh mode, we observed that both cultures never exceeded the selected transmittance thresholds specified to the VCCD software (Figs. 2.5A and B), except when refreshes were manually disabled to let them grow to their maximal density. As expected, a very high stability in cell concentration was obtained for *M. florum* continuous culture experiments (Figs. 2.5C and D). *E. coli* continuous cultures using the Threshold-activated refresh mode provided similar results (Figs. 2.5E and F). Furthermore, the VCCD allowed us to maintain a *S. cerevisiae* continuous culture in tight range of cell concentrations for almost a week (Figs. 2.5G and F), and could be used to maintain cell density for longer periods. Taken

together, these results demonstrate that the VCCD is capable of reliably maintaining cell culture densities under a narrow and stable range for various microorganisms and under different operating modes.

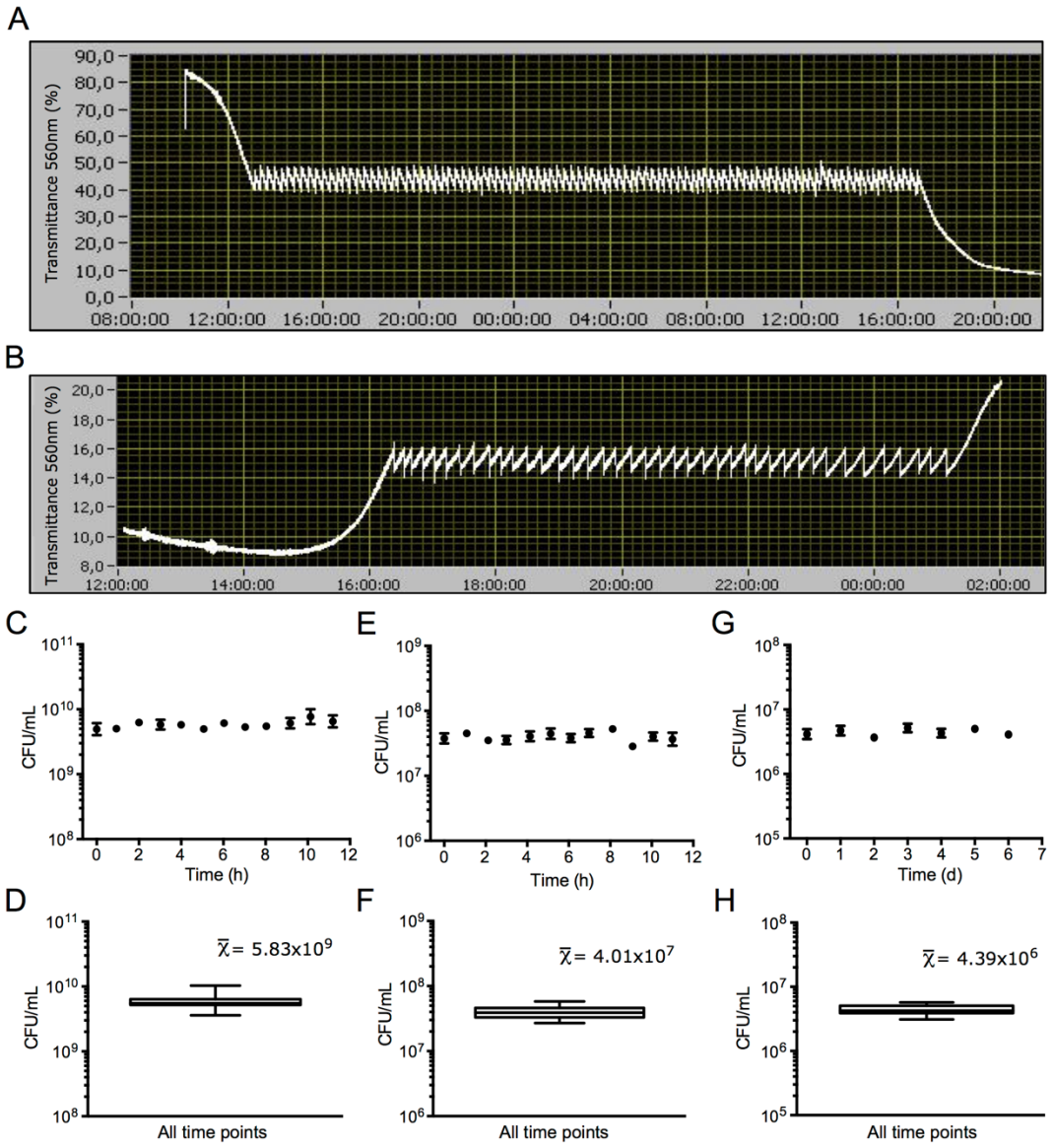

Figure 2.5. Establishment of continuous cultures using the versatile continuous culture device (VCCD). Example of transmittance curves monitored by the VCCD of an *E. coli* culture growing in LB broth (A) and a *M. florum* culture growing in ATCC 1161 medium (B) maintained for several hours at 40% and 16% of transmittance, respectively. Culture refreshes were performed using the Real-time feedback loop mode. Cell concentrations measured throughout continuous culture experiments carried out on *M. florum* (C), *E. coli* (E), and *S. cerevisiae* cultures (G) using different culture refresh modes, as well as box and whiskers plots of the whole experiments (D), (F), and (H) respectively.

2.7 Discussion

We have designed the VCCD, a continuous cultivation device that allows reaching and maintaining a constant cell density in a stable growth environment for various types of microorganisms such as bacteria and yeast. The complete documentation to build the VCCD along with the GUI software to operate the system is also provided for any potential user. The system is modular with three growth chambers that can each independently be programmed to operate as a turbidostat or chemostat. The instrument can be built at a relatively affordable cost of ~$1,400 US. We expect that biologists will be able not only to take advantage of the system to perform their experiments but also to customize it according to their experimental requirements and make the modifications publicly available. For example, different growth chamber configurations could be created and alternative sensors could be implemented such as a pH meter, thermometer, dissolved oxygen probe, light meter, etc. These modifications may require an upgrade to a data acquisition card containing additional ports but this would require a minimal effort. The VCCD software was designed with the NI Lab View platform to facilitate such changes and allow the rapid and easy integration of new features. As a first step, a fluorescence acquisition mode that could be useful for the validation of synthetic gene circuits is already included in the software although this functionality has not been tested yet (see Manual S2.1). There is clearly ample room to significantly extend the capabilities of the VCCD. This could allow the cultivation of new organisms such as photosynthetic microbes, algae, or insect and animal cells. Ultimately, the VCCD could become an interesting small-scale bioreactor used in metabolic engineering, synthetic biology and in several other disciplines.

In systems biology as well as other fields, reaching stable growth conditions is particularly important when trying to decipher the precise mechanisms of cell functioning. Steady state conditions should greatly reduce interference or noise arising from a constantly fluctuating environment. In principle, steady state conditions can be maintained indefinitely, which could be particularly useful to study the evolution of a cell population under selective conditions or to follow the fate of a group of mutants. However, in practice establishing steady state conditions

does not necessarily require extended incubations under constant growth conditions. For example, as shown in Fig. 2.5E, an *E. coli* population was maintained at maximal growth rate under controlled growth conditions for ~25 generations over a 12 hour period based on the estimated growth rate in batch culture (Fig. S2.10B). Similarly, *M. florum* was subjected to ~20 cell divisions during a 12 hour incubation in the VCCD (Fig. 2.5C and Fig. S2.10F). It is fair to assume that cells can reach a steady state during this incubation, which would be important and achievable within a manageable time frame for systems biology studies. For longer experiments, the use of continuous culture device is also interesting but often limited by biofilm formation on the growth chambers walls, which compromises optical density measurements and dilution operations. This limitation is not specific to the VCCD but could be countered by the use of new coating agents that abolish or significantly reduce biofilm formation [28-30]. Alternatively, users can simply transfer the culture to a new tube to prolong the experiment each time biofilm formation becomes problematic.

In sum, the description and publication of this initial version of the VCCD represents a first step in the development of an open-source laboratory platform for continuous cultivation that we hope will be embraced by the scientific community.

2.8 Acknowledgments

We thank Benoît Couture and Frédéric Francoeur for technical assistance, Thomas F. Knight and Marie-Eve Pepin for their appreciated suggestions in the development of the VCCD, as well as Alain Lavigueur for his insightful comments on the manuscript.

2.9 References

1. Monod J (1950) La technique de culture continue. Théorie et applications. Ann Inst Pasteur (Paris) 79: 390–410.

2. Novick A, Szilard L (1950) Description of the Chemostat. Science 112: 715–716.

3. Bull AT (2010) The renaissance of continuous culture in the post-genomics age. J Ind Microbiol Biotechnol 37: 993–1021.

4. Markx GH, Davey CL, Kell DB (1991) The permittistat: a novel type of turbidostat. J Gen Microbiol 137: 735–743.

5. De Crécy E, Metzgar D, Allen C, Pénicaud M, Lyons B, Hansen CJ, et al. (2007) Development of a novel continuous culture device for experimental evolution of bacterial populations. Appl Microbiol Biotechnol 77: 489–496.

6. Lee KS, Boccazzi P, Sinskey AJ, Ram RJ (2011) Microfluidic chemostat and turbidostat with flow rate, oxygen, and temperature control for dynamic continuous culture. Lab Chip 11: 1730–1739.

7. Toprak E, Veres A, Yildiz S, Pedraza JM, Chait R, Paulsson J, et al. (2013) Building a morbidostat: an automated continuous-culture device for studying bacterial drug resistance under dynamically sustained drug inhibition. Nat Protoc 8: 555–567.

8. Moffitt JR, Lee JB, Cluzel P (2012) The single-cell chemostat: an agarose-based, microfluidic device for high-throughput, single-cell studies of bacteria and bacterial communities. Lab Chip 12: 1487–1494.

9. Zhang Z, Boccazzi P, Choi H-G, Perozziello G, Sinskey AJ, Jensen KF (2006) Microchemostat-microbial continuous culture in a polymer-based, instrumented microbioreactor. Lab Chip 6: 906–913.

10. Tomson K, Barber J, Vanatalu K (2006) Adaptastat - A new method for optimising of bacterial growth conditions in continuous culture: Interactive substrate limitation based on dissolved oxygen measurement. J Microbiol Methods 64: 380–390.

11. Martin G, Hempfling W (1976) A method for the regulation of microbial population density during continuous culture at high growth rates. Arch Microbiol 107: 41–47.

12. Bijmans MFM, de Vries E, Yang CH, Buisman CJN, Lens PNL, Dopson M (2010) Sulfate reduction at pH 4.0 for treatment of process and wastewaters. Biotechnol Prog 26: 1029–1037.

13. Gilbert A, Sangurdekar DP, Srienc F (2009) Rapid strain improvement through optimized evolution in the cytostat. Biotechnol Bioeng 103: 500–512.

14. Winder CL, Lanthaler K (2011) The use of continuous culture in systems biology investigations. Methods in Enzymology. Elsevier Inc., Vol. 500. pp. 261–275.

15. Hoskisson P a, Hobbs G (2005) Continuous culture--making a comeback? Microbiology 151: 3153–3159.

16. Piper MDW, Daran-Lapujade P, Bro C, Regenberg B, Knudsen S, Nielsen J, et al. (2002) Reproducibility of oligonucleotide microarray transcriptome analyses. An interlaboratory comparison using chemostat cultures of Saccharomyces cerevisiae. J Biol Chem 277: 37001–37008.

17. Toda K (2003) Theoretical and methodological studies of continuous microbial bioreactors. J Gen Appl Microbiol 49: 219–233.

18. Chuang H-Y, Hofree M, Ideker T (2010) A decade of systems biology. Annu Rev Cell Dev Biol 26: 721–744.

19. Monod J (1949) The Growth of Bacterial Cultures. Annu Rev Microbiol 3: 371–394.

20. Takahashi CN, Miller AW, Ekness F, Dunham MJ, Klavins E (2015) A Low Cost, Customizable Turbidostat for Use in Synthetic Circuit Characterization. ACS Synth Biol 4: 32–38.

21. Miller AW, Befort C, Kerr EO, Dunham MJ (2013) Design and use of multiplexed chemostat arrays. J Vis Exp: e50262.

22. Esvelt KM, Carlson JC, Liu DR (2011) A system for the continuous directed evolution of biomolecules. Nature 472: 499–503.

23. Stemke GW, Robertson J a. (1982) Comparison of two methods for enumeration of mycoplasmas. J Clin Microbiol 16: 959–961.

24. Clyde WA, Senterfit LB (1985) Laboratory diagnosis of mycoplasma infections. In: Razin S, Barile F, editors. Mycoplasma Pathogenicity, Volume 6. Academic Press. pp. 391–402.

25. Grenier F, Matteau D, Baby V, Rodrigue S (2014) Complete Genome Sequence of Escherichia coli BW25113. Genome Annoucements 2: e01038–14.

26. Mccoy RE, Basham HG, Tully JG, Rose DL, Carle P, Bové JM (1984) Acholeplasma florum, a New Species Isolated from Plants. Int J Syst Bacteriol 34: 11–15.

27. Baby V, Matteau D, Knight TF, Rodrigue S (2013) Complete genome sequence of the Mesoplasma florum W37 strain. Genome Annoucements 1: e00879–13.

28. Park KD, Kim YS, Han DK, Kim YH, Lee EHB, Suh H, et al. (1998) Bacterial adhesion on PEG modified polyurethane surfaces. Biomaterials 19: 851–859.

29. Epstein a. K, Wong T-S, Belisle R a., Boggs EM, Aizenberg J (2012) Liquid-infused structured surfaces with exceptional anti-biofouling performance. Proc Natl Acad Sci 109: 13182–13187.

30. Banerjee I, Pangule RC, Kane RS (2011) Antifouling coatings: Recent developments in the design of surfaces that prevent fouling by proteins, bacteria, and marine organisms. Adv Mater 23: 690–718.

2.10 Supporting Information

Tables S2.1-S2.3 (S1-S3 Tables), Manual S2.1 (S1 Manual), Appendix S2.1 (S1 Appendix), and Files S2.1-S2.2 (S1-S2 Files) are available online at:

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0133384#sec019

Table S2.1. Frame material list. Contains detailed description for each part belonging to the VCCD frame, as well as their reference code in Fig. S2.1 and Appendix S2.1.

Table S2.2. Culture system material list. Contains detailed description for each part belonging to the VCCD culture system, as well as their reference code on Fig. S2.2.

Table S2.3. Electronics material list. Contains detailed description for each part belonging to the VCCD Electronics, as well as their reference code on Figs. S2.3-S2.8.

Manual S2.1. VCCD User Manual. Contains all construction and assembly instructions, a system and software utilization guide, as well as important operation notes

Appendix S2.1. Frame machining details.

File S2.1. Frame parts in 3D CAD format.

File S2.2. Example of a complete VCCD assembly in 3D CAD format.

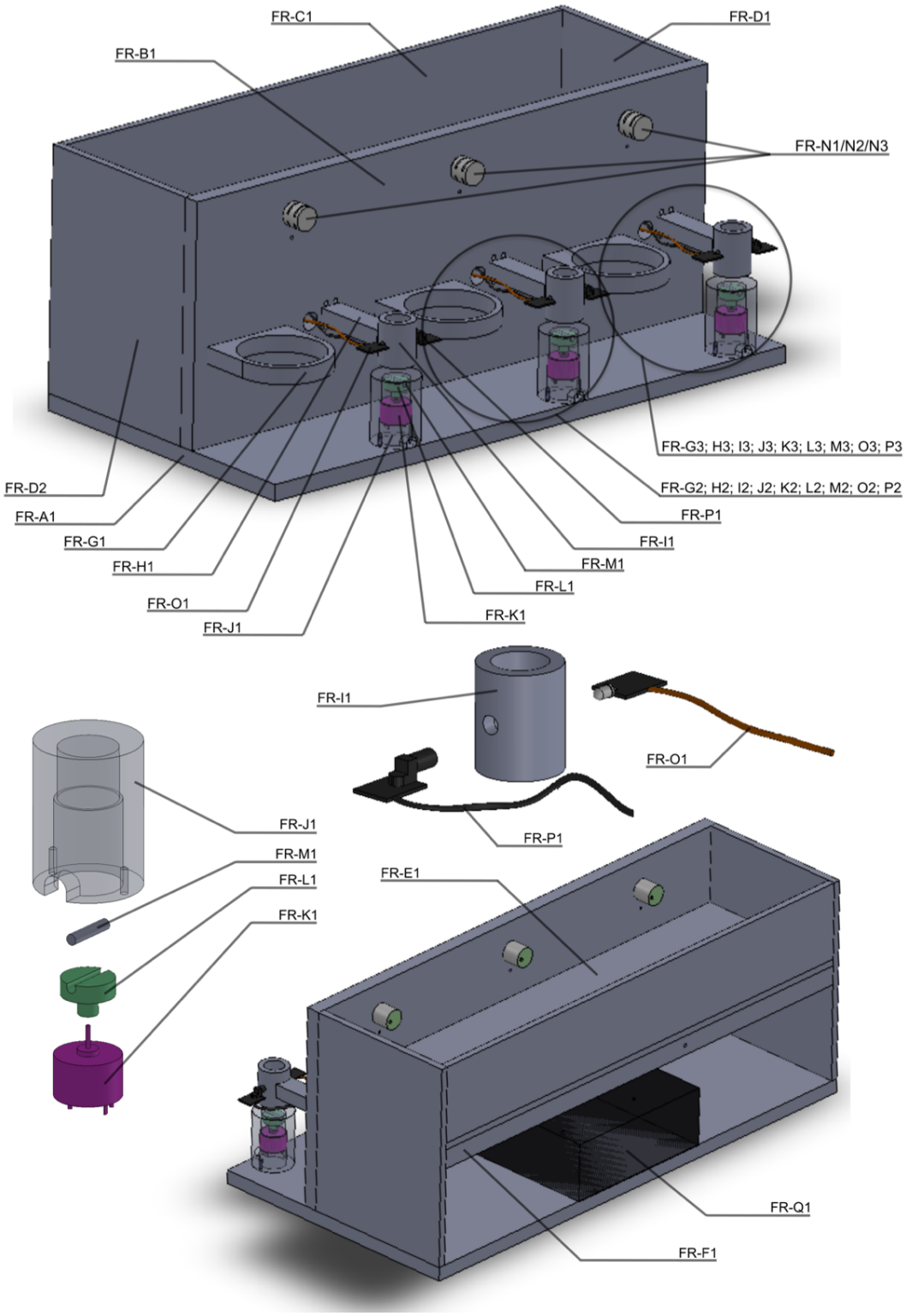

Figure S2.1. Frame assembly.

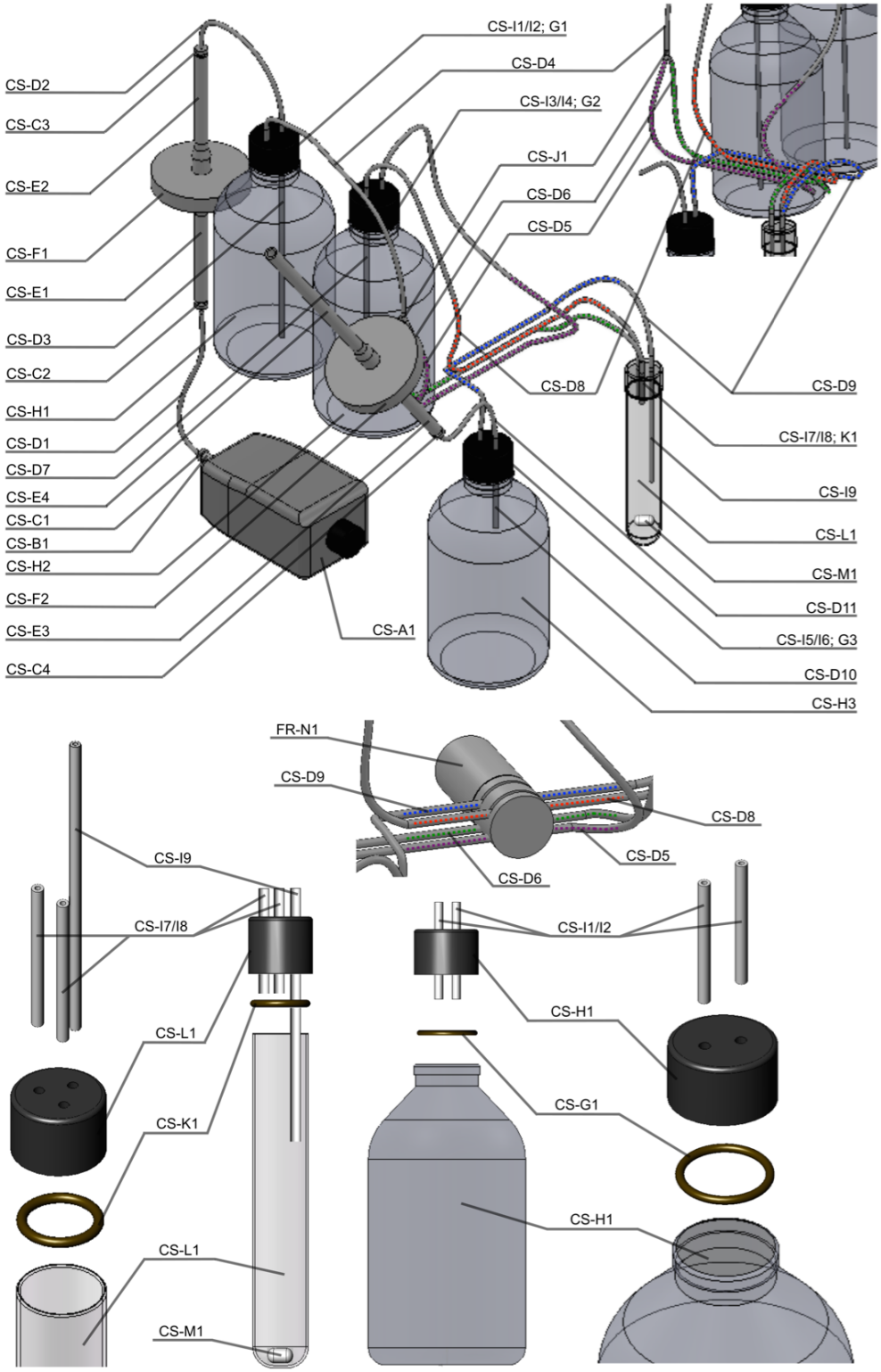

Figure S2.2 Culture system assembly.

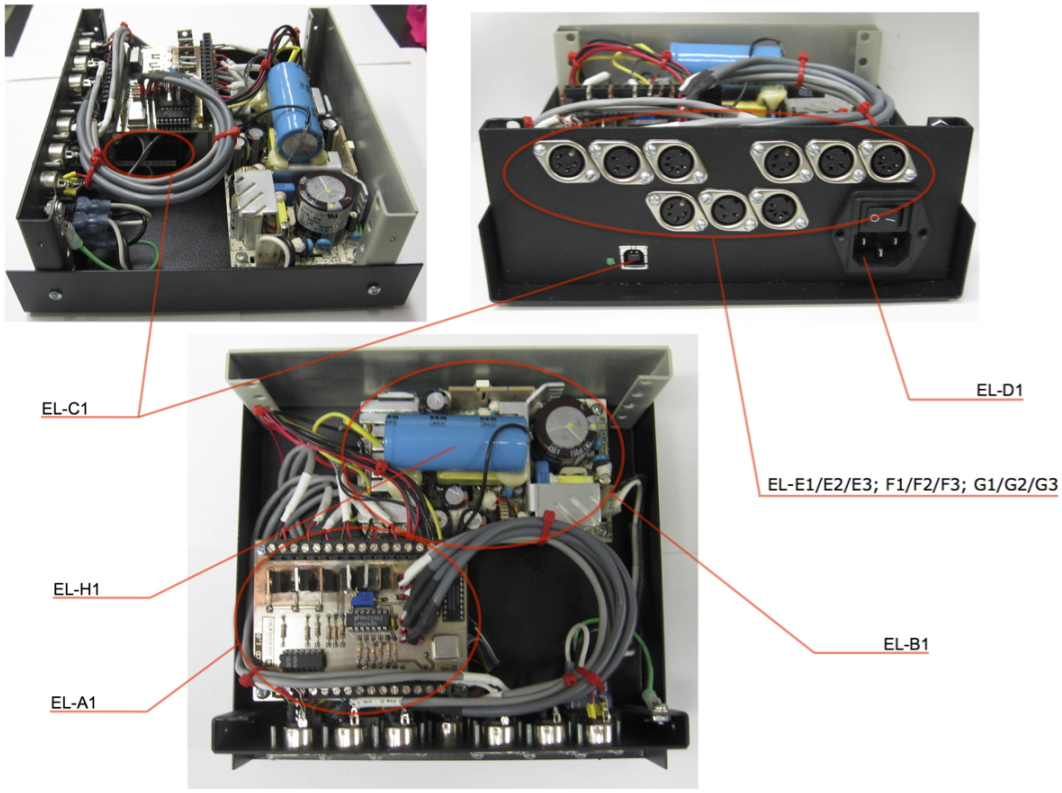

Figure S2.3. Electronics box assembly.

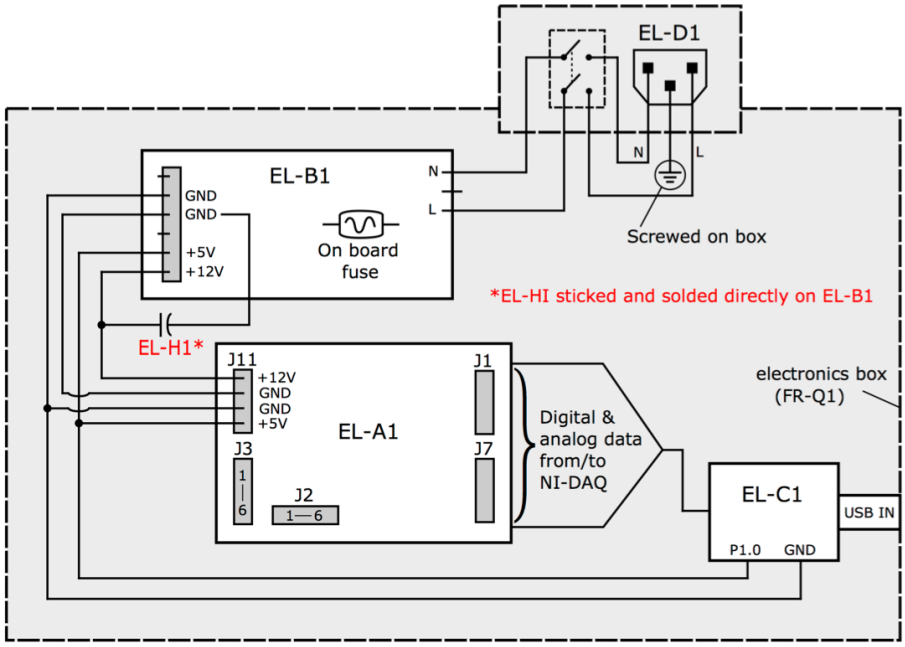

Figure S2.4. Assembly of the main electronics components.

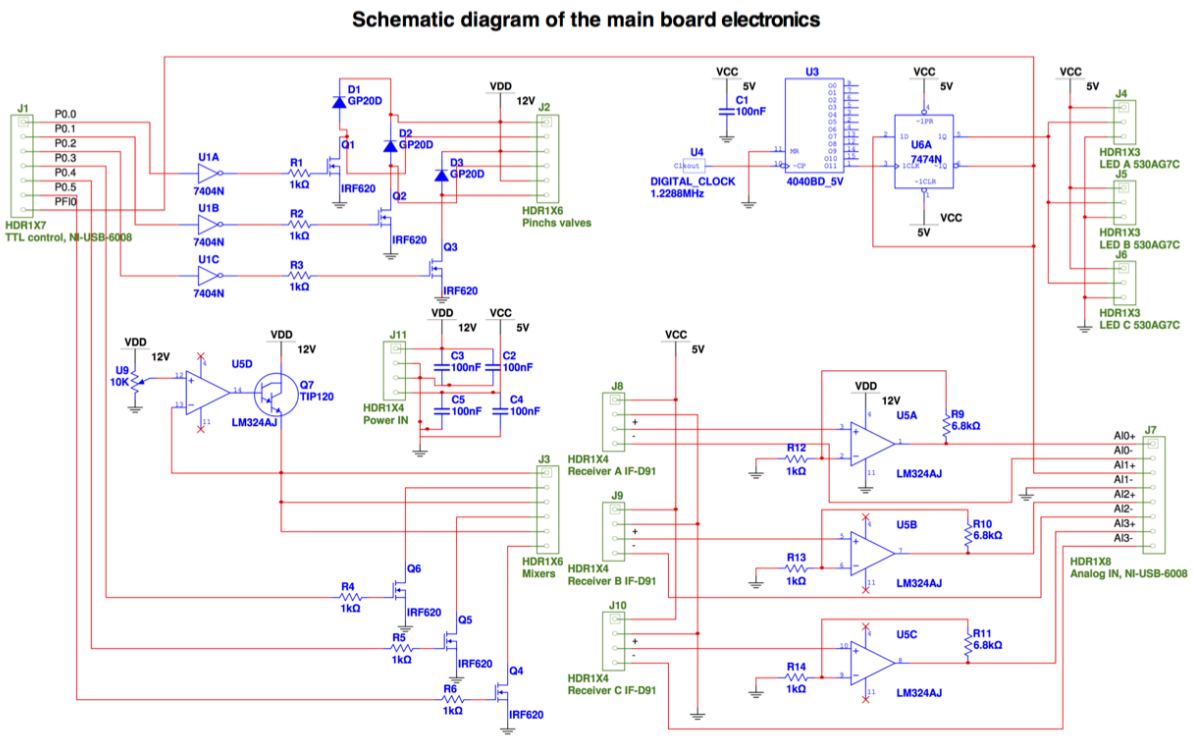

Figure S2.5. Schematic diagram of the main board electronics.

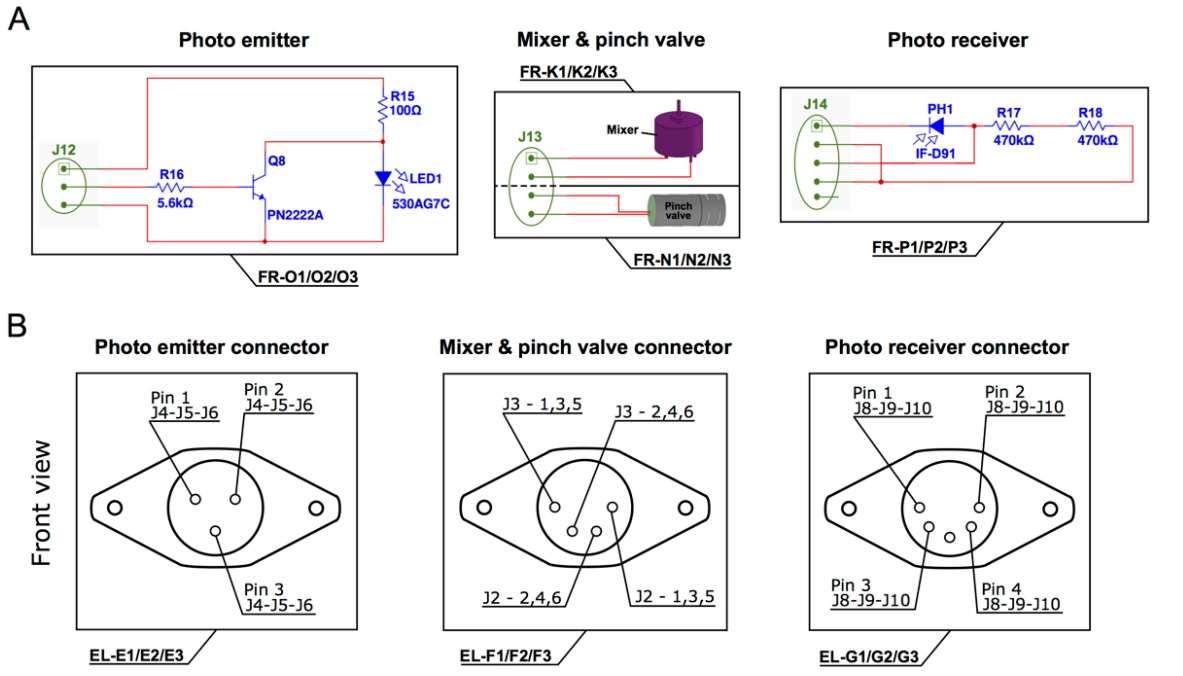

Figure S2.6. Schematic diagrams of the electronics (A), and circular connectors (B) of the photo emitter, photo receiver, mixer, and pinch valve.

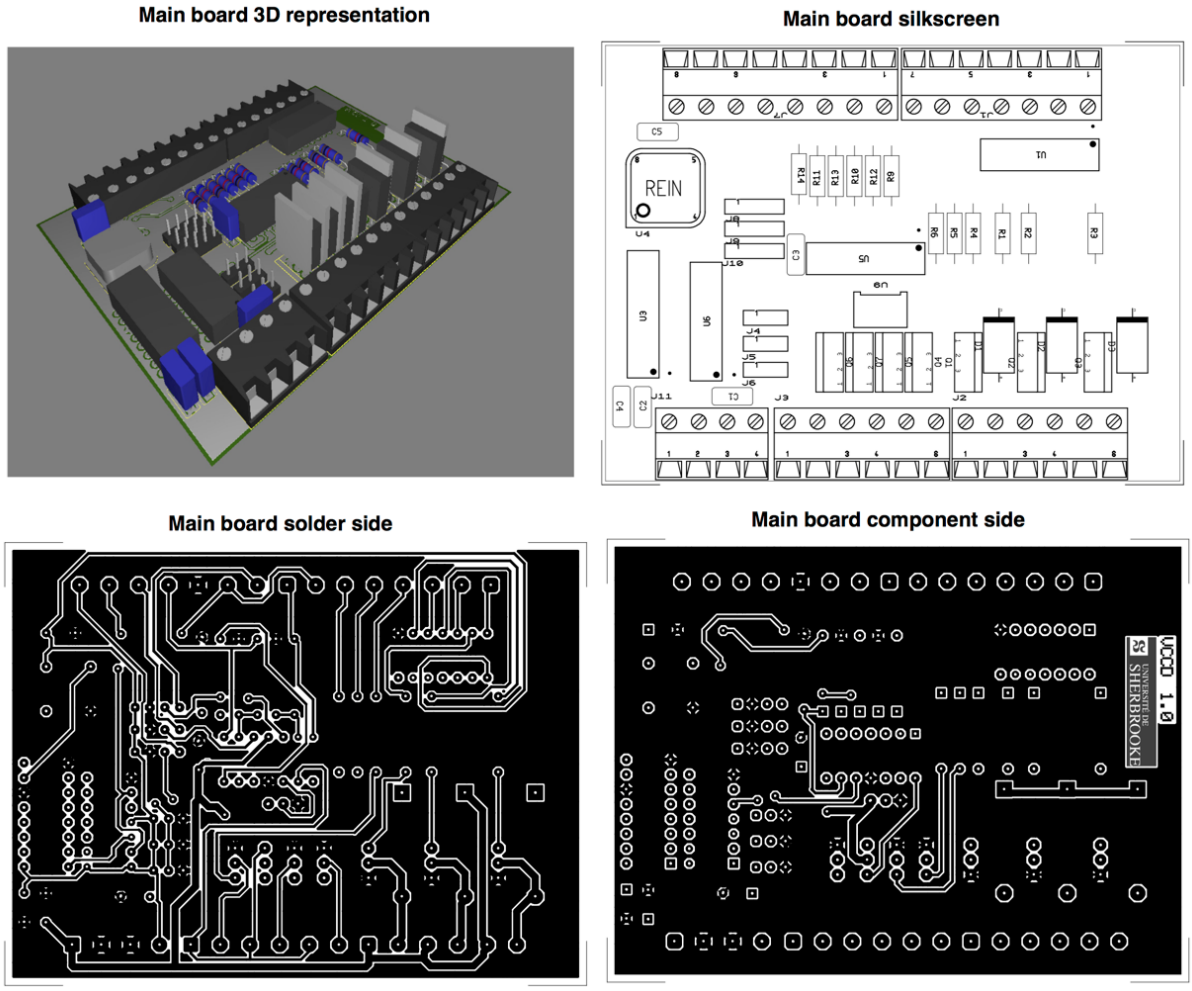

Figure S2.7. Details of the main PCB.

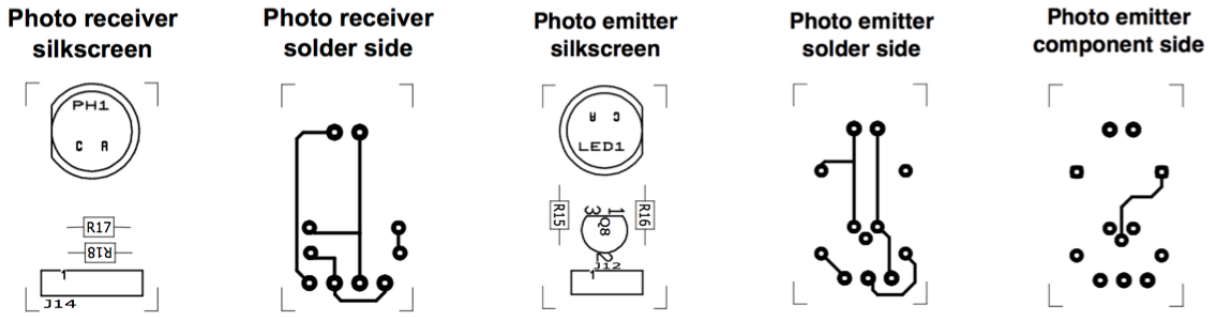

Figure S2.8. Details of the photo emitter and receiver PCBs.

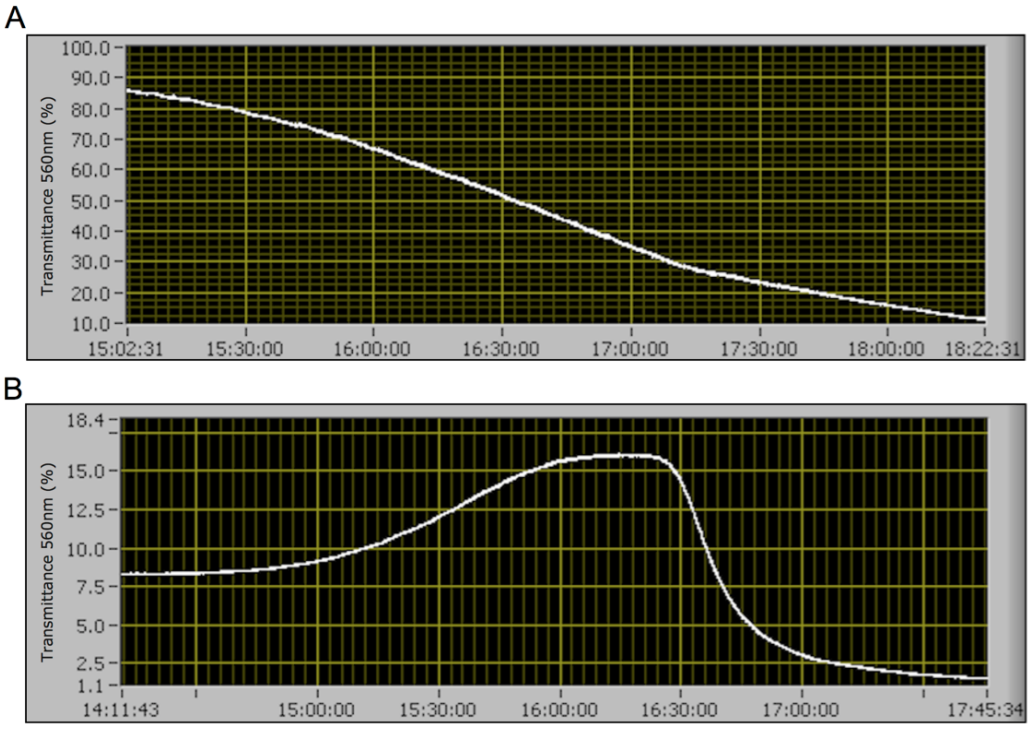

Figure S2.9. Typical batch culture growth curves displayed on the graphical user interface (GUI) software for *E. coli* (A) and *M. florum* (B).

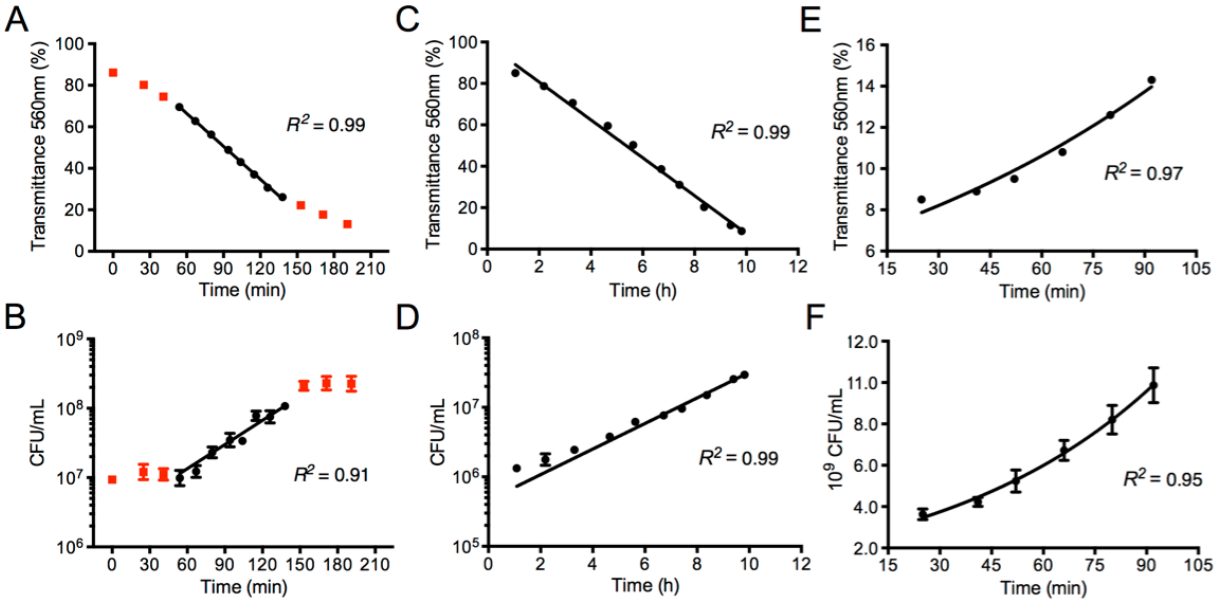

Figure S2.10. Bacterial and yeast batch culture growth curves.

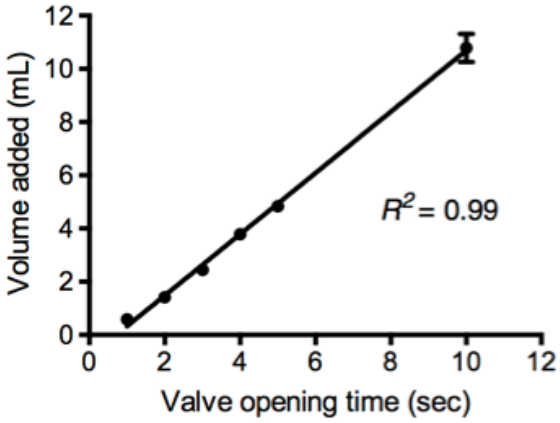

Figure S2.11. Volume of liquid added to a culture vessel during a refresh cycle of a specified time.

CHAPITRE 3

DEVELOPMENT OF *ORIC*-BASED PLASMIDS FOR *MESOPLASMA FLORUM*

3.1 Présentation de l'article et contributions

Idéalement, un châssis cellulaire spécifiquement conçu pour la biologie synthétique et la génomique synthétique devrait être simple, bien caractérisé, à croissance rapide, facile à manipuler génétiquement et ne poser aucun risque pour la santé humaine et l'environnement. Même si *M. florum* présente déjà plusieurs de ces caractéristiques, très peu voire pratiquement aucun outil moléculaire n'était initialement disponible afin de modifier génétiquement cette bactérie. Cette lacune complexifie le développement de circuits génétiques artificiels dans cette bactérie, limite l'étude de sa biologie, et restreint grandement notre capacité à développer un châssis cellulaire réduit à partir de son génome.

Les Mollicutes sont généralement déficients en éléments génétiques extrachromosomiques pouvant servir de vecteur de clonage ou d'outils de modifications génétiques (Sirand-Pugnet et al., 2007b, 2007a). Par conséquent, plusieurs groupes de recherche ont utilisé l'*oriC* de Mollicutes afin de développer des plasmides artificiels capables de réplication dans ces bactéries (Chopra-Dewasthaly et al., 2005; Cordova et al., 2002; Janis et al., 2005; Lartigue et al., 2002; Lee et al., 2008; Maglennon et al., 2013; Renaudin et al., 1995; Shahid et al., 2014). Même si *M. florum* est phylogénétiquement très proche de *M. mycoides* et *M. capricolum*, les plasmides *oriC* développés chez ces deux mycoplasmes ne peuvent se répliquer chez *M. florum* (Matteau et al., 2017). En nous inspirant de publications précédentes, nous avons donc entrepris le développement des premiers plasmides spécifiquement conçus pour se répliquer chez *M. florum,* en plus de mettre au point différentes méthodes de transformation pour cette bactérie (Matteau et al., 2017). Basés sur l'*oriC* de *M. florum*, ces plasmides ont d'ailleurs permis de montrer expérimentalement, pour la toute première fois, la conjugaison d'éléments génétiques entre *E. coli* et une espèce de Mollicutes. Ils ont également servi à tester la fonctionnalité de

plusieurs marqueurs de sélection aux antibiotiques, et sont à la base de la technique utilisée afin de cloner le génome complet de *M. florum* dans la levure *S. cerevisiae* (Baby et al., 2017). Cette souche de levure peut maintenant servir de plateforme afin de modifier efficacement le génome de *M. florum* et ensuite le transplanter dans une cellule réceptrice. Ces outils génétiques vont faciliter l'étude du fonctionnement de *M. florum* en plus de rendre possible le développement d'un châssis cellulaire simplifié basé sur le génome de *M. florum* pour les domaines de la biologie synthétique et la génomique synthétique.

L'accomplissement du projet décrit dans ce chapitre est le résultat d'une étroite collaboration entre différentes personnes au sein du laboratoire du Pr Sébastien Rodrigue. Les efforts communs et la persévérance du groupe auront permis de surmonter les défis et rebondissements rencontrés tout au long de cette étude. Je tiens à profiter de l'occasion pour souligner le travail important accompli par Marie-Eve Pepin dans le cadre de ce projet. Sa contribution a permis de défricher plusieurs aspects qui ne figurent pas nécessairement dans la publication finale. Plus spécifiquement, j'ai été chargé d'effectuer les analyses bio-informatiques des régions *oriC* des espèces apparentées à *M. florum*, la construction et la transformation des plasmides portant les régions *oriC* hétérologues, ainsi que l'optimisation des méthodes d'électroporation et de conjugaison entre *E. coli* et *M. florum*. Assistés par Joëlle Brodeur, Marie-Eve Pepin et moi-même avons optimisé et réalisé les expériences d'immunobuvardage de type *Southern*, ainsi que les essais de stabilité des plasmides *oriC* utilisant le VCCD (Matteau et al., 2015) (voir Chapitre 2). Marie-Eve Pepin a également effectué les essais de détermination des concentrations minimales inhibitrices (CMI) de la souche *M. florum* L1 de type sauvage. De plus, elle a participé, avec mon aide et celle de Mélissa Arango Giraldo, Samuel Gauthier et Vincent Baby, à la construction et à la transformation des différentes versions des plasmides *oriC* de *M. florum*, ainsi qu'à la détermination de la CMI des souches portant les différents gènes de résistance aux antibiotiques. Le Pr Sébastien Rodrigue a participé à la conception des différentes expériences et a supervisé l'ensemble des étudiants impliqués dans le projet. Thomas F. Knight a contribué aux idées originales figurant dans la publication. J'ai, avec la participation de Marie-Eve Pepin et du Pr Sébastien Rodrigue, rédigé le manuscrit décrivant les plasmides

oriC de *M. florum*. Vincent Baby et Alain Lavigueur ont révisé le manuscrit final. Je tiens également à remercier Carole Lartigue and Fabien Labroussaa pour l'ensemble des discussions enrichissantes entourant ce projet, et pour nous avoir fourni les plasmides pMYCO1, pMYSO1, pMCO3, et pSD4.

Référence bibliographique : Matteau, D.*, Pepin, M.-E.*, Baby, V., Gauthier, S., Arango Giraldo, M., Knight, T.F., Rodrigue, S. (2017). Development of *oriC*-based plasmids for *Mesoplasma florum*. Appl. Environ. Microbiol. *83*, e03374-16.

* Ces auteurs ont contribué de manière équivalente à cette publication.

3.2 Title page

Development of *oriC*-based plasmids for *Mesoplasma florum*

Dominick Matteau[a *], Marie-Eve Pepin[a *], Vincent Baby[a], Samuel Gauthier[a], Mélissa Arango Giraldo[a], Thomas F. Knight[b] & Sébastien Rodrigue[a#].

Département de biologie, Université de Sherbrooke, Sherbrooke, Québec, Canada[a];
Ginkgo Bioworks, Boston, Massachusetts, USA[b]

Running title: *Mesoplasma florum oriC* plasmids

#Address correspondence to Sébastien Rodrigue, sebastien.rodrigue@usherbrooke.ca

[*] D.M. and M-E.P. contributed equally to this work.

3.3 Abstract

The near-minimal bacterium *Mesoplasma florum* constitutes an attractive model for systems biology and for the development of a simplified cell chassis in synthetic biology. However, the lack of genetic engineering tools for this microorganism has limited our capacity to understand its basic biology and modify its genome. To address this issue, we have evaluated the susceptibility of *M. florum* to common antibiotics, and developed the first generation of artificial plasmids able to replicate in this bacterium. Selected regions of the predicted *M. florum* chromosomal origin of replication (*oriC*) were used to create different plasmid versions that were tested for their transformation frequency and stability. Using polyethylene glycol mediated transformation, we observed that plasmids harbouring both *rpmH*/*dnaA* and *dnaA*/*dnaN* intergenic regions, interspaced or not with a copy of the *dnaA* gene, resulted in a frequency of ~4.1 x 10^{-6} transformant per viable cell and were stably maintained throughout multiple generations. In contrast, plasmids containing only one *M. florum oriC* intergenic region or the heterologous *oriC* region of *Mycoplasma capricolum*, *Mycoplasma mycoides* or *Spiroplasma citri* failed to produce any detectable transformants. We also developed alternative transformation procedures based on electroporation or conjugation from *Escherichia coli*, reaching frequencies up to 7.87 x 10^{-6} and 8.44 x 10^{-7} transformant per viable cell, respectively. Finally, we demonstrated the functionality of antibiotic resistance genes active against tetracycline, puromycin, as well as spectinomycin/streptomycin in *M. florum*. Taken together, these valuable genetic tools will facilitate efforts towards building a *M. florum* based near-minimal cellular chassis for synthetic biology.

3.4 Importance

Mesoplasma florum constitutes an attractive model for systems biology, and for the development of a simplified cell chassis in synthetic biology. *M. florum* is closely related to the mycoides cluster of mycoplasmas that has become a model for whole-genome cloning, genome transplantation, and genome minimization. However, *M. florum* shows faster growth rates

compared to other Mollicutes, has no known pathogenic potential, and possesses a significantly smaller genome that positions this species among some of the simplest free-living organisms. So far, the lack of genetic engineering tools has limited our capacity to understand the basic biology of *M. florum* in order to modify its genome. To address this issue, we have evaluated the susceptibility of *M. florum* to common antibiotics, and developed the first artificial plasmids as well as transformation methods for this bacterium. This represents a strong basis for on-going genome engineering efforts using this near-minimal microorganism.

3.5 Introduction

Mollicutes are a class of bacteria mainly characterized by small genome sizes (0.58 – 2.2 Mbp), small cell dimensions (~0.2 – 0.4 μM), and the absence of a cell wall (1-3). Mollicutes are thought to have derived from low G-C content Gram-positive bacteria through genome reduction, which resulted in a significant simplification of their metabolic pathways (1-3). Consequently, many Mollicutes have evolved a parasitic life style with the ability to infect various plants and animals, including humans (1, 2). Unlike other small genome bacteria such as chlamydias and rickettsias, Mollicutes can be cultured in acellular medium, except for phytoplasmas that are obligate parasites of plants (4). The remarkable genomic simplicity of Mollicutes makes members of this class attractive candidates to develop minimal cells in which the thorough characterization of global cellular mechanisms will be more easily achievable (5, 6).

Mesoplasma florum, first described as *Acholeplasma florum* in 1984 (7), constitutes a particularly interesting member of the Mollicutes as a new model for systems and synthetic biology studies. *M. florum* is closely related to the mycoides cluster of mycoplasmas, which includes *Mycoplasma mycoides* and *Mycoplasma capricolum* that have become model organisms for whole-genome cloning (8-10), genome transplantation (8, 11, 12) and genome minimization (13). However, *M. florum* shows faster growth rates (~34 min), has no known pathogenic potential, and possesses a significantly smaller genome that positions this species

among some of the simplest free-living organisms (1, 6, 14, 15). For example, *M. florum* L1 (RefSeq NC_006055.1), the first representative of its species, has a total genome size of only ~793 kb compared to ~1.2 Mb and ~1.0 Mb for *M. mycoides* and *M. capricolum*, respectively (1). *M. florum* also uses an alternative genetic code (Mycoplasma/Spiroplasma code) in which the UGA codon signals the incorporation of a tryptophan in the nascent protein rather than a stop codon, a feature that limits horizontal gene transfer from and to other microorganisms (16, 17). Despite these advantageous characteristics, practically no genetic tools are currently available to reduce and reprogram the genome of *M. florum* as well as to build artificial gene circuits.

Many Mollicutes phylogenetically related to *M. florum,* including *M. mycoides*, *M. capriolum,* and *Spiroplasma citri* have been successfully transformed with artificial plasmids containing a chromosomal origin of replication (*oriC*) (18-26). *OriC*-based plasmids have multiple uses such as expression of exogenous genes, inactivation of target genes by recombination or complementation of chromosomal mutations. Since Mollicutes are naturally susceptible to tetracycline, the *tetM* gene derived from the Tn916 transposon of *Enterococcus faecalis* is often used as an antibiotic resistance marker for robust *oriC*-based plasmid selection (18-26). Following transformation in a recipient cell, the *oriC* plasmids can replicate due to specific interactions of the DnaA protein with sequences called DnaA boxes (24, 27). In Mollicutes, DnaA boxes have been shown to generally be located within the two A-T rich intergenic regions flanking the *dnaA* gene, with a proposed 9 bp asymmetric sequence of 5'-TT(A/T)TC(C/A)ACA-3' (21, 24). By virtue of their sequence homology, *oriC* plasmids can also integrate into the *oriC* region of the host cell chromosome by recombination events (18-24, 26).

In this work, we evaluate the susceptibility of *M. florum* L1 to common antibiotics, and describe the successful utilization of the predicted *oriC* region of *M. florum* L1 chromosome to generate the first replicable plasmids in this microorganism. These *oriC* plasmids were characterized for their transformation frequency, stability, as well as their propensity to recombine at the

chromosomal *oriC* region of *M. florum*. We also report successful *oriC* plasmids transformation using electroporation or conjugation as alternative transformation methods to the more traditional polyethylene glycol (PEG)-mediated procedure, and investigate the capacity of *M. florum* to replicate heterologous *oriC* plasmids. The genetic tools developed in this study will contribute to on-going efforts towards building a *M. florum* based near-minimal cellular chassis for synthetic biology.

3.6 Materials and Methods

3.6.1 Strains and growth conditions

Bacterial strains used in this study are described in Table 3.1. *E. coli* strains EC100D *pir*+ and MM294 were routinely grown in Luria-Bertani (LB) broth at 37°C. *E. coli* strain MFD*pir* was grown at 37°C in LB broth supplemented with 0.3 mM diaminopimelic acid (DAP) and 200 µg/ml erythromycin. *M. florum* strain L1 (ATCC 33453) was grown at 34°C in ATCC 1161 medium. All strains were grown using an orbital shaker incubator and preserved at -80°C in their respective growth medium containing 25% (v/v) glycerol. Unless specified, antibiotics were used at the following concentrations for *E. coli*: ampicillin, 100 µg/ml; chloramphenicol, 34 µg/ml; erythromycin, 200 µg/ml; streptomycin, 50 µg/ml; spectinomycin, 100 µg/ml; puromycin, 125 µg/ml. Unless specified, tetracycline was used at 15 µg/ml for either *E. coli* and *M. florum*. Penicillin was used at 200 U/ml for *M. florum.*

3.6.2 ATCC 1161 medium preparation

To prepare 1L of ATCC 1161 medium, 17.5 g of heart infusion broth, 40 g of sucrose, and 12 g of agar (for solid medium) were first mixed in 710 ml of water before being autoclaved at 121°C. After sterilization, the mix was cooled to room temperature (broth) or to 55°C (solid), and 200 ml of horse serum (Sigma H1138), 90 ml of 15% (w/v) yeast extract, 8 ml of 0.5% (w/v) phenol red, and 200 U/ml of penicillin G were added. The pH was then adjusted to 7.6 with sterile

NaOH. The final composition of ATCC 1161 medium is: heart infusion broth, 17.5 g/L; sucrose, 40 g/L; agar (for solid medium), 12 g/L; horse serum, 20% (v/v); yeast extract, 1.35% (w/v); phenol red, 0.004% (w/v) and penicillin G, 200 U/ml.

Table 3.1. Strains and plasmids used in this study.

Strain or plasmid	Relevant genotype or phenotype	Source or reference
Escherichia coli		
EC100D *pir*+	*F-, mcrA,* Δ*(mrr-hsdRMS-mcrBC), φ 80dlacZ*Δ*M15,* Δ*lacX74, recA1, endA1, araD139,* Δ*(ara, leu)7697, galU, gal K, λ-, rpsL, nupG, pir+ (DHFR)*(Sm^r)	Epicentre
MFD*pir*	MG1655 RP4-2-Tc::[ΔMu1::*aac(3)IV*-Δ*aphA*-Δ*nic35*-ΔMu2::*zeo*] Δ*dapA*::(*erm-pir*) Δ*recA* (Apra^r Zeo^r Em^r)	(41)
MM294	*F-, glnX44(AS), λ-, endA1, spoT1, thiE1,hsdR17, creC510*	*E. coli* Genetic Stock Center (strain 6315)
Mesoplasma florum		
L1		ATCC 33453
L1 clone 3632	*mfl169*::Tn-*tetM*	This study
Plasmids		
pMflT-o1	colE1, *oriT$_{RP4}$*, *M. florum oriC1, tetM* (Tc^r)	This study
pMflT-o2	colE1, *M. florum oriC2, tetM* (Tc^r)	This study
pMflT-o3	colE1, *oriT$_{RP4}$*, *M. florum oriC3, tetM* (Tc^r)	This study
pMflT-o4	colE1, *oriT$_{RP4}$*, *M. florum oriC4, tetM* (Tc^r)	This study
pMflCT-o4	colE1, *oriT$_{RP4}$*, *M. florum oriC4, tetM, cat* (Tc^r Cm^r)	This study
pMflET-o4	colE1, *oriT$_{RP4}$*, *M. florum oriC4, tetM, ereB* (Tc^r Em^r)	This study
pMflPT-o4	colE1, *oriT$_{RP4}$*, *M. florum oriC4, tetM, pac* (Tc^r Pu^r)	This study

Table 3.1. Strains and plasmids used in this study (continued).

pMflST-o4	colE1, *oriT*$_{RP4}$, *M. florum oriC4, tetM, aadA1* (Tc^r Sp^r Sm^r)	This study
pMYCO1	colE1, *Mmc oriC*, *tetM, bla* (Tc^r Ap^r)	(24)
pMYSO1	colE1, *Mmm oriC*, *tetM, bla* (Tc^r Ap^r)	(24)
pMCO3	colE1, *Mcap oriC*, *tetM, bla* (Tc^r Ap^r)	(24)
pSD4	colE1, *S. citri oriC*, *tetM, bla* (Tc^r Ap^r)	(25)
pMmcT	colE1, *oriT*$_{RP4}$, *Mmc oriC*, *tetM* (Tc^r)	This study
pMmmT	colE1, *oriT*$_{RP4}$, *Mmm oriC*, *tetM* (Tc^r)	This study
pMcapT	colE1, *oriT*$_{RP4}$, *Mcap oriC*, *tetM* (Tc^r)	This study
pSciT-o4	colE1, *oriT*$_{RP4}$, *S. citri oriC*, *tetM* (Tc^r)	This study
ereB-pUC57	colE1, *bla, ereB* (Ap^r Em^r)	This study
pTT01	colE1, *tetM* (Tc^r)	This study
pUC19	colE1, *bla,* (Ap^r)	(67)
pSW23T	R6K, *oriT*$_{RP4}$, *cat* (Cm^r)	(68)

Sm^r, streptomycin resistant; Apra^r, apramycin resistant; Zeo^r, zeocin resistant; Em^r, erythromycin resistant; Tc^r, tetracycline resistant; Cm^r, chloramphenicol resistant; Pu^r, puromycin resistant; Sp^r, spectinomycin resistant; Ap^r, ampicillin resistant; *Mmc*, *M. mycoides subsp. capri*; *Mmm*, *M. mycoides subsp. mycoides*; *Mcap, M. capricolum subsp. capricolum*.

3.6.3 Antimicrobial susceptibility assays

MIC values were determined by the growth inhibition assay according to the broth microdilution method in a 96-well microplate (28). The following antibiotics were tested for the *M. florum* L1 wild-type strain: ampicillin, chloramphenicol, erythromycin, gentamicin, kanamycin, puromycin, rifampicin, spectinomycin, streptomycin, sulfamethoxazole, tetracycline, and trimethoprim. For *M. florum* L1 carrying pMflT-o4, pMflPT-o4, pMflCT-o4 or pMflET-o4, tetracycline, puromycin, chloramphenicol, and erythromycin were respectively tested. For *M. florum* L1 carrying pMflST-o4, spectinomycin and streptomycin were tested separately. Assays

were conducted with three biological replicates in a final volume of 200 µl of ATCC 1161 medium supplemented with a decreasing concentration of the tested antibiotic. Medium was inoculated with ~1.0 x 10^7 colony-forming unit (CFU) of a log-phase batch culture for all tested strains. Microplates were next incubated at 34°C for 14 hours. Bacterial growth was assessed by measuring optical density at 560 nm every hour with a microplate reader (BioTek, Synergy HT). The metabolic activity of *M. florum* was previously shown to result in the acidification of the ATCC 1161 growth medium, causing changes in the absorbance of phenol red at 560 nm that correlate with the number of CFU (15). The MIC of each antibiotic was defined as the lowest tested concentration that inhibited the growth of *M. florum* (28).

3.6.4 Sequence analysis of the *oriC* region of the Spiroplasma group

DNA sequence of the *oriC* region of selected representative members of the Spiroplasma group (*M. florum* L1, RefSeq NC_006055.1; *M. capricolum subsp. capricolum* ATCC 27343, RefSeq NC_007633.1; *M. capricolum subsp. capripneumoniae* 9231-Abomsa, RefSeq NZ_LM995445.1; *Mycoplasma leachii* PG50, RefSeq NC_014751.1; *M. mycoides subsp. capri* GM12, RefSeq NZ_CP001621.1; *M. mycoides subsp. mycoides* PG1, RefSeq NC_005364.2; *Mycoplasma putrefaciens* KS1, RefSeq NC_015946.1; *Mycoplasma yeatsii* GM274B, RefSeq NZ_CP007520.1; *S. citri* GII3-3x, GenBank AM285301.1 and AM285302.1 (29); *Spiroplasma kunkelii* CR2-3x, RefSeq NZ_CP010899.1; *Mycoplasma feriruminatoris* G5837, GenBank ANFU01000022.1 (30)) were aligned using the Multiple Sequence Comparison by Log-Expectation (MUSCLE) tool (3.8.31) (31). Alignments were cured using Gblocks 0.91b (32) and phylogeny was assessed using PhyML 3.1/3.0 aLRT (33) with a bootstrapping procedure repeated 1000 times. Phylogenetic tree was drawn using TreeDyn (34).

The consensus sequence for DnaA boxes of the Spiroplasma group was generated by providing the intergenic regions upstream and downstream of the *dnaA* gene to the Multiple Em for Motif Elicitation (MEME) tool (35) using the “any number of repetitions” option and a maximum motif length of 15 pb. Precise locations of DnaA boxes within the *oriC* region of each Mollicute

chromosome were determined using the Motif Alignment and Search Tool (MAST) and the found MEME matrix (36). Positive and negative DNA strands were treated as separate strands, and only motifs with a p-value below 1.0×10^{-5} were considered as significant hits.

3.6.5 Plasmids construction

Plasmids and oligonucleotides used in this study are listed in Table 3.1 and Table S3.1, respectively. Detailed methodology of *oriC* plasmids construction is described in Supplementary material (Text S3.1). *M. florum oriC* plasmids were constructed as depicted in Figure 3.2. DNA fragments were amplified by PCR using VeraSeq 2.0 DNA polymerase (Enzymatics) and purified using Solid Phase Reversible Immobilization (SPRI) bead capture using Agencourt AMPure XP magnetic beads (Beckman Coulter) (37). Briefly, *M. florum oriC* fragments were amplified from *M. florum* L1 gDNA, *tetM* resistance cassette was amplified from pTT01, *colE1* replication origin was amplified from pUC19 (GenBank accession number: L09137), and $oriT_{RP4}$ was amplified from pSW23T (Genbank accession number: AY733066). PCRs fragments were assembled together using the Gibson Assembly Master Mix (New England BioLabs) to generate pMflT-o1, pMflT-o3, and pMflT-o4 plasmids. pMflT-o2 was built by circularizing the 3.6 kb fragment of pMflT-o4 ClaI digestion. pMflPT-o4, pMflST-o4, and pMflCT-o4 plasmids were generated by cloning the *pac*, *aadA1*, and *cat* resistance cassettes into the NotI site of pMflT-o4, respectively. pMflET-o4 was obtained by cloning the *ereB* resistance cassette into a pMflT-o4 derivative plasmid. pMcapT, pMmmT, pMmcT, and pSciT-o4 plasmids were created using the pMflT-o4 backbone and the heterologous *oriC* fragment of *M. capricolum*, *M. mycoides* or *S. citri* (Text S3.1 and Fig. S3.4). Plasmids were cloned in chemically competent *E. coli* strain EC100D *pir*+ cells, except for pMflPT-o4, which was cloned in *E. coli* strain MM294. Constructions were analyzed by restriction enzymes digestion, and *M. florum oriC* plasmid sequences were confirmed by paired-end Illumina sequencing at the Laboratoire de Génomique Fonctionnelle de l'Université de Sherbrooke (QC, Canada). Plasmids sequence and annotations are available in Genbank format at: http://lab-rodrigue.recherche.usherbrooke.ca/m_florum_plasmids/.

3.6.6 Polyethylene glycol transformation

M. florum L1 competent cells were prepared for PEG-mediated transformation by centrifuging 1 ml of a mid-logarithmic-phase bacterial culture at 21,100 x g for 1 min at 10°C. The cell pellet was washed with S/T buffer (10 mM Tris-HCl pH 6.5, 250 mM NaCl) and centrifuged again using the same conditions. Cells were resuspended in 200 µl of 0.1 M $CaCl_2$, incubated 30 min on ice, and then transformed using a PEG-mediated transformation procedure (3, 38). Briefly, 400 µl of modified ATCC 1161 medium (horse serum replaced by NaCl at a final concentration of 0.4% [w/v]) along with 1 µg of plasmid DNA were added to the previously resuspended cells and the solution was gently mixed by inverting the tube a few times. Then, one volume of 2X Fusion Buffer (20 mM Tris-HCl pH 6.5, 250 mM NaCl, 20 mM $MgCl_2$, 10% (w/v) PEG 8000) was immediately added and cells were gently mixed. Cells were incubated for 50 min at 34°C and then poured into 5 ml of pre-warmed ATCC 1161. The culture was gently mixed again and then incubated for 3 hours at 34°C without shaking. After, cells were centrifuged at 7,900 x g for 5 min at 10°C and the pellet was resuspended in 600 µl of ATCC 1161. Cells were serially diluted from 10^0 to 10^{-7} and plated on ATCC 1161 medium supplemented with tetracycline. To calculate transformation frequency, 5 µl of each dilution was also spotted on ATCC 1161 medium without tetracycline. Plates were incubated at 34°C, colonies were counted, and transformation frequency was calculated according to the number of transformants obtained per recipient CFU. Assays were performed using at least three independent biological replicates.

3.6.7 Southern blot hybridization

gDNA of isolated clones of *M. florum* L1 carrying pMflT-o3 or pMflT-o4 *oriC* plasmid was purified using the Quick-gDNA MiniPrep kit (Zymo Research) according to the manufacturer's specifications. 500 ng of gDNA was then digested at 37°C overnight using HindIII-HF restriction enzyme (New England BioLabs). After digestion, restriction fragments were separated on a 0.8% agarose gel, and DNA was depurinated and denaturated by soaking the gel for 15 min in 0.25 M HCl and 0.4 M NaOH, respectively. DNA was then transferred onto a

nylon membrane (Amersham Biosciences, Hybond^{TM}-XL) by capillarity using 0.4 M NaOH. DNA was fixed to the membrane by UV crosslinking (700 J) and blot prehybridized for 1 hour in Church Buffer (0.25 M $NaHPO_4$, 7% (w/v) SDS, 1X Denhardt, 1 mM EDTA). Labeled probe for *tetM* was synthesised by PCR from pMflT-o4 DNA template using OneTaq DNA polymerase (New England BioLabs), pBOT2-F/*tetM*-probe-R primer pair (Table S3.1), and 0.008 µM of EasyTide-[α-^{32}P]dCTP-3000 Ci/mmol 10 mCi/ml (PerkinElmer). The following cycling conditions were used: (i) 30 sec at 94°C; (ii) 30 cycles of 30 sec at 94°C, 30 sec at 55°C, and 45 sec at 68°C; (iii) 5 min at 68°C. Radiolabeled DNA probe was separated from unincorporated radioactive nucleotides using Bio-Spin Columns (Bio-Rad) according to manufacturer's recommendations. Purified *tetM* probe was next denatured at 95°C for 5 min, mixed with 10 ml of Church Buffer, and added to the membrane for hybridization at 65°C overnight with gentle shaking. After hybridization, the membrane was washed twice for 5 min each using 2X SSC (0.3 M NaCl, 30 mM sodium citrate) containing 1% (w/v) SDS at 50°C, and washed again using 0.2X SSC containing 1% (w/v) SDS at 55°C. Restriction fragments containing the *tetM* gene were finally visualized by autoradiography using a Typhoon FLA 9500 imaging system (GE Healthcare Life Sciences).

3.6.8 Quantification of *oriC* plasmids copy number

gDNA of isolated clones of *M. florum* L1 carrying pMflT-o3 or pMflT-o4 *oriC* plasmid, as well as wild-type *M. florum* L1 and *M. florum L1* clone 3632 (*mfl169*::Tn-*tetM*) was purified using the Quick-gDNA MiniPrep kit (Zymo Research) according to the manufacturer's specifications. Quantitative PCR (qPCR) assays targeting the *tetM* gene were performed using qPCR-*tetM*-F/qPCR-*tetM*-R primers (Table S3.1) and iQ SYBR Green Supermix (Bio-Rad) at a final concentration of 1X. Relative abundance of the *tetM* gene was calculated using the delta-delta Ct method (39) normalized to the *rpoB* (qPCR-*rpoB*-F/qPCR-*rpoB*-R) and *rpoC* (qPCR-*rpoC*-F/qPCR-*rpoC*-R) housekeeping genes (Table S3.1). qPCR amplifications were performed in triplicates using the following conditions: (i) 5 min at 95°C; (ii) 35 cycles of 15 sec at 95°C, 30 sec at 60°C, and 30 sec at 72°C; (iii) 5 min at 72°C. pMflT-o3 and pMflT-o4 copy number in

M. florum was determined by measuring the relative abundance of the *tetM* gene in 12 individual clones for each plasmid compared to the *M. florum* L1 clone 3632 control strain containing a single copy of the *tetM* gene (40).

3.6.9 Plasmids stability assays

One ml of a *M. florum* L1 log-phase culture carrying pMflT-o3 or pMflT-o4 growing in ATCC 1161 medium supplemented with tetracycline was centrifuged at 21,100 x g for 1 min at 4°C. Cell pellet was washed twice with 1 ml of ATCC 1161 medium without tetracycline, and then resuspended in 200 µl of the same medium that was used to inoculate 20 ml of ATCC 1161 medium without tetracycline. The culture was next maintained under the exponential growth phase using a Versatile Continuous Culture Device (VCCD) as previously described (15). 5 ml of culture was harvested every 12 hours for 48 hours, serially diluted from 10^0 to 10^{-7} and plated on non-selective ATCC 1161 medium and on ATCC 1161 supplemented with tetracycline. Plates were incubated at 34°C, colonies were counted, and plasmid stability was calculated according to the number of colonies growing on tetracycline divided by the number of colonies growing without tetracycline selection. Assays were performed using three independent biological replicates.

3.6.10 Conjugation assays

E. coli MFD*pir* (41) carrying pMflT-o4 and wild-type *M. florum* L1 were grown until mid-logarithmic growth phase, corresponding to ~2.5 x 10^7 CFU/ml and ~5.0 x 10^9 CFU/ml, respectively. Both cultures were centrifuged at 8,000 x g for 5 min, and cell pellets were resuspended in their original volume using fresh ATCC 1161 medium without penicillin and supplemented with 0.3 mM DAP (ATCC PEN-/DAP+). Conjugation assays were performed by mixing a variable volume of resuspended *M. florum* recipient cells with 1 ml of resuspended *E. coli* donor cells to obtain different mating ratios (see Table S3.3). For each mating ratio, mixed cells were centrifuged at 16,000 x g for 2 min and washed twice with ATCC PEN-/DAP+. Cells

were then resuspended in 30 µl of ATCC PEN-/DAP+ and the mating mixture was spotted on a 0.2 µm nitrocellulose filter (25 mm, Maine Manufacturing #1214898) laid on top of an ATCC PEN-/DAP+ plate. Conjugation plates were incubated at 30°C for 24 hours. Cells were recovered from the nitrocellulose filter using ATCC PEN-/DAP+ medium, and serially diluted from 10^0 to 10^{-7} before plating. To select exconjugants, cells were plated on ATCC 1161 medium supplemented with tetracycline and 50 µg/ml ampicillin. Recipient cells were selected by spotting 5 µl of the 10^0 to 10^{-7} dilutions on an ATCC 1161 plate supplemented with 50 µg/ml ampicillin. Plates were incubated at 34°C, colonies were counted, and conjugation frequencies were calculated according to the number of exconjugants obtained per recipient CFU. Assays were performed using three independent biological replicates.

3.6.11 Electroporation of *M. florum*

M. florum L1 cells were prepared for electroporation by centrifuging 1.0 ml of a mid-logarithmic-phase bacterial culture at 21,100 x g for 1 min at 4°C. The cell pellet was washed twice with an equal volume of electroporation buffer (272 mM sucrose, 1 mM HEPES pH 7.4). Cells were centrifuged again at 21,100 x g for 1 min at 4°C and the cell pellet was resuspended in 100 µl of electroporation buffer. 1 µg of plasmid DNA was added to 100 µl of previously prepared electrocompetent cells, and cells were transferred into a cold 1 mm electroporation cuvette. DNA was electroporated using a Gene Pulser Xcell electroporator system (Bio-Rad) set to 25 µF, 200 Ω, with a voltage varying from 0.5 to 3.0 kV. After electroporation, cells were recovered in 2 ml of ATCC 1161 medium and incubated at 34°C for 2 hours. Recovered cells were serially diluted from 10^0 to 10^{-7} and plated on ATCC 1161 medium supplemented with tetracycline. To calculate transformation frequency, 5 µl of each dilution was also spotted on ATCC 1161 medium without tetracycline. Plates were incubated at 34°C, colonies were counted, and transformation frequency was calculated according to the number of transformants obtained per recipient CFU. Assays were performed using three independent biological replicates.

3.7 Results

3.7.1 Antibiotic susceptibilities of *M. florum* L1

While several studies have established antibiotic susceptibilities in members of the Mollicutes class, the sensitivity of *M. florum* to some commonly used antibiotics was lacking. Using growth inhibition assays, we tested 12 antibiotics commonly used for genetic manipulation in bacteria (Table 3.2). We confirmed that some drugs were ineffective against *M. florum,* which could be used to eliminate contaminating bacteria when needed. As expected, *M. florum* L1 showed natural resistance to ampicillin, rifampicin, sulfamethoxazole and trimethoprim, displaying MICs above 100 µg/ml for each of these antibiotics (Table 3.2). Interestingly, *M. florum* was resistant to kanamycin and gentamicin but slightly susceptible to streptomycin and spectinomycin. *M. florum* also showed a high sensitivity to chloramphenicol, erythromycin and

Table 3.2. MICs of some common antibiotics against *M. florum* L1.

Antibiotic	MICs (µg/ml)
Ampicillin	> 100
Chloramphenicol	[5 - 8.5]
Erythromycin	[1 - 1.5]
Gentamicin	> 65
Kanamycin	> 100
Puromycin	[8 - 15.5]
Rifampicin	> 100
Spectinomycin	[25 - 50]
Streptomycin	[50 - 75]
Sulfamethoxazole	> 200
Tetracycline	≤ 10
Trimethoprim	> 100

puromycin, exhibiting MICs of 5 to 8.5 μg/ml, 1 to 1.5 μg/ml and 8 to 15.5 μg/ml, respectively (Table 3.2). Finally, *M. florum* showed a susceptibility to tetracycline with a MIC of less than 10 μg/ml.

3.7.2 Identification of putative DnaA boxes within the *oriC* region of *M. florum*

Previously, no self-replicative plasmid had been either identified in or developed for *M. florum*. The susceptibility of *M. florum* to tetracycline (Table 3.2) offers the possibility to take advantage of the widely used *tetM* resistance marker for plasmid selection. However, the localization of putative DnaA boxes in *M. florum* remains unknown, hindering our ability of developing plasmids based on the *oriC* region of the chromosome. We therefore compared the *oriC* region of eleven selected representative members of the Spiroplasma group using multiple sequence alignment (Fig. S3.1 and Table S3.4), and evaluated the phylogenetic relationships between species using sequence similarity (Fig. 3.1A). We observed that the differences in the *oriC* region sequence is consistent with the Mollicutes phylogeny based on conserved proteins (26, 42) and 16S rRNA sequences (43, 44). Mycoplasmas of the mycoides cluster (*M. leachii, M. capricolum, and M. mycoides*) shared an *oriC* region with a high percentage of nucleotide similarity (>90%), while *S. citri* and *S. kunkelii* were more phylogenetically distant and characterized by a more divergent *oriC* sequence (Fig. 3.1A and Table S3.4). As expected, *M. florum* was phylogenetically closer to the mycoides cluster than the spiroplasmas based on the *oriC* region sequence, but remained clearly separated from all analyzed mycoplasmas (Fig. 3.1A and Table S3.4).

We next hypothesized that the conservation property of the *oriC* region in the Spiroplasma group could be used to identify putative DnaA boxes in *M. florum*. We submitted the DNA sequence of the two intergenic regions flanking the *dnaA* gene found in representative species of the Spiroplasma group to the *de novo* motif discovery tool MEME (35), and detected a motif that is highly consistent with the previously proposed putative DnaA box consensus of Mollicutes (TT(A/T)TC(C/A)ACA) (21, 24) (Fig. 3.1B). We then searched the precise

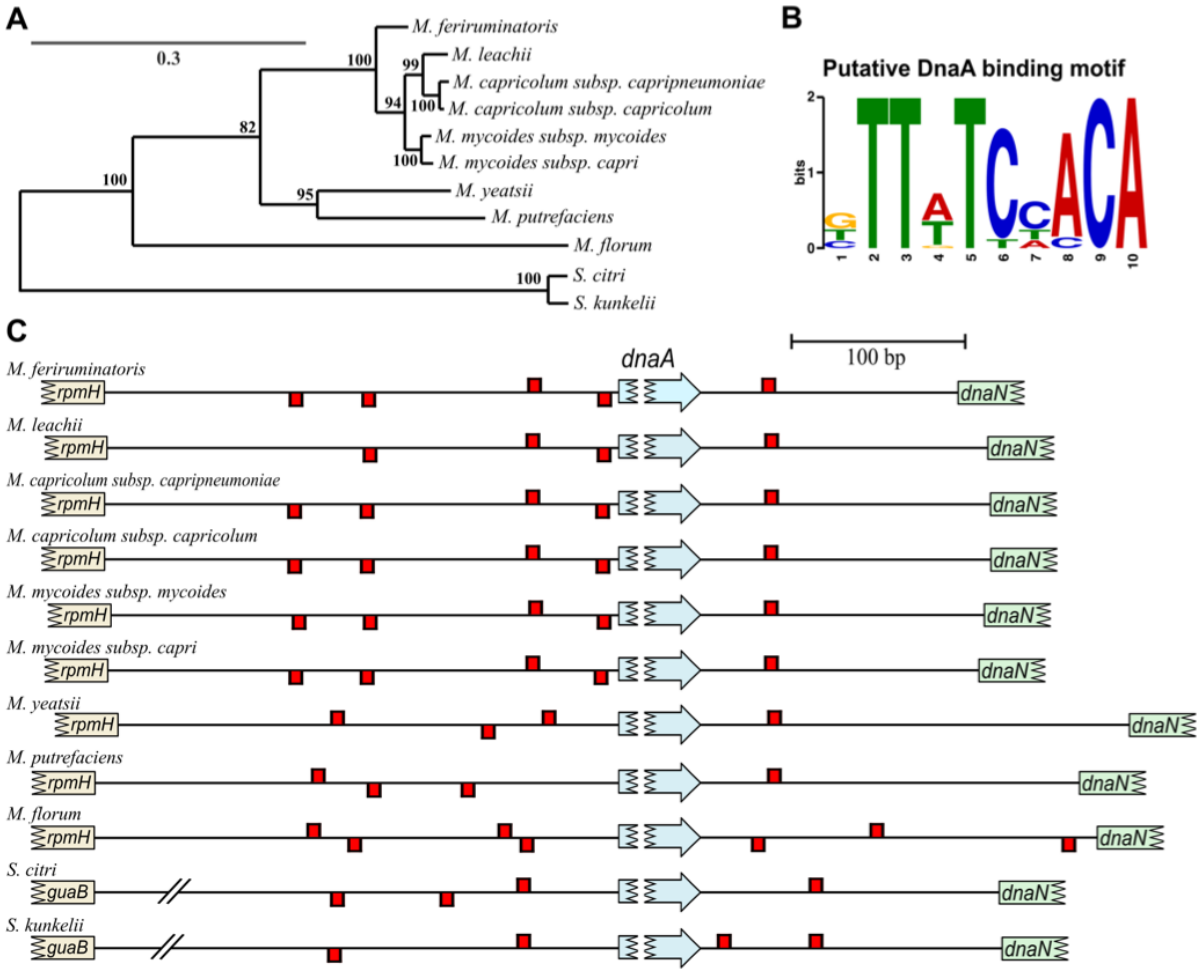

Figure 3.1. Sequence analysis of the *oriC* region of the Spiroplasma group. Sequence analysis of the predicted *oriC* region of eleven selected Mollicutes of the Spiroplasma group. (A) Phylogenetic tree based on the *oriC* region sequence of the chromosome using maximum likelihood. The number on each node indicates the percentage with which each branch topology was supported. The tree is drawn to scale, with branch lengths representing the number of substitution per site. (B) Putative DnaA binding motif found using MEME. (C) Localization of putative DnaA boxes within the intergenic regions upstream and downstream of *dnaA*. Putative DnaA boxes on positive and negative DNA strands are indicated by red rectangles positioned above and below the chromosomal line, respectively. Regions are drawn to scale. *S. citri* and *S. kunkelii guaB/dnaA* intergenic region is cut for presentation purposes, as well as represented genes.

localization of putative DnaA boxes within the two *oriC* intergenic regions using MAST (36), and observed that the number of DnaA boxes and their organization was reminiscent of the species phylogenetical relationships (Fig. 3.1A and C, Fig. S3.1, Table S3.5). For instance, members of the mycoides cluster all shared the same four putative DnaA boxes located at approximately 6 bp, 47 bp, 144 bp, 185 bp upstream of *dnaA*, with the exception of *M. leachii* in which the latter box was not detected due to a transversion mutation (C→A) at position 6 of the consensus sequence. Species of the mycoides cluster also shared a unique putative DnaA box located ~1,391 bp downstream of the start codon of *dnaA* (Fig. 3.1C, Fig. S3.1 and Table S3.5). Interestingly, this box is shared and highly conserved (7 out of 10 bp) between all eleven selected Mollicutes. *M. yeatsii*, *M. putrefaciens*, *M. florum* and spiroplasmas were distinguished from the mycoides cluster mostly by the number and position of putative DnaA boxes located upstream of *dnaA*. For example, putative DnaA boxes located ~6 bp and ~185 bp before the *dnaA* gene in the mycoides cluster were not detected in other analyzed Mollicutes (Fig. 3.1C, Fig. S3.1 and Table S3.5).

3.7.3 Development of *M. florum oriC*-based plasmids

In total, seven putative DnaA boxes were identified within the *oriC* region of *M. florum*. Four of them were located in the intergenic region between *rpmH* and *dnaA*, whereas three were found in the intergenic region between *dnaA* and *dnaN* (Fig. 3.1C, Fig. S3.1 and Table S3.5). Except for the two boxes located ~1,363 bp and ~1,545 bp downstream of the *dnaA* start codon, all DnaA boxes found in *M. florum* coincided with boxes found in one or many Mollicutes analyzed here. However, the importance of both intergenic regions for plasmid replication, as well as the presence of a copy of the *dnaA* gene remained to be established in *M. florum*. We therefore developed four different plasmids based on the localization of predicted DnaA boxes within the *oriC* region of the *M. florum* chromosome: two plasmids containing either the *rpmH*/*dnaA* or the *dnaA*/*dnaN* intergenic region (pMflT-o1 and pMflT-o2, respectively), one plasmid containing both regions but lacking the *dnaA* gene (pMflT-o3), as well as another plasmid including the whole *oriC-dnaA* locus (pMflT-o4) (Fig. 3.2). The *tetM* gene coding for

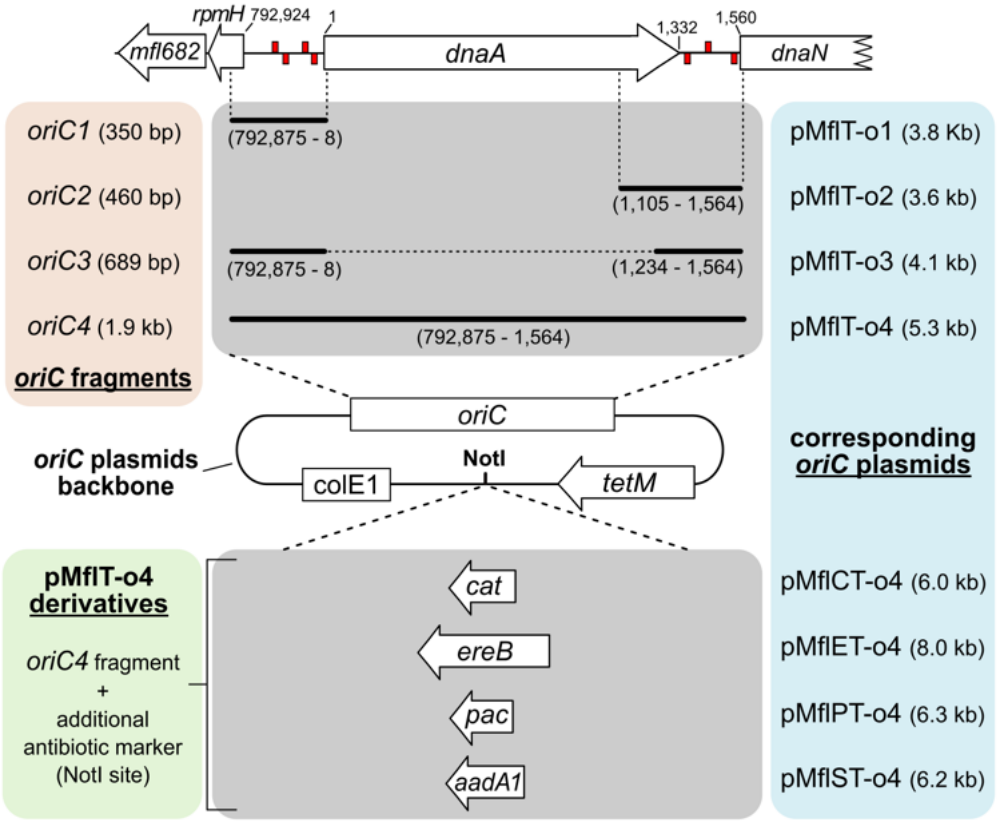

Figure 3.2. Schematic representation of *M. florum oriC*-based plasmids. *M. florum oriC* plasmids contain a variable *oriC* fragment, a colE1 replication origin, and a tetracycline resistance cassette (*tetM*). *oriC* fragments were based on the predicted *oriC* region of *M. florum* L1 chromosome, and their respective coordinates and size are indicated in brackets. Coordinates of the start codon of *rpmH* and *dnaN*, as well as the start and stop codons of *dnaA* are also indicated. Putative DnaA boxes found in the *rpmH/dnaA* and *dnaA/dnaN* intergenic regions (see Fig. 3.1C) are represented by red rectangles on positive and negative DNA strands. *oriC* fragments were assembled with *tetM* and colE1 fragments to produce pMflT-o1, -o2, -o3, and -o4 plasmids. Additional antimicrobial resistance gene cassettes were cloned in the NotI site of pMflT-o4 or a derivative plasmid to generate pMflCT-o4 (*cat*), pMflET-o4 (*ereB*), pMflPT-o4 (*pac*), or pMflST-o4 (*aadA1*).

a tetracycline ribosomal protection protein was chosen as a selectable marker in the *oriC* plasmids, and was specifically recoded to be functional in both *E. coli* and *M. florum*. Following assembly in *E. coli*, *oriC* plasmids were transformed in *M. florum* L1 by a PEG-mediated procedure (3, 38). Intriguingly, pMflT-o1 and pMflT-o2 failed to produce any detectable tetracycline resistant transformant, while pMflT-o3 and pMflT-o4 transformation resulted in

several hundreds to thousands of colonies on solid medium, with overall frequencies of 4.06 x 10^{-6} and 4.16 x 10^{-6} transformant per viable cell, respectively (Fig. 3.3A).

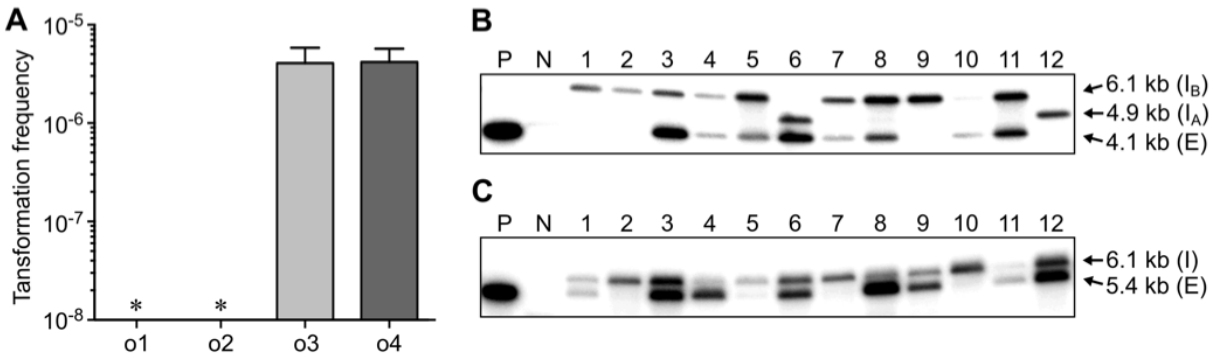

Figure 3.3. Transformation frequencies of *M. florum oriC* plasmids and recombination with the chromosome. (A) Transformation frequencies of *M. florum oriC* plasmids using polyethylene glycol (PEG)-mediated transformation procedure. o1, pMflT-o1; o2, pMflT-o2; o3, pMflT-o3; o4, pMflT-o4. Error bars indicate the standard deviation calculated from six independent biological replicates. Asterisks indicate transformation frequencies below the detection limit. Southern blot analysis of pMflT-o3 (B) and pMflT-o4 (C) recombination with the *M. florum* chromosome. Fragment size corresponding to the integrated and extrachromosomal forms of each plasmid are indicated. I: plasmid integrated at the *oriC* region of the chromosome. I_A: plasmid integrated at the *rpmH*/*dnaA* intergenic region. I_B: plasmid integrated at the *dnaA*/*dnaN* intergenic region. E: plasmid as an extrachromosomal element. Twelve isolated *M. florum* clones were analyzed for each plasmid (clone number indicated above each well). P: purified plasmid control. N: *M. florum* L1 WT genomic DNA (negative control).

Growth analysis revealed that pMflT-o4 transformants were not affected by tetracycline concentrations considerably higher than those tolerated by *M. florum* L1 (Fig. S3.2A). In fact, the *tetM* gene conferred a resistance to tetracycline concentrations exceeding 100 μg/ml (Table 3.3), a concentration at least 10 times higher than the MIC of the *M. florum* wild-type strain (Table 3.2). Similar results were also obtained for *M. florum* carrying pMflT-o3 (data not shown). Because additional selectable markers would offer a broader range of possibilities, genes conferring resistance to chloramphenicol (*cat*), erythromycin (*ereB*), puromycin (*pac*), and spectinomycin/streptomycin (*aadA1*) were introduced into pMflT-o4 to generate pMflCT-o4, pMflET-o4, pMflPT-o4, and pMflST-o4 plasmids, respectively (Fig. 3.2). We observed that

the *pac* gene included in pMflPT-o4 conferred a protection against more than 200 µg/ml of puromycin (Table 3.3 and Fig. S3.2B), a concentration 20 times higher than the MIC of the wild-type L1 strain (Table 3.2). Similar results were obtained with *M. florum* carrying pMflST-o4 growing in medium with or without spectinomycin or streptomycin (Fig. S3.2C and D, Table 3.3). For pMflET-o4, growth inhibition assays suggested a very weak protection against erythromycin that is not sufficient to be exploited robustly (data not shown). Similarly, our data suggest that the *cat* gene of the pMflCT-o4 plasmid is not functional in *M. florum* since no protection against chloramphenicol was observed (data not shown).

Table 3.3. MICs of *M. florum* carrying different antibiotic resistance markers.

Plasmid	Antibiotic	Gene conferring resistance	MICs (µg/ml)
pMflT-o4	Tetracycline	*tetM*	>100
pMflPT-o4	Puromycin	*pac*	>200
pMflST-o4	Spectinomycin	*aadA1*	>200
	Streptomycin	*aadA1*	>200

3.7.4 Homologous recombination with the host chromosome

Since *oriC*-based plasmids are known to frequently recombine at the *oriC* region of the chromosome due to sequence homology (18-24, 26), twelve *M. florum* isolated clones carrying pMflT-o3 or pMflT-o4 were analyzed by Southern blot using a radiolabeled probe targeting a region of the *tetM* gene to discriminate between the integrated and extrachromosomal forms of the plasmids (Fig. S3.3). Interestingly, all pMflT-o3 and pMflT-o4 tested clones showed the presence of recombination events with the host chromosome after an overnight growth with selective antibiotics (Fig. 3.3B and C). More specifically, the majority of pMflT-o3 tested clones exhibited a recombined form of the plasmid at the *dnaA*/*dnaN* intergenic region (10 out of 12), while only 2 clones showed a band corresponding to the recombined element at the *rpmH*/*dnaA*

region (Fig. 3.3B). In addition, a total of 17 out of 24 analyzed clones were found to carry the *oriC* plasmids as extrachromosomal elements (9/12 for pMflT-o3, and 8/12 for pMflT-o4). All clones that presented a band corresponding to the extrachromosomal form of the elements also showed a recombination event with the *oriC* region of *M. florum* chromosome (17/17), suggesting the presence of heterogeneous populations of cells deriving from a same initial colony (Fig. 3.3B and C). Taken together, these results indicate that plasmids based on the *oriC* of *M. florum* have a strong tendency to recombine with *oriC* region of the chromosome, regardless of the presence of a copy of the *dnaA* gene.

3.7.5 *oriC* plasmids copy number and stability

Using qPCR analysis, we next quantified the number of pMflT-o3 and pMflT-o4 copies per cell relatively to the *M. florum* chromosome. Plasmids copy number was determined by comparing the relative abundance of the *tetM* gene of individual pMflT-o3 and pMflT-o4 clones to the control strain *M. florum* L1 clone 3632 containing one copy of *tetM* integrated in the chromosome. We observed that the overall copy number of pMflT-o3 and pMflT-o4 was between 1 and 2 copies per *M. florum* genome (Fig. 3.4A). We then sought to determine if these *oriC* plasmids were stable over several generations by maintaining *M. florum* L1 carrying either pMflT-o3 or pMflT-o4 under continuous culture conditions for 48 hours without tetracycline. Colony counts revealed no significant reduction in tetracycline resistant colonies during and after continuous growth without selective pressure (Fig. 3.4B and C). Considering that *M. florum* has a doubling time of ~34 min in ATCC 1161 medium (14, 15), this indicates that pMflT-o3 and pMflT-o4 plasmids can be stably maintained for at least 85 generations without detectable loss.

3.7.6 Alternative transformation methods

Transformation by electroporation is generally successful with most cell types, and was previously reported for *S. citri* and *Mycoplasma genitalium* (3, 45). This method requires fewer

steps than PEG-mediated transformation, and could potentially offer higher transformation frequencies. However, we are not aware of any report documenting the successful transformation of *M. florum* using electroporation. We therefore optimized the electroporation procedure with *M. florum* using pMflT-o4 plasmid, and observed drastic effects of

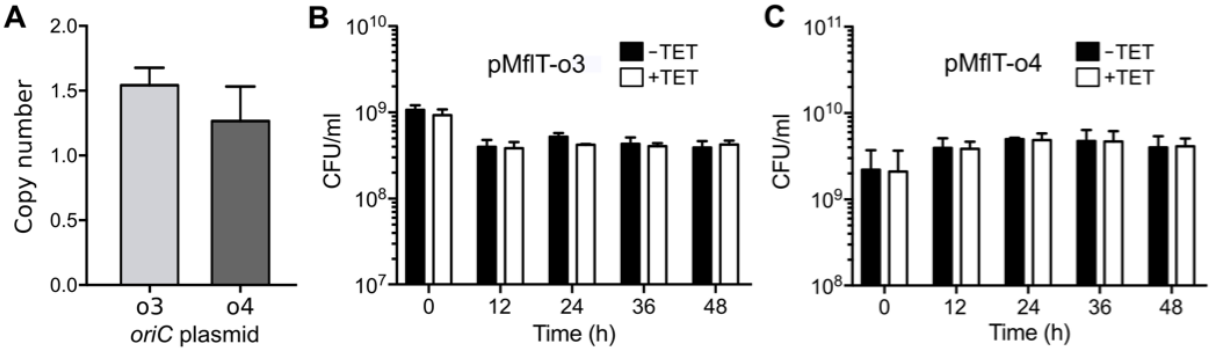

Figure 3.4. *M. florum oriC* plasmids copy number and stability. (A) Number of *oriC* plasmid per *M. florum* cell obtained by quantitative PCR targeted on the *tetM* gene. o3, pMflT-o3; o4, pMflT-o4. Evaluation of pMflT-o3 (B) and pMflT-o4 (C) stability in *M. florum* under continuous culture conditions for up to 48 hours. For each represented time point, cells were plated on ATCC 1161 solid medium with (+TET, white bars) or without (-TET, black bars) tetracycline and colony-forming units per ml (CFU/ml) were quantified. Error bars represent standard deviation from three independent biological replicates.

electroporation voltage on transformation frequency (Fig. 3.5A). Indeed, transformation frequency was just above the detection limit of approximately 1 x 10^{-9} transformant per viable cell when 0.5 kV was used (2.18 x 10^{-9} transformant per viable cell), while using 2.5 kV yielded more than 70,000 transformants per ml of *M. florum* culture (7.87 x 10^{-6} transformant per viable cell), which is comparable to the frequency observed for PEG-mediated transformation (Fig. 3.3A).

Bacterial conjugation is another common method to deliver plasmids in several species. Conjugation allows the mobilization of large DNA molecules, can reach high transfer frequencies, and is possible between phylogenetically distant organisms. However, we are not aware of any report of plasmid delivery between *E. coli* and Mollicutes. To investigate conjugation as another alternative transformation method for *M. florum*, we included the

transfer origin of broad-host range plasmid RP4 ($oriT_{RP4}$) in the backbone of our *oriC* plasmids and tested different mating ratios using the *E. coli* MFD*pir* strain (41) as a donor (Table S3.3). Our results indicate that plasmid conjugation can generate more than 400 colonies per experiment, reaching a frequency of 8.44 x 10^{-7} transformant per viable cell (Fig. 3.5B). This frequency is slightly lower than those observed for PEG-mediated transformation and electroporation (Fig. 3.3A and 3.5A). No colony was observed for controls lacking the donor or

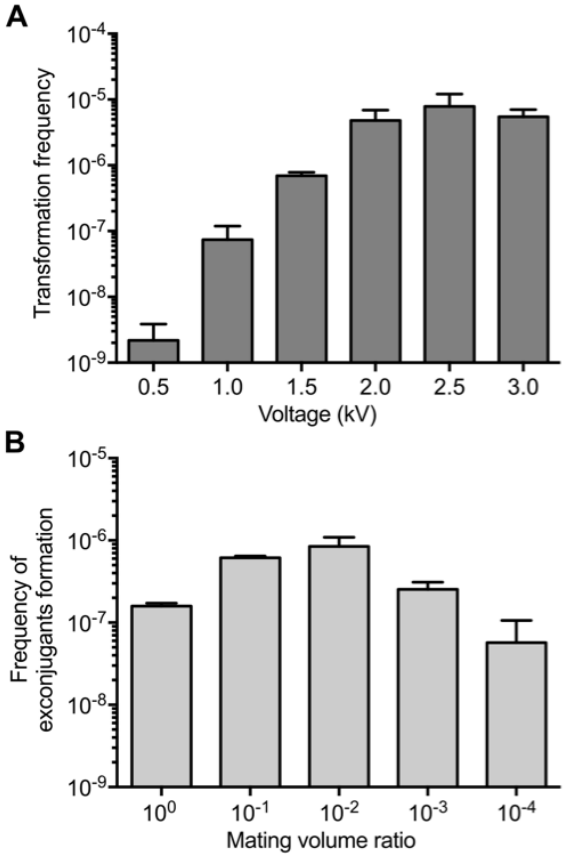

Figure 3.5. Frequencies of plasmid introduction in *M. florum* by electroporation or conjugation. (A) Transformation frequencies of pMflT-o4 in *M. florum* L1 using the electroporation procedure with 1mm cuvettes and different voltage values. Error bars indicate the standard deviation from three independent biological replicates. (B) pMflT-o4 transfer rates by conjugation using different mating volume ratios of donor (*E. coli* MFD*pir*) and recipient cells (*M. florum* L1). Indicated mating volume ratios are calculated by dividing the volume of *M. florum* culture by the volume of *E. coli* culture mixed during the conjugation process (see Table S3.3). pMflT-o4 transfer frequency is expressed as the number of exconjugants per viable recipient colony-forming unit (CFU). Error bars indicate the standard deviation from three independent biological replicates.

recipient cells. Co-incubation of *M. florum* cells with 1 µg of purified pMflT-o4 plasmid yielded only two tetracycline-resistant colonies in a single replicate out of three independent experiments, which sits right at the detection limit of our assay. PCR amplifications performed on exconjugants confirmed that the resulting clones truly harboured the pMflT-o4 plasmid and were not spontaneous mutants or contaminants (data not shown).

3.7.7 Transformation of *M. florum* with heterologous *oriC* plasmids

We previously observed that the *oriC* region of closely related Mollicutes shared some similarities relatively to their sequence and DnaA boxes organization (Fig. 3.1, Fig. S3.1, Tables S3.4 and S3.5). However, it is still unclear what degree of sequence divergence the replication machinery of Mollicutes can tolerate, and more specifically for *M. florum*. To better define these parameters, we investigated the capacity of *M. florum* to replicate heterologous *oriC* plasmids containing the *oriC* region and *dnaA* gene of closely related Mollicutes. Using PEG-mediated transformation, we first attempted to transform *M. florum* with *oriC* plasmids previously developed in *M. mycoides* (pMYCO1 and pMYSO1)*, M. capricolum* (pMCO3) and *S. citri* (pSD4) (21, 24, 25) (Fig. S3.4A and Table 3.1). Unfortunately, these plasmids failed to yield any transformant since their tetracycline resistance cassette was not properly expressed from the spiralin promoter in *M. florum* (data not shown) (19, 25). We therefore constructed four pMflT-o4 derivative plasmids in which the *oriC* region of *M. florum* was replaced by the *oriC* region of *M. mycoides* (pMmcT and pMmmT)*, M. capricolum* (pMcapT) and *S. citri* (pSciT-o4) (Fig. S3.4B and Table 3.1). These new heterologous *oriC* plasmids were all shown to confer tetracycline resistance in *E. coli.* However, in contrast to the *M. florum oriC*-based pMflT-o4 plasmid, none of the new heterologous *oriC* constructs yielded any tetracycline resistant colony when transformed in *M. florum*.

3.8 Discussion

In order to develop new genetic manipulation tools for the near-minimal bacterium *M. florum*, we investigated antibiotics susceptibility and *oriC* replication in this organism. We first validated that *M. florum* was indeed resistant to ampicillin, rifampicin, sulfamethoxaxole and trimethoprim (Table 3.2), which are class-specific resistances shared among members of the Mollicutes (46-50). We also observed that *M. florum* was resistant to kanamycin and gentamicin (Table 3.2). Interestingly, the sensitivity of Mollicutes to aminoglycosides has been reported to vary among strains and isolates (47, 50-55). Similarly to the Mollicutes rifampicin resistance (47, 48), it is likely that *M. florum* resistance to kanamycin and gentamicin depends on variations in the targeted gene products, e.g. the 16S rRNA of the 30S ribosome subunit. More importantly, we showed that *M. florum* was sensitive to antibiotics generally effective against Mollicutes (46, 47, 50, 56), i.e. tetracycline, chloramphenicol, erythromycin, and puromycin (Table 3.2). *M. florum* was also found to be relatively sensitive to streptomycin and spectinomycin (Table 3.2).

The evaluation of *M. florum* antibiotics susceptibilities allowed us to investigate the functionality of different markers frequently used in bacteria. As expected, *tetM* and *pac* genes conferred *M. florum* resistance to high concentrations of tetracycline and puromycin (Table 3.3, Fig. S3.2A and B). These markers were previously shown to be functional in several Mollicutes including *M. capricolum* and *M. mycoides* (18-26, 56). On the other hand, the functionality of the *aadA1* gene in *M. florum* was interesting since it is, to our knowledge, the first time that this genetic marker has been artificially introduced in a bacterium of the Mollicutes class (Table 3.3, Fig. S3.2C and D). The *cat* and *ereB* genes did not confer protection against their cognate antibiotics in *M. florum*. However, these markers were functional in *E. coli* carrying pMflCT-o4 and pMflET-o4 plasmids, and have been employed in other Mollicutes (57-61). Since *cat* and *ereB* were recoded to be functional in *E. coli* and in *M. florum*, it remains possible that they were not properly or sufficiently expressed in *M. florum* to confer a resistance phenotype.

Using available genomic sequences of Mollicutes closely related to *M. florum,* we constructed a putative DnaA binding motif and we identified putative DnaA boxes within previously uncharacterized *oriC* regions of members of the Spiroplasma group such as *M. leachii*, *M. putrefaciens*, and more importantly *M. florum* (Fig. 3.1B and C, Fig. S3.1, Table S3.5). Our predicted DnaA binding motif is highly consistent with the previously proposed putative DnaA box consensus of Mollicutes (TT(A/T)TC(C/A)ACA) and is reminiscent of the consensus sequence found in *E. coli* (21, 24, 27, 62). Furthermore, high-confidence putative DnaA boxes previously identified in *M. mycoides*, *M. capricolum*, and *S. citri* using *E. coli* DnaA binding consensus were successfully identified using our approach (24). Still, it is possible that more degenerate DnaA boxes exist and contribute to the chromosome replication in these bacteria, but were not detected by our motif according to our search parameters. For example, Lartigue et al. (24) proposed a degenerate putative DnaA box located ~30 bp from the start codon of *rpmH* in *M. mycoides* and *M. capricolum* that were not identified by our method (Fig. 3.1C and Fig. S3.1).

Plasmids harbouring both *M. florum oriC* intergenic regions, with or without a copy of the *dnaA* gene (pMflT-o4 and pMflT-o3, respectively), were found to transform *M. florum* at approximately the same frequency (Fig. 3.3A). These results indicate that *cis*-expression of the DnaA protein or the spacing provided by the *dnaA* gene between the two clusters of DnaA boxes is probably not essential for proper plasmid replication and maintenance in *M. florum*. Intriguingly, even if the majority of analyzed transformants showed extrachromosomal forms of the *oriC* plasmids after an overnight culture (Fig. 3.3B and C), we observed that recombination with the *M. florum* chromosome also occurred for all tested clones, corroborating previous observations indicating that *oriC* plasmids are highly recombinogenic in Mollicutes (18-24, 26). Since both the pMflT-o3 and pMflT-o4 plasmids were present in approximately one copy per cell relatively to the *M. florum* chromosome (Fig. 3.4A), this suggests that a dynamic state between the circular and the integrated forms of the plasmids may exist within a clonal population of cells. Nevertheless, we showed that both constructs were maintained for at least 85 generations (48 hours of continuous growth) without any selection (Fig. 3.4B and C).

It remains to be determined if the extrachromosomal form is disfavoured through time, and if the long-term *oriC* plasmid stability could be dependent on integration events. Additional experiments will also be necessary to engineer *M. florum oriC* plasmids to remain as extrachromosomal molecules, or conversely to perform specific gene targeting.

Using pMflT-o4, we also demonstrated that electroporation and conjugation are viable transformation methods for *M. florum* (Fig. 3.5), thus offering alternative procedures that require less material and hands-on time than the PEG-mediated transformation protocol. Interspecies conjugation from *E. faecalis* to *Mycoplasma gallisepticum* (63), *Mycoplasma arthritidis* (64), or *Mycoplasma hominis* (65) have previously been reported to deliver a Tn916 transposon. However, our results constitute the first reported example of plasmid conjugation from *E. coli* to a Mollicute species. Although the current results using the RP4 conjugation machinery showed slightly lower plasmid transfer rates than the electroporation and PEG-mediated transformation frequencies (Fig. 3.3A and Fig. 3.5), this approach might be improved with the use of alternative conjugative systems that could be better adapted for gene transfer into Mollicutes. For example, certain machineries could be better adapted for the absence of a cell wall or specific pili could stabilize the contact between *E. coli* and the comparatively small *M. florum* cells. It will be interesting to test whether our conjugation system is also working with other Mollicutes, and what factors might affect transfer frequency.

Interestingly, plasmids containing only one *oriC* intergenic region (pMflT-o1 and pMflT-o2) were not able to replicate in *M. florum* (Fig. 3.3A), while only the sole intergenic region located downstream of *dnaA* was shown to be sufficient for plasmid replication in *S. citri* (25). Unfortunately, minimization efforts have not been reported for *M. mycoides* and *M. capriolum oriC* plasmids, thus preventing any comparison with *M. florum*. We also observed that *oriC* plasmids containing the heterologous *oriC* region of *M. mycoides*, *M. capricolum*, and *S. citri* (Fig. S3.5) failed to replicate in *M. florum*. It is however unclear why *M. florum* failed to replicate these heterologous *oriC* plasmids since they contain their own heterologous *dnaA* gene. One possibility that could explain this host/plasmid incompatibility is that the heterologous

DnaA protein of *M. mycoides*, *M. capricolum*, and *S. citri* were not sufficiently expressed in *M. florum* context due to differences in the *dnaA* gene sequence, especially in the promoter region, or simply unable to interact with other proteins responsible for the DNA replication in *M. florum* (e.g. helicase). If this is the case, then the *M. florum* DnaA protein would have to properly recognize the DnaA boxes of the heterologous *oriC* regions to ensure plasmid replication. This recognition could however be impaired by divergences observed in the *oriC* region sequence and DnaA boxes organization of the Spiroplasma group. Indeed, *M. florum* and the mycoides cluster share 62% to 64% of nucleotide identity at this region, and only 57% with *S. citri* (Fig. 3.1A and C, Table S3.4). However, it was shown that even closely related species with high similarity of *oriC* region and DnaA boxes organization can fail to replicate heterologous *oriC* plasmids (18, 22, 24, 26). For instance, plasmids harbouring the *oriC* region of *M. mycoides* were shown to replicate in *M. capricolum* (92% nucleotide identity), whereas the reverse experiment was shown to be unsuccessful (24). Furthermore, *M. capricolum* was also recently shown to allow the replication of *oriC* plasmids developed from *S. citri*, *M. leachii*, *M. putrefaciens*, and more importantly *M. florum* (12). Besides *oriC* region similarities, it is clear that much remains to be understood about the factors allowing or limiting replication of heterologous *oriC* plasmids between Mollicute species. Broad host range vectors based on natural plasmid replicons could circumvent this limitation while also offering the possibility of introducing more than one plasmid per bacterium, and potentially allowing a wide range of copies per cell. So far, plasmids have been isolated from some mycoplasmas and spiroplasmas species (44, 66) but not from *Mesoplasma sp*. Additional work will be needed to experimentally test plasmids of interest in *M. florum*.

In summary, we reported the development of the first genetic tools specifically designed for the near-minimal bacterium *M. florum*: two *oriC* plasmid configurations (pMflT-o3 and pMflT-o4), three functional antibiotics resistance markers (*tetM*, *pac*, and *aadA1*) as well as three different transformation methods (PEG-mediated, electroporation, and conjugation). This initial set of genetic tools will now be available for introducing genes in *M. florum*, and will constitute a strong basis for other genetic engineering approaches. For example, *oriC* plasmids could be

used to insert genes required for whole bacterial chromosome cloning in *Saccharomyces cerevisiae*. This strategy is now possible for *M. florum* (12), which offers the opportunity to efficiently modify its genome using the powerful yeast genetic engineering tools. Whole-genome cloning and transplantation has notably been used for the creation of the first synthetic bacterial genome, and a quasi-minimal genome based on *M. mycoides subsp. capri* (8-11, 13), and will offer new opportunities for the development of a *M. florum* simplified cell chassis.

3.9 Acknowledgments

We are grateful to Carole Lartigue and Fabien Labroussaa for helpful discussions and for the kind gift of pMYCO1, pMYSO1, pMCO3, and pSD4 plasmids. We thank Joëlle Brodeur for technical assistance, and Alain Lavigueur for critical reading of the manuscript.

3.10 References

1. **Sirand-Pugnet P**, **Citti C**, **Barré A**, **Blanchard A**. 2007. Evolution of mollicutes: down a bumpy road with twists and turns. Res Microbiol **158**:754–66.
2. **Pettersson B**, **Johansson K-E**. 2002. Taxonomy of Mollicutes, p. 1–30. *In* Razin, S, Herrmann, R (eds.), Molecular Biology and pathogenicity of Mycoplasmas. Springer, New York (USA).
3. **Dybvig K**, **Voelker LL**. 1996. Molecular biology of Mycoplasmas. Annu Rev Microbiol **50**:25–57.
4. **Moran NA**. 2002. Microbial minimalism: Genome reduction in bacterial pathogens. Cell **108**:583–586.
5. **Mushegian A**, **Koonin E**. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci **93**:10268–10273.
6. **Peterson SN**, **Fraser CM**. 2001. The complexity of simplicity. Genome Biol **2**:comment2002.1-2002.8.
7. **McCoy RE**, **Basham HG**, **Tully JG**, **Rose DL**, **Carle P**, **Bové JM**. 1984. *Acholeplasma florum*, a New Species Isolated from Plants. Int J Syst Bacteriol **34**:11–15.
8. **Gibson DG**, **Glass JI**, **Lartigue C**, **Noskov VN**, **Chuang R-Y**, **Algire MA**, **Benders GA**, **Montague MG**, **Ma L**, **Moodie MM**, **Merryman C**, **Vashee S**, **Krishnakumar R**, **Assad-Garcia N**, **Andrews-Pfannkoch C**, **Denisova EA**, **Young L**, **Qi Z-Q**, **Segall-Shapiro TH**, **Calvey CH**, **Parmar PP**, **Hutchison CA**, **Smith HO**, **Venter JC**. 2010.

Creation of a bacterial cell controlled by a chemically synthesized genome. Science **329**:52–6.

9. **Benders GA**, **Noskov VN**, **Denisova EA**, **Lartigue C**, **Gibson DG**, **Assad-Garcia N**, **Chuang R-Y**, **Carrera W**, **Moodie M**, **Algire MA**, **Phan Q**, **Alperovich N**, **Vashee S**, **Merryman C**, **Venter JC**, **Smith HO**, **Glass JI**, **Hutchison CA**. 2010. Cloning whole bacterial genomes in yeast. Nucleic Acids Res **38**:2558–69.

10. **Lartigue C**, **Vashee S**, **Algire MA**, **Chuang R-Y**, **Benders GA**, **Ma L**, **Noskov VN**, **Denisova EA**, **Gibson DG**, **Assad-Garcia N**, **Alperovich N**, **Thomas DW**, **Merryman C**, **Hutchison CA**, **Smith HO**, **Venter JC**, **Glass JI**. 2009. Creating bacterial strains from genomes that have been cloned and engineered in yeast. Science **325**:1693–1696.

11. **Lartigue C**, **Glass JI**, **Alperovich N**, **Pieper R**, **Parmar PP**, **Hutchison CA**, **Smith HO**, **Venter JC**. 2007. Genome transplantation in bacteria: Changing one species to another. Science **317**:632–638.

12. **Labroussaa F**, **Lebaudy A**, **Baby V**, **Gourgues G**, **Matteau D**, **Vashee S**, **Sirand-Pugnet P**, **Rodrigue S**, **Lartigue C**. 2016. Impact of donor–recipient phylogenetic distance on bacterial genome transplantation. Nucleic Acids Res **44**:8501-8511.

13. **Hutchison CA**, **Chuang R-Y**, **Noskov VN**, **Assad-Garcia N**, **Deerinck TJ**, **Ellisman MH**, **Gill J**, **Kannan K**, **Karas BJ**, **Ma L**, **Pelletier JF**, **Qi Z-Q**, **Richter RA**, **Strychalski EA**, **Sun L**, **Suzuki Y**, **Tsvetanova B**, **Wise KS**, **Smith HO**, **Glass JI**, **Merryman C**, **Gibson DG**, **Venter JC**. 2016. Design and synthesis of a minimal bacterial genome. Science **351**:aad6253-1-aad6253-11.

14. **Baby V**, **Matteau D**, **Knight TF**, **Rodrigue S**. 2013. Complete genome sequence of the *Mesoplasma florum* W37 strain. Genome Annoucements **1**:e00879-13.

15. **Matteau D**, **Baby V**, **Pelletier S**, **Rodrigue S**. 2015. A Small-Volume, Low-Cost, and Versatile Continuous Culture Device. PLoS One **10**:e0133384.

16. **Navas-Castillo J**, **Laigret F**, **Tully J**, **Bové JM**. 1992. Mollicute *Acholeplasma florum* possesses a gene of phosphoenolpyruvate sugar phosphotransferase system and it uses UGA as tryptophan codon. C R Acad Sci III **315**:43–48.

17. **Tully JG**, **Whitcomb RF**, **Hackett KJ**, **Rose DL**, **Henegar RB**, **Bové JM**, **Carle P**, **Williamson DL**, **Clark TB**. 1994. Taxonomic descriptions of eight new non-sterol-requiring Mollicutes assigned to the genus *Mesoplasma*. Int J Syst Bacteriol **44**:685–93.

18. **Maglennon GA**, **Cook BS**, **Matthews D**, **Deeney AS**, **Bossé JT**, **Langford PR**, **Maskell DJ**, **Tucker AW**, **Wren BW**, **Rycroft AN**. 2013. Development of a self-replicating plasmid system for *Mycoplasma hyopneumoniae*. Vet Res **44**:1–10.

19. **Renaudin J**, **Marais A**, **Verdin E**, **Duret S**, **Foissac X**, **Laigret F**, **Bové JM**. 1995. Integrative and free *Spiroplasma citri oriC* plasmids: Expression of the *Spiroplasma phoeniceum* spiralin in *Spiroplasma citri.* J Bacteriol **177**:2870–2877.

20. **Chopra-Dewasthaly R**, **Marenda M**, **Rosengarten R**, **Jechlinger W**, **Citti C**. 2005. Construction of the first shuttle vectors for gene cloning and homologous recombination in *Mycoplasma agalactiae*. FEMS Microbiol Lett **253**:89–94.

21. **Cordova CMM**, **Lartigue C**, **Sirand-Pugnet P**, **Renaudin J**, **Cunha RAF**, **Blanchard A**. 2002. Identification of the origin of replication of the *Mycoplasma pulmonis* chromosome and its use in *oriC* replicative plasmids. J Bacteriol **184**:5426–5435.

22. **Lee S-W**, **Browning GF**, **Markham PF**. 2008. Development of a replicable *oriC* plasmid for *Mycoplasma gallisepticum* and *Mycoplasma imitans*, and gene disruption through homologous recombination in *M. gallisepticum*. Microbiology **154**:2571–80.

23. **Janis C**, **Lartigue C**, **Frey J**, **Wróblewski H**, **Thiaucourt F**, **Blanchard A**, **Sirand-Pugnet P**. 2005. Versatile use of *oriC* plasmids for functional genomics of *Mycoplasma capricolum* subsp. *capricolum*. Appl Environ Microbiol **71**:2888–2893.

24. **Lartigue C**, **Blanchard A**, **Renaudin J**, **Thiaucourt F**, **Sirand-Pugnet P**. 2003. Host specificity of mollicutes *oriC* plasmids: Functional analysis of replication origin. Nucleic Acids Res **31**:6610–6618.

25. **Lartigue C**, **Duret S**, **Garnier M**, **Renaudin J**. 2002. New plasmid vectors for specific gene targeting in *Spiroplasma citri*. Plasmid **48**:149–159.

26. **Sharma S**, **Citti C**, **Sagné E**, **Marenda MS**, **Markham PF**, **Browning GF**. 2015. Development and Host Compatibility of Plasmids for Two Important Ruminant Pathogens, *Mycoplasma bovis* and *Mycoplasma agalactiae*. PLoS One **10**:e0119000.

27. **Messer W**. 2002. The bacterial replication initiator DnaA. DnaA and *oriC*, the bacterial mode to initiate DNA replication. FEMS Microbiol Rev **26**:355–374.

28. **Andrews JM**. 2001. Determination of minimum inhibitory concentrations. J Antimicrob Chemother **48 Suppl 1**:5–16.

29. **Carle P**, **Saillard C**, **Carrère N**, **Carrère S**, **Duret S**, **Eveillard S**, **Gaurivaud P**, **Gourgues G**, **Gouzy J**, **Salar P**, **Verdin E**, **Breton M**, **Blanchard A**, **Laigret F**, **Bové JM**, **Renaudin J**, **Foissac X**. 2010. Partial chromosome sequence of *Spiroplasma citri* reveals extensive viral invasion and important gene decay. Appl Environ Microbiol **76**:3420–3426.

30. **Fischer A**, **Santana-Cruz I**, **Giglio M**, **Nadendla S**, **Drabek E**, **Vilei EM**, **Frey J**, **Jores J**. 2013. Genome Sequence of *Mycoplasma feriruminatoris* sp. nov., a Fast-Growing *Mycoplasma* Species. Genome Announc **1**:2012–2013.

31. **Edgar RC**. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32**:1792–1797.

32. **Castresana J**. 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Mol Biol Evol **17**:540–552.

33. **Guindon S**, **Gascuel O**. 2003. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Syst Biol **52**:696–704.

34. **Chevenet F**, **Brun C**, **Bañuls A-L**, **Jacq B**, **Christen R**. 2006. TreeDyn: towards dynamic graphics and annotations for analyses of trees. BMC Bioinformatics **7**:439.

35. **Bailey TL**, **Elkan C**. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol **2**:28–36.

36. **Bailey TL**, **Gribskov M**. 1998. Combining evidence using p-values: application to sequence homology searches. Bioinformatics **14**:48–54.

37. **Rodrigue S**, **Materna AC**, **Timberlake SC**, **Blackburn MC**, **Malmstrom RR**, **Alm EJ**, **Chisholm SW**. 2010. Unlocking Short Read Sequencing for Metagenomics. PLoS One **5**:e11840.

38. **King KW**, **Dybvig K**. 1991. Plasmid transformation of *Mycoplasma mycoides* subspecies *mycoides* is promoted by high concentrations of polyethylene glycol. Plasmid **26**:108–115.

39. **Livak KJ**, **Schmittgen TD**. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods **25**:402–408.

40. **Goryshin IY**, **Jendrisak J**, **Hoffman LM**, **Meis R**, **Reznikoff WS**. 2000. Insertional transposon mutagenesis by electroporation of released Tn5 transposition complexes. Nat Biotechnol **18**:97–100.

41. **Ferrières L**, **Hémery G**, **Nham T**, **Guérout AM**, **Mazel D**, **Beloin C**, **Ghigo JM**. 2010. Silent mischief: Bacteriophage Mu insertions contaminate products of *Escherichia coli* random mutagenesis performed using suicidal transposon delivery plasmids mobilized by broad-host-range RP4 conjugative machinery. J Bacteriol **192**:6418–6427.

42. **Bolanos LM**, **Servin-Garciduenas LE**, **Martinez-Romero E**. 2015. Arthropod-*Spiroplasma* relationship in the genomic era. FEMS Microbiol Ecol **91**:1–8.

43. **Sirand-Pugnet P**, **Lartigue C**, **Marenda M**, **Jacob D**, **Barré A**, **Barbe V**, **Schenowitz C**, **Mangenot S**, **Couloux A**, **Segurens B**, **De Daruvar A**, **Blanchard A**, **Citti C**. 2007. Being pathogenic, plastic, and sexual while living with a nearly minimal bacterial genome. PLoS Genet **3**:744–758.

44. **Breton M**, **Tardy F**, **Dordet-Frisoni E**, **Sagne E**, **Mick V**, **Renaudin J**, **Sirand-Pugnet P**, **Citti C**, **Blanchard A**. 2012. Distribution and diversity of mycoplasma plasmids: lessons from cryptic genetic elements. BMC Microbiol **12**:257.

45. **Renaudin J**, **Breton M**, **Citti C**. 2014. Molecular Genetic Tools of Mollicutes, p. 55–76. *In* Browning, GF, Citti, C (eds.), Mollicutes : Molecular Biology and Pathogenesis. Caister Academic Press, Wymondham, [England].

46. **Taylor-Robinson D**, **Bébéar C**. 1997. Antibiotic susceptibilities of mycoplasmas and treatment of mycoplasmal infections. J Antimicrob Chemother **40**:622–630.

47. **Bébéar CM**, **Bébéar C**. 2002. Antimycoplasmal Agents, p. 545–566. *In* Razin, S, Herrmann, R (eds.), Molecular Biology and pathogenicity of Mycoplasmas. Springer, New York (USA).

48. **Gaurivaud P**, **Laigret F**, **Bové JM**. 1996. Insusceptibility of members of the class Mollicutes to rifampin: Studies of the *Spiroplasma citri* RNA polymerase B-subunit gene. Antimicrob Agents Chemother **40**:858–862.

49. **Waites KB**, **Waites KB**, **Talkington DF**, **Talkington DF**. 2004. *Mycoplasma pneumoniae* and its role as human pathogen. Clin Microbiol Rev **17**:697–728.

50. **Waites KB**, **Lysnyansky I**, **Bébéar CM**. 2014. Emerging Antimicrobial Resistance in Mycoplasmas of Humans and Animals, p. 289–322. *In* Browning, GF, Citti, C (eds.), Mollicutes : Molecular Biology and Pathogenesis. BOOK, Caister Academic Press, Wymondham, [England].

51. **Uemura R**, **Sueyoshi M**, **Nagatomo H**. 2010. Antimicrobial susceptibilities of four species of *Mycoplasma* isolated in 2008 and 2009 from cattle in Japan. J Vet Med Sci **72**:1661–1663.

52. **Davis JW**, **Hanna BA**. 1981. Antimicrobial susceptibility of *Ureaplasma urealyticum*. J Clin Microbiol **13**:320–325.

53. **Hannan PCT**. 1995. Antibiotic susceptibility of *Mycoplasma fermentans* strains from various sources and the development of resistance to aminoglycosides *in vitro*. J Med Microbiol **42**:421–428.

54. **Hannan PCT**. 1997. Observations on the possible origin of *Mycoplasma fermentans* incognitus strain based on antibiotic sensitivity tests. J Antimicrob Chemother **39**:25–30.

55. **Waites KB**, **Talkington DF**. 2004. *Mycoplasma pneumoniae* and its role as a human pathogen. Clin Microbiol Rev **17**:697-728.

56. **Algire MA**, **Lartigue C**, **Thomas DW**, **Assad-Garcia N**, **Glass JI**, **Merryman C**. 2009. New selectable marker for manipulating the simple genomes of *Mycoplasma* species. Antimicrob Agents Chemother **53**:4429–4432.

57. **Duret S**, **André A**, **Renaudin J**. 2005. Specific gene targeting in *Spiroplasma citri*: improved vectors and production of unmarked mutations using site-specific recombination. Microbiology **151**:2793–2803.

58. **Dybvig K**. 1989. Transformation of *Acholeplasma laidlawii* with Streptococcal Plasmids pVA868 and pVA920. Plasmid **21**:155–160.

59. **Hahn TW**, **Mothershed EA**, **Waldo RH**, **Krause DC**. 1999. Construction and analysis of a modified Tn4001 conferring chloramphenicol resistance in *Mycoplasma pneumoniae*. Plasmid **41**:120–124.

60. **Dybvig K**, **French CT**, **Voelker LL**. 2000. Construction and use of derivatives of transposon Tn4001 that function in *Mycoplasma pulmonis* and *Mycoplasma arthritidis*. J Bacteriol **182**:4343–4347.

61. **King KW**, **Dybvig K**. 1994. Mycoplasmal cloning vector derived from plasmid pKMK1. Plasmid **31**:49–59.

62. **Robison K**, **McGuire AM**, **Church GM**. 1998. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. J Mol Biol **284**:241–254.

63. **Ruffin DC**, **van Santen VL**, **Zhang Y**, **Voelker LL**, **Panangala VS**, **Dybvig K**. 2000. Transposon mutagenesis of *Mycoplasma gallisepticum* by conjugation with *enterococcus faecalis* and determination of insertion site by direct genomic sequencing. Plasmid **44**:191–5.

64. **Voelker LL**, **Dybvig K**. 1996. Gene transfer in *Mycoplasma arthritidis*: transformation, conjugal transfer of Tn916, and evidence for a restriction system recognizing AGCT. J Bacteriol **178**:6078–81.

65. **Roberts MC**, **Kenny GE**. 1987. Conjugal Transfer of Transposon Tn916 from *Streptococcus faecalis* to *Mycoplasma hominis.* J Bacteriol **169**:3836–3839.

66. **Marenda MS**. 2014. Genomic mosaics, p. 15–54. *In* Browning, GF, Citti, C (eds.), Mollicutes : Molecular Biology and Pathogenesis. Caister Academic Press, Wymondham, [England].

67. **Yanisch-Perron C**, **Vieira J**, **Messing J**. 1985. Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. Gene **33**:103–119.

68. **Demarre G**, **Guérout AM**, **Matsumoto-Mashimo C**, **Rowe-Magnus DA**, **Marlière P**, **Mazel D**. 2005. A new family of mobilizable suicide plasmids based on broad host range R388 plasmid (IncW) and RP4 plasmid (IncPα) conjugative machineries and their cognate *Escherichia coli* host strains. Res Microbiol **156**:245–255.

3.11 Supplemental material

3.11.1 Molecular biology methods

Genes conferring resistance to tetracycline (*tetM)*, puromycin (*pac*), streptomycin and spectinomycin (*aadA1*), chloramphenicol (*cat*), and erythromycin (*ereB*) were recoded using a compromise codon table (Table S3.2) to obtain functional proteins (GenBank accession numbers: WP_000691749, AHL28657, WP_001206315, KLX70575, and WP_032488343, respectively) in *Escherichia coli* as well as in *Mesoplasma florum*. The *pac*, *aadA1*, and *cat* resistance genes were synthesized in gBlocks fragments (Integrated DNA Technologies), *ereB* resistance gene was obtained from Biobasic's gene synthesis service (*ereB*-pUC57), and *tetM* resistance cassette was amplified from the pTT01 plasmid (Table 3.1). All PCRs were performed using VeraSeq 2.0 DNA polymerase (Enzymatics) and primers listed in Table S3.1. PCR conditions were as follows: (i) 30 sec at 95°C; (ii) 30 cycles of 10 sec at 95°C, 30 sec at the appropriate annealing temperature, and 30 sec/kb at 72°C; (iii) 2 min at 72°C. PCR products were purified using Solid Phase Reversible Immobilization (SPRI) bead capture using Agencourt AMPure XP magnetic beads (Beckman Coulter) (1). Wild-type *M. florum* L1

genomic DNA (gDNA) was extracted using the Quick-gDNA MiniPrep kit (Zymo Research) according to the manufacturer's specifications. All plasmids generated in this study were cloned in chemically competent *E. coli* strain EC100D *pir*+ cells, except for pMflPT-o4 was cloned in *E. coli* strain MM294. Plasmid DNA was extracted using the EZ-10 Spin Column Plasmid DNA Minipreps kit (Biobasic) according to the manufacturer's instructions. Plasmid constructions were analyzed by restriction enzymes digestion, and *M. florum oriC* plasmids sequence was confirmed by paired-end Illumina sequencing at the Laboratoire de Génomique Fonctionnelle de l'Université de Sherbrooke (QC, Canada). Plasmids sequence and annotations are available at http://lab-rodrigue.recherche.usherbrooke.ca/m_florum_plasmids/.

3.11.2 Construction of *M. florum oriC* plasmids

Plasmids and oligonucleotides used in this study are listed in Table 3.1 and in Table S3.1, respectively. *M. florum oriC* plasmids were constructed as depicted in Figure 3.2. The different *M. florum oriC* fragments were all amplified from *M. florum* L1 gDNA. Regions upstream and downstream of *dnaA*, designated *oriC1* and *oriC2* respectively, were amplified using *oriC1*-F/*oriC1*-R and *oriC2*-F/*oriC2*-R primer pairs. To build the *oriC3* fragment, *oriC1* and *oriC2* regions containing an overlapping sequence of 40 bp were amplified using *oriC1*-F/*oriC3*-R and *oriC3*-F/*oriC2*-R primer pairs. The *oriC3* fragment was then assembled by fusion PCR using flanking primers *oriC1*-F/*oriC2*-R. *dnaA* gene along with both upstream and downstream regions, designated *oriC4*, was first amplified in four distinct fragments using *oriC1*-F/*oriC4*-1-R, *oriC4*-1-F/*oriC4*-2-R, *oriC4*-2-F/*oriC4*-3-R and *oriC4*-3-F/*oriC2*-R primer pairs. The complete *oriC4* region was next assembled by fusion PCR using flanking primers *oriC1*-F/*oriC2*-R. *tetM* resistance cassette containing the pBOT1 promoter from pBOT1 plasmid (2) was amplified from pTT01 using *tetM*-F/*tetM*-R and *tetM*-1-F/*tetM*-R primer pairs to build pMflT-o1 and pMflT-o3/-o4, respectively. ColE1 replication origin was amplified from pUC19 (GenBank accession number: L09137) using colE1-F/colE1-R primers and $oriT_{RP4}$ was amplified from pSW23T (Genbank accession number: AY733066) using RP4-F/RP4-R primers. The colE1 and $oriT_{RP4}$ fragments were then assembled by fusion PCR using colE1-

F/RP4-R flanking primers. PCR fragments *oriC1*, *oriC3,* and *oriC4* were assembled with the appropriate *tetM*, colE1, and $oriT_{RP4}$ fragments using the Gibson Assembly Master Mix (New England BioLabs) to build pMflT-o1, pMflT-o3, and pMflT-o4, respectively. To generate pMflT-o2, pMflT-o4 was digested with ClaI, the resulting 3.6 kb band was purified using the Zymoclean Gel DNA Recovery kit (Zymo Research), and the purified fragment was circularized using the T4 DNA ligase.

3.11.3 Construction of *M. florum oriC* plasmids derivatives

In order to build pMflPT-o4, pMflST-o4, and pMflCT-o4 plasmids (see Fig. 3.2 and Table 3.1), pMflT-o4 was linearized using NotI and the corresponding resistance cassette was cloned using the Gibson Assembly Master Mix (New England BioLabs). For pMflPT-o4, pBOT1 promoter was amplified from pTT01 and *pac* resistance gene was amplified from gBlocks fragments using pBOT1-F/pBOT1-R and *pac*-F/*pac*-R primer pairs (Table S3.1), respectively. To construct the *pac* resistance cassette, pBOT1 and *pac* fragments were assembled by fusion PCR using pBOT1-F*/pac*-R flanking primers. For pMflST-o4, *aadA1* resistance cassette containing P_{N25} promoter (3) was amplified from gBlocks fragment using *aadA1-F*/*aadA1*-R primers. For pMflCT-o4, P_{N25} promoter and *cat* resistance gene were amplified from gBlocks fragments using P_{N25}-F/P_{N25}-R and *cat*-F/*cat*-R primer pairs, respectively. The P_{N25} and *cat* fragments were assembled by fusion PCR using P_{N25}-*F/cat*-R flanking primers to construct the *cat* resistance cassette. pMflET-o4 was constructed by cloning the *ereB* resistance cassette, obtained from *ereB*-pUC57 AscI/XhoI digestion, into a pMflT-o4 derivative plasmid.

3.11.4 Construction of heterologous *oriC* plasmids

M. florum heterologous *oriC* plasmids were constructed as depicted in Figure S3.4B. pMflT-o4 backbone was amplified using pMfl-F/pMfl-R primer pair (Table S3.1), digested using XhoI/PvuI, and dephosphorylated with Antarctic phosphatase (New England BioLabs). Heterologous *dnaA*/*oriC* regions were amplified from pMCO3 (*Mycoplasma capricolum subsp.*

capricolum), pMYSO1 (*Mycoplasma mycoides subsp. mycoides*), pMYCO1 (*M. mycoides subsp. capri*), and pSD4 (*Spiroplasma citri*) plasmids (Table 3.1 and Fig. S3.4A) using *Mcap*-F/*Mcap*-R, *Mm*-F/*Mmm*-R, *Mm*-F/*Mmc*-R, and *Sci*-F/*Sci*-R primers pairs (Table S3.1), respectively. *OriC* fragments were then digested with XhoI/PvuI and phosphorylated using T4 phosphonucleotide kinase (Enzymatics). To construct pMcapT, pMmmT, pMmcT, and pSciT-o4 (Table 3.1 and Fig. S3.4B), pMflT-o4 backbone and the corresponding heterologous *oriC* fragment were ligated using T7 DNA ligase (New England BioLabs).

3.11.5 Supplementary Figures

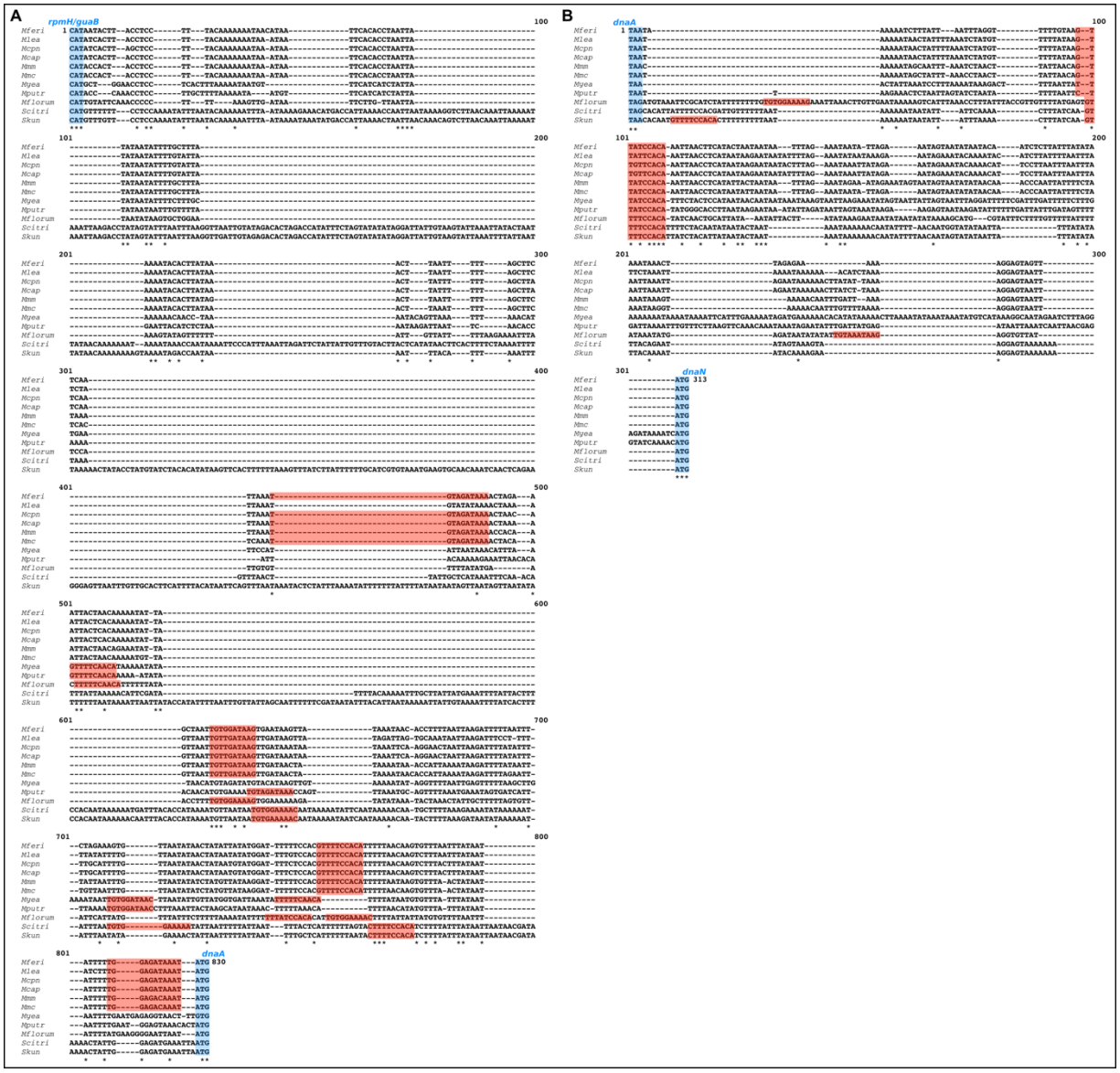

Figure S3.1. DNA sequence alignment of the intergenic regions upstream (A) and downstream (B) of *dnaA* in selected species of the Spiroplasma group. Start and stop codons of surrounding genes are highlighted in blue, and putative DnaA boxes (see Fig. 3.1C) are highlighted in red. Perfectly conserved nucleotides are indicated with asterisks. *Mferi*, *Mycoplasma feriruminatoris*; *Mlea*, *Mycoplasma leachii*; *Mcpn*, *M. capricolum subsp. capripneumoniae*; *Mcap*, *M. capricolum subsp. capricolum*; *Mmm*, *M. mycoides subsp. mycoides*; *Mmc*, *M. mycoides subsp. capri*; *Myea*, *Mycoplasma yeatsii*; *Mputr*, *Mycoplasma putrefaciens*; *Mflorum*, *M. florum*; *Scitri*, *S. citri*; *Skun*, *Spiroplasma kunkelii.*

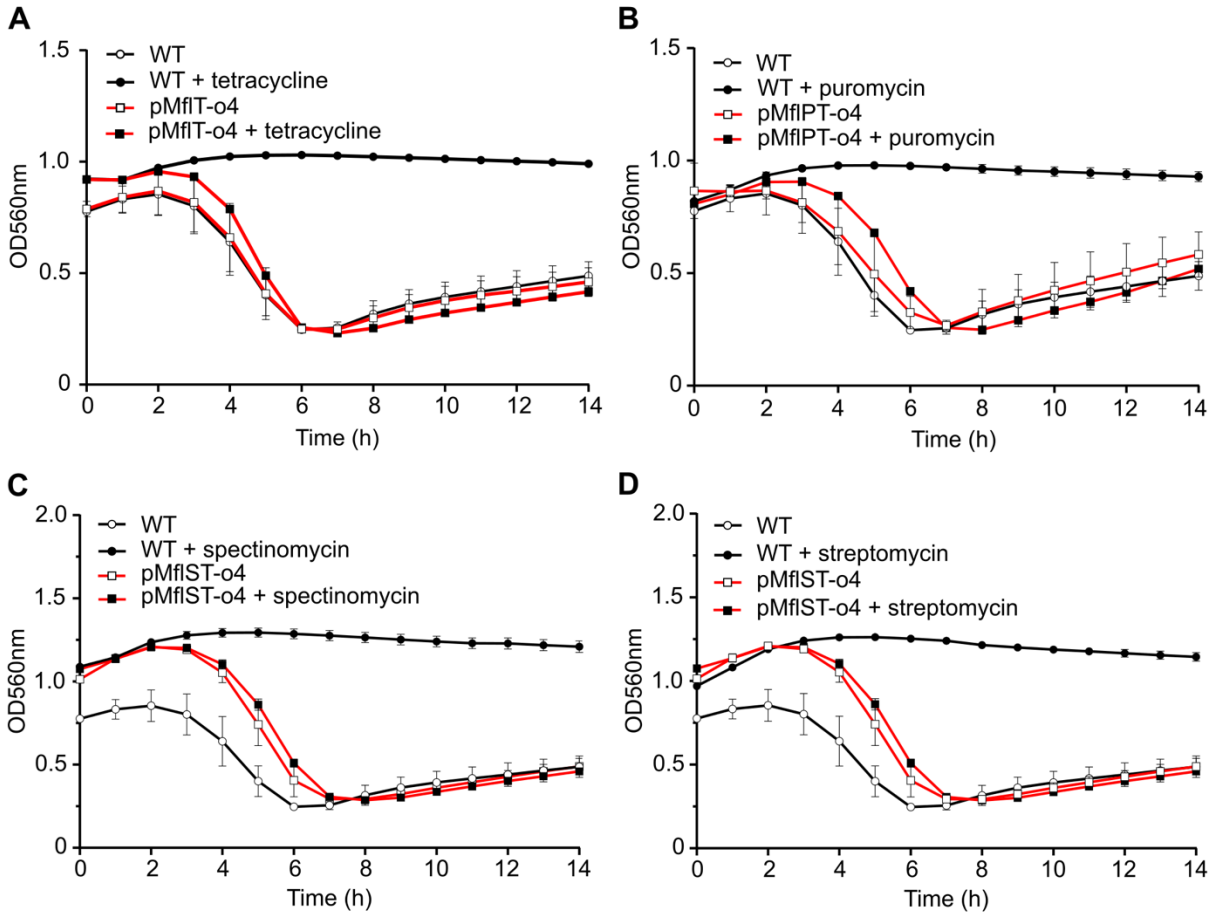

Figure S3.2. Growth curves of *M. florum* L1 wild-type strain (WT) and *M. florum* L1 carrying pMflT-o4 (A), pMflPT-o4 (B) or pMflST-o4 (C and D) in ATCC 1161 medium with or without the indicated antibiotics. Growth was monitored by a decrease in absorbance of phenol red at 560 nm caused by the acidification of the culture medium that correlates with *M. florum* cell concentrations (4). Antibiotics were used at the following concentrations: tetracycline, 15 µg/ml; puromycin, 30 µg/ml; spectinomycin and streptomycin, 100 µg/ml. Each data point represents the mean and the standard deviation of values obtained from three independent experiments.

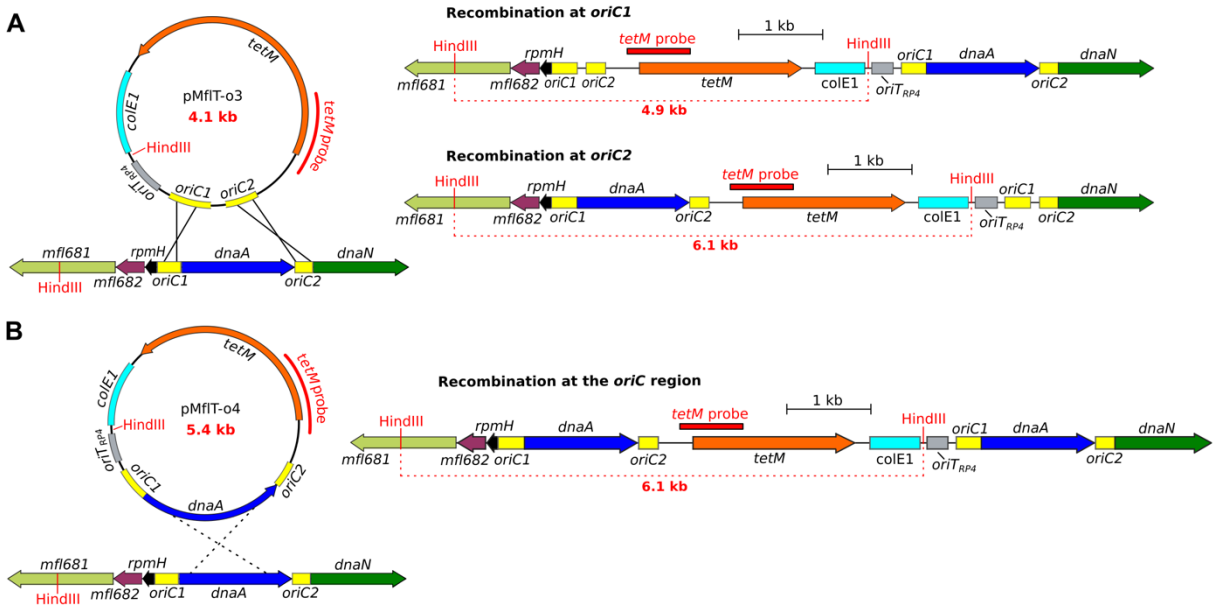

Figure S3.3. Schematic representation of pMflT-o3 (A) and pMflT-o4 (B) plasmids recombination with the *oriC* region of the *M. florum* chromosome. The localization of HindIII sites and *tetM* probe used for Southern blot analysis are indicated in red, as well as the resulting fragment size. Genes are drawn to scale.

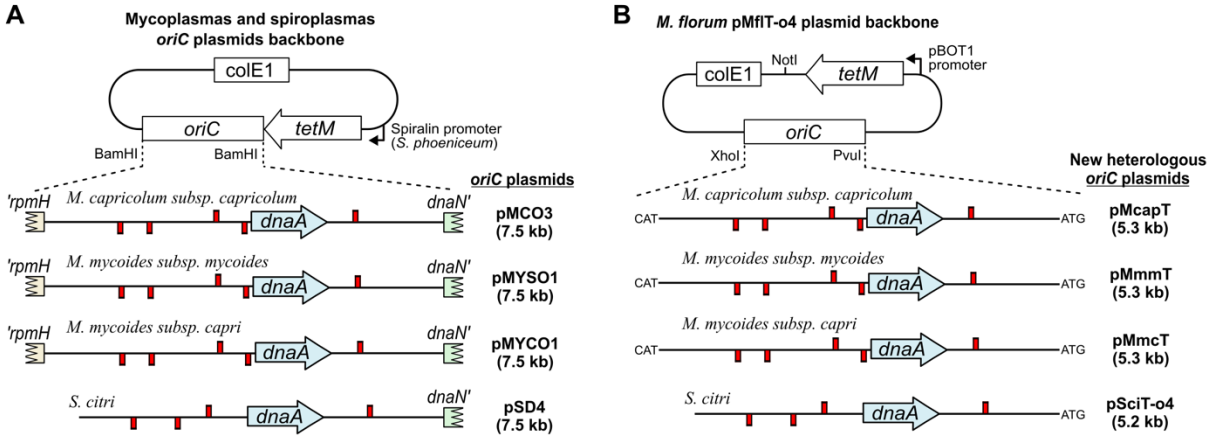

Figure S3.4. Schematic representation of *M. florum* heterologous *oriC* plasmids. Putative DnaA boxes found within the intergenic regions upstream and downstream of *dnaA* are represented by red boxes as shown in Figure 3.1C. (A) Principal features of *M. capricolum*, *M. mycoides*, and *S. citri oriC* plasmids developed by Lartigue et al. and used in this study (2, 5, 6). (B) Depiction of the new *M. florum* heterologous *oriC* plasmids developed in this study (see Text S3.1 for construction details). Each plasmid contains a pMflT-o4 backbone in which the complete *M. florum oriC* region was replaced by a heterologous *oriC* region depicted in A.

3.11.6 Supplementary Tables

Table S3.1. Primers used in this study.

Name	Nucleotide sequence (5' to 3')
oriC1-F	GCGATGAGGATGAACCTACTCAGGGACCTGATGTGACGTGCCGTGAACACGAGCGTGTTT
oriC1-R	ATTAACCCTCACTAAAGGGAGTTTCCATATTAATTCCCCT
oriC2-F	GCGATGAGGATGAACCTACTCAGGGACCTGATGTGACGTGGTTTTAAGTGCTATTTCGCG
oriC2-R	ATTAACCCTCACTAAAGGGACTCATATAACACCTCTTATTTAC
oriC3-R	CGCGAAATAGCACTTAAAACGTTTCCATATTAATTCCCCT
oriC3-F	AGGGGAATTAATATGGAAACGTTTTAAGTGCTATTTCGCG
oriC4-1-R	CAATATATTCTTCAATAATGTCTTGGTCTATTAATTTTTCCTTCTTTAACTTA
oriC4-1-F	TAAGTTAAAGAAGGAAAAATTAATAGACCAAGACATTATTGAAGAATATATTG
oriC4-2-R	TAAAATAGTTACACCAAGATTACTTCTAACAAGGATTACGAACTCTG
oriC4-2-F	CAGAGTTCGTAATCCTTGTTAGAAGTAATCTTGGTGTAACTATTTTA
oriC4-3-R	CGCGAAATAGCACTTAAAACTGTACTATGGTCTCTTCCACCAAATTCTGCACC
oriC4-3-F	GGTGCAGAATTTGGTGGAAGAGACCATAGTACAGTTTTAAGTGCTATTTCGCG
tetM-F	AGGGGAATTAATATGGAAACTCCCTTTAGTGAGGGTTAAT
tetM-R	TATGCATCAGTATCGCATAACGACCATATAGCCCTACCTACTAAATTACCCTGTTATCCC
tetM-1-F	GTAAATAAGAGGTGTTATATGAGTCCCTTTAGTGAGGGTTAAT
colE1-F	TAGGTAGGGCTATATGGTCGTTATGCGATACTGATGCATAAAGGCCGCGTTGCTGGCGTT
colE1-R	GCTCCAGCTTTTGTTCCCTTTAGTGAGGGTTAATTGCGCGTTGAGATCCTTTTTTTCTGC
RP4-F	CGCGCAATTAACCCTCACTAAAGGGAACAAAAGCTGGAGCGCTGCAGGAATTCGATATCA
RP4-R	CACGTCACATCAGGTCCCTGAGTAGGTTCATCCTCATCGCTGGGTACCAGCGCTTTTCCG
pBOT1-F	CGTTATATGTTCAACAAAATCACATAATAATACTAGTAGCTCCCTTTAGTGAGGGTTAAT
pBOT1-R	GTTGGTTTGTATTCAGTCATCTAGATTTCCTCCATTCAAA
pBOT2-F	GTCGTGACTGGGAAAACCCT
tetM-probe-R	TCCGTTTTGGTCAATTTTGTT
pac-F	TTTGAATGGAGGAAATCTAGATGACTGAATACAAACCAAC
pac-R	CCTAAATCGTATGCCCTATAGTGAGTCGTATTACTGCAGCTTATGCACCTGGTTTTCTTG
aadA1-F	CAAGAAAACCAGGTGCATAATCCCTTTAGTGAGGGTTAAT

Table S3.1. Primers used in this study (continued).

aadA1-R	ATTAACCCTCACTAAAGGGATTATGCACCTGGTTTTCTTG
P_{N25}-*F*	CGTTATATGTTCAACAAAATCACATAATAATACTAGTAGCTCATAAAAAATTTATTTGCT
P_{N25}-*R*	CATCTAGATTTCCTCCATATAGT
cat-F	GATTCATACGACTCACTATATGGAGGAAATCTAGATGGAAAAAAAGATAACTGGGTACAC
cat-R	CCTAAATCGTATGCCCTATAGTGAGTCGTATTACTGCAGCTCGAGTTAAGCTCCACCTTG
qPCR-*tetM*-F	TGACCGTGCATATTCAGGTG
qPCR-*tetM*-R	TCACGTTGTTCAGGTTTGCT
qPCR-*rpoB*-F	CATGGCTGAAGCTGGAATGGAAAACTATGG
qPCR-*rpoB*-R	CGTTGTCCCCCGTTTTGTGC
qPCR-*rpoC*-F	CCTAAAGATGGAAAAGCGATTG
qPCR-*rpoC*-R	TCAACTGCAATCCCAATAACTG
pMfl-F	CTTATCTTAAAAGGAGTAATTATGCGATCGTCCCTTTAGTGAGGGTTAATGTCGTGACTG
pMfl-R	ATTTTTTTGTAAAGGAGGTAAGTGATATGCTCGAGCACGTCACATCAGGTCCCTGAGTAG
Mcap-F	CCTGATGTGACGTGCTCGAGCATATCACTTACCTCCTTTACAAAAAAATAAATAATTCAC
Mcap-R	ATTAACCCTCACTAAAGGGACGATCGCATAATTACTCCTTTTAAGATAAGTTTTTTATTC
Mm-F	TACTCAGGGACCTGATGTGACGTGCTCGAGCATACCACTACCTCCTTTACAAAAAAATAA
Mmm-R	CATTAACCCTCACTAAAGGGACGATCGCATAATTACTCCTTTTAATCAAATTGTTTTTAC
Mmc-R	CGACATTAACCCTCACTAAAGGGACGATCGCATAATTACTCCTTTTTTAAAACAAATTGT
Sci-F	TACTCAGGGACCTGATGTGACGTGCTCGAGTCGATATTTTACAAAAATTTGCTTATTATG
Sci-R	CATTAACCCTCACTAAAGGGACGATCGCATTTTTTTTACTCCTTACTTTAGTATATTCTG

Table S3.2. Compromise codon table for *M. florum* and *E. coli.*

Amino acid	Codon	Number in *M. florum*	Number / 1000 codons	*M. florum* Probability	*E. coli* Probability	Compromise probability[a]
Gly	GGG	848	3.45	0.06	.15	.12
Gly	GGA	5717	23.26	0.43	.11	.26
Gly	GGT	6147	25.01	0.47	.34	.49
Gly	GGC	461	1.88	0.03	.40	.13
Glu	GAG	1543	6.28	0.09	.31	.17
Glu	GAA	16303	66.34	0.91	.69	.83
Asp	GAT	10918	44.43	0.83	.63	.74
Asp	GAC	2203	8.96	0.17	.37	.26
Val	GTG	632	2.57	0.04	.37	.15
Val	GTA	4237	17.24	0.28	.15	.25
Val	GTT	9660	39.31	0.65	.26	.52
Val	GTC	344	1.40	0.02	.22	.08
Ala	GCG	519	2.11	0.04	.36	.15
Ala	GCA	6198	25.22	0.46	.21	.38
Ala	GCT	6297	25.62	0.47	.16	.34
Ala	GCC	522	2.12	0.04	.27	.13
Arg	AGG	194	0.79	0.03	.02	.05
Arg	AGA	5448	22.17	0.80	.04	.33
Ser	AGT	3752	15.27	0.23	.15	.23
Ser	AGC	935	3.80	0.06	.28	.16
Lys	AAG	2110	8.59	0.09	.23	.15
Lys	AAA	21721	88.38	0.91	.77	.85
Asn	AAT	13612	55.39	0.77	.45	.62
Asn	AAC	4161	16.93	0.23	.55	.38
Met	ATG	5660	23.03	1.00	1.00	1.00
Ile	ATA	7748	31.53	0.32	.07	.17
Ile	ATT	14701	59.82	0.60	.51	.62
Ile	ATC	1968	8.01	0.08	.42	.21
Thr	ACG	211	0.86	0.02	.27	.10
Thr	ACA	7243	29.47	0.55	.13	.38
Thr	ACT	5558	22.62	0.42	.17	.38
Thr	ACC	274	1.11	0.02	.44	.14
Trp	TGG	101	0.41	0.04	1.00	1.00
Trp	TGA	2489	10.13	0.96	0.0	0.0
Cys	TGT	1275	5.19	0.85	.45	.68
Cys	TGC	228	0.93	0.15	.55	.32

Table S3.2. Compromise codon table for *M. florum* and *E. coli* (continued).

End	TAG	132	0.54	0.19	.07	.14
End	TAA	551	2.24	0.81	.64	.86
Tyr	TAT	7191	29.26	0.78	.57	.68
Tyr	TAC	1972	8.02	0.22	.43	.32
Leu	TTG	1765	7.18	0.08	.13	.16
Leu	TTA	15736	64.03	0.71	.13	.49
Phe	TTT	10293	41.88	0.81	.57	.70
Phe	TTC	2391	9.73	0.19	.43	.30
Ser	TCG	206	0.84	0.01	.16	.05
Ser	TCA	7413	30.16	0.46	.12	.29
Ser	TCT	3585	14.59	0.22	.15	.22
Ser	TCC	92	0.37	0.01	.15	.05
Arg	CGG	6	0.02	0.00	.10	0.0
Arg	CGA	128	0.52	0.02	.06	.06
Arg	CGT	998	4.06	0.15	.38	.44
Arg	CGC	67	0.27	0.01	.40	.12
Gln	CAG	404	1.64	0.05	.65	.24
Gln	CAA	7058	28.72	0.95	.35	.76
His	CAT	2211	9.00	0.73	.57	.65
His	CAC	836	3.40	0.27	.43	.35
Leu	CTG	182	0.74	0.01	.50	.11
Leu	CTA	1690	6.88	0.08	.04	.08
Leu	CTT	2682	10.91	0.12	.10	.17
Leu	CTC	52	0.21	0.00	.10	0.0
Pro	CCG	203	0.83	0.03	.52	.16
Pro	CCA	3516	14.31	0.57	.19	.44
Pro	CCT	2306	9.38	0.37	.16	.32
Pro	CCC	157	0.64	0.03	.12	.08

[a]The compromise probability is calculated by taking the geometric mean of the two genome codon probabilities and renormalizing such that the sum of probabilities of each codon for a specific amino acid is unity.

Table S3.3. *E. coli* and *M. florum* mating ratios for pMflT-o4 conjugation.

E. coli MFD*pir*			*M. florum* L1			Mating volume ratio (*M. florum* / *E. coli*)
Volume[a] (ml)	Dilution	Approx. CFU	Volume[b] (ml)	Dilution	Approx. CFU	
1	undiluted	2.5×10^7	1	undiluted	5×10^9	10^0
1	undiluted	2.5×10^7	0.1	undiluted	5×10^8	10^{-1}
1	undiluted	2.5×10^7	0.01	undiluted	5×10^7	10^{-2}
1	undiluted	2.5×10^7	0.01	10^{-1}	5×10^6	10^{-3}
1	undiluted	2.5×10^7	0.01	10^{-2}	5×10^5	10^{-4}
1	undiluted	2.5×10^7	-	-	-	No recipient control
-	-	-	1	undiluted	5×10^9	No donor control
-	-	-	1	undiluted	5×10^9	Purified plasmid control (1 µg of pMflT-o4)

[a]Washed *E. coli* cells from a ~2.5×10^7 CFU/ml culture.
[b]Washed *M. florum* cells from a ~5×10^9 CFU/ml culture.

Table S3.4. *OriC* region percentage identity matrix of selected species of the Spiroplasma group.

Species	*Mferi*	*Mlea*	*Mcpn*	*Mcap*	*Mmm*	*Mmc*	*Myea*	*Mputr*	*Mflorum*	*Scitri*	*Skun*
Mferi	100%	90%	90%	90%	90%	90%	74%	73%	64%	62%	62%
Mlea		100%	94%	94%	91%	91%	73%	72%	63%	62%	62%
Mcpn			100%	98%	91%	92%	73%	72%	64%	62%	62%
Mcap				100%	92%	92%	73%	72%	63%	62%	63%
Mmm					100%	97%	72%	72%	62%	62%	62%
Mmc						100%	72%	73%	62%	62%	62%
Myea							100%	74%	61%	59%	60%
Mputr								100%	61%	59%	59%
Mflorum									100%	57%	57%
Scitri										100%	94%
Skun											100%

Mferi, *M. feriruminatoris*; *Mlea*, *M. leachii*; *Mcpn*, *M. capricolum subsp. capripneumoniae*; *Mcap*, *M. capricolum subsp. capricolum*; *Mmm*, *M. mycoides subsp. mycoides*; *Mmc*, *M. mycoides subsp. capri*; *Myea*, *M. yeatsii*; *Mputr*, *M. putrefaciens*; *Mflorum*, *M. florum*; *Scitri*, *S. citri*; *Skun*, *S. kunkelii.*

Table S3.5. Putative DnaA boxes found within the *oriC* intergenic regions of selected species of the Spiroplasma group.

DnaA box sequence	Strand	*p*-value	Genomic position (bp)	Position relative to *dnaA* start codon (bp)
M. feriruminatoris				
rpmH/dnaA				
TTTATCTACA	-	4.8E-5	4,332 – 4,341	(-182) – (-191)
CTTATCCACA	-	6.0E-6	4,374 – 4,383	(-140) – (-149)
GTTTTCCACA	+	2.2E-6	4,471 – 4,480	(-43) – (-52)
TTTATCTCCA	-	6.3E-5	4,512 – 4,521	(-2) – (-11)
dnaA/dnaN				
GTTATCCACA	+	1.1E-6	5,908 – 5,917	1,386 – 1,395
M. leachii				
rpmH/dnaA				
CTTATCAACA	-	3.4E-5	79 – 88	(-139) – (-148)
GTTTTCCACA	+	2.2E-6	175 – 184	(-43) – (-52)
TTTATCTCCA	-	6.3E-5	216 – 225	(-2) – (-11)
dnaA/dnaN				
GTTATTCACA	+	4.2E-5	1,614 – 1,623	1,388 – 1,397
M. capricolum subsp. capripneumoniae				
rpmH/dnaA				
TTTATCTACA	-	4.8E-5	1,017,103 – 1,017,112	(-182) – (-191)
CTTATCAACA	-	3.4E-5	1,017,145 – 1,017,154	(-140) – (-149)
GTTTTCCACA	+	2.2E-6	1,017,242 – 1,017,251	(-43) – (-52)
TTTATCTCCA	-	6.3E-5	1,017,283 – 1,017,292	(-2) – (-11)
dnaA/dnaN				
GTTGTTCACA	+	5.6E-5	1,388 – 1,397	1,388 – 1,397
M. capricolum subsp. capricolum				
rpmH/dnaA				
TTTATCTACA	-	4.8E-5	1,009,833 – 1,009,842	(-182) – (-191)
CTTATCAACA	-	3.4E-5	1,009,875 – 1,009,884	(-140) – (-149)
GTTTTCCACA	+	2.2E-6	1,009,972 – 1,009,981	(-43) – (-52)
TTTATCTCCA	-	6.3E-5	1,010,013 – 1,010,022	(-2) – (-11)
dnaA/dnaN				
GTTGTTCACA	+	5.6E-5	1,388 – 1,397	1,388 – 1,397
M. mycoides subsp. mycoides[a]				
rpmH/dnaA				
TTTATCTACA	-	4.8E-5	1,211,602 – 1,211,611	(-180) – (-189)
CTTATCAACA	-	3.4E-5	1,211,644 – 1,211,653	(-138) – (-147)
GTTTTCCACA	+	2.2E-6	37 – 46	(-42) – (-51)
TTTGTCTCCA	-	7.4E-5	77 – 86	(-2) – (-11)
dnaA/dnaN				
GTTATCCACA	+	1.1E-6	1,474 – 1,483	1,387 – 1,396
M. mycoides subsp. capri				
rpmH/dnaA				
TTTATCTACA	-	4.8E-5	1,078,118 – 1,078,127	(-181) – (-190)

Table S3.5. Putative DnaA boxes found within the *oriC* intergenic regions of selected species of the Spiroplasma group (continued).

CTTATCAACA	-	3.4E-5	1,078,160 – 1,078,169	(-139) – (-148)
GTTTTCCACA	+	2.2E-6	1,078,257 – 1,078,266	(-42) – (-51)
TTTGTCTCCA	-	7.4E-5	1,078,297 – 1,078,306	(-2) – (-11)
dnaA/dnaN				
GTTATCCACA	+	1.1E-6	1,387 – 1,396	1,387 – 1,396
M. yeatsii[b]				
rpmH/dnaA				
GTTTTCAACA	+	2.3E-5	894,886 – 894,895	(-172) – (-181)
GTTATCCACA	-	1.1E-6	894,975 – 894,984	(-83) – (-92)
TTTTTCAACA	+	5.6E-5	895,010 – 895,019	(-48) – (-57)
dnaA/dnaN				
TTTATCCACA	+	1.2E-5	1,425 – 1,434	1,410 – 1,419
M. putrefaciens				
rpmH/dnaA				
GTTTTCAACA	+	2.3E-5	832,426 – 832,435	(-169) – (-178)
TTTATCTACA	-	4.8E-5	832,459 – 832,468	(-136) – (-145)
GTTATCCACA	-	1.1E-6	832,513 – 832,522	(-82) – (-91)
dnaA/dnaN				
CTTATCCACA	+	6.0E-6	1,392 – 1,401	1,392 – 1,401
M. florum				
rpmH/dnaA				
TTTTTCAACA	+	5.6E-5	793,045 – 793,054	(-171) – (-180)
CTTTTCCACA	-	7.1E-6	793,070 – 793,079	(-146) – (-155)
TTTATCCACA	+	1.2E-5	793,157 – 793,166	(-59) – (-68)
GTTTTCCACA	-	2.2E-6	793,170 – 793,179	(-46) – (-55)
dnaA/dnaN				
CTTTTCCACA	-	7.1E-6	1,359 – 1,368	1,359 – 1,368
GTTTTCCACA	+	2.2E-6	1,428 – 1,437	1,428 – 1,437
CTTATTTACA	-	9.6E-5	1,541 – 1,550	1,541 – 1,550
S. citri				
guaB/dnaA				
GTTTTCCACA	-	2.2E-6	52,010 – 52,019	(-159) – (-168)
TTTTTCCACA	-	1.3E-5	52,075 – 52,084	(-94) – (-103)
CTTTTCCACA	+	7.1E-6	52,120 – 52,129	(-49) – (-58)
dnaA/dnaN				
GTTTTCCACA	+	2.2E-6	1,413 – 1,422	1,413 – 1,422
S. kunkelii				
guaB/dnaA				
GTTTTTCACA	-	4.5E-5	1,463,816 – 1,463,825	(-159) – (-168)
CTTTTCCACA	+	7.1E-6	1,463,926 – 9	(-49) – (-58)
dnaA/dnaN				
GTTTTCCACA	+	2.2E-6	1,417 – 1,426	1,360 – 1,369
GTTTTCCACA	+	2.2E-6	1,470 – 1,479	1,413 – 1,422

[a]According to multiple DNA sequence alignment, protein blast, and absence of putative Shine-Dalgarno sequence, the start codon of *dnaA* was considered to be located at position 88 bp instead of position 1 bp, and the start codon of *rpmH* was considered to be located at position 1,211,498 bp instead of 1,211,534 bp, relatively to the available Genbank sequence (RefSeq NC_005364.2).

[b]According to multiple DNA sequence alignment, protein blast, and absence of putative Shine-Dalgarno sequence, the start codon of *dnaA* was considered to be located at position 16 bp instead of position 1 bp, relatively to the available Genbank sequence (RefSeq NZ_CP007520.1).

3.11.7 Supplementary References

1. **Rodrigue S**, **Materna AC**, **Timberlake SC**, **Blackburn MC**, **Malmstrom RR**, **Alm EJ**, **Chisholm SW**. 2010. Unlocking Short Read Sequencing for Metagenomics. PLoS One **5**:e11840.
2. **Renaudin J**, **Marais A**, **Verdin E**, **Duret S**, **Foissac X**, **Laigret F**, **Bové JM**. 1995. Integrative and free *Spiroplasma citri oriC* plasmids: Expression of the *Spiroplasma phoeniceum* spiralin in *Spiroplasma citri*. J Bacteriol **177**:2870–2877.
3. **Brunner M**, **Bujard H**. 1987. Promoter recognition and promoter strength in the *Escherichia coli* system. EMBO J **6**:3139–3144.
4. **Matteau D**, **Baby V**, **Pelletier S**, **Rodrigue S**. 2015. A Small-Volume, Low-Cost, and Versatile Continuous Culture Device. PLoS One **10**:e0133384.
5. **Lartigue C**, **Blanchard A**, **Renaudin J**, **Thiaucourt F**, **Sirand-Pugnet P**. 2003. Host specificity of mollicutes *oriC* plasmids: Functional analysis of replication origin. Nucleic Acids Res **31**:6610–6618.
6. **Lartigue C**, **Duret S**, **Garnier M**, **Renaudin J**. 2002. New plasmid vectors for specific gene targeting in *Spiroplasma citri*. Plasmid **48**:149–159.

CHAPITRE 4

INTEGRATIVE CHARACTERIZATION OF THE NEAR-MINIMAL BACTERIUM *MESOPLASMA FLORUM*

4.1 Présentation de l'article et contributions

En raison de son petit génome, son absence de pouvoir pathogène et sa croissance rapide en laboratoire, *M. florum* constitue un modèle particulièrement intéressant pour la biologie des systèmes et la génomique synthétique. Par sa simplicité exceptionnelle, *M. florum* offre une opportunité sans pareil de décortiquer de manière approfondie le fonctionnement intégral d'une cellule microbienne par l'intégration d'une quantité massive de données à l'échelle du génome. De plus, maintenant que nous avons à notre disposition plusieurs outils moléculaires afin de modifier génétiquement *M. florum* (Baby et al., 2017; Matteau et al., 2017), ces données pourront être utilisées afin de guider la modification et la réorganisation du génome de cette bactérie dans le but de développer une plateforme optimisée pour la programmation de génomes.

Alors que les récentes technologies de génomique fonctionnelle permettent de générer des quantités impressionnantes de données en une seule expérience, l'intégration de différents ensembles de données dans le contexte de la physiologie cellulaire peut s'avérer plutôt difficile. Les modèles informatiques tels que les GEMs représentent cependant des outils efficaces pour y parvenir (Bordbar et al., 2014; Ebrahim et al., 2016; Gu et al., 2019; Kim et al., 2016). Ceux-ci permettent notamment de générer des prédictions phénotypiques à partir de l'information génomique disponible pour un organisme donné. À titre d'exemple, il est possible d'estimer l'impact de la délétion de plusieurs gènes sur les flux métaboliques et le taux de croissance d'un organisme. Au fur et à mesure que les GEMs gagnent en précision et complexité, modélisant ainsi une quantité croissante de processus cellulaires, ceux-ci pourraient rapidement devenir des outils particulièrement puissants dans le contexte de la génomique synthétique. En effet, même pour des génomes particulièrement petits, le nombre de configurations possibles est pratiquement illimité. L'utilisation de GEMs de haute qualité constituerait alors un avantage

fort considérable afin de prédire systématiquement l'impact de réarrangements génomiques chez *M. florum* (Chalkley et al., 2019; Rees et al., 2018), ce qui permettrait de choisir les configurations génomiques les plus pertinentes à tester à l'aide des méthodes de synthèse disponibles. Cette stratégie pourra être appliquée dans un premier temps pour guider la minimisation du génome de cette bactérie, et pourra ensuite servir afin d'étudier les règles régissant la conception des génomes qui demeurent, à l'heure actuelle, quasiment inexplorées.

La fiabilité des prédictions générées par les GEMs dépend grandement du niveau d'exactitude de plusieurs contraintes définies dans le modèle telles que le taux de croissance, la composition macromoléculaire de la cellule et la masse sèche de la cellule (Feist and Palsson, 2010; Lachance et al., 2019b). Afin de permettre le développement d'un GEM reproduisant le plus précisément possible la physiologie cellulaire de *M. florum*, j'ai procédé à la caractérisation expérimentale de cette bactérie en combinant différentes méthodes et techniques. Le manuscrit qui suit décrit l'ensemble de cette caractérisation. Celui-ci regroupe notamment la mesure de plusieurs aspects physiques et physiologiques incluant le temps de doublement, le diamètre cellulaire, la masse cellulaire sèche, ainsi que les fractions macromoléculaires de la cellule. Nous montrons également la première caractérisation du transcriptome et du protéome de *M. florum* grâce aux techniques de 5' -RACE (Matteau and Rodrigue, 2015a), RNA-seq, et MS/MS. Finalement, nous utilisons les fractions macromoléculaires de la cellule pour convertir les niveaux de transcription et d'expression des gènes en abondances moléculaires absolues et ainsi brosser un portrait sans précédent de la composition intracellulaire de *M. florum*. En plus de servir de fondation pour le développement d'un GEM de haute qualité pour *M. florum*, ces efforts de caractérisation vont aider à acquérir une compréhension détaillée du fonctionnement global de cette bactérie.

Initié au tout début de mes études graduées au sein des laboratoires des Prs Sébastien Rodrigue et Pierre-Étienne Jacques, ce projet intégratif de longue haleine m'a permis d'acquérir une expertise tant au niveau de la réalisation d'expériences à l'échelle du génome que dans l'analyse des données produites par celles-ci. Initialement jumelé à un manuscrit (actuellement en

préparation) décrivant la reconstruction, description et validation d'un GEM pour *M. florum*, le manuscrit portant sur la caractérisation intégrative de *M. florum* a finalement été séparé de celui-ci en raison de l'ampleur des deux projets respectifs. En effet, l'article présenté dans ce chapitre combine de nombreuses méthodes et intègre une quantité importante d'information. Par conséquent, cet article a nécessité la collaboration de plusieurs personnes possédant des expertises relativement diverses. Plus spécifiquement, Samuel Gauthier a réalisé les courbes de croissance de *M. florum* à différentes températures, alors que j'ai effectué les expériences de suivi de la concentration cellulaire lors des différentes phases de croissance de la bactérie. J'ai également réalisé les essais de quantification de la biomasse et de mesure de densité cellulaire de *M. florum*, et Jean-Christophe Lachance a participé à la représentation graphique des équations de la masse cellulaire. Les essais de microscopie ont été effectués par Charles Bertrand, Daniel Garneau et moi-même, avec l'assistance technique de la Plateforme de microscopie photonique de l'Université de Sherbrooke. Les librairies de séquençage Illumina de type 5' -RACE et RNA-seq ont été générées par moi-même, alors que les expériences de quantification des lipides et des protéines par spectrométrie de masse ont été exécutées par PhenoSwitch Bioscience. Sous la supervision du Pr Pierre-Étienne Jacques et assisté par Frédéric Grenier et Jean-François Lucier, j'ai réalisé l'ensemble des analyses bio-informatiques pour les données de spectrométrie de masse et de séquençage Illumina. James Daubenspeck et Kevin Dybvig ont effectué les expériences de quantification des glucides par chromatographie en phase gazeuse couplée à la spectrométrie de masse, ainsi que l'analyse des données s'y rattachant. J'ai, avec l'aide de Jean-Christophe Lachance, réalisé les analyses d'allocation des catégories fonctionnelles pour le transcriptome et le protéome de *M. florum*. Finalement, les Prs Sébastien Rodrigue et Pierre-Étienne Jacques ont supervisé l'ensemble des étudiants impliqués dans ce projet, et ont, avec la participation de Jean-Christophe Lachance, contribué aux différentes idées figurant dans le manuscrit. J'ai, avec l'aide du Pr Sébastien Rodrigue, rédigé l'ensemble du manuscrit décrivant la caractérisation intégrative de *M. florum*. Ce dernier est présentement à un stade très avancé de préparation et sera soumis sous peu dans la revue *Molecular Systems Biology*.

Référence bibliographique : Matteau, D., Lachance, J.-C., Grenier, F., Gauthier, S., Daubenspeck, J. M., Dybvig, K., Garneau, D., Knight, T.F., Jacques, P.-É., Rodrigue, S. (2020). Integrative characterization of the near-minimal bacterium *Mesoplasma florum*. En préparation de soumission pour Mol. Syst. Biol.

4.2 Title page

Integrative characterization of the near-minimal bacterium *Mesoplasma florum*

Dominick Matteau[a], Jean-Christophe Lachance[a], Frédéric Grenier[a], Samuel Gauthier[a], James M. Daubenspeck[b], Kevin Dybvig[b], Daniel Garneau[a], Thomas F. Knight[c], Pierre-Étienne Jacques[a], & Sébastien Rodrigue[a#].

[a]Département de biologie, Université de Sherbrooke, Sherbrooke, Québec, Canada.
[b]Department of Genetics, University of Alabama at Birmingham, Birmingham, Alabama, USA.
[c]Ginkgo Bioworks, Boston, Massachusetts, USA.

#Address correspondence to Sébastien Rodrigue, sebastien.rodrigue@usherbrooke.ca

4.3 Abstract

The near-minimal bacterium *Mesoplasma florum* constitutes an interesting model for synthetic genomics and systems biology studies due to its small genome, fast growth rate, and lack of pathogenic potential. However, practically no quantitative data about the physiology of *M. florum* is available, and some fundamental aspects of its biology remain largely unexplored. Here, we report a broad yet remarkably detailed characterization of *M. florum* using a various number of approaches and techniques, and integrate the generated data to gain additional knowledge on parameters difficult to evaluate experimentally. More specifically, we precisely measured several physical and physiological aspects of this bacterium, including the doubling time, growth kinetics, cell diameter, cell buoyant density, dry mass, and use these data to infer the most probable *M. florum* cell mass, volume, and surface area. We also performed the first transcriptome and proteome characterization of this microorganism using a combination of genome-wide 5'-rapid amplification of cDNA ends (5'-RACE), RNA sequencing, and protein tandem mass spectrometry. These analyses notably revealed the position of more than 400 transcription start sites associated to a conserved promoter sequence, the transcription and expression levels of all annotated genes, the relative importance of predicted cellular functions, as well as the first experimental cartography of *M. florum* transcription units and untranslated regions. Finally, we measured the macromolecular mass fractions of the cell to convert gene transcription and expression levels into absolute molecular species abundances and generate an unprecedented view of the intracellular composition of the *M. florum* cell. In addition to provide an experimental foundation for the development of a genome-scale metabolic model, these characterization efforts will help acquiring a detailed understanding of global cell functioning in a near-minimal bacterium and will guide future genome engineering designs for *M. florum*.

4.4 Introduction

Acquiring a deep and quantitative understanding of biological systems through holistic approaches represents a major goal of systems biology. As available omics technologies

continue to become cheaper, more powerful, and more diversified, cells can be systematically characterized at an unprecedented level of details. However, while recent omics technologies can generate massive amounts of data in a single experiment, the integration of different omics datasets in the context of cellular physiology can be challenging. Computational models, such as genome-scale metabolic models (GEMs), constitute an interesting tool for this task (1-4). GEMs consist of mathematically structured knowledge frameworks describing the metabolism of organisms, where each metabolic reaction is stored in a stochiometric matrix and linked to their corresponding enzymes and associated genes (4-7). This gene-protein-reaction association allows phenotypic predictions to be made from available genomic information using optimization techniques such as flux balance analysis (FBA) (8). For instance, the impact of multiple gene deletions, environmental stresses or nutritional changes on metabolic fluxes and predicted growth rate can be computed, providing context-specific hypotheses for experimental testing. To perform accurate predictions, GEMs must be highly constrained and validated by experimental data such as growth rate, cell dry mass, and macromolecular composition of the cell (% of DNA, RNA, proteins, etc.) (9, 10). Nevertheless, the experimental measurement of these parameters is not commonplace; most GEMs reconstructed so far rely on biomass data gathered for model organisms (e.g. *Escherichia coli*), which might affect their overall quality and most importantly the accuracy of their predictions.

Recently, GEMs have been extended to include cellular processes beyond metabolism, thereby increasing their capabilities and breadth of applications (11, 12). Genome-scale models of metabolism and macromolecular expression (ME-models), for example, account for the synthesis of the gene expression machinery, and can explain up to 80% of the cell proteome by mass (13-16). As GEMs continue to become increasingly refined by covering more and more biochemical data and cellular processes, these frameworks could soon become powerful tools for the rational design of synthetic or semi-synthetic organisms, an emerging field commonly referred as synthetic genomics (17-20). Given proper design, such organisms could play a very important role in addressing some of most critical challenges of the 21st century such as the

development of sustainable energy sources, the fight against antibiotic resistance, and the treatment of diseases such as cancer and diabetes (21-23).

Although DNA synthesis is gradually becoming cheaper, not even a handful of significatively modified synthetic genomes have been reported (24-26), and our ability to design complete genomes from scratch is extremely poor at best. Consequently, little is still truly understood about genome design principles. This is mainly explained by the overwhelming complexity of common model organisms, which outstrips our current modeling capacities and inhibits our ability to rationally evaluate genome designs. The most recent GEM describing *E. coli* (*i*JL1678-ME), for example, contains only about one third of the total number of genes predicted in this bacterium (13). Moreover, since most of these model organisms possess relatively large chromosomes (≥4 Mb), the DNA synthesis cost required to explore artificial genome architectures can reach prohibitive levels, and the number of possible genome reconfigurations can quickly become overwhelming. In contrast, small genome bacteria constitute attractive substrates for synthetic genomics efforts. Their greater simplicity offers the possibility to achieve a more exhaustive and accurate description of cellular processes using omics technologies and high-quality GEMs (27, 28), reduces the number of genome organisation possibilities to be tested using synthetic genomics approaches, as well as decrease the costs related to chromosome synthesis.

Because of their exceptionally small genomes (0.58 to 2.2 Mbp)(29), near-minimal bacteria of the Mollicutes class have long been proposed as models to study the basic principles of life and constitute very interesting platforms for synthetic genomics studies (30). Although exceptionally simple, these wall-less and very small bacteria (~0.2 to 0.4 μm) are not ancient or primitive forms of life. Mollicutes rather evolved from low G-C content Gram positive bacteria through a process of massive gene loss (31, 32). This genomic reduction resulted in a drastic simplification of their metabolism, with many metabolic pathways missing or incomplete (33, 34). Hence, many Mollicutes have adopted parasitic lifestyles to palliate their metabolic deficiencies, infecting various animals and plants (32, 33). Still, most of them are capable of

autonomous growth in axenic culture using standard conditions. Amongst all Mollicutes, members of the *Mycoplasma* genus are the most extensively studied, some of which have become model organisms in the fields of synthetic genomics and systems biology. The genome of *Mycoplasma mycoides*, for example, have been entirely chemically synthesized and cloned in yeast, where it could be easily modified and next transplanted into a recipient bacterium, namely *Mycoplasma capricolum* (35). This impressive *tour de force* recently culminated with the creation of the first artificial "working approximation" of a minimal cell, JCVI-syn3.0 (24). This minimal bacterium harbours a reduced and synthetic version of the *M. mycoides* genome totalizing ~531 kb, making it the smallest genome ever observed in any autonomously replicating cell (24, 36). The JCVI-syn3.0 strain however showed altered morphological traits and impaired growth rates compared to the *M. mycoides* parent strain (doubling time of ~2-3 hrs vs ~1 hr). Interestingly, these features were restored by the incorporation of 19 additional *M. mycoides* genes into the JCVI-syn3.0 genome (37). The new strain, named JCVI-syn3A, was used for the reconstruction and validation of the first GEM describing the metabolism of a synthetic and minimal cell (37).

Beside mycoplasmas, another particularly interesting member of the Mollicutes class for synthetic genomics and systems biology studies is the near-minimal bacterium *Mesoplasma florum*. *M. florum* was first described in 1984 as *Acholeplasma florum* and is closely related to members of the mycoides cluster (29, 38). *M. florum* however has a smaller genome compared to *M. mycoides* and *M. capricolum* (~793 kb vs ~1.2 Mb and ~1.0 Mb, respectively) and shows faster growth rates (29, 35, 39, 40). Unlike mycoplasmas, *M. florum* has no pathogenic potential, which facilitates its manipulation and its distribution throughout the scientific community. Moreover, genetic manipulation tools have recently been developed for this microorganism, including procedures for whole genome cloning in yeast and genome transplantation (41, 42). Different genome reduction scenarios based on gene conservation and essentiality have also been recently proposed, providing starting points for genome minimization efforts (43). Altogether, these characteristics position *M. florum* as a prime candidate as a cellular chassis specifically designed for systems biology studies and synthetic

genomics approaches. While the creation of the first artificial minimal cell represents a major breakthrough in synthetic genomics, we recently showed that 57 putatively essential *M. florum* genes have no homolog in the synthetic JCVI-syn3.0 strain (43). This suggests that different minimal genome compositions and configurations probably exist, even within closely related species. Similarly to what the comparison of the two first complete bacterial genome sequences revealed on the minimal gene set required for cellular life (44), the comparison of the JCVI-syn3.0 genome with other minimal genomes offers a unique opportunity to decipher genome design principles and some of the most fundamental principles of life.

Here, we report the first integrative characterization of the near-minimal bacterium *M. florum* to advance knowledge on this emerging model for systems and synthetic biology and to provide an experimental foundation for the development of a high-quality GEM. More specifically, we accurately measure several physical and physiological parameters of *M. florum* growing in rich medium, including the cell diameter, buoyant density, dry mass, optimum growth temperature, growth rate, and growth kinetics. We also define the macromolecular composition of the cell, identify and characterize more than 400 active promoters, and proceed to the reconstruction of *M. florum* transcription units (TUs). Finally, we use transcriptomics and proteomics expression datasets to estimate RNA and protein species abundances, revealing the relative importance of the different cellular processes of a near-minimal cell. In combination with a GEM, these extensive characterization efforts aim at acquiring a detailed understanding of global cell functioning in a simple organism, which could be used to guide future *M. florum* genome engineering efforts.

4.5 Results

4.5.1 *M. florum* optimal growth temperature and growth kinetics

The assessment of the doubling time and the optimal growth temperature represent fundamental steps for the characterization of a bacterial strain. Moreover, the doubling time is critical for the

definition of growth- and non-growth associated maintenance (GAM and NGAM, respectively) parameters in GEMs (9, 10). However, accurate measurement of the doubling time and optimal growth temperature have never been reported specifically for the *M. florum* L1 type strain. We therefore evaluated the doubling time of *M. florum* L1 in the ATCC1161 rich medium at different temperatures normally encompassing the optimal growth temperature of Mollicutes (~30-38°C) (45). Doubling times were measured using a 2-fold microplate dilution doubling time assay (2F-DT), a technique based on growth assays previously developed for spiroplasmas (46). The 2F-DT assay relies on the presence of phenol red in the culture medium, a pH indicator, to measure the amount of time separating 2-fold culture dilutions for them to reach the same optical density at 560nm (OD_{560nm}) and therefore approximately the same medium pH. Raw *M. florum* growth curves are presented in Figure S4.1. The smallest doubling time was observed at a temperature of 34°C (38 ± 5 min) while no growth was observed at a temperature higher than 36°C (Fig. 4.1A). These results are consistent with previous observations concerning the temperatures allowing the growth of different *M. florum* strains (38) and other members of the *Mesoplasma* genus (47).

We then used flow cytometry (FCM) and colony forming units (CFUs) to precisely measure the growth kinetics of *M. florum* in batch culture incubated at the optimal growth temperature (34°C). We first validated that cell concentrations measured by FCM were well correlated with culture dilutions (Fig. S4.2). By following cell concentrations over ~24 hrs, we could observe an overall pattern typical of the four bacterial growth phases, namely the lag, log, stationary, and death phases (Fig. 4.1B). The log phase coincided with an important decrease in OD_{560nm} easily noticeable by a growth medium color change from red to orange, corresponding to a drop in medium pH (from ~8.0 to 6.5). Using exponential curve fitting performed on FCM and CFUs data of the log phase, we determined a doubling time of 31 min and 33 min, respectively (Fig. 4.1C), which is similar to the value obtained using the 2F-DT assay on microplate cultures (Fig. 4.1A). CFU and FCM cell concentrations were highly consistent with each other until late stationary phase, where they culminated at ~$1x10^{10}$ cells/ml. The stationary phase was also marked by the lowest OD_{560nm} value observed for the entire experiment, corresponding to a

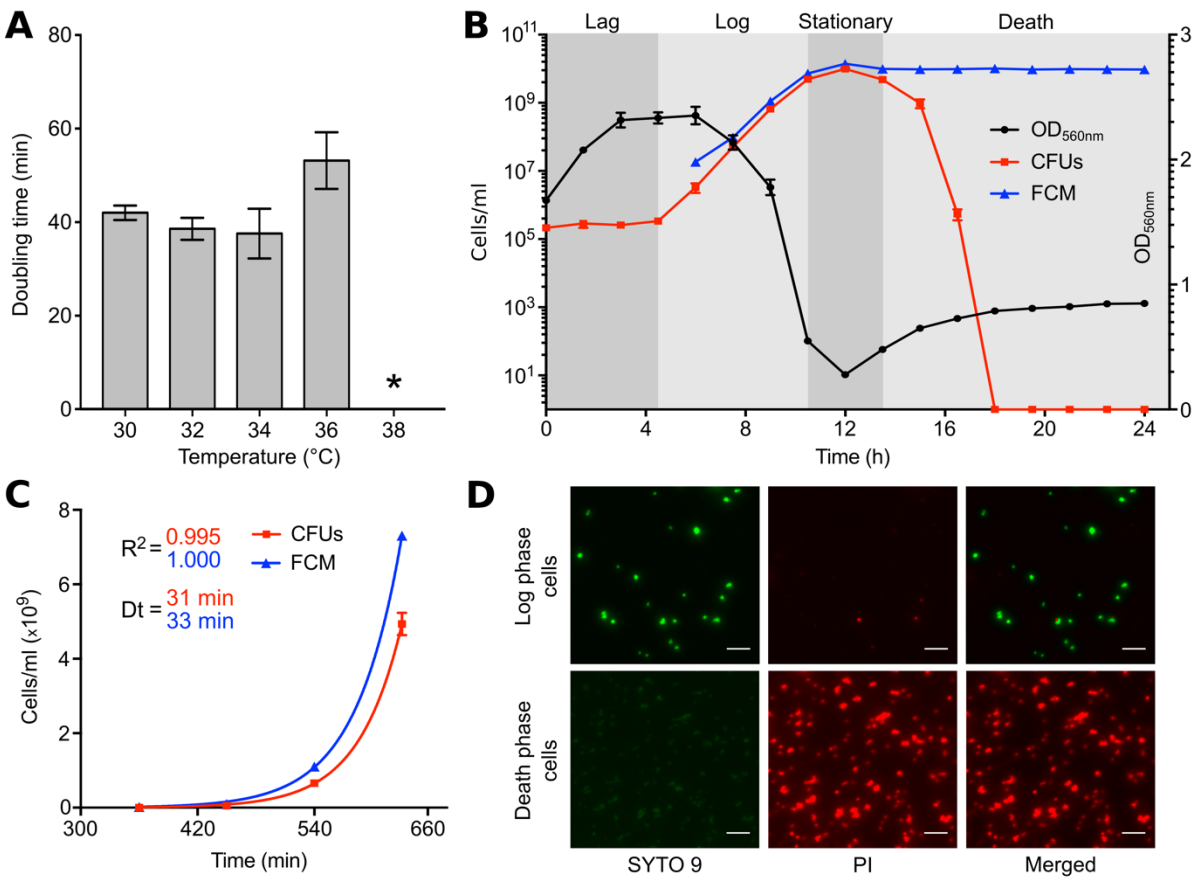

Figure 4.1. **Analysis of *M. florum* growth in ATCC 1161 medium.** A) *M. florum* doubling time at different incubation temperatures. Doubling times were calculated according to the 2-fold microplate dilution doubling time assay (2F-DT). The bars represent the mean and standard deviation values obtained from three technical replicates. The asterisk indicates the absence of significant growth, therefore a non-calculable doubling time. B) *M. florum* growth kinetics at 34°C. Growth was monitored for 24 hrs by measuring the optical density at 560nm (black circles) as well as the cell concentrations using two different methods, colony forming units (CFUs, red squares) and flow cytometer (FCM, blue triangles). The four bacterial growth phases (lag, log, stationary, and death) are represented by gray shading. The dots and error bars indicate the mean and standard deviation values obtained from three independent biological replicates. CFU data points superimposed to the x-axis represent values below the limit of detection ($2x10^{-2}$). C) Exponential growth fit on CFU (red squares) and FCM (blue triangles) counts shown in B. Calculated doubling times (Dt) and correlation coefficients (R^2) are shown. Dots and error bars are as in B. D) Representative images of SYTO 9 and propidium iodide (PI) double stained *M. florum* cells, harvested from a log- or death-phase culture, observed by widefield fluorescence microscopy. The brightness of each channel was adjusted equally between conditions. Scale bar: 5 µm.

yellow medium color and a medium pH around 6.0, followed by a gradual increase in the measured OD_{560nm} caused by the apparition of cell aggregates in the culture. This observation was concomitant with a rapid diminution of CFU counts but not in the FCM cell counts which remained stable for the rest of the experiment, suggesting an important loss in cell viability reminiscent of the death phase (Fig. 4.1B). We validated that the decrease in CFU counts was effectively due to an altered cell viability by fluorescence microscopy using SYTO 9 and propidium iodide (PI) dual staining (Fig. 4.1D). As expected, *M. florum* cells harvested at the death phase showed an intense PI signal and practically no SYTO 9 fluorescence, indicating a significantly compromised cell membrane integrity. Similar signals were observed for formaldehyde fixed and permeabilized cells (Fig. S4.3), whereas log-phase cells rather showed a strong SYTO 9 fluorescence and almost no PI signal, typical of healthy cells (Fig. 4.1D).

4.5.2 Physical characteristics and macromolecular composition of the cell

We next wanted to better define the physical constraints shaping the biology of *M. florum* to provide high-quality and species-specific data for GEM reconstruction. We first proceeded to the precise evaluation of its cell diameter since the only quantitative data available for the cell size of this species relies on filtration studies (38). Filtration constitute an indirect approach that can be subjected to different sources of variation such as pore size heterogeneity and deformation of cellular morphology, especially for wall-less bacteria. To accurately measure the *M. florum* cell diameter, we analyzed log-phase cells using two different techniques, transmission electron microscopy (TEM) and stimulated emission depletion (STED) microscopy. Cells were stained with PicoGreen and mCLING (48), respectively targeting the DNA and the cellular membrane, prior to STED microscopy examination. Representative images obtained from both techniques are shown in Figures 4.2A and 4.2B. Both TEM and STED microscopy showed cells predominantly ovoid, with a cell diameter ranging from approximately 300 to 600 nm and 500 to 1,000 nm, respectively (Fig. 4.2C). An average cell diameter of 434 nm was observed for TEM and 741 nm for STED microscopy (Fig. 4.2D). Interestingly, TEM pictures also showed evidences of a polysaccharidic layer surrounding

M. florum cells, a morphological feature shared by many Mollicutes including the closely related *M. mycoides* and *M. capricolum* (Bertin et al., 2013, 2015; Daubenspeck et al., 2014; Gaurivaud et al., 2014).

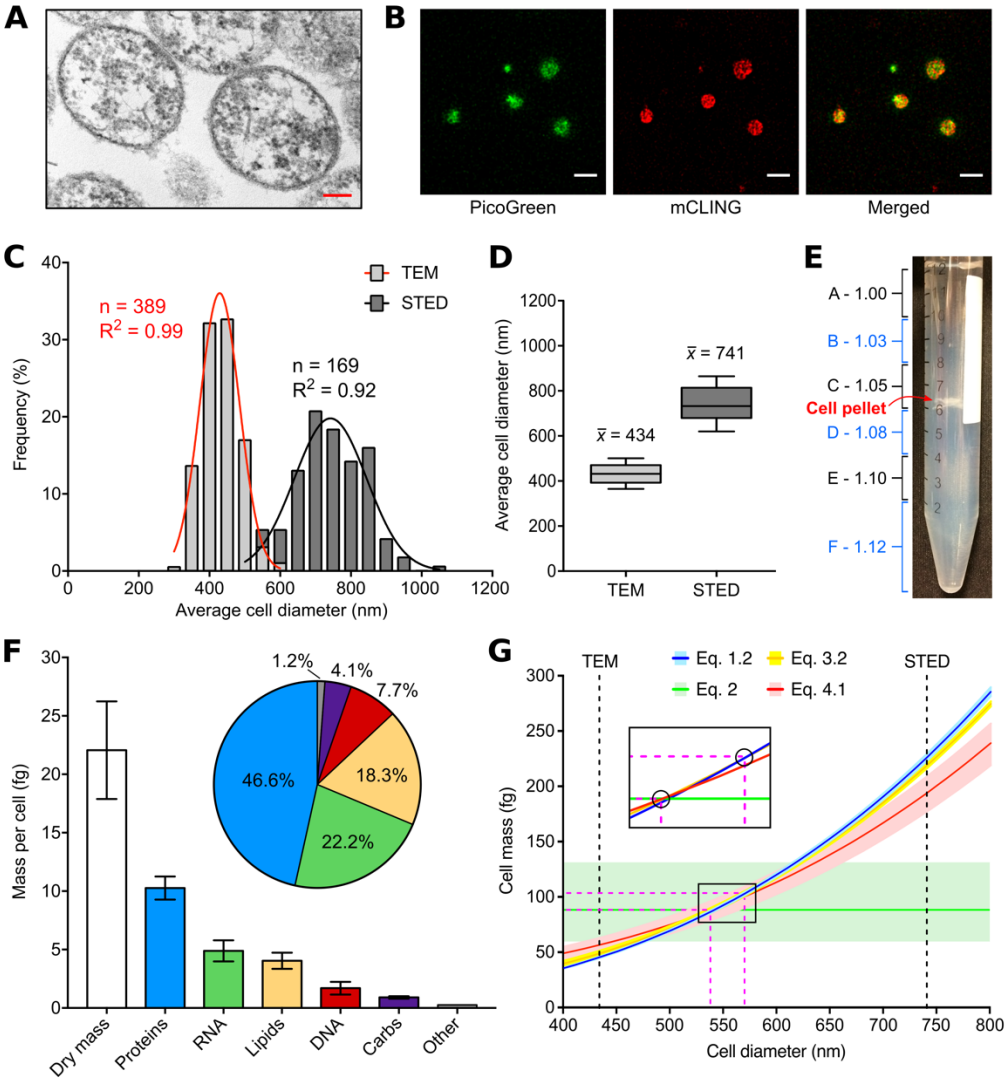

Figure 4.2. *M. florum* physical characteristics. A) Representative image of *M. florum* cells observed by transmission electronic microscopy (TEM). Scale: 100 nm. B) Representative image of PicoGreen (DNA) and mCLING (cellular membrane) double stained *M. florum* cells observed by stimulated emission depletion (STED) microscopy. Scale: 1 µm. C) Frequency distribution of *M. florum* average cell diameter measured by TEM and STED as shown in A and B, respectively. The average cell diameter was obtained by averaging the minor and major axis values measured for each cell. A Gaussian curve fit is indicated for each method, and the calculated correlation coefficients are shown. Bins: 50 nm. D) Box and whiskers plots showing the mean and 10-90 percentile range of the average cell diameter calculated from cells analyzed

in C. E) Picture of *M. florum* cells analyzed by discontinuous density gradient centrifugation in Percoll. The approximative density of each Percoll layer is indicated and colored in blue if trypan blue was added. A, 0%; B, 20%; C, 40%; D, 60%; E, 80%; F, 100% Stock Isotonic Percoll (SIP) solution. The position of the cell pellet is marked. F) *M. florum* biomass quantification. The mass of the principal macromolecular constituents of the cell is shown as well as their relative fraction in the quantified cellular dry mass. Bars represent the mean and standard deviation values obtained from three independent biological replicates (dry mass) or four technical replicates (proteins, RNA, lipids, DNA, carbs). The "Other" category bar represents the residual mass obtained by the subtraction of all quantified macromolecule masses from the total dry mass value. G) Graph showing the relation between the *M. florum* cell diameter (d) and its cell mass (CM) according to cell mass equations 1.2, 2, 3.2, and 4.1 (see Materials and Methods). For each equation, the mean cell mass (CM_{mean}) is indicated by a colored line, and the range of probable values (CM_{min} – CM_{max}) is shown by a light-colored shading. The mean values of the average cell diameter measured by TEM and STED (see panel D) are indicated by black dashed lines. The portion of the graph where all the CM_{mean} curves converge is enlarged and devoid of colored shadings for representation purposes. CM_{mean} interception points encompassing all other interception points are encircled, and their corresponding x and y coordinates are indicated by fuchsia dashed lines (most probable cell diameter and most probable cell mass ranges).

Although very informative, measuring the total mass of a cell can be very challenging, especially for very small cells like Mollicutes. This often requires complex and specialized equipment not necessarily available in conventional biology laboratories (53-55). The cell mass can however be estimated using different mathematical equations that involve only a limited number of variables more easily amenable to quantification, including the cell diameter, cell buoyant density, as well as the cell dry mass. Since we already measured the cell diameter of *M. florum* using TEM and STED microscopy, we then sought to evaluate its cell buoyant density by discontinuous Percoll density gradient centrifugation. After one round of centrifugation, the *M. florum* cell pellet was located at the bottom of the 1.05 g/ml Percoll layer (Fig. 4.2E). The position of the cell pellet also remained the same after a second round of centrifugation, indicating a cell buoyant density lying between 1.05 and 1.08 g/ml (Fig. 4.2E and Table 4.1). We next determined the *M. florum* cell dry mass using conventional weighting procedures performed on log-phase batch cultures (see Materials and Methods and Fig. S4.4), and observed a total cell dry mass of 22.1 ± 4.2 fg (Fig. 4.2F and Table 4.1). The measured cell buoyant density and cell dry mass were then used to infer the most probable *M. florum* cell mass using four different equations (see Equations 1.2, 2, 3.2, and 4.1). Three of those equations also require

the total dry mass fraction and the dry mass density to estimate the total mass of the cell, which were assumed to be within typical ranges found in bacteria, i.e. 20 – 30% and 1.3 – 1.5 g/ml, respectively (56-60). Interestingly, all four equations converged to a relatively tight range of cellular mass (88.2 – 103.3 fg), which corresponds to a cell diameter (538-570 nm) positioned in-between average values obtained by TEM and STED microscopy and within the overlapping portion of their relative distribution (Fig. 4.2C, 4.2F, and Table 4.1). Refining the cell diameter also allowed us to estimate the most probable cell volume (0.082 – 0.097 μm^3), cell surface area (0.911 – 1.021 μm^2), and surface area to volume ratio (SA:V; 10.5 – 11.1 μm^{-1}) using Equations 1.1, 5, and 5.1, respectively (Table 4.1).

Table 4.1. Summary of *M. florum* biomass composition and physical characteristics measured or estimated in this study.

Cellular biomass	Mean ± SD (fg)	Physical parameters	Most probable values
Dry mass	22.1 ± 4.2	Density	1.05 – 1.08 g/ml[(a)]
Proteins	10.3 ± 1.0	Cell diameter	538 – 570 nm[(b)]
RNA	4.9 ± 0.9	Cell mass	88.2 – 103.3 fg[(b)]
Lipids	4.0 ± 0.7	Cell volume	0.082 – 0.097 μm^{3}[©]
DNA	1.7 ± 0.5	Cell surface area	0.911 – 1.021 μm^{2}[©]
Carbohydrates	0.9 ± 0.1	SA:V	10.5 – 11.1 μm^{-1}[©]

[(a)]Measured by discontinuous Percoll density gradient centrifugation.
[(b)]Estimated using cell mass equations (see figure 4.2G and equations 1.2, 2, 3.2, and 4.1).
[(c)]Inferred from the most probable cell diameter (see equations 1.1, 5, and 5.1).

The vast majority of the cell dry mass can be decomposed into four classes of macromolecules: proteins, lipids, nucleic acids, and carbohydrates (61). To better define the *M. florum* dry mass we quantified each of these macromolecules using high sensitivity commercial kits and mass spectrometry methods (see Materials and Methods and Fig. S4.4). According to our analysis, nearly two third of the total dry mass was occupied by proteins and RNA, with a relative abundance of approximately 46.6% and 22.2%, respectively (Fig. 4.2F and Table 4.1). The remaining fraction of the dry mass was divided as follows: 18.3% for lipids, 7.7% for DNA, and 4.1% for carbohydrates. The majority of the carbohydrate fraction most probably account for

the polysaccharidic layer observed by TEM. Carbohydrates detected by mass spectrometry were mainly composed of galactose (0.50 ± 0.07 fg), glucose (0.19 ± 0.03 fg), rhamnose (0.18 ± 0.01 fg), and mannose (0.04 ± 0.01 fg), representing approximately 54.9%, 20.6%, 20.0%, and 4.5% of the total carbohydrate mass, respectively. Interestingly, the residual dry mass, i.e. the difference between the quantified dry mass and the sum of all quantified macromolecules, represented only 1.2% (0.26 fg) of the total dry mass, most likely accounting for small molecules, metabolites, cofactors, and ions (Fig. 4.2F).

4.5.3 Genome-wide identification of promoters

Transposon mutagenesis and gene conservation datasets have recently been published for *M. florum*, and allowed the design of different genome reduction scenarios for this bacterium (43). However, these predictions did not account for promoter organization, and therefore retained all intergenic regions in the reduced genome designs. The identification of all *M. florum* promoters and corresponding TUs would certainly improve the quality and accuracy of these predictions, in addition to providing invaluable information about the transcriptome of this near-minimal cell. We therefore proceeded to the cartography of all *M. florum* transcription start sites (TSSs) at single nucleotide resolution using a previously described genome-wide 5'-rapid amplification of cDNA ends (5'-RACE) method (62, 63). Following Illumina sequencing (see Table S4.1 for a summary of library statistics), the number of read starts per million of mapped reads (RSPM) was calculated for each genomic position in a strand specific manner, resulting in a frequency distribution reminiscent of a Poisson distribution (Fig. S4.5A). Out of 1,586,448 possible sites (genome size multiplied by 2 to account for both strands), a total of 68,650 sites had a non-null TSS signal, of which 1,514 (<0.1% of all sites) displayed a significant intensity (see Fig. S4.5B and Materials and Methods for further details). This resulted in the identification of 605 candidate TSSs distributed throughout the *M. florum* chromosome (Fig. 4.3A). Interestingly, a conserved promoter characterized by a -10 box typical of the σ^{70} family (TAWAAT [36]), the only σ factor family predicted in *M. florum*, could be identified for 422

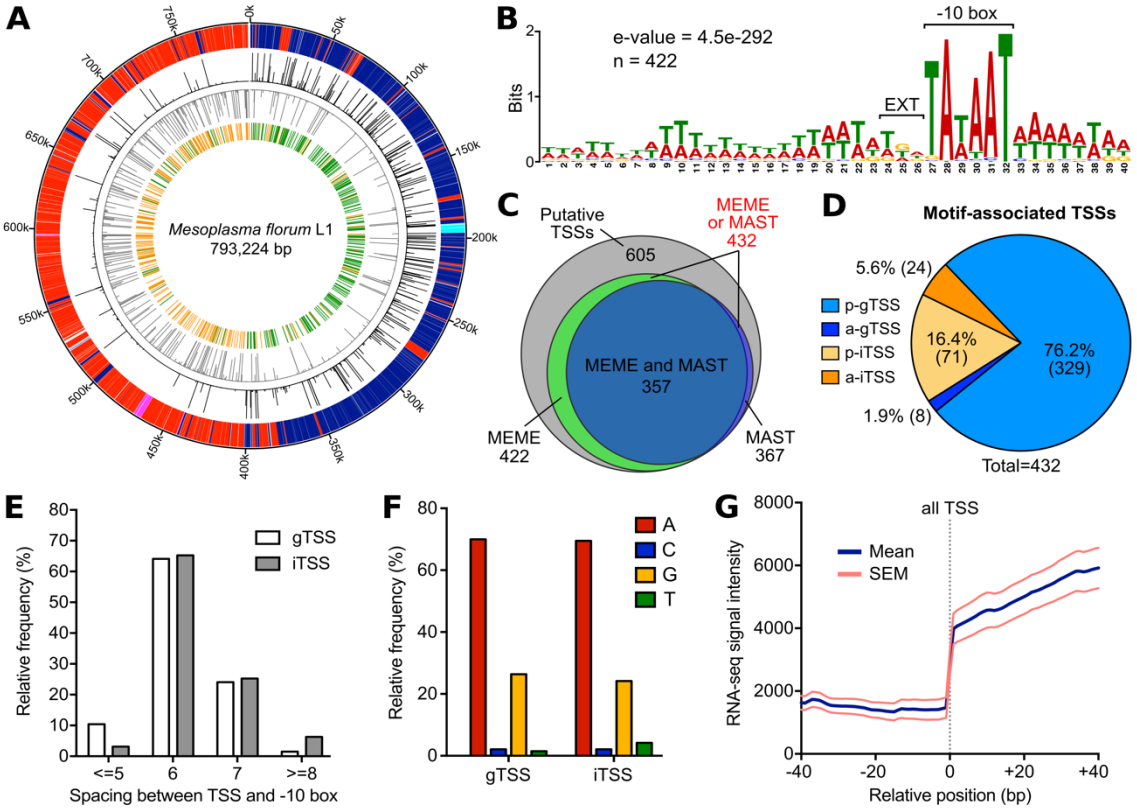

Figure 4.3. Identification and analysis of *M. florum* promoters. A) Circular representation of the *M. florum* L1 chromosome enhanced with 5'-RACE data generated in this study. Outer to inner circle: genomic coordinates (kbp); genes encoded on the positive (blue for coding sequences, turquoise for RNAs) and negative (red for coding sequences, fuchsia for RNAs) DNA strands; raw 5'-RACE signal (0-1,000 read starts scale) observed at each genomic position for the positive (black) and negative (gray) DNA strands; putative transcription start sites (TSSs) identified on the positive (green) and negative (orange) DNA strands from significant 5'-RACE peaks. B) *M. florum* consensus promoter generated using MEME software (65) from all 605 putative TSSs identified by 5'-RACE. A total of 422 sites across the genome were included in the motif. The position of the -10 box (TAWAAT) and the extended element (EXT) is indicated. C) Venn diagram illustrating the number of TSSs associated with a conserved promoter motif (see panel B) found by MEME, MAST or both software compared to the total number of putative TSSs passing filters. D) Localization and orientation of TSSs associated with a MEME or MAST promoter motif. p-gTSS, parallel intergenic TSS; a-gTSS, antiparallel intergenic TSS; p-iTSS, parallel intragenic TSS; a-iTSS, antiparallel intragenic TSS. For gTSSs, the orientation was defined according to the closest downstream gene, while the overlapping gene was used in the case of iTSSs. E) Relative frequency distribution of the spacing between TSSs and their associated promoter -10 box. F) Nucleotide identity at the transcription initiation site (+1) for gTSSs and iTSSs associated to a promoter motif. G) Aggregate profile showing the mean RNA-seq read coverage observed at and surrounding all motif-associated TSSs identified in this study. The calculated SEM is also shown. The aggregate profile was centered on the TSSs coordinates (relative position 0 bp), indicated by a gray dashed line.

sites using the MEME (65) software (Fig. 4.3B and 4.3C). This conserved promoter also contained a partially degenerated TGN extension of the -10 box (EXT element), whereas no clear evidence of a -35 box emerged from the analysis (Fig. 4.3B). The occurrence of this promoter was validated in ~85% (357) of cases using the MAST software (66), which also provided evidences for an additional 10 sites not initially included in the MEME constructed motif, for a grand total of 432 motif-associated TSSs (Fig. 4.3C and Dataset S4.1). No promoter consensus could be identified for the remaining TSS candidates, suggesting a higher sequence variability at these sites. As expected, the vast majority (78.0%) of motif-associated TSSs were located within intergenic regions of the chromosome (gTSSs), even though these regions occupy only ~6.1% of the genome (Fig. 4.3D) (43). In almost all cases, motif-associated gTSSs were in the same orientation (parallel) as their closest downstream gene (p-gTSS), with only a few cases of antiparallel downstream associated gene (a-gTSS) (Fig. 4.3D). The remaining TSSs (22.0%) were found to be positioned within coding regions of the genome (iTSS). Interestingly, putative TSSs devoid of promoter motif were located within coding regions in more than 90% of all instances (Fig. S4.6A).

Intergenic regions can be divided into three types according to the topology of the neighbouring genes; divergent, convergent, and parallel (Fig. S4.7A). As expected, intergenic regions hit by gTSSs were significantly larger than those without any gTSS (Fig. S4.7B). Most of gTSSs (71.5%) were comprised within parallel intergenic regions as they constitute the most abundant type present in the genome (Fig. S4.7C). Conversely, only one case of gTSS was observed in convergent intergenic regions (0.3%), the rest of gTSSs being located within divergent counterparts (28.2%). Nonetheless, divergent intergenic regions were the most frequently hit by gTSSs (96.2%) relatively to their total number of instances in the genome (Fig. S4.7D). In contrast, only about the half (43.5%) of parallel intergenic regions were positive for gTSSs. As expected from their respective configuration, divergent intergenic regions positive for gTSSs contained most of the time two instances per region, generally disposed back-to-back, with some cases remarkably displaying two overlapping -10 promoter boxes (Fig. S4.7E and S4.7F). In

comparison, more than 95% of positive parallel regions showed only a single gTSS occurrence, similarly to iTSSs found within coding regions of the genome (Fig. S4.7E).

Nearly 12% (86) of all *M. florum* genes contained at least one motif-associated iTSS (Fig. S4.7D). iTSSs can be separated in two distinct groups according to their orientation relatively to the overlapping gene: p-iTSSs, initiating transcription on the same strand, and a-iTSSs, initiating transcription on the opposite strand (Fig. S4.8A). The majority of motif-associated iTSSs identified in this study consisted of p-iTSSs (74.7%, 16.4% of total TSSs), a-iTSSs representing only 5.6% of all TSSs (25.3% of iTSSs)(Fig. 4.3D). iTSSs can be further categorized relatively to the orientation of the most immediate downstream gene, i.e. whether or not a gene is appropriately oriented to be expressed from a given iTSS (Fig. S4.8A). Interestingly, most p-iTSSs were located upstream genes transcribed on the same strand (88.7%), contrasting with a-iTSSs predominantly facing their nearest downstream gene (83.3%; Fig. S4.8B). p-iTSSs were also found to be enriched at the very beginning as well as near the end of their overlapping gene, a tendency not observed for a-iTSSs (Fig. S4.8C). In addition, several cases of p-iTSSs located directly on the first base of translation start codons were observed, suggesting the existence of leaderless mRNA in *M. florum* as observed in other bacteria (67-70). When we compared the spacing between TSSs and their associated promoter -10 box, both gTSSs and iTSSs shared approximately the same distribution, predominantly being separated by 6 or 7 bases from the -10 box most proximal extremity (Fig. 4.3E). Both TSS types were also located preferentially on coordinates corresponding to purine nucleotides (A or G), yet with an important bias for adenine (~70% of cases), reflecting the low G-C nature of the *M. florum* genome (Fig. 4.3F). Despite these similarities, motif-associated gTSSs displayed a significantly higher signal intensity compared to motif-associated iTSSs, the latter group being not significantly different from TSSs without promoter motif (Fig. S4.6B). TSSs lacking the *M. florum* promoter motif were however not enriched for purine nucleotides like motif-associated gTSSs and iTSSs (Fig. S4.6C).

To validate the identified promoters, we performed directional RNA sequencing (RNA-seq) on three log-phase *M. florum* steady-state cultures and evaluated read coverage across the genome. RNA-seq libraries were prepared in duplicate for each biological replicate, resulting in a total of six replicates. A statistics summary of RNA-seq libraries is presented in Table S4.1. We observed excellent correlations between the read coverage of the different replicates calculated on non-overlapping 1 kb windows, indicating a very good reproducibility of the method (Fig. S4.9A). More importantly, coordinates of motif-associated TSSs coincided with a sharp increase in RNA-seq signal intensity calculated over the merged replicates, corroborating 5'-RACE identification results (Fig. 4.3G). This feature was also visible for gTSSs and iTSSs analyzed independently, but to a much lesser extent in the case of iTSSs because of their intragenic context (Fig. S4.10). Taken together, these results showed that motif-associated iTSSs and gTSSs share similar features and could both be responsible for the transcription of downstream genes.

4.5.4 Reconstruction of transcription units

Having identified the key features of the *M. florum* promoters as well as the genomic coordinates of TSSs, we wanted to use this information to reconstruct TUs in this quasi-minimal bacterium. A TU consists of a DNA segment transcribed into a single mRNA molecule from one promoter to a transcription termination site (TTS) and encoding for zero, one or many open reading frames (ORFs). In Mollicutes, termination of transcription is believed to occur through a Rho-independent mechanism since no Rho protein homologue is detected in their genomes (71, 72). This mechanism involves structured terminators that can be reliably predicted from the DNA sequence and genes annotation, reaching excellent sensitivity for certain species, especially for *M. florum* (72). We therefore used an updated version of an algorithm developed by De Hoon and colleagues to predict the position of terminators in *M. florum* according to our current genome annotation (43, 72). In total, 298 different Rho-independent terminators were predicted for the entire genome (Dataset S4.2). As expected, the positions of the predicted terminators matched with an important decrease in the RNA-seq signal intensity, supporting the predictions

made by the algorithm (Fig. S4.11). We then used the 432 motif-associated TSSs (gTSSs and iTSSs) identified by 5'-RACE along with the predicted transcription terminators to reconstruct all possible TUs (see Fig. S4.12 and Materials and Methods for a detailed description of the procedure). After manual curation, a total of 387 TUs, each responsible for the expression of at least one gene, were reconstructed (Dataset S4.3). These TUs encompassed more than 90% of all annotated *M. florum* genes (652), including all rRNA and tRNA genes, thus leaving only 68 genes out of 720 without an associated promoter. TUs start and stop coordinates coincided with a steep increase and decrease in the average RNA-seq read coverage (Fig. 4.4A). Almost half of the reconstructed TUs contained only a single gene, the other half transcribing up to 21 genes simultaneously, for an average of approximately 2.2 genes per TU (Fig. 4.4B). The gene-associated TUs were ranging from 112 bp to 12.5 kb and were characterized by an average length of ~2.4 kb (Fig. 4.4C), as well as by a 5' and 3' untranslated region (UTR) length of 58 and 51 bp, respectively (Fig. 4.4D). Representative *M. florum* TUs are depicted in Figure 4.4E along with their associated 5'-RACE, terminators, and RNA-seq data.

As expected, most of the gene-encoding TUs were transcribed from gTSSs (86.6%) since they constitute the majority of TSSs identified in *M. florum* (Fig. 4.3D and Fig. 4.4F). The remaining TUs were associated to p-iTSS (12.9%) and a-iTSS (0.5%). Both gTSS and iTSS driven TUs showed enrichment of RNA-seq coverage, yet with a less defined 5' border for iTSS TUs (Fig. S4.13). The TU fractions originating from p-iTSS (12.9%) and a-iTSS (0.5%) were also lower compared to their corresponding TSS proportions (Fig. 4.3D, Fig. 4.4F and Dataset S4.3). Indeed, a small number of mapped TSSs (56), principally iTSSs (45 out of 56), could not be attributed to any downstream gene according to their genetic context (Dataset S4.4). These TSSs were either 1) located within an intergenic region immediately upstream a predicted terminator; 2) located within a gene positioned at the end of a TU, so once again positioned before a terminator; or 3) facing a gene in the opposite direction. Nonetheless, orphan TSSs facing a gene in the opposite direction or located within an intergenic region before a terminator concurred with a small (~50-75 bp) RNA-seq signal enrichment (Fig. S4.14). Some of these TSSs could be responsible for the expression of small non-coding RNAs (sRNAs) or antisense RNAs

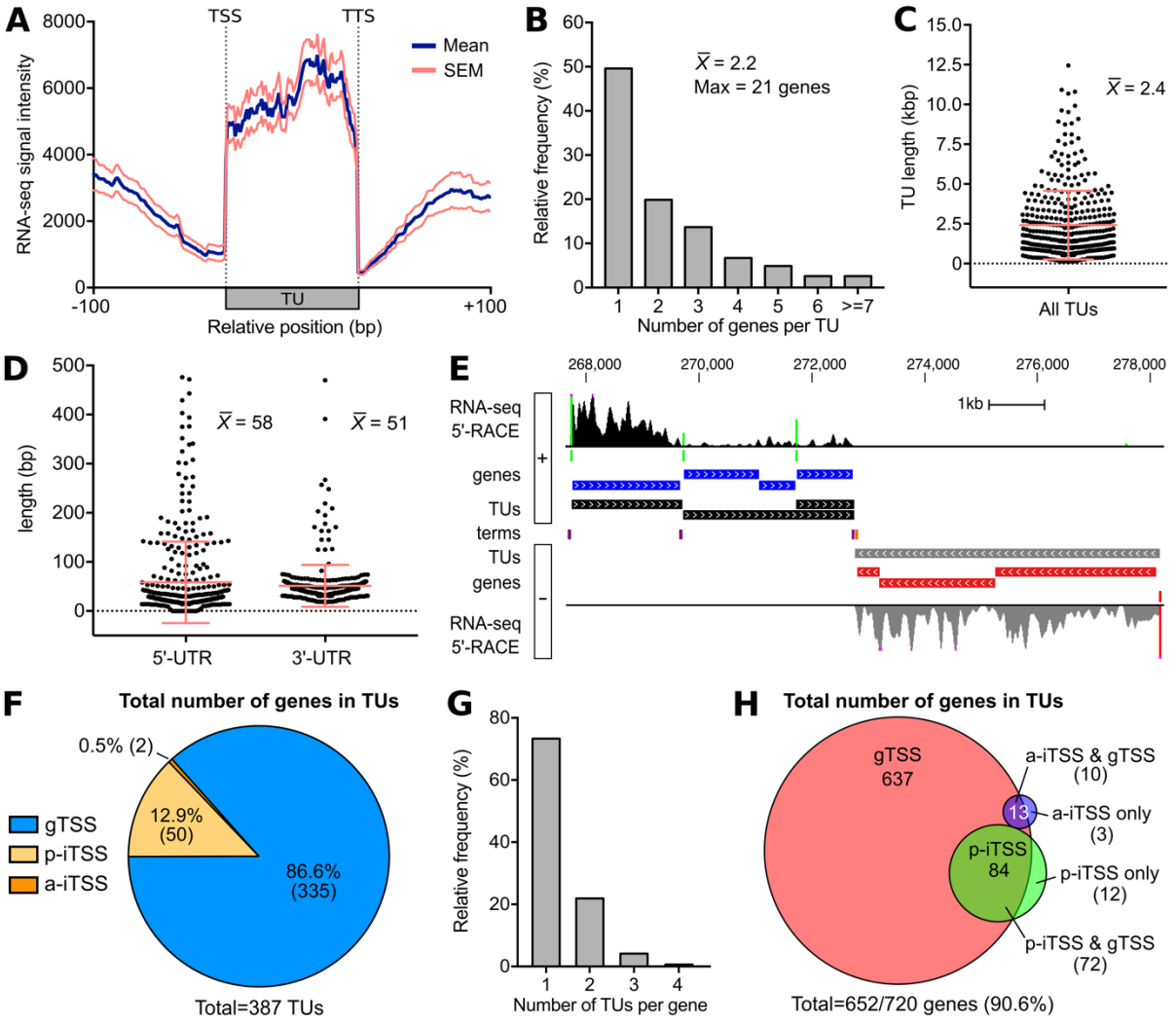

Figure 4.4. Analysis of reconstructed *M. florum* transcription units (TUs). A) Aggregate profile showing the mean RNA-seq read coverage observed for all reconstructed TUs and their surrounding DNA regions. The calculated SEM is also shown. The aggregate profile was centered on the TUs start and stop coordinates, corresponding to transcription start site (TSS) and termination site (TTS), respectively. B) Relative frequency distribution of the number of genes per TU. The average as well as the maximal number of genes per TU are indicated. C) Scatter plot showing the length of all reconstructed TUs. The mean and associated SD are shown. D) Scatter plot showing the 5' and 3' untranslated regions (UTR) length of reconstructed TUs. The mean and associated SD are shown for UTR type. E) Genomic locus showing a representative example of reconstructed TUs. Genomic coordinates are indicated at the top of the panel. From innermost to outermost tracks: terminators predicted on the positive (purple) and negative (orange) DNA strands; coordinates of TUs on the positive (black) and negative (gray) DNA strands; *M. florum* genes encoded on the positive (blue) and negative (red) DNA strands; position of motif-associated TSSs identified on the positive (green) and negative (red) DNA strands; RNA-seq and 5'-RACE signals observed on the positive and negative DNA

strands, colored-coded identically to TUs and identified TSSs, respectively. Illustrated RNA-seq and 5'-RACE signals represent the number of read and read starts observed for a given position, respectively. RNA-seq signal was smoothed using a 5 pixels window (UCSC Genome Browser integrated function). RNA-seq and 5'-RACE peaks above 20,000 reads and 1,000 read starts are cut and marked by fuchsia dots, respectively. F) Proportion of TUs per TSS type. a-gTSS are by definition excluded from the analysis since they are facing the nearest downstream gene. G) Relative frequency distribution of the number of TUs per *M. florum* gene. H) Venn diagram showing the total number of genes included in TUs generated from the different TSS types.

(asRNAs). Of the 652 genes covered by TUs, nearly two-thirds were individually included in only one TU, each thus being transcribed from a single promoter (Fig. 4.4G). The remaining genes were found to be comprised in up to four different TUs each. Interestingly, the vast majority of genes associated to an iTSS were also found to be transcribed from a gTSS, revealing only 15 genes exclusively transcribed from iTSSs (Fig. 4.4H). In fact, every gene associated to more than one TUs was part of a gTSS TU, and only about half of them (45.4%) were also transcribed from an iTSS TU. Overall, this suggests that iTSSs might have only a secondary role in the transcription of downstream genes. Yet, iTSSs could be involved in the transcription of other elements such as sRNAs.

4.5.5 Estimation of intracellular levels of protein and nucleic acid species

In order to obtain an even more detailed overview of a near-minimal cell, we then used our macromolecular biomass quantification data to estimate the intracellular levels of the different nucleic acid and protein species of *M. florum*. In *M. florum* L1, the genome is organized as a single and circular chromosome of 793,224 bp (43, 73). Based on its sequence, this chromosome has an estimated molecular weight of 489,954.48 kDa. The number of chromosome copies can then be directly estimated from the DNA mass per cell in respect with its molecular weight. Given that *M. florum* contains 1.70 ± 0.54 fg of DNA per cell during the log phase (Table 4.1), we estimated that the average *M. florum* cell should contain ~2 chromosome copies under these growth conditions.

In cells, RNA can be decomposed into three major classes, i.e. rRNA, tRNA, and mRNA. In both bacteria and eukaryotes, rRNA and tRNA constitute the predominant forms of cellular RNA, occupying approximately 80% and 15% of the total RNA mass, respectively (59, 74). Prokaryotes rRNA is composed of three different molecules, the 5S, 16S, and the 23S rRNA, which are typically organized as a co-transcribed operon and produced by the cleavage of a long precursor transcript. In *M. florum*, the genomic locus coding for the RNA genes is found at two copies, and our 5'-RACE results showed that they are indeed transcribed as single polycistronic transcripts corresponding to TUs TU_090 and TU_229 (Datasets S4.3 and S4.5). The remaining proportion of cellular RNA principally accounts for mRNA (~5%), as well as other less abundant types of RNA such as sRNA (<1%) (74). According to our macromolecular quantification results (see Table 4.1) and supposing that the proportions of RNA classes are conserved in *M. florum*, rRNA, tRNA, and mRNA have a total mass of 3.91, 0.73, and 0.24 fg, respectively (Dataset S4.5). If we assume that the 5S, 16S, and 23S rRNAs are found at equimolar ratios, the calculated rRNA mass and estimated molecular weight suggest that roughly 4,900 rRNA molecules are present in a single *M. florum* cell (see Dataset S4.5). Using the same assumption for the different tRNAs, this cell would contain a total of approximately 18,000 tRNA molecules. Given the most probable *M. florum* cell volume (Table 4.1), we estimate that rRNAs and tRNAs are present at concentrations of ~5.4×10^4 rRNAs/μm^3 and ~2.0×10^5 tRNAs/μm^3, respectively (Table S4.2). tRNAs are thus almost four times more abundant than rRNA molecules even though they occupy only ~15% of the total RNA mass.

We then used our RNA-seq data to estimate the intracellular abundance of each *M. florum* mRNA species (Dataset S4.5). We observed excellent correlations between the number of fragments per kilobase per million of mapped reads (FPKM) of *M. florum* coding sequences (CDS) calculated from the different replicates (Fig. S4.9B). The FPKM values averaged over all replicates followed a typical Poisson distribution, with two-thirds of all CDS (453/685) siting between 0 and 1,000 FPKM (Fig. 4.5A and Fig. S4.9C). A total of 660 CDS showed a detectable expression level (FPKM>0), and 314 of these were expressed at a higher level than if all the reads were equally distributed among *M. florum* genes (FPKM>630) (Fig. 4.5A and Fig. S4.9D).

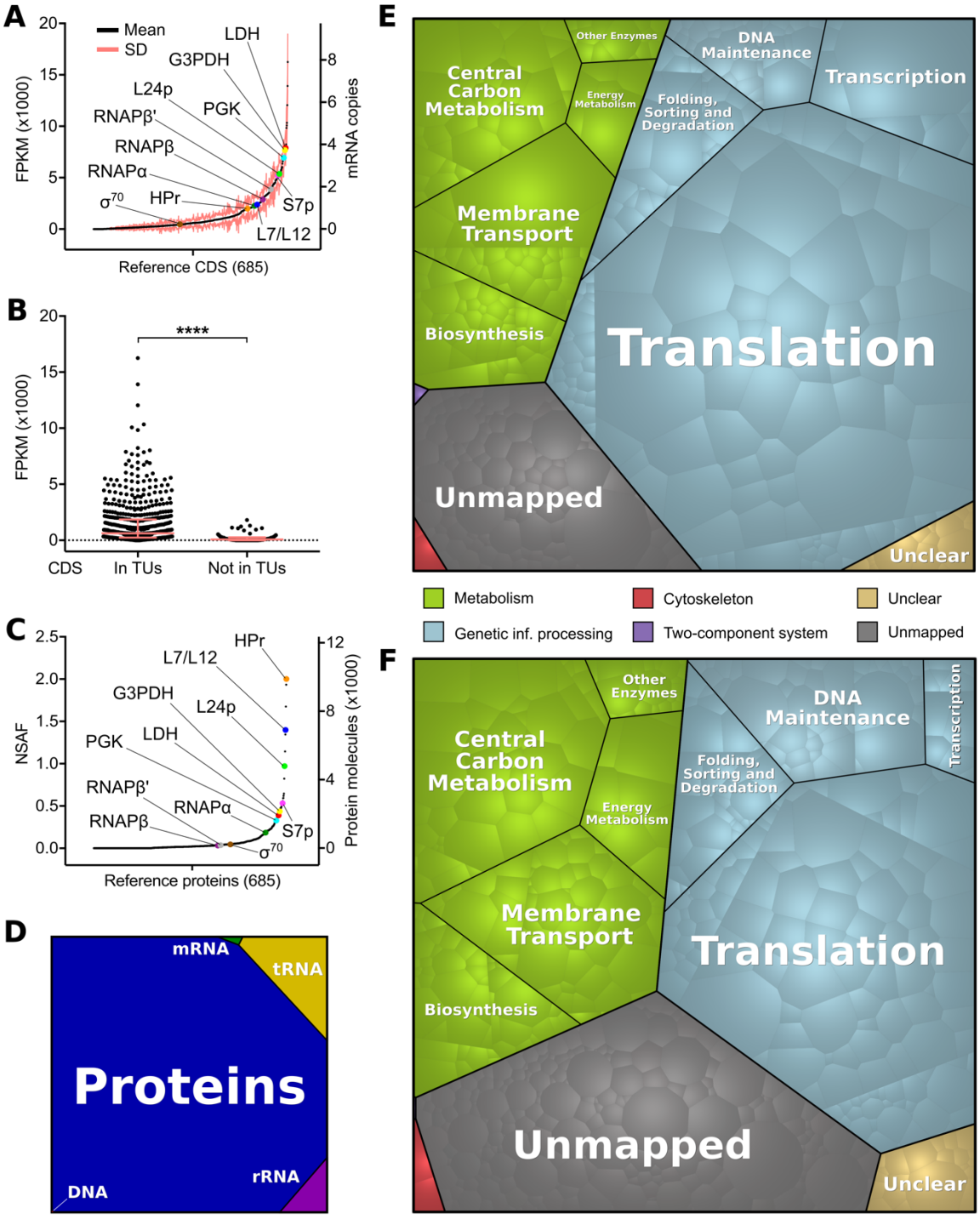

Figure 4.5. Expression levels of *M. florum* protein-coding genes and enrichment of functional categories. A) Transcription levels of all *M. florum* coding sequences (CDS) quantified by RNA-seq. Transcription levels were calculated according to the number of fragments per kilobase per million of mapped reads (FPKM) observed over six replicates. The

corresponding number of mRNA copies per cell, estimated from the measured *M. florum* RNA mass, are also indicated. CDS were sorted from least to most transcribed. The transcription level of selected genes of importance are presented. LDH, L-lactate dehydrogenase (peg.600); G3PDH, glyceraldehyde-3-phosphate dehydrogenase (peg.583); PGK, phosphoglycerate kinase (peg.582); L24p and L7/L12, large subunit ribosomal proteins L24p (peg.133) and L7/L12 (peg.605); S7p, small subunit ribosomal protein S7p (peg.626); RNAPβ, RNAPβ', and RNAPα, RNA polymerase subunits β, β', and α (peg.601, peg.602, and peg.149); HPr, phosphotransferase system phosphocarrier protein HPr (peg.570); σ^{70}, RNA polymerase sigma factor RpoD (peg.269). B) Transcription level of CDS included in transcription units (TUs) compared to CDS not associated to any TU. The median and interquartile range are shown for both groups. The mean rank of each group was compared using a Mann-Whitney test (p-value <0.0001). C) Abundance of all *M. florum* reference proteins quantified by two-dimensional liquid chromatography-tandem mass spectrometry (2D LC-MS/MS). Abundance was estimated according to the normalized spectral abundance factor (NSAF) associated to each protein. A NSAF value of 0 was assigned to undetected proteins. The corresponding number of protein molecules per cell (derived from the biomass data) are indicated. Proteins were sorted from least to most abundant. The expression level associated to the same genes of interest presented in panel A are indicated. D) Overall DNA, tRNA, rRNA, mRNA, and protein proportions in terms of intracellular abundances in *M. florum*. E) Voronoi diagram illustrating the relative abundance of *M. florum* reference proteins associated to different functional categories. Each polygon represents a different protein weighted by its expression level quantified by 2D LC-MS/MS. Functions were attributed based on the KEGG Orthology (KO) database (79). The unmapped category regroups proteins for which no KO identifier could be assigned, while the unclear category contains proteins with KO numbers matching to unclear functions. F) As panel E but for mRNA abundances quantified by RNA-seq.

Many metabolic genes involved in glycolysis showed particularly high expression levels, notably peg.600 (L-lactate dehydrogenase), peg.583 (glyceraldehyde-3-phosphate dehydrogenase), and peg.582 (phosphoglycerate kinase) (Fig. 4.5A and Dataset S4.5). Interestingly, three of the ten most expressed genes were annotated as hypothetical proteins, suggesting that important cellular functions are still missing in the current genome annotation. We also observed a striking difference in the transcription levels of CDS included in TUs compared to those for which no TSS could be attributed (Fig. 4.5B). Indeed, protein-coding genes without an associated promoter displayed significatively lower expression values, which could explain why no TSS was identified for these CDS. We did however not observe any clear correlation between the TSS signal intensity of a TU and the expression of its associated genes (data not shown). According to the measured RNA mass (Table 4.1) and calculated FPKM values, we estimated that a total of approximately 420 mRNA molecules are expected to be

present at any moment within a log-phase *M. florum* cell growing in rich medium (Dataset S4.5). If we normalize this value according to the most probable *M. florum* volume (Table 4.1), this represents ~4.7x10^3 mRNAs per μm^3 of cell volume (Table S4.2). The expression value of most CDS (553/685) corresponded to less than one mRNA copy per cell, suggesting heterogenous expression levels between cells of the population and dynamic control of gene expression. Considering that *M. florum* has a doubling time of 31-33 min (see Fig. 1C) and that most mRNA in bacteria have a very short half-life (less than 7 min in *B. subtilis* (145) and between 3 and 8 min in *E. coli* (146)), it is fair to assume that the entire *M. florum* mRNA pool of is almost completely renewed after one generation. In fact, more than 1,000 mRNA molecules are expected to be synthetized during a single cell cycle. Consequently, even mRNA expressed at less than one copy per cell could still be expressed at substantial levels at some points during the cell cycle.

Proteins occupy nearly the half (46.6%) of the total *M. florum* dry mass (Fig. 4.2F and Table 4.1). However, this macromolecular quantification does not provide information about the identity and specific abundance of the different proteins produced by the cell. This information is highly relevant in the context of whole-cell modelling approaches such as GEMs, which can be used to validate the predicted metabolic fluxes of the reconstructed network. We therefore performed two-dimensional liquid chromatography-tandem mass spectrometry (2D LC-MS/MS) on a log-phase *M. florum* culture, and analyzed the obtained spectra using three different search engines to maximize the identification of peptides matching the genome annotation (see Materials and Methods). With the applied filters and parameters, more than 6,400 different peptides were identified, all together supported by more than 40,000 validated spectra at 1% false-discovery rate (FDR). Both the identified peptides and matching spectra showed very high average confidence rates (98.9%). More importantly, the detected peptides matched with 481 different *M. florum* ORFs, each protein supported by an average of 84.3 peptides, for an average protein coverage of ~33.0% (Dataset S4.6). The detected proteins also showed a very high average confidence rate (99.8%), and similarly to the estimated transcription levels, the normalized spectral abundance factor (NSAF) associated to each protein followed a

Poisson distribution (Fig. 4.5C and Dataset S4.6). Indeed, a very low numbers of proteins were detected at strikingly high levels, principally ribosomal proteins, with most proteins showing mid to relatively low expression levels. Nonetheless, the correlation between transcription (FPKM) and expression (NSAF) levels was shown to be relatively modest (Spearman r = 0.61), a tendency also observed in other organisms (75-78). Using the calculated molecular weight of each proteins and the measured total protein mass (Table 4.1), we converted the associated NSAF into absolute molecular quantities. According to our data, the average *M. florum* cell should contain approximately 250,000 protein molecules, with the most abundant protein present at almost 10,000 copies (peg. 570, HPr PTS phosphocarrier protein) (Fig. 4.5C and Dataset S4.6). This represents more than ten times more molecules compared to the RNA fraction of the cell, for roughly twice the mass (Fig. 4.5D). If we normalize the number of protein molecules per unit of cell volume, this represents roughly $2.8x10^6$ proteins/μm^3 (Table S4.2).

4.5.6 Overview of expressed cellular functions

We finally wanted to use our proteomic quantification data to visualize what cellular functions are predominantly expressed by *M. florum*. We therefore assigned KEGG Orthology (KO) identifiers (79) to *M. florum* annotated ORFs and retrieved the associated functional categories. A KO number was successfully attributed to a total of 435 *M. florum* proteins, of which 22 showed unclear function (Dataset S4.7). Since the same protein can be assigned to multiple functional categories, we then curated the assigned categories based on the non-redundant Proteomap functional hierarchy (80). This allowed us to create a curated tree-like functional hierarchy for 413 different *M. florum* annotated proteins (Table 4.2 and Dataset S4.7). The predicted functions of these proteins could be regrouped in just 27 different functional categories, illustrating the striking simplicity of this near-minimal organism. We then used weighted Voronoi diagrams to visualize the relative importance of the assigned functional categories (80). Unsurprisingly, the largest portion of the *M. florum* proteome was occupied by proteins implicated in translation processes, representing almost half (49.0%) of the total protein molecules of the cell and 33.5% of the total protein mass (Fig. 4.5E, Dataset S4.6 and S4.7).

Table 4.2. Curated functional hierarchy tree of *M. florum* annotated ORFs.

Functional category	Subcategory	Sub subcategory	Number of ORFs	% of total ORFs
Cellular Processes	Cytoskeleton	Cytoskeleton proteins	2	0.3
Environmental Information Processing	Signal Transduction	Two-component system	1	0.1
Genetic Information Processing	DNA Maintenance	DNA repair	23	3.4
		DNA replication and partition	30	4.4
		Subtotal	53	7.7
	Folding, Sorting and Degradation	Chaperones and folding catalysts	7	1.0
		Nucleases	11	1.6
		Peptidases	9	1.3
		Protein export	7	1.0
		Sulfur relay system	2	0.3
		Subtotal	36	5.3
	Transcription	RNA polymerase	5	0.7
		Transcription factors	6	0.9
		Subtotal	11	1.6
	Translation	Ribosome	51	7.4
		Ribosome biogenesis	29	4.2
		Translation factors	11	1.6
		tRNA loading and maturation	30	4.4
		Subtotal	121	17.7
	Total		221	32.3
Metabolism	Biosynthesis	Amino acid metabolism	5	0.7
		Cofactor biosynthesis	16	2.3
		Lipid and steroid metabolism	8	1.2
		Purine and pyrimidine metabolism	23	3.4
		Subtotal	52	7.6
	Central Carbon Metabolism	Glycolysis and carbohydrate metabolism	35	5.1
		Other central metabolism enzymes	6	0.9
		Pentose phosphate metabolism	8	1.2
		Subtotal	49	7.2
	Energy Metabolism	Oxidative phosphorylation	9	1.3
	Membrane Transport	PTS system	13	1.9
		Secretion system	2	0.3
		Transport	42	6.1
		Subtotal	57	8.3
	Other Enzymes	Other enzymes	22	3.2
	Total		189	27.6
Not mapped	-	-	250	36.5
Unclear	-	-	22	3.2
Grand total			685	100.0

Central carbon metabolism and membrane transport categories also displayed particularly important proteome fractions, accounting for 7.5% and 7.4% of the *M. florum* protein diversity, respectively (Fig. 4.5E). On the opposite, only very limited proteome allocation (<1%) was shown to be associated with cytoskeleton and two-component system functional categories. More importantly, proteins assigned to functional categories comprised 86.0% of the total estimated protein molecules per cell, excluding those with unclear function, which represents 82.1% of the *M. florum* protein mass (Fig. 4.5E, Dataset S4.6 and S4.7). Functional categories weighted with the estimated mRNA abundances also showed the same overall picture, with however a slightly larger portion occupied by metabolism and unmapped categories (Fig. 4.5F). Additional experiments will be required to determine the role of proteins with unknown or hypothetical function, and therefore assign the remaining protein fraction to the appropriate functional categories. Interestingly, our protein quantification data and functional category assignments could be used to estimate the abundance of conserved protein complexes, the bacterial ribosome for example. According to our analysis, we estimated that each *M. florum* cell should contain between 1,600-2,100 ribosomes, which represents approximately 18,000-24,000 ribosomes per μm^3 of cell volume (Table S4.2). We also estimated that ~270 core RNA polymerase (RNAP) should be present in the average *M. florum* cell (~3,000 RNAP/μm^3), which nearly matches the number of σ^{70} factor per cell (~230).

4.6 Discussion

Due to its peculiar characteristics, *M. florum* constitutes an attractive candidate to become a model organism for synthetic genomics and systems biology studies. This near-minimal bacterium possesses a genome smaller than those of most model organisms currently used in that context (e.g. *E. coli, Mycoplasma pneumoniae, M. mycoides*), have never been associated to any disease and is therefore classified as a BSL-1 organism, and grows rapidly in conditions that do not require expensive or specialized equipment. The flip side of being non-pathogenic is that up to very recently, only little attention had been given to *M. florum*, even though it was isolated almost forty years ago (38, 81, 82). In fact, only a few studies specifically investigated

the biology of this microorganism. Consequently, practically no quantitative data about the physiology of *M. florum* is available in the literature, and many important aspects of its cellular mechanisms and metabolism are yet to be characterized. Here, we measure several physical, physiological, and molecular characteristics of *M. florum* and integrate the generated data to propose estimates on parameters difficult to measure using conventional laboratory equipment. A summary of the characterization performed in this study is presented in Figure 4.6. More specifically, we precisely measure the *M. florum* growth kinetics in rich medium (Fig. 4.1), and use the measured cell diameter, buoyant density, and dry mass to infer the most probable cell mass, volume, and surface area (Fig. 4.2 and Table 4.1). We also quantify the macromolecular mass fractions of the cell (Fig. 4.2F and Fig. S4.4) and proceed to the first experimental cartography of *M. florum* TUs based on 5'-RACE TSSs identification results and Rho-independent terminator predictions (Figs. 4.3-4.4, Fig. S4.5-S4.14, Datasets S4.1-S4.4). Finally, we quantify the transcription and expression levels of all *M. florum* reference CDS, use the macromolecular quantification results to estimate absolute mRNA and protein abundances, and use these estimations to evaluate the relative importance of protein functional categories (Fig. 4.5, Table 4.2, and Datasets S4.5-S4.7).

Using the 2F-DT assay, we showed that *M. florum* L1 has an optimal growth temperature of 34°C and exhibits impaired growth phenotypes at temperatures equal or above 36°C (Fig. 4.1A). Similar results were observed in 1984 by McCoy et al. when they first described several *M. florum* strains (38), including L1, although the permissive temperature ranges and optimum growing temperatures were not specifically reported for each strain included in the study. While *M. florum* has never been associated to any disease, this does not completely rule out the possibility that this bacterium could be pathogenic to some organisms similarly to certain mycoplasmas and spiroplasmas. However, since the growth of *M. florum* L1 is completely abolished at 38°C (and impaired at 36°C), the probability that it parasitizes warm-blooded animals is extremely low. In addition, *M. florum* has typically been isolated from plant flowers and insects (38, 43, 81, 82), and no apparent virulence factor is predicted from its genome sequence. Yet, the exact nature of its primary host remains unclear, but the previous isolation of

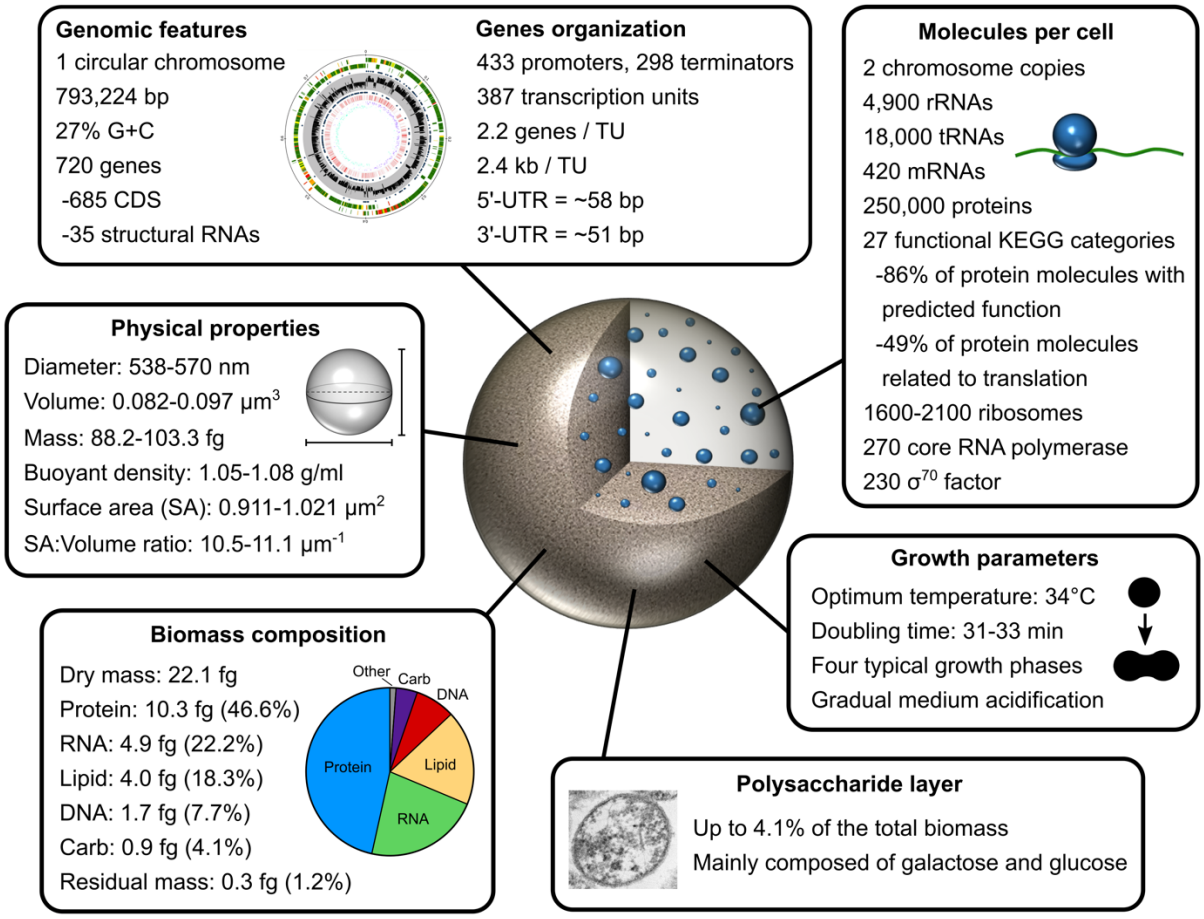

Figure 4.6. Overview of the *M. florum* characterization reported in this study.

various *M. florum* strains from insects gut suggest that it could be an commensal of the digestive tract of these organisms (83). This is supported by the fact that many members of the *Entomoplasmatales* group, including several species of the *Mesoplasma*, *Spiroplasma* and *Entomoplasma* genera, have been isolated from or are associated with arthropods (47, 84-87). Some of these Mollicutes even share particularly interesting mutualistic relationships with their hosts. This lifestyle would also explain the presence of *M. florum* on plant flowers given that insects are responsible for most of the pollination. The unique environment provided by digestive tract of insects would allow *M. florum* to have continuous access to complex nutrients such as lipids and peptides to palliate for its metabolic deficiencies, as well as to various sugar

sources depending of the diet of its host. Further data is however required to confirm that hypothesis.

By following the CFU and FCM counts of a *M. florum* batch culture incubated at its optimal growing temperature (34°C), we showed that this bacterium goes through the four typical bacterial growth phases (lag, log, stationary, death) (Fig. 4.1B-D). The measured OD_{560nm} signal, which is correlated with the medium pH, showed a progressive medium acidification during the log phase. Since the main route for energy production in *M. florum* is predicted to be the glycolysis pathway and no tricarboxylic acid (TCA) cycle is present, this gradual acidification is most probably caused by the accumulation of fermentation products (lactate and acetate) in the medium (34, 88-90). The observed decrease in OD_{560nm} eventually reached a plateau, corresponding to a medium pH of ~6.0, which also coincided with the beginning of the death phase. At that point, the high concentration of H^+ in the medium is likely toxic for *M. florum*, but the exact causes are currently unknown. Nonetheless, it would be interesting to measure the production rate of lactate and acetate by *M. florum* and integrate the data into a GEM. Given the simplicity of the *M. florum* metabolism, the definition of this constraint will have an important impact on the quality of GEM fluxes predictions for metabolic pathways linked to acetate and lactate production, for example the glycolysis pathway. The accurate prediction of the overall glycolysis flux is particularly informative since it allows to link the ATP production to the *M. florum* doubling time, which we showed to be remarkably small (~31-33 min, Fig. 4.1C) compared to other Mollicutes. *M. mycoides*, for example, exhibits a doubling time of ~60 min in rich medium (24, 35)(Gibson et al., 2010a; Hutchison et al., 2016), while it is estimated to be around 90 min for *M. capricolum* (40) and 8-20 hrs for *M. pneumoniae* (91, 92). Intriguingly, *Mycoplasma genitalium*, who possesses the smallest genome amongst all Mollicutes (~580 kb), has an extremely slow growth rate corresponding to a doubling time of ~16 hrs (24, 93). Clearly, the doubling time of Mollicutes is not correlated with their genome size. At this time, the factors contributing to the fast-growing phenotype that characterizes *M. florum* are still elusive. Perhaps this distinctive trait is rather related to the nature of the natural habitat in which this bacterium has evolved, as well as its interaction with it, that for

now remain only speculative for *M. florum*. Still, if this bacterium happens to be a commensal of insect guts, it is plausible that the fast-growing phenotype might be a major factor contributing to its fitness given that this environment is characterized by a continuous flow of nutrients and a constant elimination of bacteria through the feces. In fact, these characteristics resemble to those artificially produced by chemostat systems, which have been demonstrated to allow the selection of fast-growing *E. coli* mutants under constant glucose excess (94). The utilization of a GEM that integrates the metabolic fluxes, nutrients availability, growth rate, and ATP production rate of *M. florum*, and more importantly its comparison with other Mollicutes GEMs, might yield more specific hypotheses on the underlying genetic factors contributing to the fast-growing phenotype of *M. florum*.

Using TEM and STED microscopy, we showed that *M. florum* has an average cell diameter comprised within 434 and 741 nm (Fig. 4.2A-D). The significant difference observed by the two methods is most likely caused by biases associated with sample preparation. TEM, for example, implies a dehydration of the cells with ethanol, which can cause cell shrinkage and therefore a reduction in their apparent diameter (95). STED, on the other hand, requires the use of a mounting media during slide preparation that can cause sample distortion and alteration of morphological features (96, 97). Nevertheless, we confirmed that the *M. florum* cell diameter was effectively within the range obtained by TEM and STED microscopy using a mathematical approach that integrates other physical parameters such as the cell buoyant density and cell dry mass (Fig. 4.2E-G). Surprisingly, the most probable cell diameter range inferred by the integrative approach (538-570 nm) corresponded almost exactly to the overlapping portion of the distributions observed with the two microscopy methods. The determination of the most probable cell diameter allowed us to estimate the most probable cell mass, volume, surface area, and SA:V ratio of *M. florum* (Table 4.1). According to our analysis, *M. florum* is expected to have a volume between 0.082 and 0.097 μm^3 during the log phase, which is nearly 50 times smaller than *E. coli* growing in similar conditions (~4 μm^3) (98, 99). This important difference in cell volume also results in the augmentation of its SA:V ratio relatively to *E. coli*, with values approaching 10 μm^{-1} for *M. florum* compared to ~4 μm^{-1} for *E. coli*. Recent publications

demonstrated that bacteria exhibit robust SA:V ratio homeostasis in response to different types of perturbations, including nutritional shifts and genetic alterations (100-102). Since Mollicutes have lost the ability to synthetize many important metabolites through massive gene loss, their high SA:V ratios could represent a physical adaptation to increase their capacity of importing complex nutrients from the environment. Interestingly, this difference in SA:V ratios between *M. florum* and *E. coli* is also apparent when comparing the macromolecular mass fractions associated to each bacterium (Fig. 4.2F). In *M. florum*, we showed that approximately 18% of the dry mass comes from lipids and 47% from proteins, whereas these fractions typically represent ~9% and ~55% of the *E. coli* dry mass (59, 103, 104). Overall, the *M. florum* macromolecular mass fractions were comparable to fractions observed for other Mollicutes, with slight differences depending on which species are compared (105).

Likewise *M. mycoides* (105), a notable fraction of the *M. florum* biomass is allocated to carbohydrates (4.1%). This suggests the presence of a polysaccharide capsule similarly to phylogenetically related mycoplasmas of the mycoides cluster (49-52). TEM pictures strongly suggest the presence of a thin polysaccharide layer at the periphery of *M. florum* cells (Fig. 4.2A), a morphological feature previously observed by our group using scanning electron microscopy (data not shown). According to the gas chromatography-mass spectrometry (GC-MS) results used to quantify the total amount of carbohydrate per cell, this layer would be primarily composed of galactose (54.9%) and glucose (20.6%). This suggests the presence of a biosynthesis pathway similar to what is found in *M. mycoides* and *M. capricolum* (49-52). However, the genetic determinants responsible for the synthesis of this polysaccharide layer, its biological function, as well as the precise organization of its sugar monomers remains to be identified in *M. florum*. Additionally, it is still unclear whether this thin layer constitutes capsular polysaccharides (CPS) covalently bound to the cell surface or exopolysaccharides (EPS) secreted in the culture medium that passively coat *M. florum* cells. In fact, both forms could exist and be subjected to regulation depending on specific environmental signals or conditions. Since *M. florum* cells were washed several times prior biomass quantification and TEM examination, the sole presence of EPS would be quite surprising. Nevertheless, the

production of this polysaccharide layer could be responsible for the aggregation of cells under low pH conditions observed during the stationary phase (Fig. 4.1B). In the environment, this layer could potentially serve as a protection against desiccation outside of its host. This would provide increased survivability on plant surfaces and contribute to its dissemination across insect populations.

Even though the genome sequence of *M. florum* L1 is publicly available since 2004 and figures amongst the smallest genomes found in free-living bacteria, no studies have experimentally explored the transcriptomic complexity of this bacterium. Here, we combined 5'-RACE and RNA-seq methodologies to dress a first portray of its transcriptome. The analysis of 5'-RACE reads revealed 432 TSSs associated to a highly conserved Pribnow box (TAWAAT [36]) and a relatively conserved EXT element (Fig. 4.3A-C and Dataset S4.1). These TSSs were also confirmed by RNA-seq methodology (Fig. 4.3G). Transcription was shown to be predominantly initiated on purine nucleotides (A or G) and separated by 6 or 7 nucleotides from the -10 box (Fig. 4.3E-F). Interestingly, no clear sequence enrichment was visible around position -35, suggesting a highly degenerated -35 box as previously observed in other Mollicutes (67, 106, 107). The absence of a -35 box could however be compensated by the EXT element, which was also shown to be conserved in Mollicutes lacking the -35 box (106, 107). Furthermore, EXT element is conserved and important for transcription initiation in many gram-positive bacteria such as *Bacillus subtilis*, and was shown to compensate the absence of the -35 box in *Streptococcus pneumoniae* (108, 109). The remaining positions of the *M. florum* consensus promoter were mostly enriched for A or T nucleotides. It would be interesting to experimentally investigate the relative importance of the -10 box and the EXT element in *M. florum*, and verify the impact of the addition of the σ^{70} -35 box consensus (TTGACA) on promoter strength. High-throughput approaches using randomized promoter libraries could be an efficient strategy to explore the diversity of sequence enabling transcription initiation in *M. florum* and systematically determine their promoting strength (110, 111). This method could also be exploited to test transcriptional regulators as well as study the dependency of *M. florum* upon ribosome binding sites in order to efficiently initiate the translation process. This would be

particularly interesting since our results showed evidences of leaderless mRNAs in *M. florum* (Fig. S4.8), although a highly conserved Shine-Dalgarno motif can be found in 80% of all CDS instances (Fig. S4.15). Investigating the impact of mutations in the Shine-Dalgarno motif on translation would however require the development of reporter systems which are currently not available in *M. florum*. Overall, such efforts could help establish a collection of highly characterized regulatory pieces specially developed for *M. florum*, including promoters encompassing a wide range of transcriptional activity, which will facilitate the design of artificial gene circuits in this bacterium.

With such an A-T rich genome (~73%), paired with the absence of a conserved -35 box, the number of spurious promoters giving rise to transcriptional noise is expected to be relatively high in *M. florum* (107, 112). Our 5'-RACE results revealed 181 putative TSSs not associated to the *M. florum* consensus promoter sequence (Fig. S4.6). Contrary to motif-associated TSSs, these TSSs were mostly located within coding regions of the genome, displayed a weaker signal intensity, and initiated transcription predominantly on C or T nucleotides instead of A or G. Additional efforts to search for promoter sequence similarities between these TSSs were proved to be unsuccessful. Considering these striking differences, putative TSSs not associated to the *M. florum* promoter motif were not considered for the reconstruction of TUs. These 5'-RACE peaks are probably the result of low affinity binding events of the σ^{70} to sequences faintly resembling to promoter elements, resulting in the initiation of transcription at spurious sites. However, since the intensity of these TSSs is globally very low, their potential impact on the normal transcription of overlapping genes is likely negligible. Even though the phenomenon of spurious transcription seems to be widespread across bacterial species, its putative biological function is still controversial. Recent studies in *B. subtilis* and *M. pneumoniae* showed that spurious promoters might be involved in the production of sRNAs (112, 113). Intriguingly, the conservation of sRNAs tends to be limited, which implies a dynamic process of creation and elimination throughout the course of evolution (107, 112, 114-116). Spurious promoters might in fact serve as a reservoir on which natural selection can operate to produce functional sRNAs, thus participating to the overall transcriptome plasticity of cells. Such plasticity is crucial for the

survival of bacterial populations when facing selective pressures such as environmental stresses or genetic perturbations (e.g. insertion of mobile genetic elements) (116). Furthermore, the energetic cost related to the transcription of sRNAs in *M. pneumoniae* was estimated to be extremely small, representing less than 3% of the total energy spent on transcription (112). The low associated cost, absence of deleterious effect, and bias towards A-T mutations may have led to the accumulation of spurious promoters in bacterial genomes, which is exacerbated in bacteria containing A-T rich genomes such as the Mollicutes (112). Using our 5'-RACE and RNA-seq data, we found contextual evidences suggesting the expression of sRNAs in *M. florum* (e.g. motif-associated TSSs placed before predicted terminators), but additional experiments are required to confirm their existence as well as determine their potential function (Fig. S4.14 and Dataset S4.4).

Apart from sRNAs, it is quite reasonable to presume that spurious promoters could also evolve and be selected to drive the expression of other type of functional RNAs, mRNAs for instance. Interestingly, we found that a small proportion (22%) of identified motif-associated TSSs were in fact located within coding regions of the *M. florum* chromosome (iTSSs), and many of them were properly oriented to drive the expression of downstream genes (Fig. 4.3D, Fig. S4.8, and Dataset S4.1). Excepted for their weaker signal intensity, iTSSs displayed the same characteristics as intergenic TSSs (gTSSs), and were found to be enriched near the end of their overlapping gene (Fig. 4.3E-G, Fig. S4.10). Yet, the majority of genes located downstream of iTSSs were apparently also controlled by a gTSS (Fig. 4.4H), which raises questions about their actual role in *M. florum*. Are iTSS really important for the correct expression of certain genes via the production of supplementary mRNA isoforms, or they simply are the results of mutations within spurious promoters bringing them closer to the *M. florum* promoter consensus without significatively interfering with the normal transcription of genes? Could mRNA isoforms generated from iTSSs be used to produce shorter, or conversely, longer form of proteins via alternative translation initiation codons? Or maybe certain iTSS-associated promoters constitute a regulatory platform for the biding of transcriptional factors activating or repression transcription upon specific signals, while its cognate intergenic promoter provide a more

constitutive expression under normal conditions. The biological functions of iTSSs could possibly be quite diverse and are most likely highly dependent on the context in which each is specifically involved.

Using the identified motif-associated TSSs and the predicted Rho-independent terminators, we reconstructed 387 TUs in *M. florum*, encompassing for more than 90% of all annotated genes (Fig. 4.4, Fig. S4.12, and Dataset S4.3). Although about half of TUs were shown to contain only a single gene, many TUs were polycistronic, and about 25% of all *M. florum* genes were included in more than one TU (Fig. 4.4B and 4.4G). This resulted in many overlapping TUs, illustrating once again the impressive genomic complexity residing in cells that are presumed to be the simplest among all self-replicative organisms (117). Recent studies showed that transcription terminators are often not entirely efficient, allowing transcriptional readthrough and thus contributing to the transcription of downstream elements (113, 114, 118). For now, our TU reconstructions were based on the assumption that all predicted terminators were 100% efficient, which almost certainly underestimates the full transcriptome diversity in *M. florum.* Nevertheless, our RNA-seq data correlates very well with the reconstructed TU boundaries as well as terminator predictions (Fig. 4.4A, Fig. S4.11 and Fig. S4.13), suggesting that transcriptional readthrough is not predominant in *M. florum*. Termination readthrough could still be responsible for the very low expression of genes not associated to any of the identified promoters, which represent roughly 10% of all *M. florum* genes (Fig. 4.5B). Of course, as this represents the very first characterization of the *M. florum* transcriptome, it will be possible to integrate additional datasets to improve its overall precision and breadth. For example, methods that inform about the 3' end coordinates of transcripts such as the Rend-seq (118) could be used to validate and improve the current terminator predictions, in addition to potentially highlight occurrences of leaking terminators. The incorporation of the regulatory networks of all predicted transcriptional regulators would also represent an important step towards a deep and complete characterization of the mechanisms governing the *M. florum* global cell functioning. This could be accomplished by systematically tagging each predicted transcriptional regulator and performing chromatin immunoprecipitation coupled to exonucleases (ChIP-exo) (119-121).

Achieving a complete and quantitative description of all the constituents of a cell constitute one of the most important goals of systems biology. To understand the global properties of complex biological systems such as cells, one must clearly define the identity of their components, and ultimately know the specific abundance at which each of these components is found within the system. According to our biomass quantification results (Fig. 4.2F and Table 4.1), the transcription and expression levels of each gene (Fig. 4.5A-C and Datasets S4.5-S4.6), as well as the calculated molecular weight respective to each molecular species, we estimated that the average *M. florum* cell should contain approximately 250,000 proteins, 4,900 rRNAs, 18,000 tRNAs, 420 mRNAs, and ~2 copies of the chromosome (Fig. 4.5D, Table S4.2, and Datasets S4.5-S4.6). Considering the functional categories assigned by the KO database, we further estimated that about 1,600-2,100 ribosomes, 270 core RNAP, and 230 σ^{70} factor are expected to be present in the average *M. florum* cell (Table S4.2 and Datasets S4.6-S4.7). Overall, the number of RNA and protein molecules per cell is comparable to what is estimated in other Mollicutes but roughly ten times lower than for *E. coli*, which is not surprising considering the large difference in respective cell volumes (Table S4.2). Still, amongst the two other Mollicutes species we selected for comparison, *M. florum* shows the highest number of proteins and ribosomes per cell but is also the species with the highest cell volume with almost three times more cytoplasmic space than *M. mycoides* (Table S4.2). Interestingly, if we normalize the intracellular levels of RNAs and proteins by the cell volume, *M. florum* and *E. coli* show very similar concentrations. The total number of proteins and ribosomes per unit of volume is also very consistent between all the species compared, with the exception of *M. pneumoniae* which has the lowest concentration of proteins and nearly ten times less ribosomes per μm^3 than *M. florum*, *M. mycoides*, and *E. coli* (Table S4.2). This disparity between *M. pneumoniae* and *M. florum* is also apparent when we compare the relative importance of protein functional categories in both species, with *M. pneumoniae* displaying significatively reduced investments in translation processes at the benefits of other processes such as cell motility and cytoskeleton (Fig. 4.5E)(80, 122). Consistently, the overall RNA levels of *M. pneumoniae* are also remarkably lower compared to *M. florum* and *E. coli*, which is not surprising since *M. pneumoniae* has only one rRNA operon per genome versus two and seven for *M. florum* and

E. coli, respectively. These observations are in agreement with the important difference between the growth rate of *M. florum* (~31-33 min) and *M. pneumoniae* (~8-20 hrs), supporting the idea that *M. pneumoniae* is not optimized for biomass production but rather depends on more complex strategies for fitness and competition in its natural environment (91). Furthermore, GEM reconstruction for *M. pneumoniae* revealed that most of the energy produced by this pathogenic bacterium is used for maintenance tasks (NGAM) instead of growth (GAM), strongly contrasting with *M. mycoides* for which the complete opposite was observed (37, 92).

Since *M. florum* and *M. mycoides* share similar numbers of ribosomes per unit of volume, it would be interesting to compare overall aspects of their metabolism using GEMs to see if they also share comparable GAM and NGAM values. If they do, will we observe a small shift towards the GAM parameter for *M. florum* that might explain the small gap in their respective growth rates (~31-33 min for *M. florum* vs ~60 min for *M. mycoides*)? Or perhaps this difference resides in other parameters such as the overall efficiency of the glycolysis pathway or the efficiency of the gene expression machinery. By reconstructing whole-cell models for *M. florum*, it will be possible to integrate the data generated in this study to investigate these questions and gain additional knowledge about the global cell functioning of this near-minimal bacterium. Moreover, since we generated a first draft of the *M. florum* TUs, we now have the data required to use whole-cell modelling algorithms such as MinGenome (123) to improve the initial genome reduction scenarios based on gene essentiality and conservation (43). MinGenome identifies in size descending order all dispensable contiguous sequences and preserves needed promoter regions for proper transcription of retained genes (123). The minimal genome designs inferred by this method could then be systematically analyzed using modelling approaches and compared to the synthetic minimal organism JCVI-syn3.0 to highlight differences in their genome composition and retained protein functions. Interesting genome architectures emerging from these analyses could next be subjected to total DNA synthesis and assembly in yeast followed by transplantation in a recipient bacterium. If successful, the transplanted synthetic genomes could be analyzed using the methods described in this study to

potentially acquire new knowledge about genome design principles, which are currently lacking and restraining the rational design of synthetic organisms.

4.7 Materials and Methods

4.7.1 Bacterial strains and growth conditions

All experiments reported in this study were performed using *M. florum* strain L1 (ATCC 33453) grown with shaking at a temperature of 34°C (unless stated otherwise) in ATCC 1161 medium (1.75% (w/v) heart infusion broth, 4% (w/v) sucrose, 20% (v/v) horse serum, 1.35% (w/v) yeast extract, 0.004% (w/v) phenol red, 200 U/ml penicillin G) (41).

4.7.2 2-fold microplate dilution doubling time assays (2F-DT)

The 2F-DT is a doubling time measurement method based on growth assays previously developed for spiroplasmas (46). Briefly, ATCC 1161 medium was inoculated with a log-phase *M. florum* preculture to obtain a final concentration of ~$1x10^5$ CFU/ml. The inoculated medium was then serial diluted using 2-fold dilutions to obtain a total of four dilutions (1:1, 1:2, 1:4, and 1:8). Each dilution was transferred in triplicate into a 96-well microplate, and the plate was incubated with shaking at the desired temperature (30, 32, 34, 36 or 38°C) in a Multiskan GO microplate reader (Thermo Scientific). Bacterial growth was monitored by measuring the optical density at 560 nm every 10 min for ~16 hrs. The metabolic activity of *M. florum* was previously shown to result in the acidification of the ATCC 1161 growth medium, causing changes in the absorbance of phenol red at 560 nm that correlate with the number CFUs (39). To calculate doubling times, linear regressions ($R^2 > 0.999$) were traced on the linear portion of the 560 nm optical density curves, and the amount of time separating each dilution curve was calculated according to the linear regression equations.

4.7.3 Growth kinetics assays

Growth kinetics assays were performed in triplicate by monitoring the cell concentration of three independent *M. florum* cultures using CFU counts and flow cytometry. Briefly, ATCC 1161 medium was inoculated with a log-phase *M. florum* preculture to obtain a final concentration of ~1x10^5 CFU/ml. Inoculated medium was incubated at 34°C with shaking for ~24 hrs in an orbital shaker incubator. Aliquots were harvested every ~2 hrs and optical density at 560 nm was immediately measured in duplicate using a Multiskan GO microplate reader (Thermo Scientific). CFUs were evaluated in duplicate by spotting serial dilutions of the aliquots (in PBS1X) on ATCC 1161 solid medium and counting colonies after an incubation of 24-48 hrs at 34°C. 37% (w/v) formaldehyde was then added and mixed to each dilution to obtain a final concentration of 1% (w/v) and the plate was incubated at RT for ~25 min. SYBR Green I (Invitrogen) dye was added to a final concentration of 1X, mixed, and samples were incubated again at RT for ~25 min. Cell concentration was finally measured in duplicate by flow cytometry using a BD Accuri C6 Plus flow cytometer (BD Biosciences) equipped with a 488 nm laser. FSC-H and FL1-H (FITC) channel thresholds were set at 100 and 750, respectively. Fluidics were set to high speed, and a maximum of 40 µl or 1x10^6 events were collected for each sample. We validated that cell concentrations were well correlated with culture dilutions diluted in PBS1X (Fig. S4.2), and appropriate controls were performed (PBS1X without cells, unstained cells, etc.).

4.7.4 Cell viability assay

Cell viability of *M. florum* was assessed by SYTO 9 and PI double staining (124). *M. florum* cells were centrifuged at 10°C for 2 min at 21,100 x *g*, and washed once with cold PBS1X. Cells were centrifuged again and then resuspended in PBS1X containing 5 µM SYTO 9 (Molecular Probes) and 10 µg/ml PI (Biotium). Cells were stained at RT for ~20 min. A fixed-cells control was also performed by incubating a *M. florum* washed cell aliquot with 1% (w/v) formaldehyde at RT for ~25 min. Fixed cells were centrifuged at 10°C for 2 min at 21,100 x *g*, resuspended in

PBS1X containing 0.1% (v/v) Triton X-100, and incubated at RT for 2 min. Cells were centrifuged again, and finally resuspended in PBS1X containing 5 μM SYTO 9 (Thermo Fisher Scientific) and 10 μg/ml PI (Biotium). Samples were immobilized on agarose pad slides and examined by widefield fluorescence microscopy using an Axio Observer Z1 inverted microscope (Zeiss) equipped with a AxioCam 506 mono (Zeiss) camera and a 100X/NA1.4 Plan-Apochromat oil immersion objective. SYTO 9 and PI were excited and acquisitioned using GFP and Cy3 excitation/emission filters, respectively. Images were captured with Zeiss Zen 2.0 imaging software and analyzed using Fiji (125).

4.7.5 Stimulated emission depletion (STED) microscopy

STED microscopy was performed using double stained (membrane and DNA) *M. florum* cells. Briefly, a log-phase *M. florum* culture was centrifuged at 10°C for 2 min at 21,100 x *g* and washed twice with cold electroporation buffer (272 mM sucrose, 1 mM HEPES [pH 7.4]). Washed cells were then immobilized on a poly-L-lysine coated glass slide (Poly-Prep Slide, Sigma-Aldrich) and incubated at RT for 5 min. Cells were washed on slide twice with PBS1X, and then stained, fixed, permeabilized, and stained again for 5 min each at RT using the following solutions (all reagents diluted in PBS1X, with two PBS1X washes between each step): 1) 0.5 μM mCLING-ATTO 647N-labeled dye (Synaptic Systems); 2) 4% (w/v) formaldehyde and 0.2% (w/v) glutaraldehyde; 3) 0.1% (v/v) Triton X-100; 4) 1/100 dilution (100X) of PicoGreen concentrate reagent (Molecular Probes). Cells were washed twice again with PBS1X, and then finally mounted for STED microscopy using ProLong Diamond Mountant (Molecular Probes). Two-color STED microscopy was performed using a DMi8 STED microscope (Leica TCS SP8) equipped with a 100X/NA1.4 HC Plan-Apochromat CS2 oil immersion objective and operated with the LAS X imaging software (version 3.1.1.15751, Leica). mCLING-ATTO 647N and PicoGreen were excited using a pulsed white light laser set at 646 and 488 nm, respectively, and depleted using 775 and 592 nm depletion lasers. Signals were acquisitioned using HyD SMD hybrid detectors (Leica) set at 658-698 nm for the ATTO 647N channel and 505-565 nm for the PicoGreen channel. Images were acquisitioned using a 4X zoom factor and deconvolved

using Huygens Professional with STED optical option (version 18.04, Scientific Volume Imaging). Images and cell diameter were analyzed using Fiji (125). Since cells displayed a variable morphology from ovoid to spherical, minor and major axes were measured and averaged to obtain a single representative cell diameter for each cell. Only cells exhibiting both signals were considered in the analysis.

4.7.6 Transmission electron microscopy (TEM)

A log-phase *M. florum* culture was centrifuged at 10°C for 15 min at 7,900 x *g*, and then washed three times with cold PBS1X. The supernatant was discarded, and cells were fixed at RT for 45 min and then overnight at 4°C by adding 1 ml of 2.5 % (w/v) glutaraldehyde on top of the cell pellet. Cells were then washed twice with PBS1X, post-fixed at RT for 90 min using a 1% (w/v) osmium tetroxide solution, and washed twice with water. Cells were then dehydrated through a series of washes (5 min each) with 30%, 50%, 70%, 85%, 95%, and three times 100% (v/v) ethanol. Cells were washed again three times using propylene oxide, with a 5 min incubation at RT after each wash. Cells were then incubated at RT for 1 hr with 1:1 propylene oxide:Epon, incubated two times at RT for 180 min with pure Epon, and then overnight at RT with pure Epon. The Epon and cells mixture was embedded within a polyethylene capsule (BEEM) and polymerized by baking at 70°C for 48 hrs. The block was cut into thin sections (~80 nm) and placed on a copper grid, stained sequentially with uranyl acetate and lead citrate (~10 min each), and finally examined under a Hitachi H-7500 TEM microscope operating at an accelerating voltage of 80 kV. Images and cell diameter were analyzed using Fiji (125). Only cells with a clearly distinguishable cellular membrane, as shown in Figure 4.2A, were selected for diameter measurement. Since cells displayed a variable morphology from ovoid to spherical, minor and major axes were measured and averaged to obtain a single representative cell diameter for each cell.

4.7.7 Measurement of buoyant cell density

M. florum buoyant cell density was assessed by discontinuous density gradient centrifugation in Percoll (GE Healthcare). First, a Stock Isotonic Percoll (SIP) solution was prepared by mixing 9 parts (v/v) of Percoll (GE Healthcare) to 1 part (v/v) of 1.5M NaCl. The 100% (v/v) SIP solution was then diluted with 0.15M NaCl to obtain 80%, 60%, 40%, and 20% (v/v) SIP solutions. Trypan blue was added to half of the dilutions (100%, 60%, and 20%) to a final concentration of 0.0008%. 2 ml of each dilution was then slowly layered from most concentrated to less into a 15 ml conical tube to create a discontinuous density gradient varying from 1.12 to 1.03 g/ml. 20 ml of a log-phase *M. florum* culture was centrifuged at 10°C for 15 min at 7,900 x *g*, and then washed twice with cold PBS1X. Cells were resuspended in 2 ml of NaCl 0.15M, and slowly loaded on the top of the density gradient. Cells were then centrifuged two times at 7,900 x *g* (10°C) for 30 min each, and the position of the cell pellet was noted after each centrifugation.

4.7.8 Dry mass quantification

Dry mass quantification of *M. florum* was performed in quadruplicate and repeated three times using 20 ml log-phase cultures. A summary of the procedure is shown in Figure S4.4. Briefly, cultures were centrifuged at 10°C for 15 min at 7,900 x *g*, washed twice with cold PBS1X, and then transferred into 1.7 ml microtubes pre-weighted using a Sartorius ME235P analytical scale. Microtubes containing cells were centrifuged at 10°C for 2 min at 21,100 x *g* and cell pellets were resuspended in PBS1X. Resuspended cells were then serial diluted in triplicate with PBS1X in a 96-well microplate and cell concentration was measured by flow cytometry as described previously in the Growth kinetics assays section of Materials and Methods. Undiluted cell suspensions were then centrifuged at 10°C for 2 min at 21,100 x *g*, supernatants were removed, and cell pellets were baked at 80°C for ~36 hrs. Dried cell pellets were then weighted using a Sartorius ME235P analytical scale. The *M. florum* dry mass per cell was determined by

dividing the mass of the dried cell pellet by the total number of cells present in the sample measured by flow cytometry.

4.7.9 Protein mass quantification

Protein mass quantification of *M. florum* was performed in quadruplicate by fluorescence-based protein quantification of whole-cell lysates (see Figure S4.4). Briefly, whole-cell lysates were prepared by centrifuging log-phase *M. florum* cultures at 10°C for 15 min at 7,900 x *g*. Cells were washed twice with cold PBS1X, and then resuspended in PBS2X. CFU counts were measured in triplicate by spotting serial dilutions of the samples on ATCC 1161 solid medium and counting colonies after an incubation of 24-48 hrs at 34°C. Sodium deoxycholate was then added to the cell suspensions to obtain a final concentration of 0.4% (w/v) in PBS1X, and cells were lysed using a Bioruptor UCD-200 sonication system (Diagenode) set at high intensity and 4°C for 35 cycles (30 sec on, 30 sec off). Protein concentration was measured using the CBQCA Protein Quantitation Kit (Molecular Probes, C-6667) according to the manufacturer's specifications. Fluorescence was measured using a Synergy HT microplate reader (BioTek) with the 485/20 and 528/20 nm excitation and emission filters, respectively. The total mass of protein per cell was determined by dividing the protein concentration of the sample by the cell concentration measured by CFU counts.

4.7.10 DNA mass quantification

DNA mass quantification of *M. florum* was performed in quadruplicate by fluorescence-based nucleic acid quantification of purified genomic DNA (gDNA). A summary of the procedure is shown in Figure S4.4. gDNA was extracted from log-phase *M. florum* cultures using the Zymo Quick-DNA MiniPrep Kit (Zymo Research, D3025) according to the manufacturer's specifications, excepted that cells were sonicated in genomic lysis buffer using a Bioruptor UCD-200 sonication system (Diagenode) set at medium intensity and 4°C for 5 cycles (30 sec on, 30 sec off) prior to the column purification step. A purification control consisting of

previously purified *M. florum* gDNA of known concentration (measured using Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher Scientific, P7589)) was also performed in quadruplicate to evaluate purification efficiency. The DNA concentration of purified gDNA samples and controls was then measured by fluorescence-based quantification using the Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher Scientific, P7589). Fluorescence was measured using a Synergy HT microplate reader (BioTek) with the 485/20 and 528/20 nm excitation and emission filters, respectively. The total mass of DNA per cell was determined by first normalizing the concentration of the purified *M. florum* gDNA by the purification efficiency, and then by dividing the normalized DNA concentration by the initial culture cell concentration measured in triplicate by spotting serial dilutions on ATCC 1161 solid medium and counting colonies after an incubation of 24-48 hrs at 34°C.

4.7.11 RNA mass quantification

RNA mass per *M. florum* cell was quantified in quadruplicate as described in the DNA mass quantification section of Materials and Methods (see above and Fig. S4.4), with the exception that cells were sonicated in QIAzol (QIAGEN) reagent, RNA was purified and treated with DNase I using the Direct-zol RNA MiniPrep Kit (Zymo Research, R2052), and that RNA was quantified using Quant-iT RiboGreen RNA Assay Kit (Thermo Fisher Scientific, R11490) according to the manufacturer's specifications.

4.7.12 Carbohydrate mass quantification and monosaccharide composition analysis

The monosaccharide composition and mass quantification of *M. florum* carbohydrates was determined in quadruplicate by GC-MS performed on whole-cell lysates (see Fig. S4.4). Briefly, log-phase *M. florum* cultures were centrifuged at 10°C for 2 min at 21,100 x *g*, and then washed twice with cold PBS1X. Cells were centrifuged again, resuspended in molecular grade water, and CFUs were evaluated in triplicate by spotting serial dilutions on ATCC 1161 solid medium and counting colonies after a 24-48 hrs incubation at 34°C (in triplicates). Resuspended cells

were then lysed using a Bioruptor UCD-200 sonication system (Diagenode) set at high intensity and 4°C for 35 cycles (30 sec on, 30 sec off). Protein concentration of whole cell lysates was calculated by multiplying the number of CFUs present in the cell resuspension before the lysis step by the total protein mass per cell evaluated previously (see Protein quantification section of Materials and Methods). Whole cell lysates were then dried by vacuum centrifugation, resuspended in 400 µl of 1.45 N methanolic HCl, and treated at 80°C overnight to generate the methyl glycosides. The methanolic HCl was removed by vacuum centrifugation, and samples were resuspended in 200 µl of methanol, followed by the addition of 25 µl of acetic anhydride and 25 µl of pyridine. The mixture was allowed to react for 30 min at RT and then evaporated under vacuum centrifugation. Samples were sealed under argon and then trimethylsilylated using 50 µl of Tri-Sil (Fisher). *M. florum* whole cell lysates were finally analyzed using a Varian GC-MS in the electron ionization mode. The monosaccharide composition and concentration were determined by comparison with known standards ran as a standard curve (Sigma-Aldrich), and then normalized using the protein concentration of the analyzed samples.

4.7.13 Lipid mass quantification

Lipid mass quantification of *M. florum* was performed in quadruplicate by fluorescence-based phospholipid quantification of whole-cell lysates. A summary of the procedure is shown in Figure S4.4. Whole-cell lysates were prepared as described in the Protein mass quantification section of Materials and Methods (see above). Samples phospholipid concentration (molarity) was measured based on choline quantification using the Phospholipid Assay Kit (Sigma-Aldrich, MAK122) according to the manufacturer's specifications. Fluorescence was measured using a Synergy HT microplate reader (BioTek) with the 530/25 and 590/35 nm excitation and emission filters, respectively. The number of moles of choline-positive lipids per *M. florum* cell was calculated by dividing the measured concentration of whole-cell extracts by the cell concentration evaluated by CFU counts. The total mass of lipids per cell was then inferred based on the lipidomic profile of *M. florum* (see Dataset S4.8) determined by direct infusion-tandem mass spectrometry (see below). Briefly, identified lipid species were categorized as either

choline-positive or choline-negative species (126), and the average molecular weight of each category was calculated from the relative abundance and theoretical molecular weight of each included species. The quantity of moles of choline-negative lipids was then calculated according to the abundance fraction of each category (~47% and ~53%, respectively), and the total mass per cell of choline-positive and choline-negative lipids was calculated by multiplying the number of moles of each category by their respective average molecular weight. The total lipid mass per *M. florum* cell was finally given by adding up the mass per cell of both lipid categories.

4.7.14 Lipid mass spectrometry

The lipid composition of *M. florum* was determined by direct infusion-tandem mass spectrometry (DI-MS/MS). Sample preparation and analysis was executed by PhenoSwitch Bioscience (Sherbrooke, Canada). Briefly, a log-phase *M. florum* culture was centrifuged at 10°C for 2 min at 21,100 x *g* and washed three times with cold electroporation buffer (272 mM sucrose, 1 mM HEPES [pH 7.4]). Cells were centrifuged again, the supernatant was discarded, and lipids were extracted from the cell pellet by liquid-liquid extraction. Cells were resuspended in 640 µl of ethanol, vortexed for 10 min, and 320 µl of chloroform was added (ethanol/chloroform 2:1 [v/v]). The mixture was vortexed again for 10 min and the insoluble material was removed by centrifugation. The supernatant was transferred into a new microtube, 400 µl of water was added, and the mixture was vortexed for 10 min. Phases were separated by centrifugation and the bottom phase was transferred into a new microtube and washed with 500 µl of chloroform/methanol/water 3:48:47 (v/v/v). The washed bottom phase was then dried and reconstituted in a 1:1 dichloromethane/methanol solution containing 2 mM ammonium acetate, diluted 10 fold, and analyzed on a TripleTOF 5600 mass spectrometer (SCIEX) by direct sample infusion (25 µl) in the mobile phase (1:1 dichloromethane/methanol, 2 mM ammonium acetate). Lipids were analyzed in positive and negative modes using a MS/MS all method (1 m/z windows). Lipid species were identified using LipidView version 1.2 (SCIEX). Only species belonging to the confirmed and common lipid group with an abundance of at least 5% relatively

to the most abundant identified species were considered significant and used in the determination of the total lipid mass per cell (see Dataset S4.8).

4.7.15 Protein mass spectrometry

The protein composition of *M. florum* was determined by 2D LC-MS/MS performed on whole-cell lysate. Sample preparation and analysis was executed by PhenoSwitch Bioscience (Sherbrooke, Canada). Briefly, a log-phase *M. florum* culture was centrifuged at 10°C for 2 min at 21,100 x *g* and washed twice with cold electroporation buffer (272 mM sucrose, 1 mM HEPES [pH 7.4]). Cells were then resuspended in 0.4% (w/v) sodium deoxycholate and lysed using a Bioruptor UCD-200 sonication system (Diagenode) set at high intensity and 4°C for 35 cycles (30 sec on, 30 sec off). Insoluble material was removed by centrifuging the cell lysate at 16,000 x *g* for 10 min at 4°C and the supernatant was recovered. Protein concentration was measured using the Bio-Rad Protein Assay (Bio-Rad) according to the manufacturer's specifications and absorbance at 595 nm was measured using a Synergy HT microplate reader (BioTek). The cell lysate was then reduced at 65°C for 15 min with 10 mM dithiothreitol (DTT) in a final pH of 8.0, and then alkylated at RT in the dark for 30 min with 15 mM iodoacetamide. 10 mM of DTT was then added to quench residual iodoacetamide and proteins (~200 µg) were digested at 37°C overnight with shaking using 1 µg of trypsin per 30 µg of proteins. The resulting peptides were first separated using a polymeric reversed phase column (Phenomenex, 8E-S100-AGB) and eluted into eight fractions with increasing concentration of acetonitrile. ~ 5 µg of each fraction was then injected into a TripleTOF 5600 mass spectrometer (SCIEX) equipped with a HALO ES-C18 column (0.5 x 150 mm). Peptides were separated with a 60 min gradient of the following two mobile phases: A) 0.2 % (v/v) formic acid and 3% (v/v) DMSO in water; B) 0.2 % (v/v) formic acid and 3% (v/v) DMSO in ethanol. Peptides were analyzed in information dependant acquisition (IDA) mode. Raw MS files were analyzed using PeptideShaker software version 1.13.4 (127) configured to run three different search engines (MS-GF+, Comet, and OMSSA) via SearchGUI (version 3.1.0) (128). SearchGUI parameters were set as follows: maximum precursor charge, 5; maximum number of post-translational

modification per peptide, 4; precursor ion m/z tolerance, 0.006 Da; fragment ion m/z tolerance, 0.1 Da; maximum missed cleavages, 2; minimal peptide length, 8; maximal peptide length, 30. Carbamidomethylation of C was set as a fixed modification. Acetylation of K, Acetylation of protein N-term, FormylMet of protein N-term, Oxidation of M, Phosphorylation of S, Phosphorylation of T, and Phosphorylation of Y were set as variable modifications. Protein search database was defined according to the published *M. florum* L1 RAST genome annotation (43). Peptide spectrum matches, peptides and proteins were validated using a 1% FDR cut-off.

4.7.16 Cell mass equations

For simplification purposes we assumed *M. florum* cells to be of spherical form in all cell mass equations described in this study since the exact morphology is expected to vary from ovoid to spherical due to the lack of a cell wall. Therefore, given a spherical *M. florum* cell with a certain diameter (d), its cell mass (CM) can be described as the product of its volume (V) and its buoyant density (D):

$$CM = V \times D \tag{1}$$

Since the volume of a sphere (V) with a certain diameter (d) is given by the following equation:

$$V = \frac{\pi d^3}{6} \tag{1.1}$$

The cell mass (CM) of *M. florum* can thus be described as follows:

$$CM = \frac{\pi d^3}{6} \times D \tag{1.2}$$

Alternatively, the mass of a cell (CM) can also be expressed as the ratio of its dry mass (DM) and its dry mass fraction (DF), the latter given by subtracting the water mass fraction (WF) of a cell from its total mass fraction, i.e. 1:

$$CM = \frac{DM}{DF} \tag{2}$$

or

$$CM = \frac{DM}{1 - WF} \tag{2.1}$$

If we separate the dry mass (DM) of a spherical cell from its water content, then the cell mass (CM) can be written as the cell volume (V) minus the volume occupied by its dry mass (V_{DM}), to which we multiply the density of water (approximated to 1.00 g/ml) and finally add the said dry mass (DM):

$$CM = (V - V_{DM}) \times 1 + DM \tag{3}$$

Since the dry mass volume (V_{DM}) can be particularly difficult to measure, this variable can be substituted by the ratio of the dry mass (DM) and its specific density (D_{DM}), which gives the following equation:

$$CM = \left(V - \frac{DM}{D_{DM}}\right) \times 1 + DM \tag{3.1}$$

Or, if we develop the cell volume (V) as given by equation 1.1:

$$CM = \left(\frac{\pi d^3}{6} - \frac{DM}{D_{DM}}\right) \times 1 + DM \tag{3.2}$$

Conversely, if we replace the cell dry mass (DM) in equation 2 by the product of its volume (V_{DM}) and its specific density (D_{DM}), we obtain:

$$CM = \frac{D_{DM} \times V_{DM}}{DF} \tag{2.2}$$

From this formula, the dry mass volume (V_{DM}) can be isolated and substituted in equation 3:

$$V_{DM} = \frac{CM \times DF}{D_{DM}} \tag{2.3}$$

and

$$CM = \left(V - \frac{CM \times DF}{D_{DM}}\right) \times 1 + DM \tag{4}$$

Finally, we can substitute one of the cell mass (CM) of equation 4 by the cell mass expression of equation 1.2 and develop the cell volume (V) as in equation 1.1, which generates a formula unifying the *M. florum* cell diameter (d), buoyant density (D), dry mass fraction (DF), total dry mass (DM), as well as the dry mass specific density (D_{DM}):

$$CM = \left(\frac{\pi d^3}{6} - \frac{\frac{\pi d^3}{6} \times D \times DF}{D_{DM}} \right) \times 1 + DM \qquad (4.1)$$

The *M. florum* cell mass (CM) was then calculated using equations 1.2, 2, 3.2, 4.1 from the measured *M. florum* dry mass and buoyant density, assuming an estimated dry mass fraction between 20 – 30% (56-59), an estimated dry mass between 1.3 – 1.5 g/ml (59, 60), and a variable cell diameter. For each equation, the mean cell mass (CM_{mean}) was defined as the mass obtained by using the mean value associated to each measured or estimated parameter of the equation. The range of probable cell mass values was defined as the minimal (CM_{min}) and maximal (CM_{max}) masses calculable from a given equation using the mean ± SD or the range associated to each parameter included in the equation. For example, using equation 1.2 and considering a cell buoyant density between 1.05 and 1.08 g/ml, the range of probable cell mass values is given by the following expressions:

$$CM_{min} = \frac{\pi d^3}{6} \times 1.05 \qquad (1.2.1)$$

$$CM_{max} = \frac{\pi d^3}{6} \times 1.08 \qquad (1.2.2)$$

And the mean cell mass value is defined as follows:

$$CM_{mean} = \frac{\pi d^3}{6} \times 1.065 \qquad (1.2.3)$$

The range of cell mass and cell diameter values given by the interception points of CM_{mean} curves encompassing all other interception points were defined as the most probable *M. florum* cell mass and cell diameter ranges. The most probable cell diameter range was finally used to infer the most probable cell volume (V) using equation 1.1, as well as the most probable surface

area (A) and surface area to volume ratio (SA:V) ranges using the following equations, respectively:

$$A = \pi d^2 \quad (5)$$

$$SA{:}V = \frac{A}{V} \quad (5.1)$$

4.7.17 5'-RACE library preparation and analysis

The 5'-RACE sequencing library was prepared on a *M. florum* log-phase culture as previously described (62, 63, 129). Library quality and concentration were evaluated using a 2100 Bioanalyzer instrument (Agilent Technologies). Single-end Illumina sequencing (40 bp) was performed on a Illumina Genome Analyzer IIx at the BioMicroCenter of the Massachusetts Institute of Technology (Cambridge, USA). Reads were quality trimmed using Trimmomatic version 0.32 (130) and aligned on *M. florum* L1 genome (NC_006055.1) with Bowtie 2 version 2.3.3.1 (131). A summary of the 5'-RACE library statistics is shown in Table S4.1. Reads with a MAPQ below 10 were discarded using samtools version 1.5 (132), and the remaining reads were clipped to keep only a single base at the 5' extremity, corresponding to putative start of transcripts. The strand-specific coverage at each genomic position was calculated and normalized according to the number of millions of mapped reads using Bedtools genomecov version 2.27.1 (133), resulting in RSPM values. 5'-RACE peaks with a RSPM signal equal or higher than the average plus one standard deviation single base signal calculated over the entire genome (>=10.92, obtained using 1 kb windows sliding over 100 bp) were considered significant and kept for further analysis. Significant peaks located at 10 bp or less of each other were merged to retain only the peak with the highest associated RSPM signal, corresponding to a putative TSS. Promoter motifs were searched by extracting the DNA sequence surrounding each putative TSS (-45 to +5 bp relative coordinates) and submitting it to MEME version 5.0.3 (65). The motif was generated using the zero or one motif per sequence option. The presence of motifs nearby putative TSSs was further analyzed using MAST version 5.0.3 (66). Only MAST hits separated by 3 to 9 bp from a putative TSS were kept. To circumvent the misalignment of

reads at the chromosome start position, the 5'-RACE reads were realigned on the L1 chromosome sequence linearized at position 397,159 instead of 0, and the whole analysis procedure was repeated. This allowed us to identify an additional TSS located in the intergenic region upstream the *dnaA* gene (peg.1/*mfl001*). Strand-specific 1 bp resolution genome coverage tracks were generated using Bedtools genomecov version 2.27.1 (133).

4.7.18 RNA-seq libraries preparation and analysis

Total RNA-seq libraries were prepared in biological triplicate from *M. florum* steady-state cultures grown using the Versatile Continuous Culture Device (39). Total RNA was extracted in technical duplicate from each culture replicate using the Direct-zol RNA MiniPrep Kit (Zymo Research, R2052) as described previously (63), for a total of six RNA-seq libraries. RNA-seq libraries were prepared and depleted from ribosomal RNA as described previously (63), excepted that 200 µg/ml of actinomycin D was added to the reverse transcription reaction. Library quality and concentration were evaluated using a 2100 Bioanalyzer instrument (Agilent Technologies). Paired-end Illumina sequencing (2x50 bp) was performed on a HiSeq 2000 Illumina instrument at the Plateau de biologie moléculaire et génomique fonctionnelle of the Institut de recherches cliniques de Montréal (Montréal, Québec, Canada). Reads were quality trimmed using Trimmomatic version 0.32 (130) and aligned in a strand-specific manner on the *M. florum* L1 genome (NC_006055.1) with Bowtie 2 version 2.3.3.1 (131). Reads with a MAPQ below 10 were discarded using samtools version 1.5 (132). A summary of the RNA-seq library statistics is shown in Table S4.1. FPKM intensity was calculated for each *M. florum* L1 annotated gene (RAST annotation, see (43)) using the GenomicAlignments R package version 1.10.1 (134). Strand-specific 1 bp resolution genome coverage tracks were generated using Bedtools genomecov version 2.27.1 (133). Bedtools makewindows and multicov (version 2.27.1) were used to calculate the RNA-seq coverage on non-overlapping 1 kb windows for each replicate. Pearson's correlation coefficients between replicates (1 kb windows coverage as well as gene FPKM) were calculated using GraphPad Prism integrated function (version 7.0a).

4.7.19 Reconstruction of transcription units

Rho-independent terminators were predicted from *M. florum* L1 DNA sequence and genes annotation (RAST annotation, see (43)) using an updated version of the in-house Python script developed by de Hoon et al. (72). The main difference between the updated version and the original one is the replacement of the Mfold package (135, 136) by the ViennaRNA package (137) (version 2.4.11) to calculate the RNA secondary structure. The Python script is available upon request from the author. Only terminators with a calculated score above 0 were considered significant. For each predicted terminator, the TTS was defined as the last base forming the stem-loop structure since the termination was shown to occur at or near the T-stretch following the stem-loop (72, 138). Strand-specific term-to-term scaffolds were then created according to the genomic position of the TTSs, and the coordinates of the motif-associated TSSs were used to generate all possible TUs for each scaffold. Genes were attributed to a given TU only if the calculated (5'-UTR) length was =< 500 bp and their coordinates were completely included within the TU, meaning that genes intersected with iTSSs were excluded from the iTSSs derived TUs. Generated TUs were manually inspected using the UCSC genome browser (139) to correct for different scenarios such as the presence of predicted riboswitches (140, 141), the circular topology of the chromosome or the occasional overlap between TSSs and terminator sequences. In the rare cases where no motif-associated TSSs could be attributed to a gene (orphan gene), the identified TSSs without promoter motif were considered for the expression of a TU encompassing that gene, as long as they initiated transcription on a purine nucleotide and fulfilled the other criteria described previously (signal intensity threshold and 5'-UTR length). If still no TSS without promoter motif could be found, then TSSs located at the end of the previous term-to-term scaffold (thus separated from the orphan gene by a predicted terminator) were considered as putative candidates for the expression of the gene, provided that its expression was non-null and the 5'-UTR length was =< 500 bp. See the manual curation notes column in Dataset S4.3 for further details.

4.7.20 Aggregate profiles

RNA-seq aggregate profiles were generated using the Versatile Aggregate Profiler (VAP) version 1.0.0 (142). Aggregate profiles were calculated for each DNA strand independently using the RNA-seq genome coverage calculated at single bp resolution on all the RNA-seq replicates merged together. The relative analysis method was used for all cases, along with two reference points and a 1 bp window size. The number of windows for the reference feature was set to 1 in the case of TSSs, whereas this parameter was set to 100 and 40 for TUs and terminators, respectively.

4.7.21 Determination of Shine-Dalgarno consensus sequence

Shine-Dalgarno consensus sequence was searched by extracting the DNA sequence (20 bp) upstream the translation initiation codon of each *M. florum* reference ORF and submitting it to MEME version 5.0.3 (65). The zero or one motif per sequence option was used.

4.7.22 Estimation of molecular abundances

The number of *M. florum* chromosome copies per cell was estimated from the measured DNA mass and the estimated molecular weight of the chromosome (see Table 4.1 for the measured DNA, RNA and protein). The molecular weight of the *M. florum* L1 chromosome (NC_006055.1) was estimated using the Sequence Manipulation Suite server (https://www.bioinformatics.org/sms2/dna_mw.html) (143). The intracellular abundance of RNA species was calculated from the estimated molecular weight and measured RNA mass by assuming that rRNA, tRNA, and mRNA totalize 80%, 15%, and 5% of the total RNA mass of the cell (59, 74). The molecular weight of RNA species was estimated using in-house Python scripts. The intracellular levels of protein species were calculated from the estimated molecular weight and the measured protein mass. The molecular weight of proteins was either estimated by PeptideShaker software version 1.13.4 (127) for proteins detected by mass spectrometry or

using the Sequence Manipulation Suite server (https://www.bioinformatics.org/sms2/protein_mw.html) for proteins not detected by mass spectrometry (143). For rRNAs and tRNAs, the total number of copies per cell was calculated by assuming that each species is found at equimolar ratios. For mRNAs and proteins, molar ratios were normalized according to the expression value of each species, i.e. using the associated FPKM and NSAF values, respectively. Calculation details can be found in Datasets S4.5 and S4.6. The number of ribosomes per cell was estimated using two different approaches: 1) From the average number of protein per cell calculated for all predicted (KO) ribosomal proteins. 2) From the assumption that all rRNA molecules are incorporated into ribosomes, meaning that the estimated number of ribosomes per cell is equivalent to one third of the total number of rRNA molecules per cell (three rRNA molecules per ribosome). The number of RNAP complexes per cell was estimated according to the average number of protein per cell calculated for the α, β, and β' subunits (see Dataset S4.6). The protein stichometry of the RNAP complex was taken into account in the calculations (two α, one β, and one β' subunits per RNAP).

4.7.23 Analysis of functional categories expression

The KO Database was used to assign functional categories to *M. florum* reference proteins because of its clearly layered structure, and because major efforts were made to associate each KO entry with experimental evidences (79). Moreover, since proteins are assigned to functions via KO identifiers, the comparison between organisms is relatively straightforward. Briefly, the BlastKOALA server (https://www.kegg.jp/blastkoala/) (144) was used to assign KO identifiers to *M. florum* reference ORFs and retrieve associated functional categories. Since the same protein can be assigned to multiple functional categories, we then curated the assigned categories based on the non-redundant Proteomap functional hierarchy (80). ORFs not matching to any KO identifiers were assigned to the unmapped category. KO entries matching to unclear functions were regrouped into the unclear category. Assigned KO identifiers and functional categories can be found in Dataset S4.8. The Proteomap server

(https://www.proteomaps.net/index.html) was used to visualize the relative expression of functional categories using either protein or mRNA expression datasets (80).

4.7.24 Data availability

5'-RACE, RNA-seq, and mass spectrometry data files are available upon request. 5'-RACE and RNA-seq Fastq files will be deposited at the NCBI Sequence Read Archive upon acceptance of the manuscript. Protein mass spectrometry data will also be deposited at the Mass Spectrometry Interactive Virtual Environment (MassIVE) Repository (member of the ProteomeXchange consortium) after manuscript acceptance. A UCSC genome browser hub containing the complete genomic information reported in this study can be provided upon request and will be made publicly available for publication.

4.8 Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada and the Fonds de Recherche du Québec Nature et technologies.

4.9 Acknowledgements

We thank members of the Rodrigue and Jacques laboratories for helpful discussions, Joëlle Brodeur and Alain Lavigueur for critical reading of the manuscript, as well as Jean-François Lucier and the Centre de Calcul Scientifique at the Université de Sherbrooke for technical assistance. We are also grateful to Charles Bertrand and the Plateforme de microscopie photonique at the Université de Sherbrooke for technical assistance on TEM and STED microscopy, respectively. Access to computational resources was provided in part by Calcul Québec (http://www.calculquebec.ca) and Compute Canada (http://www.computecanada.ca).

4.10 Author contributions

D.M. wrote the manuscript and prepared the figures; S.R. edited the manuscript and the figures; D.M., D.G., S.G., J.M.D. and K.D. performed the experiments; D.M., F.G., and J.-C.L. performed the analyses; D.M., J.-C.L., and S.R. designed the project.

4.11 Conflict of interest

The authors declare no conflict of interest.

4.12 References

1. Ebrahim,A., Brunk,E., Tan,J., O'Brien,E.J., Kim,D., Szubin,R., Lerman,J.A., Lechner,A., Sastry,A., Bordbar,A., *et al.* (2016) Multi-omic data integration enables discovery of hidden biological regularities. *Nat. Commun.*, **7**, 13091.
2. Kim,M., Rai,N., Zorraquino,V. and Tagkopoulos,I. (2016) Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli. *Nat. Commun.*, **7**, 1–12.
3. Gu,C., Kim,G.B., Kim,W.J., Kim,H.U. and Lee,S.Y. (2019) Current status and applications of genome-scale metabolic models. *Genome Biol.*, **20**, 1–18.
4. Bordbar,A., Monk,J.M., King,Z.A. and Palsson,B.O. (2014) Constraint-based models predict metabolic and associated cellular functions. *Nat. Rev. Genet.*, **15**, 107–120.
5. O'Brien,E.J., Monk,J.M. and Palsson,B.O. (2015) Using genome-scale models to predict biological capabilities. *Cell*, **161**, 971–987.
6. Oberhardt,M.A., Palsson,B. and Papin,J.A. (2009) Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.*, **5**, 1–15.
7. Durot,M., Bourguignon,P.Y. and Schachter,V. (2009) Genome-scale models of bacterial metabolism: Reconstruction and applications. *FEMS Microbiol. Rev.*, **33**, 164–190.
8. Orth, Jeffrey D., Ines Thiele,B.Ø.P. (2010) What is flux balance analysis? *Nat. Biotechnol.*, **28**, 245–248.
9. Lachance,J.C., Lloyd,C.J., Monk,J.M., Yang,L., Sastry,A. V., Seif,Y., Palsson,B.O., Rodrigue,S., Feist,A.M., King,Z.A., *et al.* (2019) BOFdat: Generating biomass objective functions for genome-scale metabolic models from experimental data. *PLoS Comput. Biol.*, **15**, e1006971.

10. Feist,A.M. and Palsson,B.O. (2010) The biomass objective function. *Curr. Opin. Microbiol.*, **13**, 344–349.
11. King,Z.A., Lloyd,C.J., Feist,A.M. and Palsson,B.O. (2015) Next-generation genome-scale models for metabolic engineering. *Curr. Opin. Biotechnol.*, **35**, 23–29.
12. O'Brien,E.J. and Palsson,B.O. (2015) Computing the functional proteome: Recent progress and future prospects for genome-scale models. *Curr. Opin. Biotechnol.*, **34**, 125–134.
13. Lloyd,C.J., Ebrahim,A., Yang,L., King,Z.A., Catoiu,E., O'Brien,E.J., Liu,J.K. and Palsson,B.O. (2018) COBRAme: A computational framework for genome-scale models of metabolism and gene expression. *PLoS Comput. Biol.*, **14**, 1–14.
14. O'Brien,E.J., Lerman,J.A., Chang,R.L., Hyduke,D.R. and Palsson,B. (2013) Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.*, **9**.
15. Lerman,J.A., Hyduke,D.R., Latif,H., Portnoy,V.A., Lewis,N.E., Orth,J.D., Schrimpe-Rutledge,A.C., Smith,R.D., Adkins,J.N., Zengler,K., *et al.* (2012) In silico method for modelling metabolism and gene product expression at genome scale. *Nat. Commun.*, **3**.
16. Thiele,I., Fleming,R.M.T., Que,R., Bordbar,A., Diep,D. and Palsson,B.O. (2012) Multiscale Modeling of Metabolism and Macromolecular Synthesis in E. coli and Its Application to the Evolution of Codon Usage. *PLoS One*, **7**.
17. Montague,M.G., Lartigue,C. and Vashee,S. (2012) Synthetic genomics: Potential and limitations. *Curr. Opin. Biotechnol.*, **23**, 659–665.
18. Schindler,D., Dai,J. and Cai,Y. (2018) Synthetic genomics: a new venture to dissect genome fundamentals and engineer new functions. *Curr. Opin. Chem. Biol.*, **46**, 56–62.
19. van der Sloot,A. and Tyers,M. (2017) Synthetic Genomics: Rewriting the Genome Chromosome by Chromosome. *Mol. Cell*, **66**, 441–443.
20. Mitchell,L.A. and Ellis,T. (2017) Synthetic genome engineering gets infectious. *Proc. Natl. Acad. Sci. U. S. A.*, **114**, 11006–11008.
21. Khalil,A.S. and Collins,J.J. (2010) Synthetic biology: applications come of age. *Nat. Rev. Genet.*, **11**, 367–79.
22. Alper,H., Cirino,P., Nevoigt,E. and Sriram,G. (2010) Applications of synthetic biology in microbial biotechnology. *J. Biomed. Biotechnol.*, **2010**, 1–2.
23. Cambray,G., Mutalik,V.K. and Arkin,A.P. (2011) Toward rational design of bacterial genomes. *Curr. Opin. Microbiol.*
24. Hutchison,C.A., Chuang,R.Y., Noskov,V.N., Assad-Garcia,N., Deerinck,T.J., Ellisman,M.H., Gill,J., Kannan,K., Karas,B.J., Ma,L., *et al.* (2016) Design and synthesis of a minimal bacterial genome. *Science (80-.).*, **351**, aad6253-1-aad6253-11.
25. Richardson,S.M., Mitchell,L.A., Stracquadanio,G., Yang,K., Dymond,J.S., DiCarlo,J.E., Lee,D., Huang,C.L.V., Chandrasegaran,S., Cai,Y., *et al.* (2017) Design of a synthetic yeast genome. *Science (80-.).*, **355**, 1040–1044.

26. Fredens,J., Wang,K., de la Torre,D., Funke,L.F.H., Robertson,W.E., Christova,Y., Chia,T., Schmied,W.H., Dunkelmann,D.L., Beránek,V., *et al.* (2019) Total synthesis of Escherichia coli with a recoded genome. *Nature*, 10.1038/s41586-019-1192-5.

27. Xavier,J.C., Patil,K.R. and Rocha,I. (2014) Systems Biology Perspectives on Minimal and Simpler Cells. *Microbiol. Mol. Biol. Rev.*, **78**, 487–509.

28. Lachance,J.-C., Rodrigue,S. and Palsson,B.O. (2019) Synthetic Biology: Minimal cells, maximal knowledge. *Elife*, **8**, 1–4.

29. Sirand-Pugnet,P., Citti,C., Barré,A. and Blanchard,A. (2007) Evolution of mollicutes: down a bumpy road with twists and turns. *Res. Microbiol.*, **158**, 754–66.

30. Morowitz,H.J. The completeness of molecular biology. *Isr. J. Med. Sci.*, **20**, 750–753.

31. Pettersson,B. and Johansson,K.-E. (2002) Taxonomy of Mollicutes. In Razin,S., Herrmann,R. (eds), *Molecular Biology and pathogenicity of Mycoplasmas*. Springer, New York (USA), pp. 1–30.

32. Maniloff,J. (2002) Phylogeny and evolution. In *Molecular Biology and pathogenicity of Mycoplasmas*.pp. 31–43.

33. Dybvig,K. and Voelker,L.L. (1996) Molecular biology of Mycoplasmas. *Annu. Rev. Microbiol.*

34. Pollack,J.D., Williams,M. V. and McElhaney,R.N. (1997) The comparative metabolism of the mollicutes (Mycoplasmas): The utility for taxonomic classification and the relationship of putative gene annotation and phylogeny to enzymatic function in the smallest free-living cells. *Crit. Rev. Microbiol.*, **23**, 269–354.

35. Gibson,D.G., Glass,J.I., Lartigue,C., Noskov,V.N., Chuang,R.-Y., Algire,M.A., Benders,G.A., Montague,M.G., Ma,L., Moodie,M.M., *et al.* (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. *Science*, **329**, 52–6.

36. Glass,J.I., Merryman,C., Wise,K.S., Iii,C.A.H. and Smith,H.O. (2017) Minimal Cells — Real and Imagined. *Cold Spring Harb Perspect Biol.*, **1**, 1–12.

37. Breuer,M., Earnest,T.M., Merryman,C., Wise,K.S., Sun,L., Lynott,M.R., Hutchison,C.A., Smith,H.O., Lapek,J.D., Gonzalez,D.J., *et al.* (2019) Essential metabolism for a minimal cell. *Elife*, **8**, 1–77.

38. McCoy,R.E., Basham,H.G., Tully,J.G., Rose,D.L., Carle,P. and Bové,J.M. (1984) Acholeplasma florum, a New Species Isolated from Plants. *Int. J. Syst. Bacteriol.*, **34**, 11–15.

39. Matteau,D., Baby,V., Pelletier,S. and Rodrigue,S. (2015) A Small-Volume, Low-Cost, and Versatile Continuous Culture Device. *PLoS One*, **10**, e0133384.

40. Seto,S. and Miyata,M. (1998) Cell reproduction and morphological changes in Mycoplasma capricolum. *J. Bacteriol.*, **180**, 256–264.

41. Matteau,D., Pepin,M., Baby,V., Gauthier,S., Arango Giraldo,M., Knight,T.F. and Rodrigue,S. (2017) Development of oriC -based plasmids for Mesoplasma florum. *Appl. Environ. Microbiol.*, **83**, 1–16.

42. Baby,V., Labroussaa,F., Brodeur,J., Matteau,D., Gourgues,G., Lartigue,C. and Rodrigue,S. (2017) Cloning and transplantation of the Mesoplasma florum genome. *ACS Synth. Biol.*, 10.1021/acssynbio.7b00279.

43. Baby,V., Lachance,J.-C., Gagnon,J., Lucier,J.-F., Matteau,D., Knight,T.F. and Rodrigue,S. (2018) Inferring the Minimal Genome of Mesoplasma florum by Comparative Genomics and Transposon Mutagenesis. *mSystems*, **3**, e00198-17.

44. Mushegian,A. and Koonin,E. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci.*, **93**, 10268–10273.

45. Brown,D.R., Whitcomb,R.F. and Bradbury,J.M. (2007) Revised minimal standards for description of new species of the class Mollicutes (division Tenericutes). *Int. J. Syst. Evol. Microbiol.*, **57**, 2703–2719.

46. Konai,M., Clark,E.A., Camp,M., Koeh,A.L. and Whitcomb,R.F. (1996) Temperature ranges, growth optima, and growth rates of Spiroplasma (Spiroplasmataceae, class Mollicutes) species. *Curr. Microbiol.*, **32**, 314–319.

47. Tully,J.G., Whitcomb,R.F., Hackett,K.J., Rose,D.L., Henegar,R.B., Bové,J.M., Carle,P., Williamson,D.L. and Clark,T.B. (1994) Taxonomic descriptions of eight new non-sterol-requiring Mollicutes assigned to the genus Mesoplasma. *Int. J. Syst. Bacteriol.*, **44**, 685–93.

48. Revelo,N.H., Kamin,D., Truckenbrodt,S., Wong,A.B., Reuter-Jessen,K., Reisinger,E., Moser,T. and Rizzoli,S.O. (2014) A new probe for super-resolution imaging of membranes elucidates trafficking pathways. *J. Cell Biol.*, **205**, 591–606.

49. Bertin,C., Pau-Roblot,C., Courtois,J., Manso-Silván,L., Thiaucourt,F., Tardy,F., Le Grand,D., Poumarat,F. and Gaurivaud,P. (2013) Characterization of Free Exopolysaccharides Secreted by Mycoplasma mycoides Subsp. mycoides. *PLoS One*, **8**.

50. Gaurivaud,P., Lakhdar,L., Le Grand,D., Poumarat,F. and Tardy,F. (2014) Comparison of in vivo and in vitro properties of capsulated and noncapsulated variants of Mycoplasma mycoides subsp. Mycoides strain Afadé: A potential new insight into the biology of contagious bovine pleuropneumonia. *FEMS Microbiol. Lett.*, **359**, 42–49.

51. Bertin,C., Pau-Roblot,C., Courtois,J., Manso-Silván,L., Tardy,F., Poumarat,F., Citti,C., Sirand-Pugnet,P., Gaurivaud,P. and Thiaucourt,F. (2015) Highly dynamic genomic loci drive the synthesis of two types of capsular or secreted polysaccharides within the Mycoplasma mycoides cluster. *Appl. Environ. Microbiol.*, **81**, 676–687.

52. Daubenspeck,J.M., Jordan,D.S. and Dybvig,K. (2014) The Glycocalyx of Mollicutes. In Browning,G., Citti,C. (eds), *Mollicutes: molecular biology and pathogenesis*. Caister Academic Press, Norfolk, pp. 131–147.

53. Bryan,A.K., Hecht,V.C., Shen,W., Payer,K., Grover,W.H. and Manalis,S.R. (2014) Measuring single cell mass, volume, and density with dual suspended microchannel resonators. *Lab Chip*, **14**, 569–576.

54. Zhao,Y., Lai,H.S.S., Zhang,G., Lee,G. Bin and Li,W.J. (2014) Rapid determination of cell mass and density using digitally controlled electric field in a microfluidic chip. *Lab Chip*,

14, 4426–4434.

55. Rahman,M.H., Ahmad,M.R., Takeuchi,M., Nakajima,M., Hasegawa,Y. and Fukuda,T. (2015) Single Cell Mass Measurement Using Drag Force Inside Lab-on-Chip Microfluidics System. *IEEE Trans. Nanobioscience*, **14**, 927–934.
56. Bakken,L.R. and Olsen,R.A. (1983) Buoyant densities and dry-matter contents of microorganisms: Conversion of a measured biovolume into biomass. *Appl. Environ. Microbiol.*, **45**, 1188–1195.
57. Bratbak,G. and Dundas,I. (1984) Bacterial dry matter content and biomass estimations. *Appl. Environ. Microbiol.*, **48**, 755–757.
58. Bratbak,G. (1985) Biovolume and Biomass Estimations. *Microbiology*, **49**, 1488–1493.
59. Bionumbers (2015) What is the macromolecular composition of the cell.
60. Fischer,H., Polikarpov,I. and Craievich,A.F. (2009) Average protein density is a molecular-weight-dependent function. *Protein Sci.*, **13**, 2825–2828.
61. Cooper,G.M. (2000) The Molecular Composition of Cells. In Sunderland,M.A. (ed), *The Cell: A Molecular Approch*. Sinauer Associates.
62. Matteau,D. and Rodrigue,S. (2015) Precise Identification of Genome-Wide Transcription Start Sites in Bacteria by 5′-Rapid Amplification of cDNA Ends (5′-RACE). In Leblanc,B.P., Rodrigue,S. (eds), *DNA-Protein Interactions SE - 9*, Methods in Molecular Biology. Springer New York, Vol. 1334, pp. 143–159.
63. Carraro,N., Matteau,D., Luo,P., Burrus,V. and Rodrigue,S. (2014) The Master Activator of IncA/C Conjugative Plasmids Stimulates Genomic Islands and Multidrug Resistance Dissemination. *PLoS Genet.*, **10**, e1004714.
64. Shultzaberger,R.K., Chen,Z., Lewis,K.A. and Schneider,T.D. (2007) Anatomy of Escherichia coli σ70 promoters. *Nucleic Acids Res.*, **35**, 771–788.
65. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
66. Bailey,T.L. and Gribskov,M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.
67. Weiner III,J. (2000) Transcription in Mycoplasma pneumoniae. *Nucleic Acids Res.*, **28**, 4488–4496.
68. Zheng,X., Hu,G.Q., She,Z.S. and Zhu,H. (2011) Leaderless genes in bacteria: Clue to the evolution of translation initiation mechanisms in prokaryotes. *BMC Genomics*, **12**.
69. Nakagawa,S., Niimura,Y. and Gojobori,T. (2017) Comparative genomic analysis of translation initiation mechanisms for genes lacking the Shine-Dalgarno sequence in prokaryotes. *Nucleic Acids Res.*, **45**, 3922–3931.
70. Moll,I., Grill,S., Gualerzi,C.O. and Bläsi,U. (2002) Leaderless mRNAs in bacteria: Surprises in ribosomal recruitment and translational control. *Mol. Microbiol.*, **43**, 239–246.
71. D'Heygère,F., Rabhi,M. and Boudvillain,M. (2013) Phyletic distribution and conservation

of the bacterial transcription termination factor rho. *Microbiol. (United Kingdom)*, **159**, 1423–1436.

72. de Hoon,M.J.L., Makita,Y., Nakai,K. and Miyano,S. (2005) Prediction of transcriptional terminators in Bacillus subtilis and related species. *PLoS Comput. Biol.*, **1**, e25.
73. Baby,V., Matteau,D., Knight,T.F. and Rodrigue,S. (2013) Complete genome sequence of the Mesoplasma florum W37 strain. *Genome Annoucements*, **1**, e00879-13.
74. Westermann,A.J., Gorski,S.A. and Vogel,J. (2012) Dual RNA-seq of pathogen and host. *Nat. Rev. Microbiol.*, **10**, 618–630.
75. Greenbaum,D., Colangelo,C., Williams,K. and Gerstein,M. (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.*, **4**, 117.
76. Kuchta,K., Towpik,J., Biernacka,A., Kutner,J., Kudlicki,A., Ginalski,K. and Rowicka,M. (2018) Predicting proteome dynamics using gene expression data. *Sci. Rep.*, **8**, 1–13.
77. Maier,T., Güell,M. and Serrano,L. (2009) Correlation of mRNA and protein in complex biological samples. *FEBS Lett.*, **583**, 3966–3973.
78. Yang,M., Yang,Y., Chen,Z., Zhang,J., Lin,Y., Wang,Y., Xiong,Q., Li,T., Ge,F., Bryant,D.A., *et al.* (2014) Proteogenomic analysis and global discovery of posttranslational modifications in prokaryotes. *Proc. Natl. Acad. Sci.*, **111**, E5633–E5642.
79. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
80. Liebermeister,W., Noor,E., Flamholz,A., Davidi,D., Bernhardt,J. and Milo,R. (2014) Visual account of protein investment in cellular functions. *Proc. Natl. Acad. Sci.*, **111**, 8488–8493.
81. Robert F. Whitcomb, Joseph G. Tully, David L. Rose, Edward B. Stephens, w Alexis Smith,R.E.M. and Barilew, and M.F. (1982) Wall-Less Prokaryotes from Fall Flowers in Central United States and Maryland. *Curr. Microbiol.*, **7**, 285–290.
82. McCoy,R.E., Basham,H.G., Tully,J.G. and Rose,D.L. (1980) Isolation of a new acholeplasma from flowers in Florida. In *Third Conference of the International Organization for Mycoplasmology*.
83. Tully,J., Rose,D., Whitcomb,R.F., Hackett,K.J., Clark,T.B., Henegar,R.B., Clark,E., Carle,P. and Bove,J.M. (1987) Characterization of some new insect-derived acholeplasmas. *Isr. J. Med. Sci.*, **23**, 699–703.
84. Sapountzis,P., Zhukova,M., Shik,J.Z., Schiott,M. and Boomsma,J.J. (2018) Reconstructing the functions of endosymbiotic mollicutes in fungus-growing ants. *Elife*, **7**, 1–31.
85. Funaro,C.F., Kronauer,D.J.C., Moreau,C.S., Goldman-Huertas,B., Pierce,N.E. and Russell,J.A. (2011) Army ants harbor a host-specific clade of Entomoplasmatales bacteria. In *Applied and Environmental Microbiology*.Vol. 77, pp. 346–350.
86. Tully,J.G., Bove,J.M., Laigret,F. and Whitcomb,R.F. (1993) Revised Taxonomy of the Class Mollicutes : Proposed Elevation of a Monophyletic Cluster of Arthropod-Associated

Mollicutes to Ordinal Rank (Entomoplasmatales ord . nov .), with Provision for Familial Rank To Separate Species with Descriptions of the Orde. *Int. J. Syst. Bacteriol.*, **43**, 378–385.

87. Brown,D.R. and Bradbury,J.M. (2014) The Contentious Taxonomy of Mollicutes. In *Mollicutes: Molecular Biology and Pathogenesis*.pp. 1–14.
88. Halbedel,S., Hames,C. and Stülke,J. (2007) Regulation of carbon metabolism in the mollicutes and its relation to virulence. *J. Mol. Microbiol. Biotechnol.*, **12**, 147–54.
89. Caspi,R., Altman,T., Billington,R., Dreher,K., Foerster,H., Fulcher,C. a, Holland,T. a, Keseler,I.M., Kothari,A., Kubo,A., *et al.* (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **42**, D459-71.
90. Caspi,R., Billington,R., Ferrer,L., Foerster,H., Fulcher,C.A., Keseler,I.M., Kothari,A., Krummenacker,M., Latendresse,M., Mueller,L.A., *et al.* (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **44**, D471–D480.
91. Yus,E., Maier,T., Michalodimitrakis,K., van Noort,V., Yamada,T., Chen,W.-H.W.-H., Wodke,J.A.H.J.A.H., Güell,M., Martínez,S., Bourgeois,R., *et al.* (2009) Impact of Genome Reduction on Bacterial Metabolism and Its Regulation. *Science (80-.).*, **326**, 1263–1268.
92. Wodke,J.A.H., Pucha ka,J., Lluch-Senar,M., Marcos,J., Yus,E., Godinho,M., Gutierrez-Gallego,R., dos Santos,V.A.P.M., Serrano,L., Klipp,E., *et al.* (2013) Dissecting the energy metabolism in Mycoplasma pneumoniae through genome-scale metabolic modeling. *Mol. Syst. Biol.*, **9**, 653–653.
93. Jensen,J.S., Hansen,H.T. and Lind,K. (1996) Isolation of Mycoplasma genitalium Strains from the Male Urethra. *J. Clin. Microbiol.*, **34**, 286–291.
94. LaCroix,R.A., Sandberg,T.E., O'Brien,E.J., Utrilla,J., Ebrahim,A., Guzman,G.I., Szubin,R., Palsson,B.O. and Feist,A.M. (2015) Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. *Appl. Environ. Microbiol.*, **81**, 17–30.
95. Zhang,Y., Huang,T., Jorgens,D.M., Nickerson,A., Lin,L.J., Pelz,J., Gray,J.W., López,C.S. and Nan,X. (2017) Quantitating morphological changes in biological samples during scanning electron microscopy sample preparation with correlative super-resolution microscopy. *PLoS One*, **12**, 1–15.
96. Peterson,B.M., Mermelstein,P.G. and Meisel,R.L. (2015) Impact of immersion oils and mounting media on the confocal imaging of dendritic spines. *J. Neurosci. Methods*, **242**, 106–111.
97. Fouquet,C., Gilles,J.F., Heck,N., Santos,M. Dos, Schwartzmann,R., Cannaya,V., Morel,M.P., Davidson,R.S., Trembleau,A. and Bolte,S. (2015) Improving axial resolution in confocal microscopy with new high refractive index mounting media. *PLoS One*, **10**, 1–17.
98. Dai,X. and Zhu,M. (2018) High Osmolarity Modulates Bacterial Cell Size through

Reducing Initiation Volume in Escherichia coli. *mSphere*, **3**, e00430-18.

99. Volkmer,B. and Heinemann,M. (2011) Condition-Dependent cell volume and concentration of Escherichia coli to facilitate data conversion for systems biology modeling. *PLoS One*, **6**, 1–6.
100. Ojkic,N., Serbanescu,D. and Banerjee,S. (2019) Surface-to-volume scaling and aspect ratio preservation in rod-shaped bacteria. *Elife*, **8**, 1–11.
101. Harris,L.K. and Theriot,J.A. (2016) Relative rates of surface and volume synthesis set bacterial cell size. *Cell*, **165**, 1479–1492.
102. Harris,L.K. and Theriot,J.A. (2018) Surface Area to Volume Ratio: A Natural Variable for Bacterial Morphogenesis. *Trends Microbiol.*, **26**, 815–832.
103. Dennis,P.P. and Bremer,H. (1974) Macromolecular composition during steady-state growth of Escherichia coli B-r. *J. Bacteriol.*, **119**, 270–281.
104. Feist,A.M., Henry,C.S., Reed,J.L., Krummenacker,M., Joyce,A.R., Karp,P.D., Broadbelt,L.J., Hatzimanikatis,V. and Palsson,B. (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.*, **3**, 1–18.
105. RAZIN,S., ARGAMAN,M. and AVIGAN,J. (1963) Chemical Composition of Mycoplasma Cells and Membranes. *J. Gen. Microbiol.*, **33**, 477–487.
106. Lloréns-Rico,V., Lluch-Senar,M. and Serrano,L. (2015) Distinguishing between productive and abortive promoters using a random forest classifier in Mycoplasma pneumoniae. *Nucleic Acids Res.*, **43**, 3442–3453.
107. Fisunov,G.Y., Garanina,I.A., Evsyutina,D. V., Semashko,T.A., Nikitina,A.S. and Govorun,V.M. (2016) Reconstruction of transcription control networks in mollicutes by high-throughput identification of promoters. *Front. Microbiol.*, **7**, 1–15.
108. Sabelnikov,A.G., Greenberg,B. and Lacks,S.A. (1995) An Extended -10 Promoter Alone Directs Transcription of the DpnII Operon of Streptococcus pneumoniae. *J. Mol. Biol.*, **250**, 144–155.
109. Voskuil,M.I. and Chambliss,G.H. (1998) The -16 region of Bacillus subtilis and other gram-positive bacterial promoters. *Nucleic Acids Res.*, **26**, 3584–3590.
110. Mutalik,V.K., Guimaraes,J.C., Cambray,G., Lam,C., Christoffersen,M.J., Mai,Q.-A., Tran,A.B., Paull,M., Keasling,J.D., Arkin,A.P., *et al.* (2013) Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods*, **10**, 354–60.
111. Guiziou,S., Sauveplane,V., Chang,H.J., Clerté,C., Declerck,N., Jules,M. and Bonnet,J. (2016) A part toolbox to tune genetic expression in Bacillus subtilis. *Nucleic Acids Res.*, **44**, 7495–7508.
112. Lloréns-Rico,V., Cano,J., Kamminga,T., Gil,R., Latorre,A., Chen,W.-H.H., Bork,P., Glass,J.I., Serrano,L. and Lluch-Senar,M. (2016) Bacterial antisense RNAs are mainly the product of transcriptional noise. *Sci. Adv.*, **2**, 1–10.

113. Nicolas,P., Mäder,U., Dervyn,E., Rochat,T., Leduc,A., Pigeonneau,N., Bidnenko,E., Marchadier,E., Hoebeke,M., Aymerich,S., *et al.* (2012) Condition-Dependent Transcriptome Reveals High-Level Regulatory Architecture in Bacillus subtilis. *Science (80-.).*, **335**, 1103–1106.

114. Wade,J.T. and Grainger,D.C. (2014) Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nat. Rev. Microbiol.*, **12**, 647–653.

115. Raghavan,R., Sloan,D.B. and Ochman,H. (2012) Pervasive transcription is widespread but rarely conserved in Enteric bacteria. *MBio*, **3**, 1–7.

116. Jose,B.R., Gardner,P.P. and Barquist,L. (2019) Transcriptional noise and exaptation as sources for bacterial sRNAs. *Biochem. Soc. Trans.*, **47**, 527–539.

117. Güell,M., van Noort,V., Yus,E., Chen,W.-H., Leigh-Bell,J., Michalodimitrakis,K., Yamada,T., Arumugam,M., Doerks,T., Kühner,S., *et al.* (2009) Transcriptome complexity in a genome-reduced bacterium. *Science*, **326**, 1268–71.

118. Lalanne,J.B., Taggart,J.C., Guo,M.S., Herzel,L., Schieler,A. and Li,G.W. (2018) Evolutionary Convergence of Pathway-Specific Enzyme Expression Stoichiometry. *Cell*, **173**, 749-761.e38.

119. Matteau,D. and Rodrigue,S. (2015) Precise Identification of DNA-Binding Proteins Genomic Location by Exonuclease Coupled Chromatin Immunoprecipitation (ChIP-exo). In Leblanc,B.P., Rodrigue,S. (eds), *DNA-Protein Interactions SE - 11*, Methods in Molecular Biology. Springer New York, Vol. 1334, pp. 173–193.

120. Rossi,M.J., Lai,W.K.M. and Pugh,B.F. (2018) Simplified ChIP-exo assays. *Nat. Commun.*, **9**, 1–13.

121. Rhee,H.S. and Pugh,B.F. (2012) ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. In *Current Protocols in Molecular Biology*. John Wiley & Sons, Inc., Vol. Chapter 21, p. Unit21.24.

122. Kühner,S., Van Noort,V., Betts,M.J., Leo-Madas,A., Batisse,C., Rode,M., Yamada,T., Maier,T., Bader,S., Beltran-Alvarez,P., *et al.* (2009) Proteome organization in a genome-reduced bacterium. *Science (80-.).*, **326**, 1235–1240.

123. Wang,L. and Maranas,C.D. (2018) MinGenome: An in Silico Top-Down Approach for the Synthesis of Minimized Genomes. *ACS Synth. Biol.*, **7**, 462–473.

124. Boulos,L., Prévost,M., Barbeau,B., Coallier,J. and Desjardins,R. (1999) LIVE/DEAD(®) BacLight(TM): Application of a new rapid staining method for direct enumeration of viable and total bacteria in drinking water. *J. Microbiol. Methods*, **37**, 77–86.

125. Schindelin,J., Arganda-Carreras,I., Frise,E., Kaynig,V., Longair,M., Pietzsch,T., Preibisch,S., Rueden,C., Saalfeld,S., Schmid,B., *et al.* (2012) Fiji: an open-source platform for biological-image analysis. *Nat. Methods*, **9**, 676–82.

126. Fahy,E., Subramaniam,S., Murphy,R.C., Nishijima,M., Raetz,C.R.H., Shimizu,T., Spener,F., van Meer,G., Wakelam,M.J.O. and Dennis,E.A. (2009) Update of the LIPID MAPS comprehensive classification system for lipids. *J. Lipid Res.*, **50**, S9–S14.

127. Vaudel,M., Burkhart,J.M., Zahedi,R.P., Oveland,E., Berven,F.S., Sickmann,A., Martens,L. and Barsnes,H. (2015) PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.*, **33**, 22–24.

128. Barsnes,H. and Vaudel,M. (2018) SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. *J. Proteome Res.*, **17**, 2552–2555.

129. Poulin-Laprade,D., Matteau,D., Jacques,P.-É., Rodrigue,S. and Burrus,V. (2015) Transfer activation of SXT/R391 integrative and conjugative elements: unraveling the SetCD regulon. *Nucleic Acids Res.*, **43**, 2045–56.

130. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

131. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–9.

132. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

133. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

134. Lawrence,M., Huber,W., Pagès,H., Aboyoun,P., Carlson,M., Gentleman,R., Morgan,M.T. and Carey,V.J. (2013) Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.*, **9**, 1–10.

135. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–15.

136. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure.pdf. *J. Mol. Biol.*, **288**, 911–940.

137. Lorenz,R., Bernhart,S.H., Höner zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 122–128.

138. Gusarov,I. and Nudler,E. (1999) The Mechanism of Intrinsic Transcription Termination. *Mol. Cell*, **3**, 495–504.

139. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

140. Kim,J.N., Roth,A. and Breaker,R.R. (2007) Guanine riboswitch variants from Mesoplasma florum selectively recognize 2'-deoxyguanosine. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 16092–7.

141. Kalvari,I., Argasinska,J., Quinones-Olvera,N., Nawrocki,E.P., Rivas,E., Eddy,S.R., Bateman,A., Finn,R.D. and Petrov,A.I. (2018) Rfam 13.0: Shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.*, **46**, D335–D342.

142. Coulombe,C., Poitras,C., Nordell-Markovits,A., Brunelle,M., Lavoie,M.-A., Robert,F. and Jacques,P.-É. (2014) VAP: a versatile aggregate profiler for efficient genome-wide data

representation and discovery. *Nucleic Acids Res.*, **42**, W485-93.

143. Stothard,P. (2000) The Sequence Manipulation Suite. *Biotechniques*, **28**.
144. Kanehisa,M., Sato,Y. and Morishima,K. (2016) BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.*, 428, 726–731.
145. Hambraeus, G., Von Wachenfeldt, C., and Hederstedt, L. (2003). Genome-wide survey of mRNA half-lives in Bacillus subtilis identifies extremely stable mRNAs. Mol. Genet. Genomics *269*, 706–714.
146. Bernstein, J.A., Khodursky, A.B., Lin, P.H., Lin-Chao, S., and Cohen, S.N. (2002). Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. Proc. Natl. Acad. Sci. U. S. A. *99*, 9697–9702.

4.13 Supplementary Material

4.13.1 Supplementary Figures

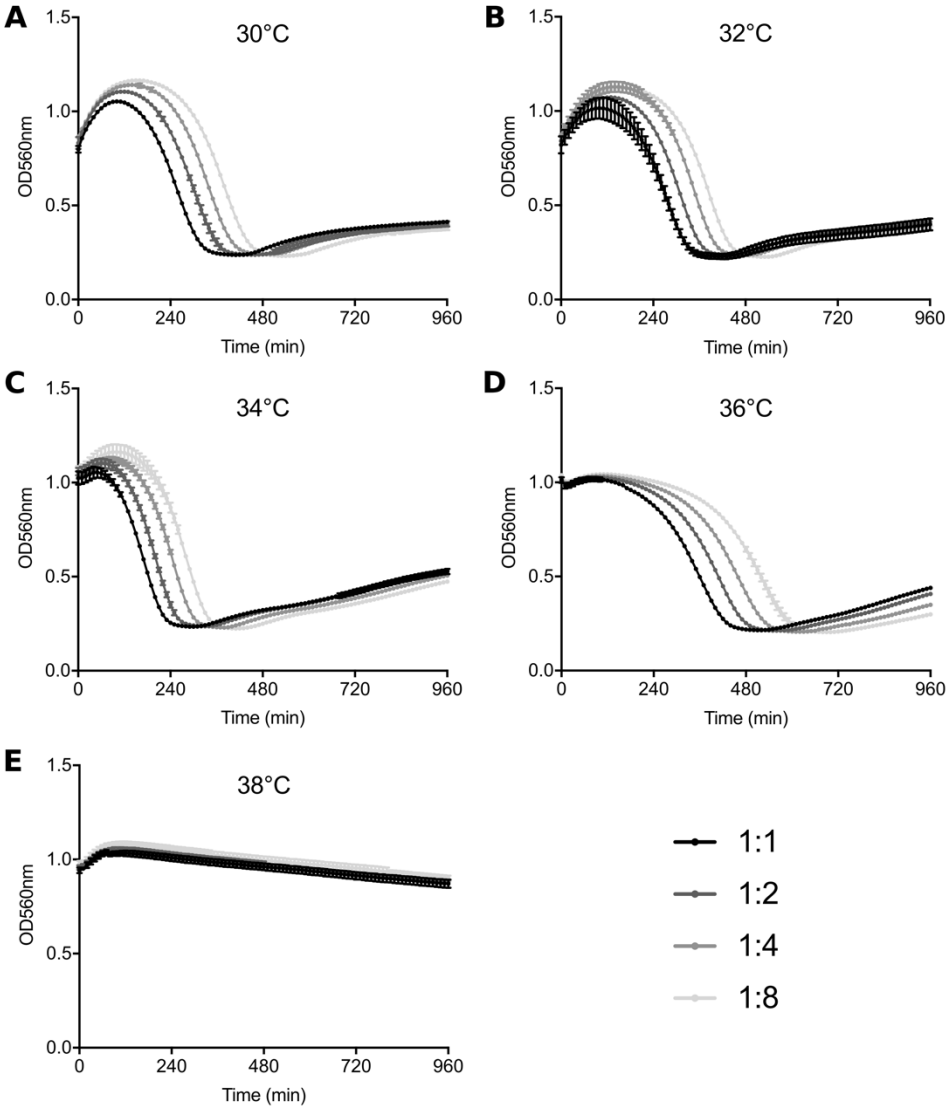

Figure S4.1. Raw *M. florum* growth curves of the 2-fold microplate dilution doubling time assay (2F-DT) performed at A) 30°C, B) 32°C, C) 34°C, D) 36°C, and E) 38°C in ATCC 1161 medium. The Dots and error bars represent the mean and standard deviation values obtained from three technical replicates.

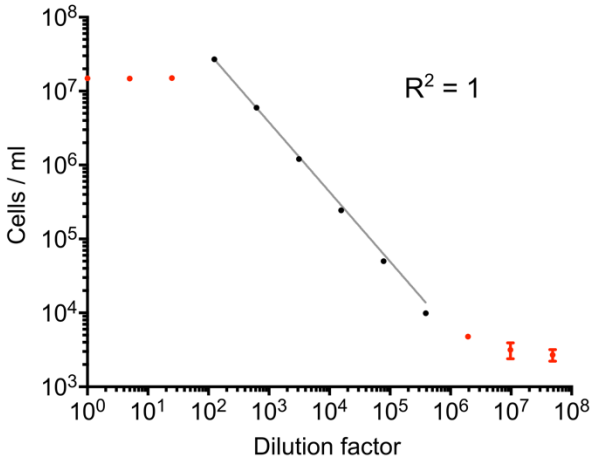

Figure S4.2. Relationship between *M. florum* cell concentrations measured by flow cytometry (FCM) and culture dilutions performed in PBS1X. A log-log nonlinear regression is shown (gray line), as well as the associated correlation coefficient (R^2). Data points outside the nonlinear regression are colored in red. The Dots and error bars represent the mean and standard deviation values obtained from technical duplicates.

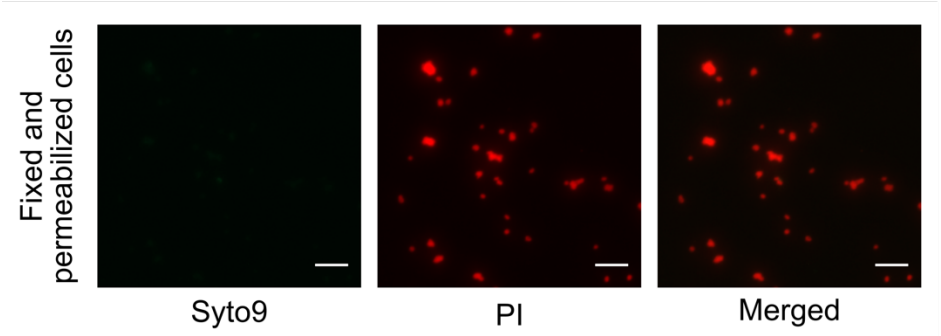

Figure S4.3. Representative image of fixed and permeabilized *M. florum* cells, double stained with SYTO 9 and propidium iodide (PI), observed by widefield fluorescence microscopy. Scale bar: 5 µm.

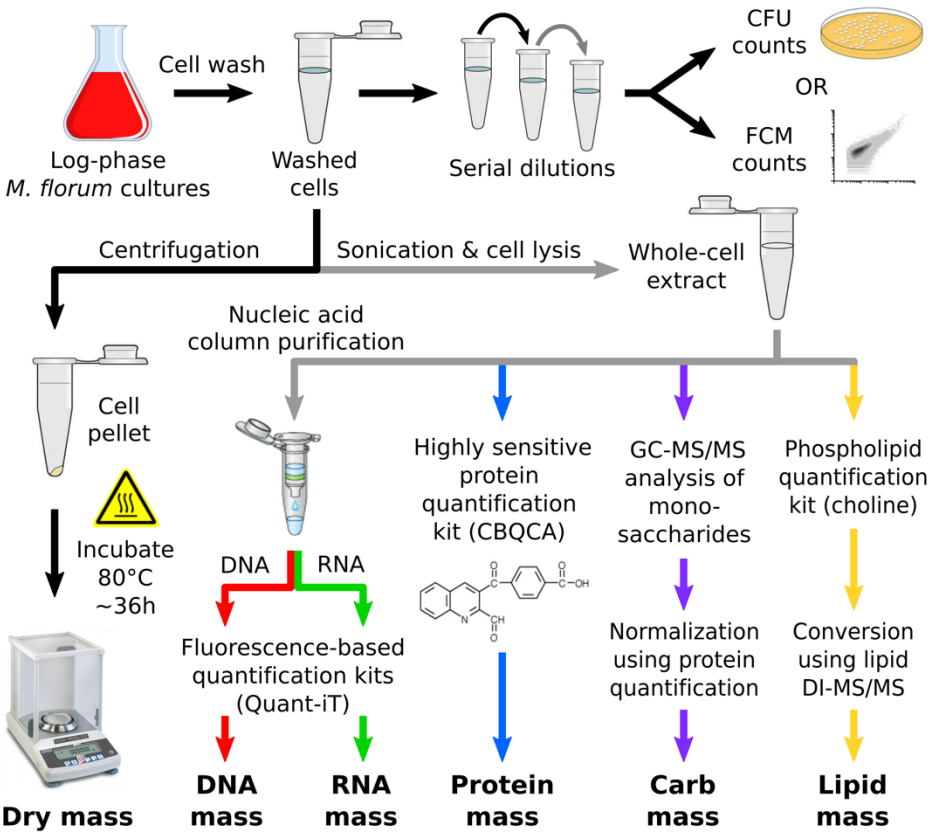

Figure S4.4. Overview of the experimental procedures used to determine the mass of the principal macromolecules contained in a *M. florum* cell as well as the total dry mass of the cell. Each constituent is quantified using high sensitivity kits, and then normalized by the number of cells used for each experiment. See Materials and Methods for further details.

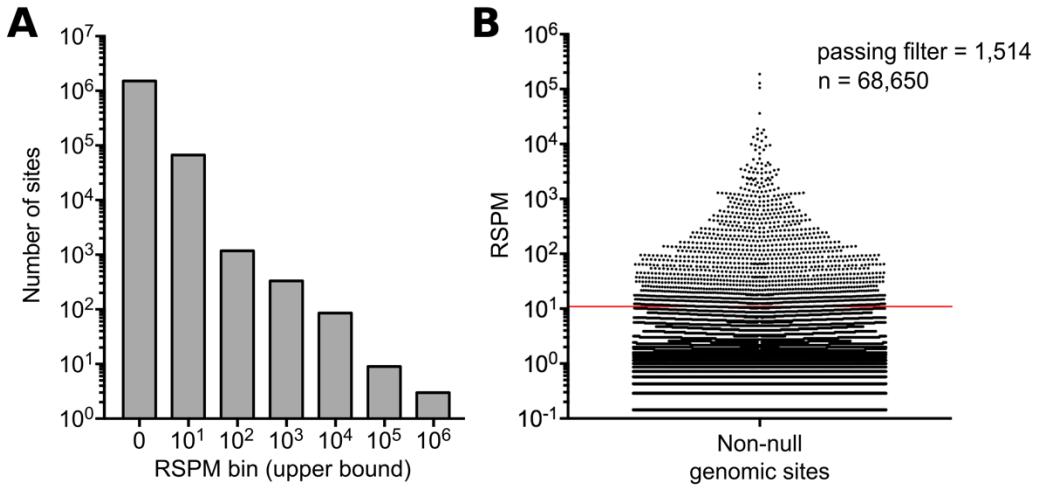

Figure S4.5. Analysis of 5'-RACE signal intensity. A) Frequency distribution of the 5'-RACE signal intensity observed at each genomic position for both DNA strands. Signal intensity was calculated according to the number of read starts per million of mapped reads (RSPM). RSPM bins are log-scale, and the upper bound value of each bin is shown. B) RSPM signal intensity of all non-null genomic positions (68,650 sites). The threshold value (10.92) used to discriminate significant 5'-RACE peaks from background noise is shown by a red line (see Material and Methods for further details). A total of 1,514 sites were considered significant.

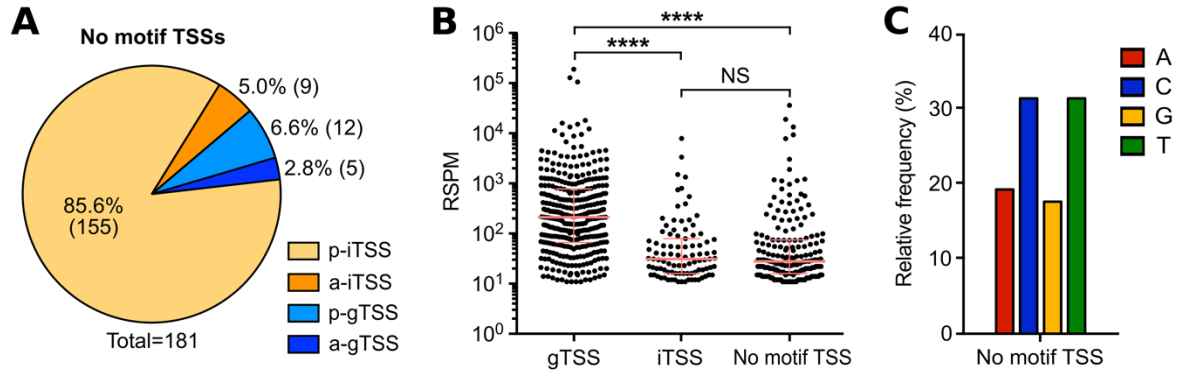

Figure S4.6. Principal characteristics of transcription start sites (TSSs) not associated to the *M. florum* promoter motif. A) Localization and orientation of TSSs without a MEME or MAST promoter motif. p-gTSS, parallel intergenic TSS; a-gTSS, antiparallel intergenic TSS; p-iTSS, parallel intragenic TSS; a-iTSS, antiparallel intragenic TSS. For gTSSs, the orientation was defined according to the closest downstream gene, while the overlapping gene was used in the case of iTSSs. B) Comparison of the read start per million of mapped reads (RSPM) signal intensity associated to gTSSs, iTSSs and TSSs without any promoter motif. The median and interquartile range are shown for each group. Distributions were compared using a Kruskal-Wallis test. C) Nucleotide identity at the transcription initiation site (+1) for TSSs not associated to a promoter motif.

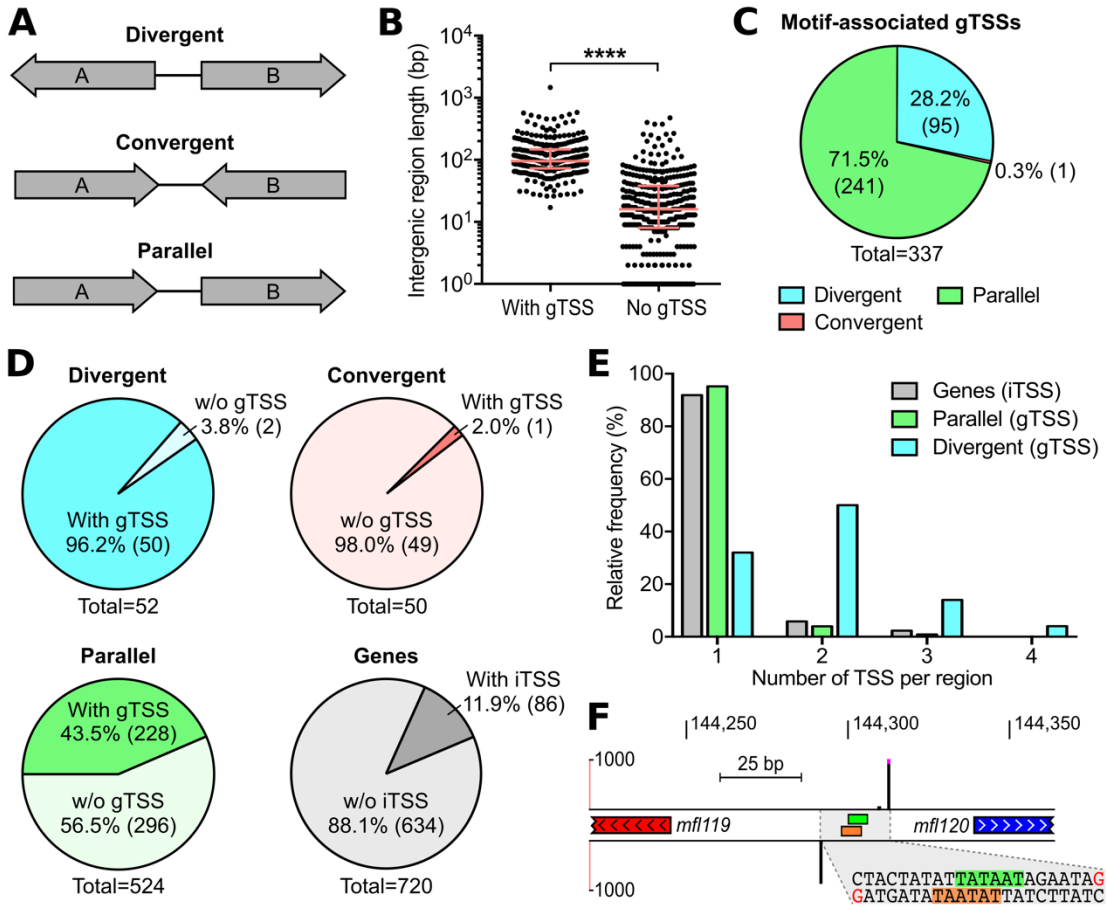

Figure S4.7. Additional information concerning the genetic context of motif-associated TSSs. A) Types of intergenic regions based on surrounding genes orientation. B) Length of intergenic regions associated or not to at least one gTSS. The median and interquartile range are shown for each group. Distributions were compared using a Mann-Whitney test (p-value <0.0001). C) Total number of gTSSs for each of the three intergenic region groups depicted in A. D) Relative proportion of divergent, convergent, and parallel intergenic regions hit by at least one gTSS out of their total respective number found across the genome. The proportion of genes hit by iTSSs is also shown. E). Relative frequency distribution of the number of motif-associated TSSs detected per gene, parallel intergenic region or divergent intergenic region. F) Genomic locus showing a representative case of two divergent genes expressed from two back-to-back overlapping promoters identified by 5'-RACE. Genomic coordinates are indicated at the top of the panel. Strand-specific 5'-RACE signals are shown by black bars (0-1,000 read starts scale). Peaks above 1,000 read starts are cut and marked by fuchsia dots. The position of -10 boxes associated to 5'-RACE peaks are indicated by green and orange rectangles for positive and negative DNA strands, respectively. The genomic coordinates containing the identified TSSs and -10 boxes is enlarged and its corresponding DNA sequence is illustrated. TSS bases are colored in red. Bases corresponding to the -10 boxes are highlighted in green and orange for positive and negative DNA strands.

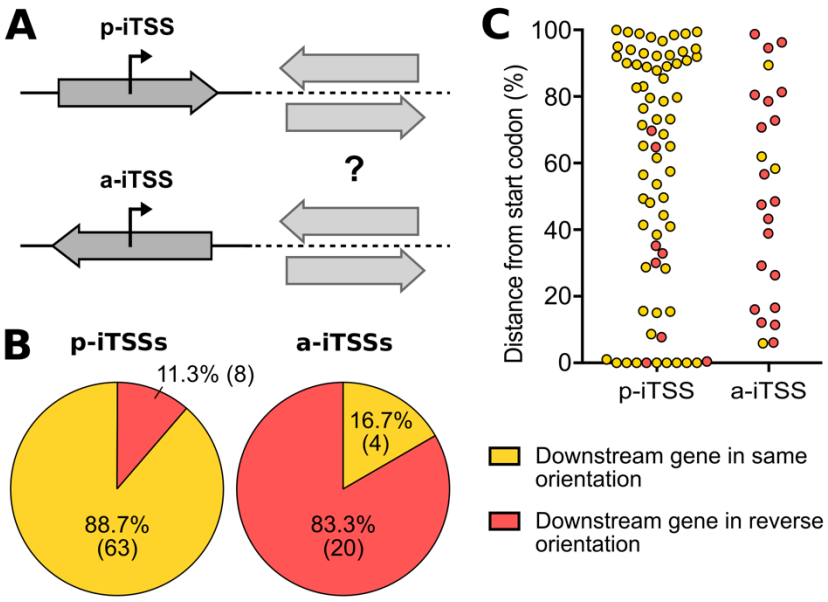

Figure S4.8. Additional information about the genetic context of motif-associated iTSSs. A) Classification of iTSSs according to their orientation relatively to the overlapping gene. p-iTSS, parallel intragenic TSS; a-iTSS, antiparallel intragenic TSS. If the most immediate downstream gene is transcribed on the same strand, both iTSS types can also be adequately oriented to drive its expression. B) p-iTSSs and a-iTSSs orientation relatively to the nearest downstream gene. C) Distance from overlapping gene start codon for p-iTSSs and a-iTSSs. Distance was normalized according to the overlapping gene length. Yellow and red dots indicate iTSSs located upstream genes of the same and reverse orientation, respectively.

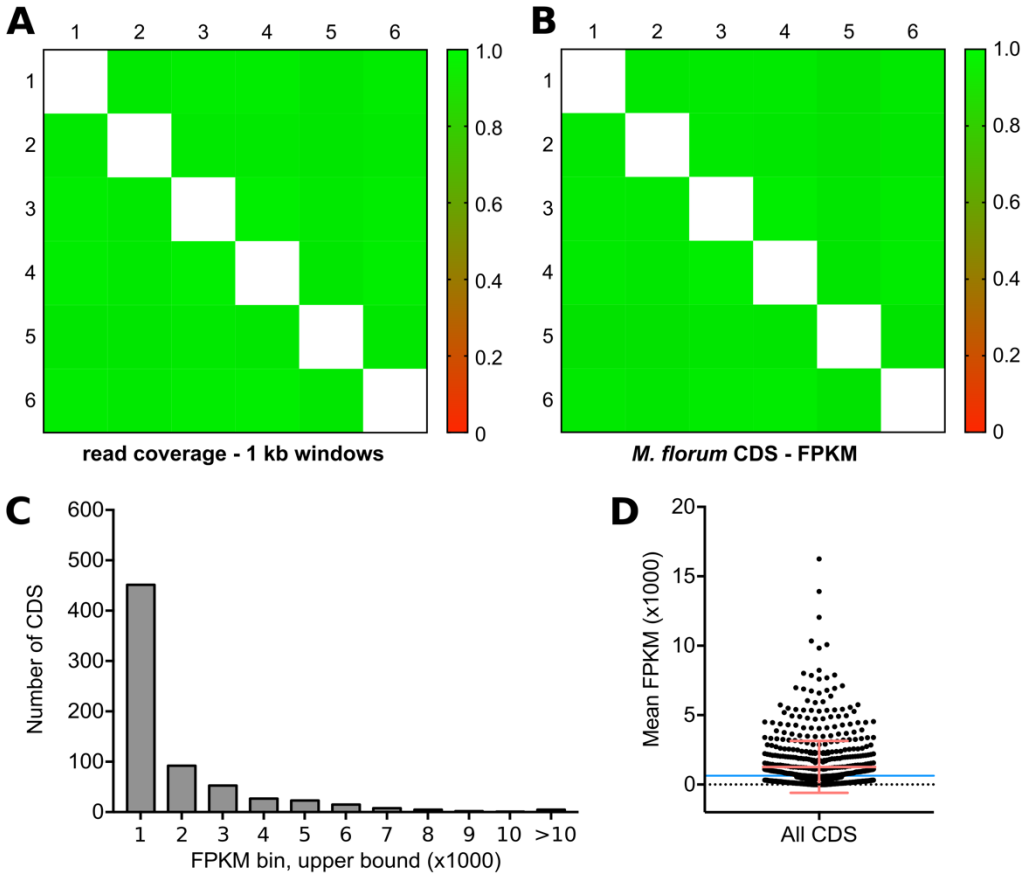

Figure S4.9. RNA-seq related correlations and distributions. A) Pearson correlation heatmap of RNA-seq read coverage calculated from the different library replicates using non-overlapping 1 kb windows. B) Same as A but using the number of fragments per kilobase per million of mapped reads (FPKM) calculated for each *M. florum* reference gene (n=720). C) Frequency distribution of the mean FPKM values of *M. florum* coding sequences (n=685). The upper bound value of each FPKM bin is shown. D) Scatter plot showing the mean FPKM value calculated for each *M. florum* coding sequence. The mean and associated SD are shown. The blue line indicates the theoretical FPKM value obtained if all the reads were equally distributed across the genome (FPKM=630).

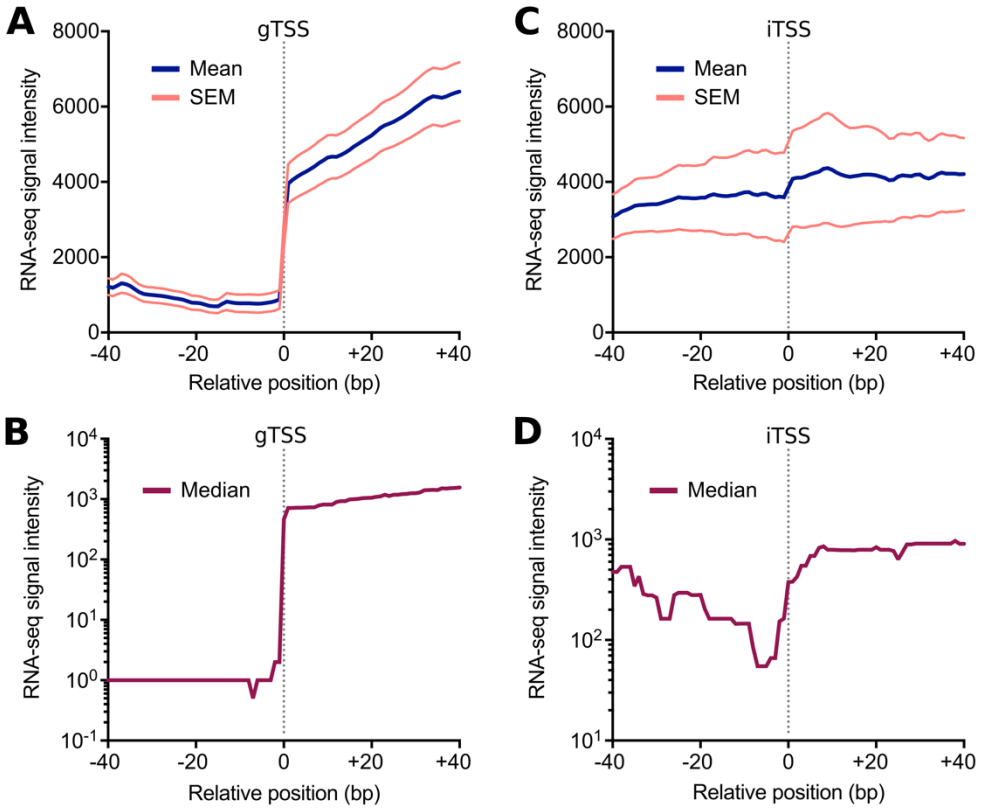

Figure S4.10. RNA-seq aggregate profiles of identified TSS types. A) Aggregate profile showing the mean RNA-seq read coverage observed at and surrounding all motif-associated gTSSs identified in this study. The calculated SEM is also shown. The aggregate profile was centered on the gTSSs coordinates (relative position 0 bp), indicated by a gray dashed line. B) Same as A but showing the median value at each position instead of the mean and SEM. C) and D) Identical to A and B, but for motif-associated iTSSs.

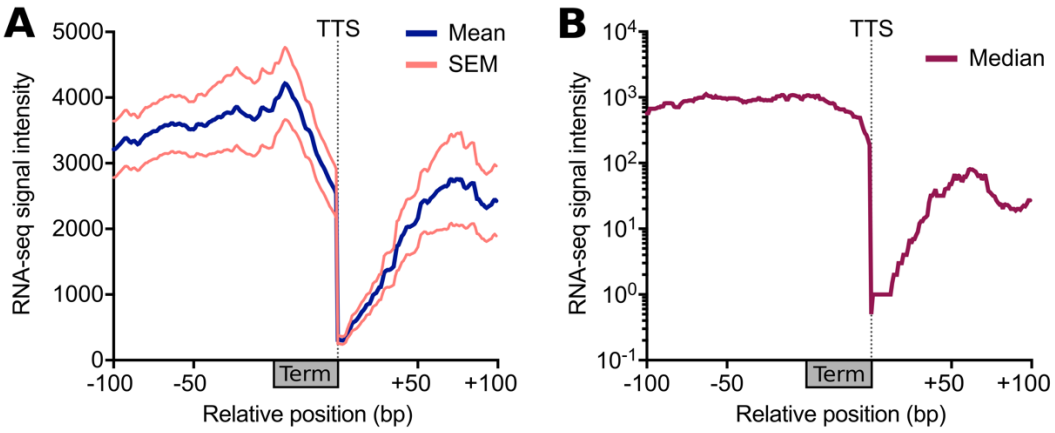

Figure S4.11. RNA-seq aggregate profiles of Rho-independent terminators predicted in this study. A) Aggregate profile showing the mean RNA-seq read coverage observed for all predicted terminators and their surrounding DNA regions. The calculated SEM is also shown. The aggregate profile was centered on the terminators start and stop coordinates. The predicted transcription termination site (TTS) is indicated by a gray dashed line. B) Same as A but showing the median value at each position instead of the mean and SEM.

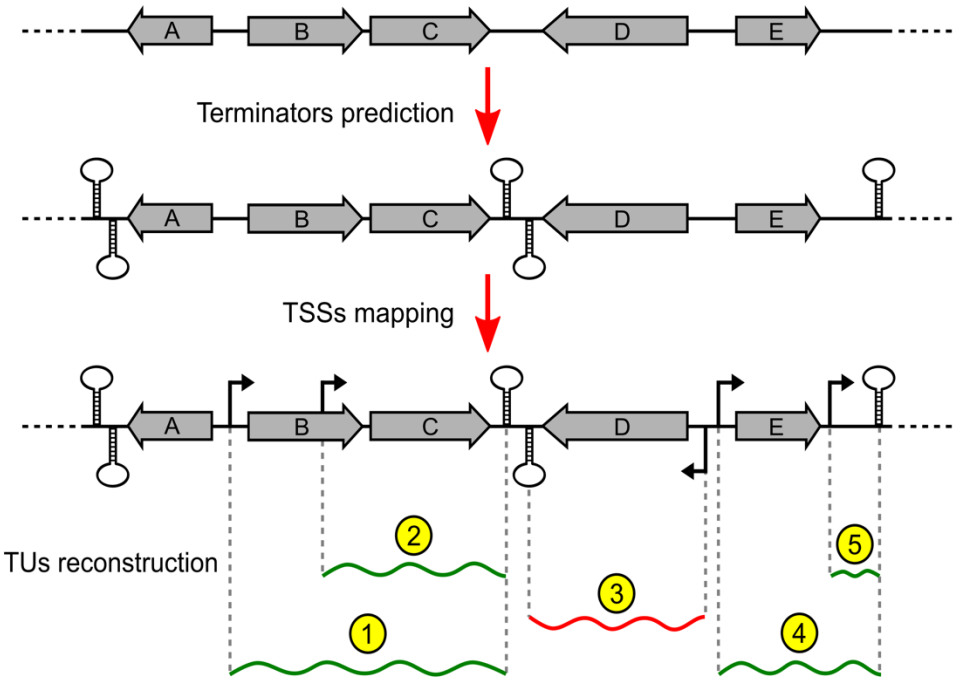

Figure S4.12. Summary of the transcription units reconstruction procedure. First, Rho-independent terminators were predicted from the DNA sequence and genes annotation as described previously (1), creating strand-specific term-to-term scaffolds. Motif-associated TSSs were then mapped onto the scaffolds, and all possible transcription units (TUs) were reconstructed. Depending on the context, some TUs may contain a single gene (TUs 2, 3, and 4), many genes (TU 1), or no gene at all (TU 5). Some TUs may also partially overlap other genes if they originate from iTSSs (TU 2). Genes not included in at least one TU and therefore not associated to any TSS are classified as orphan genes (gene A).

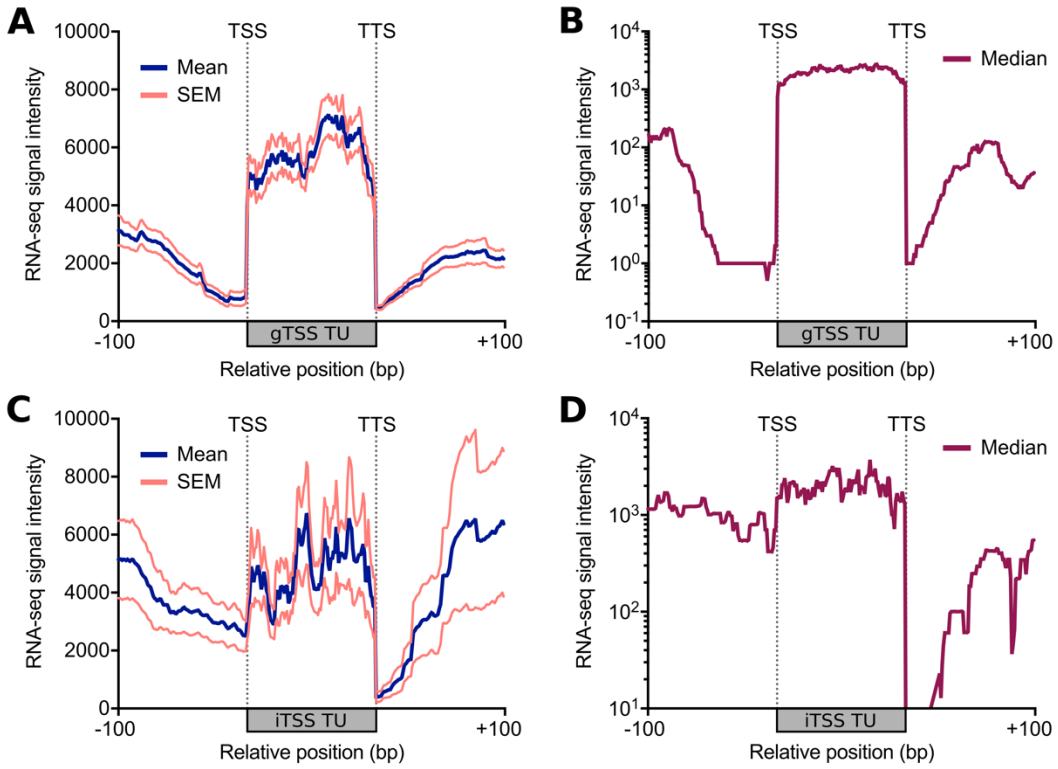

Figure S4.13. RNA-seq aggregate profiles of gTSS and iTSS transcription units (TUs). A) Aggregate profile showing the mean RNA-seq read coverage observed for all gTSS TUs and their surrounding DNA regions. The calculated SEM is also shown. The aggregate profile was centered on the TUs start and stop coordinates, corresponding to transcription start site (TSS) and termination site (TTS), respectively. B) Same as A but showing the median value at each position instead of the mean and SEM. C) and D) Identical to A and B, but for iTSS TUs.

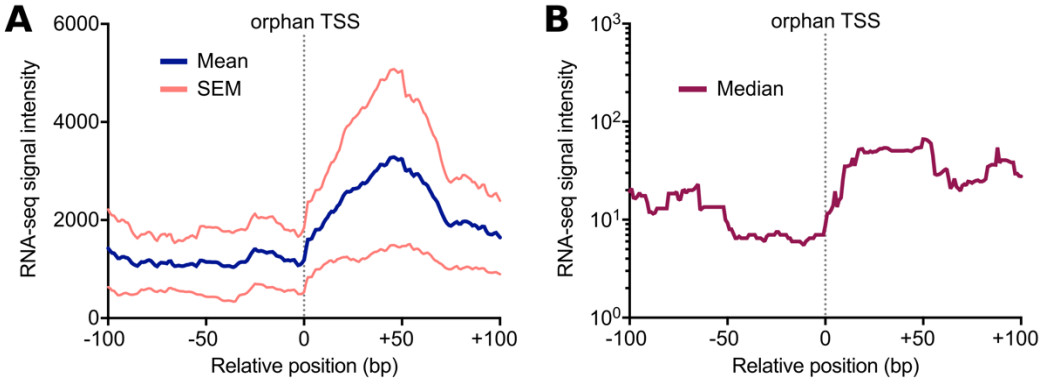

Figure S4.14. RNA-seq aggregate profiles of intergenic motif-associated TSSs not associated to any downstream gene (orphan TSSs). A) Aggregate profile showing the mean RNA-seq read coverage and the associated SEM values. The aggregate profile was centered on the TSSs coordinates (relative position 0 bp), indicated by a gray dashed line. B) Same as A but showing the median value at each position instead of the mean and SEM.

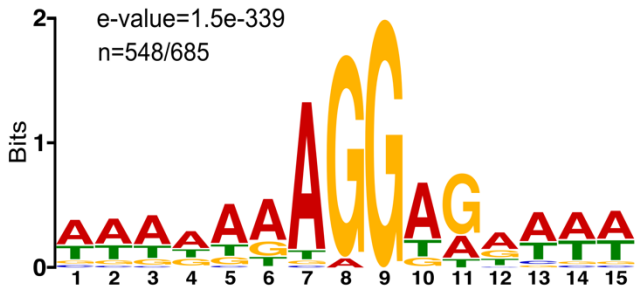

Figure S4.15. *M. florum* Shine-Dalgarno consensus sequence generated using MEME software. The region located immediately upstream (≤20 bp) a total of 548 coding sequences (out of 685) were included in the motif.

4.13.2 Supplementary Tables

Table S4.1. Statistics summary of Illumina sequencing libraries prepared in this study.

Library type	Sequencing type	Replicate	Total reads (single)	Reads passing quality filters	Aligned reads (MAPQ>=10)	Genome coverage
5'-RACE	SE 40 bp	1	10,234,272	9,442,841 (92%)	6,961,595 (74%)	~350X
RNAseq	PE 50 bp	1	16,089,680	14,003,252 (87%)	13,049,819 (93%)	~820X
		2	16,531,090	14,234,385 (86%)	12,649,001 (89%)	~800X
		3	16,788,638	14,493,548 (86%)	13,605,303 (94%)	~860X
		4	17,389,570	15,067,903 (87%)	14,377,039 (95%)	~910X
		5	18,566,270	15,929,927 (86%)	14,980,959 (94%)	~940X
		6	15,247,438	13,105,485 (86%)	12,160,110 (93%)	~770X

Table S4.2. Comparison of the intracellular levels of important molecules and complexes between *M. florum* and other selected species.

	M. florum	*M. mycoides*	*M. pneumoniae*	*E. coli*
Molecules per cell				
Total RNA	23,320	NA	4,430	261,400[(a)] 258,000[(b)]
rRNA	4,900	NA	900[©]	60,000[(a)] 54,000[(b)]
tRNA	18,000	NA	3,300[©]	200,000[(a,b)]

Table S4.2. Comparison of the intracellular levels of important molecules and complexes between *M. florum* and other selected species (continued).

mRNA	420	NA	230[(d)]	1,400[(a)] 4,000[(b)]
Protein	250,000	77,000[©]	130,000[©]	3,000,000[(a)] 3,600,000[(b)] 3,000,000-4,000,000[(f)]
Ribosome	1,600-2,100	340-670[©]	300[©] 190[(g)] 140[(h)]	6,800-72,000[(a)] 18,000[(b)] 30,000-70,000[(i)]
Core RNA polymerase	270	380[©]	300[(h)]	1,500-11,400[(a)] 2,000-10,000[(i)] 2,600-13,000[(j)]
σ^{70}	230	230	NA	4,700-17,000[(j)]
Average cell volume (µm^3)	0.090	0.034[©]	0.067[©]	1.0[(a,b)] 1.3-3.0[(i)]
Molecules per µm^3				
Total RNA	260,000	NA	65,400	261,400[(a)] 258,000[(b)]
rRNA	54,000	NA	13,000[©]	60,000[(a)] 54,000[(b)]
tRNA	200,000	NA	49,000[©]	200,000[(a,b)]
mRNA	4,700	NA	3,400[(d)]	1,400[(a)] 4,000[(b)]
Protein	2,800,000	2,260,000[©]	1,900,000[©]	3,000,000[(a)] 3,600,000[(b)] 3,000,000-4,000,000[(f)]
Ribosome	18,000-24,000	10,000-20,000[©]	4,500[©] 2,800[(g)] 2,100[(h)]	18,000[(b)] 25,000-31,000[(i)]
Core RNA polymerase	3,000	11,000[©]	4,500[(h)]	1,500-11,400[(a)] 1,800-3,500[(i)] 2,600-13,000[(j)]
σ^{70}	2,600	6,800[©]	NA	4,700-17,000[(j)]

[(a)]Bionumbers (3)
[(b)]CCDB database
[©]Yus *et al.* 2009 (4)
[(d)]Weiner *et al.* 2003 (5)
[€]Breuer *et al.* 2019 (6)
[(f)]Milo 2013 (7)
[(g)]Wodke *et al.* 2013 (8)
[(h)]Kuhner *et al.* 2009 (9)
[(i)]Bakshi *et al.* 2012 (10)
[(j)]Grigorova *et al.* 2006 (11)

4.13.3 Supplementary Datasets

Supplementary Datasets are available at:
http://lab-rodrigue.recherche.usherbrooke.ca/integrative-characterization-of-the-near-minimal-bacterium-mesoplasma-florum/

Dataset S4.1. List of motif-associated TSSs identified in this study.

Dataset S4.2. List of Rho-independent terminators predicted in this study.

Dataset S4.3. Reconstructed *M. florum* gene-associated transcription units.

Dataset S4.4. List of orphan TSSs.

Dataset S4.5. RNA-seq quantification results and intracellular abundance of RNA species.

Dataset S4.6. 2D LC-MS/MS protein quantification results and corresponding intracellular abundances.

Dataset S4.7. KO numbers and functional categories assigned to *M. florum* protein-coding genes.

Dataset S4.8. *M. florum* lipidomic profile determined by DI-MS/MS.

4.13.4 Supplementary References

1. de Hoon,M.J.L., Makita,Y., Nakai,K. and Miyano,S. (2005) Prediction of transcriptional terminators in Bacillus subtilis and related species. *PLoS Comput. Biol.*, **1**, e25.
2. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
3. Bionumbers (2015) What is the macromolecular composition of the cell.
4. Yus,E., Maier,T., Michalodimitrakis,K., van Noort,V., Yamada,T., Chen,W.-H.W.-H., Wodke,J.A.H.J.A.H., Güell,M., Martínez,S., Bourgeois,R., *et al.* (2009) Impact of Genome Reduction on Bacterial Metabolism and Its Regulation. *Science (80-.).*, **326**, 1263–1268.
5. Weiner,J., Zimmerman,C.U., Göhlmann,H.W.H. and Herrmann,R. (2003) Transcription profiles of the bacterium Mycoplasma pneumoniae grown at different temperatures. *Nucleic Acids Res.*, **31**, 6306–6320.

6. Breuer,M., Earnest,T.M., Merryman,C., Wise,K.S., Sun,L., Lynott,M.R., Hutchison,C.A., Smith,H.O., Lapek,J.D., Gonzalez,D.J., *et al.* (2019) Essential metabolism for a minimal cell. *Elife*, **8**, 1–77.
7. Milo,R. (2013) What is the total number of protein molecules per cell volume? A call to rethink some published values. *BioEssays*, **35**, 1050–1055.
8. Wodke,J.A.H., Pucha ka,J., Lluch-Senar,M., Marcos,J., Yus,E., Godinho,M., Gutierrez-Gallego,R., dos Santos,V.A.P.M., Serrano,L., Klipp,E., *et al.* (2014) Dissecting the energy metabolism in Mycoplasma pneumoniae through genome-scale metabolic modeling. *Mol. Syst. Biol.*, **9**, 653–653.
9. Kühner,S., Van Noort,V., Betts,M.J., Leo-Madas,A., Batisse,C., Rode,M., Yamada,T., Maier,T., Bader,S., Beltran-Alvarez,P., *et al.* (2009) Proteome organization in a genome-reduced bacterium. *Science (80-.).*, **326**, 1235–1240.
10. Bakshi,S., Siryaporn,A., Goulian,M. and Weisshaar,J.C. (2012) Superresolution Imaging of Ribosomes and RNA Polymerase in Live Escherichia coli Cells. *Mol. Microbiol.*, **85**, 21–38.
11. Grigorova,I.L., Phleger,N.J., Mutalik,V.K. and Gross,C.A. (2006) Insights into transcriptional regulation and σ competition from an equilibrium model of RNA polymerase binding to DNA. *Proc. Natl. Acad. Sci. U. S. A.*, 103, 5332–5337.

CHAPITRE 5
DISCUSSION ET CONCLUSION GÉNÉRALE

5.1 Résumé du projet de recherche

Alors que les technologies récentes de la génomique synthétique permettent désormais de synthétiser des chromosomes entiers sur mesure, notre capacité à prédire les impacts de modifications génomiques importantes (réductions, réorganisation, etc.) sur le comportement des cellules demeure à un niveau rudimentaire. Ceci est dû d'une part à la trop grande complexité des organismes couramment utilisés dans ce contexte (p. ex. *E. coli* et *S. cerevisiae*), et d'autre part au fait que les mécanismes du fonctionnement global des cellules demeurent mal compris. En dépit des efforts impressionnants des dernières années, par exemple la création de la souche réduite JCVI-syn3.0 (Hutchison et al., 2016), la création d'une souche d'*E. coli* ne possédant que 57 codons (Fredens et al., 2019) ou la synthèse complète de chromosomes synthétiques de la levure (Richardson et al., 2017), très peu d'architectures génomiques significativement modifiées ont été jusqu'à maintenant explorées. Par conséquent, les règles sous-tendant la programmation des génomes demeurent largement incomprises, ce qui restreint le potentiel immense de la biologie synthétique.

En utilisant l'organisme quasi minimal *M. florum* comme modèle d'étude, nous croyons qu'il sera possible d'atteindre un niveau de compréhension sans précédent du fonctionnement global d'une cellule autonome grâce à une démarche novatrice basée sur l'intégration d'une quantité massive de données obtenues à l'échelle du génome. Plus spécifiquement, cette démarche repose sur la combinaison de données expérimentales provenant des technologies de la génomique fonctionnelle, des prédictions issues d'approches bio-informatiques, ainsi que d'efforts de réorganisation du génome via des approches de génomique synthétique. Dans ce contexte, mon projet de recherche visait premièrement à développer un système de culture en continu simple, flexible et abordable afin de faire croître *M. florum* dans des conditions stables, contrôlées et hautement reproductibles (voir Chapitre 2). J'ai ensuite procédé au développement

des premiers plasmides capables de se répliquer chez ce microorganisme afin d'établir une base d'outils moléculaires spécifiquement conçus pour modifier le génome de *M. florum* (voir Chapitre 3). Finalement, j'ai entrepris la première caractérisation intégrative de cette bactérie en combinant différentes approches et méthodologies afin de mieux définir plusieurs aspects relatifs à sa morphologie, physiologie et composition moléculaire (voir Chapitre 4). Les prochaines sections porteront sur l'analyse critique des accomplissements réalisés durant mon parcours doctoral, sur leurs améliorations possibles, ainsi que sur les perspectives associées à ceux-ci.

5.2 Vers un VCCD 2.0 ?

Les premiers travaux entrepris dans le cadre de mon doctorat visaient à développer le VCCD, un système de culture en continu à la fois simple, flexible et peu dispendieux (voir Chapitre 2) (Matteau et al., 2015). Le faible volume des chambres de croissance (~20-40 ml), ses différents modes de rafraichissement des cultures et sa simplicité d'opération permettent son utilisation pour une grande variété de besoins expérimentaux. L'ensemble des matériaux et composantes du système sont facilement accessibles commercialement ou peuvent être fabriqués à l'aide d'approches conventionnelles. De plus, les informations requises pour construire et opérer l'appareil sont entièrement disponibles sous forme d'un manuel d'utilisateur complet et détaillé (Matteau et al., 2015).

Alors que le VCCD représente un appareil intéressant pour maintenir des conditions stables de croissance pour différents organismes, incluant *E. coli* et *M. florum*, certains aspects relatifs à sa conception pourraient être sujets à des améliorations. En effet, le système actuel ne comporte pas de module de contrôle de la température pour les chambres de croissance, ce qui implique que la plupart des expériences doivent être réalisées à l'intérieur d'un incubateur ou une chambre à température contrôlée. Toutefois, certains incubateurs commerciaux ne possèdent pas un volume intérieur suffisant pour abriter le système en entier, ce qui peut constituer une limitation importante. Dans ce cas, il est possible de modifier le support physique actuel du VCCD pour

une conception plus compacte, ou tout simplement rendre les chambres de croissance physiquement indépendantes, ce qui offre plus de flexibilité. L'ajout de modules thermoélectriques de type Peltier pourrait également constituer une solution simple à ce problème (https://www.digikey.fr/fr/ptm/c/cui-inc/understanding-and-using-peltier-modules-for-thermal-management/tutorial), ce qui permettrait d'effectuer la majorité des expériences à température ambiante. Ceci faciliterait aussi grandement les manipulations lors de l'utilisation du système. En effet, celles-ci peuvent s'avérer particulièrement difficiles lorsqu'elles sont effectuées dans un espace restreint tel qu'un incubateur. En dehors d'un espace confiné, il serait aussi beaucoup plus envisageable d'adapter le système pour que le prélèvement d'échantillons soit exécuté de manière automatique, par exemple avec l'utilisation de robots pipetteurs à bras articulés. Ceci constituerait un réel avantage pour la tenue d'expériences de longue durée où le prélèvement d'échantillons doit se faire de manière constante, par exemple.

Dans sa conception actuelle, le VCCD comporte trois chambres de croissance contrôlées de manière indépendante, ce qui représente le standard dans beaucoup d'expériences de croissance en laboratoire. Toutefois, advenant que des modules de contrôle de la température soient implémentés dans une version 2.0 de l'appareil, il serait possible de multiplier ce nombre par deux ou même par quatre, puisque l'espace ne serait plus une contrainte importante. Dans ce cas, il faudrait néanmoins remplacer la carte d'acquisition des données présentement proposée (NI USB-6008 DAQ) par un modèle offrant un plus grand nombre de ports d'acquisition et adapter les circuits électroniques en conséquence. De plus, il serait nécessaire d'adapter le support physique de l'appareil, et peut-être même de revoir la conception du système de rafraichissement des cultures afin d'éviter que le tout ne devienne trop encombrant. Ceci dit, il serait aussi possible de remplacer certaines pièces physiques de l'appareil présentement fabriquées par usinage traditionnel par des pièces produites à l'aide d'imprimantes 3D. Dans certains cas, cela simplifierait grandement leur fabrication, en plus de diminuer les coûts associés aux matériaux. Ces technologies sont d'ailleurs de plus en plus utilisées dans la conception d'appareils de culture en continu en raison de la diminution récente des prix reliés aux systèmes d'impression 3D (Hoffmann et al., 2017; Pilizota and Yang, 2018). Des modules

de filtres optiques, adaptés pour mesurer la fluorescence émise par les cellules en culture, pourraient notamment être fabriqués à l'aide de ce type d'imprimante.

Une des plus grandes limitations du système VCCD actuel est le logiciel utilisé pour contrôler l'appareil. Quoique fonctionnel, simple et facile d'utilisation, ce logiciel a été développé sous la plateforme de développement *LabVIEW* (*National Instruments*), une plateforme graphique spécialisée dans la conception et la programmation de logiciels d'acquisition de données. Alors que *LabVIEW* permet la distribution du code source et des fichiers exécutables développés sous sa bannière (comme le logiciel utilisé par le VCCD), la modification du code source est impossible sans licence d'utilisation de la compagnie *National Instruments.* Le système VCCD constitue tout de même un appareil à libre accès puisque nous rendons publiques toutes les informations nécessaires pour construire et opérer l'appareil, incluant le logiciel d'utilisation et son code source, mais la modification du logiciel, elle, ne l'est pas. Idéalement, le logiciel de contrôle de l'appareil devrait être développé dans un langage informatique ouvert et facilement interprétable, *Python* par exemple. Par contre, cela demanderait un effort considérable compte tenu de la complexité du programme actuel, de ses différents modes, options et paramètres, en plus de son affichage graphique en temps réel des mesures de turbidité enregistrées. Il serait cependant facile pour l'ensemble de la communauté scientifique d'apporter des modifications au logiciel en fonction de leurs besoins expérimentaux spécifiques, comme l'implémentation d'une fonction de calcul automatique du temps de doublement. Quoi qu'il en soit, le système VCCD dans sa version actuelle demeure un appareil très intéressant et pratique dans un bon nombre de situations. Cet appareil a d'ailleurs été utilisé dans le laboratoire du Pr Sébastien Rodrigue afin d'évaluer la stabilité des plasmides *oriC* développés chez *M. florum* (voir Chapitre 3) ainsi que pour la préparation de librairies de RNA-seq (voir Chapitre 4). Cet appareil subira fort probablement de nombreuses améliorations au cours des prochaines années et servira assurément dans le contexte de plusieurs autres expériences futures, telles que des expériences de mutagenèse par transposons et des essais d'évolution accélérée en laboratoire.

5.3 Développement d'outils moléculaires pour *M. florum*

Notre capacité à modifier génétiquement *M. florum* est extrêmement déterminante pour le développement d'une plateforme simplifiée de prototypage de génomes à partir de cet organisme. Cependant, au commencement de mon doctorat, pratiquement aucun outil moléculaire n'avait été développé ou testé dans cette bactérie. Le seul outil que nous savions fonctionnel chez *M. florum* était une cassette de résistance à la tétracycline provenant d'un transposon (Tn*916*) naturellement retrouvé chez *Enterococcus faecalis* (Rice, 1998). En effet, cette cassette avait été utilisée vers la fin des années 2000 par un de nos collaborateurs pionniers dans le projet *M. florum*, Thomas F. Knight, afin d'effectuer les premiers essais de mutagenèse par transposons chez cette bactérie en utilisant la transposase Tn*5* purifiée commercialement (EZ-Tn*5* ; Epicentre). Ces expériences ont notamment permis de produire une librairie contenant tout près de 3000 mutants d'insertion individuels. Ces résultats furent d'ailleurs récemment combinés à des analyses de génomique comparative afin de proposer différents scénarios de réduction du génome de *M. florum* (Baby et al., 2018). Quoi qu'il en soit, les efficacités originalement observées à l'aide du protocole de mutagenèse basé sur la transposase Tn*5* purifiée n'ont jamais pu être répétées, ni par notre laboratoire ni par Thomas F. Knight lui-même. En fait, les expériences subséquentes produisaient généralement tout au plus quelques colonies par transformation, ce qui est problématique dans la plupart des contextes, d'autant plus que le prix de la Tn*5* commerciale avoisine les 100 $ par µl. De plus, l'utilisation de transposons afin de tester de nouveaux outils moléculaires n'est pas toujours souhaitable selon les cas, principalement parce que le locus d'insertion peut affecter le comportement des éléments génétiques clonés (en raison, par exemple, d'effets polaires de la transcription). Le développement de plasmides réplicatifs transformables à haute efficacité constituait donc une première étape logique afin de développer et tester de nouveaux outils moléculaires chez *M. florum*.

Comme la plupart des Mollicutes, *M. florum* est déficient en éléments extrachromosomiques potentiellement utilisables comme vecteur de clonage. En nous inspirant de publications

antérieures décrivant la construction de plasmides artificiels basés sur l'*oriC* de différents Mollicutes, incluant *M. mycoides* et *M. capricolum* (Janis et al., 2005; Lartigue et al., 2003; Renaudin et al., 1995), nous avons entrepris le développement des tout premiers plasmides spécifiquement conçus pour se répliquer chez *M. florum* (voir Chapitre 3) (Matteau et al., 2017). En incluant différentes régions de l'*oriC* prédite de cette bactérie, nous avons pu montrer que les régions intergéniques de part et d'autre du gène *dnaA* étaient suffisantes et nécessaires à la réplication des plasmides par la machinerie de réplication de *M. florum*, ce qui n'est pas nécessairement le cas pour tous les Mollicutes (Lartigue et al., 2002). La transformation au polyéthylène glycol (PEG) des versions comprenant ces deux régions (pMflT-o3 pMflT-o4), ainsi que la cassette de résistance *tetM* provenant du transposon Tn*916*, a permis d'obtenir plusieurs centaines de colonies par transformation, ce qui a été exploité afin de tester de nouveaux marqueurs de résistance aux antibiotiques. Parmi les quatre testés (excluant *tetM*), seules les cassettes conférant une résistance à la puromycine (*pac*) et à la spectinomycine/streptomycine (*aadA1*) se sont avérées fonctionnelles chez *M. florum*, ce qui a néanmoins permis de tripler le nombre de gènes de sélection utilisables chez cette bactérie. Ces plasmides ont aussi été utilisés afin de développer et optimiser deux approches alternatives à la transformation médiée par le PEG, soit l'électroporation et la conjugaison. Montrant des efficacités semblables à la transformation au PEG, ces méthodes s'avèrent toutefois relativement plus simples à réaliser que celle-ci, et pourront être exploitées dans différents contextes intéressants. La conjugaison, par exemple, pourrait servir afin de transférer des molécules d'ADN de grande taille d'*E. coli* vers *M. florum* et procéder à un échange de fragments génomiques (voir section 5.6).

Puisque les plasmides *oriC* possèdent une grande région d'homologie avec le chromosome de la cellule hôte (~1-2 kb), leur recombinaison avec celui-ci n'est pas rare, ce qui amène un phénomène de duplication de la région *oriC*. Cette caractéristique a également pu être observée avec les plasmides *oriC* de *M. florum* (voir Chapitre 3). Même si cette recombinaison ne semble pas affecter la stabilité des plasmides à long terme, c'est-à-dire que les cellules n'ont pas tendance à perdre les plasmides lorsqu'aucune sélection n'est appliquée (voir Figure 3.4), celle-

ci a probablement des répercussions pour la réplication du chromosome de la cellule hôte. Toutefois, l'impact de la duplication de l'*oriC* demeure jusqu'à présent grandement inexploré. Certaines observations effectuées par notre laboratoire laissent croire que cette duplication affecterait la viabilité des cellules, du moins à court terme. En effet, les colonies de *M. florum* obtenues suite à la transformation des plasmides *oriC* requièrent généralement une à deux journées d'incubation supplémentaires avant d'être visibles à l'œil nu, comparativement à des cellules n'ayant reçu aucun plasmide. De plus, la morphologie des colonies est également atypique : le contour de celles-ci est difforme, leur taille est plus petite et l'aspect de « beigne » (voir [McCoy et al., 1984] pour une image représentative) habituellement observé n'est plus apparent. Les cultures inoculées par ces colonies prennent aussi 2 à 3 jours de plus avant de montrer des signes de croissance, c'est-à-dire une acidification du milieu de culture utilisé. Curieusement, les courbes de croissance de cultures de *M. florum* possédant un plasmide *oriC* ne semblent pas montrer de retard de croissance important (voir Figure S3.2). Toutefois, ces expériences ne consistaient pas en une mesure précise du temps de doublement des cultures. Une autre observation intéressante corrobore également l'hypothèse que les plasmides *oriC* pourraient déstabiliser le processus de réplication du chromosome de *M. florum* lorsqu'ils s'y recombinent : alors que différentes cassettes de résistance ont pu être testées avec ces plasmides, la surexpression de protéines fluorescentes à partir de ceux-ci a toujours échoué. En effet, même si les constructions plasmidiques montraient un signal fluorescent fort lorsque clonées dans *E. coli*, ce signal était complètement aboli chez *M. florum*. Étonnement, les PCR de vérification montraient que les plasmides avaient perdu la cassette responsable de l'expression de la protéine fluorescente. Une explication possible est que les promoteurs choisis afin de surexprimer les protéines amenaient un recrutement trop élevé de la machinerie transcriptionnelle à l'*oriC* dupliqué, ce qui nuisait ou empêchait le processus de réplication du chromosome. Les cellules recueillies et analysées suite à la transformation étaient donc majoritairement celles qui avaient procédé à l'exclusion de la cassette de fluorescence par un évènement de recombinaison. Certes, cette hypothèse reste à vérifier, mais les résultats récents d'expression de protéines fluorescentes à partir d'un transposon Tn*4001* contenant le gène de sélection *aadA1* semblent pointer dans cette direction (Chamberland et al., résultats non publiés). Ce problème de recombinaison avec le chromosome pourrait en principe être contourné en utilisant des plasmides *oriC* développés

chez d'autres espèces de Mollicutes apparentées. Cependant, nous avons montré que la transformation de *M. florum* avec des plasmides développés à partir de l'*oriC* de *M. mycoides*, *M. capricolum* ou *Spiroplasma citri* ne permettaient pas d'obtenir de transformant viable (Matteau et al., 2017). Étonnamment, l'opération inverse, c'est-à-dire la transformation de *M. capricolum* avec des plasmides portant l'*oriC* de *M. florum*, résulte en des colonies viables (Labroussaa et al., 2016). Une autre alternative serait d'utiliser des plasmides reconnus pour leur capacité à se répliquer chez une grande variété de bactéries, incluant des bactéries phylogénétiquement proches de *M. florum*. Cependant, les efforts effectués jusqu'à maintenant en ce sens se sont avérés infructueux. L'utilisation de plasmides naturels retrouvés chez *M. mycoides* (King and Dybvig, 1994) et *Mycoplasma yeatsii* (Breton et al., 2012) pourrait également s'avérer une avenue fort intéressante à explorer.

Dans certaines circonstances, la recombinaison des plasmides à l'*oriC* peut s'avérer avantageuse. En effet, ce phénomène fut récemment exploité afin d'ajouter au chromosome de *M. florum* les éléments génétiques requis pour permettre son clonage entier dans la levure *S. cerevisiae* (Baby et al., 2017). Alors que la levure représente un outil extrêmement puissant pour la modification des génomes, les modifications effectuées ne peuvent seulement être testées que si la transplantation vers une cellule réceptrice est possible. Malheureusement, pour beaucoup de génomes clonés dans la levure, cette étape est encore très limitante. En effet, la transplantation de génome n'a jusqu'à présent été démontrée qu'avec un nombre très restreint d'espèces de Mollicutes, tous phylogénétiquement très proches de la souche réceptrice *M. capricolum* sous-espèce *capricolum* (Labroussaa et al., 2019). De ce fait, cette méthode représente à l'heure actuelle le principal goulot d'étranglement des techniques modernes de la génomique synthétique et demeure, sous plusieurs angles, largement incomprise. Nonobstant son caractère encore mystérieux, certains facteurs influençant l'efficacité de la transplantation de génome ont tout de même pu être identifiés au fil des ans. Parmi ceux-ci, on compte notamment la présence d'éléments génétiques mobiles, l'expression de nucléases membranaires ou extracellulaires ainsi que la présence de systèmes de reconnaissance du soi et du non-soi, tels que les systèmes de restriction-modification (Gibson et al., 2010a; Labroussaa et al., 2016, 2019;

Lartigue et al., 2007, 2009). Outre ces facteurs, la distance phylogénétique entre la bactérie réceptrice et le génome à transplanter semble jouer un rôle déterminant dans la réussite de la procédure (Baby et al., 2017; Labroussaa et al., 2016, 2019). Curieusement, *M. florum* représente l'espèce la plus distante phylogénétiquement de *M. capricolum* pour laquelle le génome demeure transplantable. Cependant, cette distance phylogénétique relativement importante affecte grandement l'efficacité de la transplantation : seulement quelques colonies viables sont généralement obtenues par expérience de transplantation. Ceci suggère que la similarité entre les protéines de la cellule receveuse et la cellule donneuse serait décisive pour l'initialisation du génome donneur, par exemple pour les processus initiaux de réplication et ségrégation du génome acquis. Cette similarité protéique n'est toutefois pas suffisante pour expliquer les résultats obtenus. En effet, même si la transplantation du génome de *M. florum* permet d'obtenir des transplants viables, la transformation de *M. florum* avec des plasmides portant l'*oriC* de *M. capricolum* sous espèce *capricolum* résulte en un échec (Labroussaa et al., 2016; Matteau et al., 2017). Certains facteurs encore méconnus semble donc conférer à *M. capricolum* sous espèce *capricolum* une capacité unique à répliquer et partitionner adéquatement les plasmides *oriC* et les génomes provenant d'espèces étroitement apparentées. Or, le succès de la transplantation de génome ne semble pas entièrement reposer sur cette capacité. En effet, *M. capricolum* sous espèce *capricolum* a été montrée comme capable de répliquer les plasmides portant l'*oriC* de *S. citri*, alors que la transplantation du génome de ce spiroplasme n'est pas possible (Labroussaa et al., 2016; Lartigue et al., 2003). D'autres facteurs tels que la compatibilité entre la machinerie transcriptionnelle et traductionnelle de la cellule hôte et le génome transplanté – et donc la capacité à synthétiser les premiers ARN et les premières protéines encodés par celui-ci – ainsi que la capacité de la cellule hôte à adopter la morphologie conférée par le génome de la cellule donneuse pourraient également occuper une place importante dans la réussite de la méthode. Une meilleure compréhension des mécanismes sous-jacents à la transplantation de génome permettrait possiblement d'étendre cette procédure à d'autres organismes au-delà de la classe des Mollicutes.

Somme toute, la stratégie de clonage de génome complet dans la levure constitue une approche très intéressante pour modifier les génomes d'organismes réfractaires aux modifications génétiques mais possède tout de même d'importantes limitations en raison du manque de connaissances en lien avec la procédure de transplantation de génome en plus du manque d'espèces réceptrices utilisables. La transplantation de génome demeure à l'heure actuelle une méthode relativement laborieuse et extrêmement sensible à différents facteurs ; il n'est d'ailleurs pas rare de n'obtenir aucun transformant de *M. florum* suite à la transplantation de son génome (Chamberland et al., résultats non publiés). De plus, les résultats récents de stabilité du génome de *M. hominis* cloné dans la levure montrent également que les génomes clonés sous forme de plasmides centromériques peuvent être sujets à d'importants réarrangements après un certain nombre de générations (Rideau et al., 2017). Cela peut constituer un problème majeur dans le cas où les modifications doivent être accomplies de manière successive. Pour arriver à une compréhension approfondie des lois qui gouvernent la programmation des génomes, il nous faudra développer des approches complémentaires qui permettront d'explorer les différentes architectures génomiques possibles beaucoup plus facilement et efficacement. L'utilisation de sérine intégrases hautement spécifiques pourrait constituer un avantage dans ce contexte (Brown et al., 2011; Stark, 2017) (voir section 5.6).

5.4 Caractérisation intégrative et annotation génomique expérimentale

Le troisième objectif à atteindre dans le cadre de ma thèse consistait à mesurer et intégrer différents aspects physiques, physiologiques et moléculaires de *M. florum* afin de procéder à une caractérisation intégrative sans précédent de cet organisme quasi minimal. En plus d'augmenter nos connaissances fondamentales sur le fonctionnement de ce microorganisme, cette caractérisation visait à établir une base de résultats expérimentaux pour le développement d'un GEM décrivant le métabolisme de cette bactérie (voir section 5.5). Plus spécifiquement, j'ai utilisé différentes techniques et approches afin de déterminer la température optimale et la cinétique de croissance de *M. florum*, son temps de doublement, son diamètre cellulaire, son volume, sa masse cellulaire sèche, sa masse totale, ainsi que les fractions cellulaires de ses

principales macromolécules. J'ai également brossé le premier portait global du transcriptome et du protéome de cette bactérie à l'aide d'expériences de 5' -RACE, RNA-seq et MS/MS. Ceci a entre autres permis de reconstruire tout près de 400 TU distinctes, ainsi que de quantifier les niveaux de transcription et d'expression de chacun des gènes prédits de *M. florum*. En intégrant les données de RNA-seq et de MS/MS avec les fractions macromoléculaires de la cellule, il a aussi été possible d'estimer les abondances absolues de chacune des espèces moléculaires d'ARN et de protéines, d'estimer leur nombre total dans la cellule, et d'évaluer l'importance relative des différentes catégories fonctionnelles prédites chez cet organisme. Ces données ont également été utilisées afin d'estimer le nombre total de ribosomes, d'ARN polymérases et de facteurs σ^{70} typiquement retrouvés dans une cellule de *M. florum*.

Les méthodes utilisées pour mesurer les fractions des principales macromolécules de *M. florum* reposaient sur des techniques ou des trousses commerciales sélectionnées pour leur très grande sensibilité. Par conséquent, la majeure partie de la variabilité observée dans les quantifications provenait de l'étape de normalisation de la masse mesurée par le nombre de cellules analysées. En effet, avec les méthodes traditionnelles de quantification comme le décompte d'unités formatrices de colonies (UFC), il n'est pas rare d'observer des variations du compte de cellules par unité de volume de l'ordre de deux fois, voire plus. Cette variation peut avoir des effets dramatiques sur la détermination de la masse des constituants par cellule, et plus spécialement sur la détermination des fractions de masse cellulaire sèche occupées par ceux-ci. Puisque ces fractions dépendent non seulement de la normalisation par le nombre de cellules analysées, mais également du rapport avec la masse cellulaire sèche mesurée par cellule, une variation de l'ordre de deux fois des comptes cellulaires peut faire varier les portions finales de manière très importante. Pour minimiser ce problème, la méthode de dénombrement des UFC a été remplacée par une méthode plus sensible basée sur la cytométrie en flux afin de normaliser les résultats de quantification de masse cellulaire sèche totale. De plus, nous avons répété cette quantification plus de trois fois, en utilisant chaque fois quatre réplicats biologiques distincts (mesures issues de différentes cultures de départ). Idéalement, la méthode du décompte des cellules basée sur la cytométrie en flux aurait également dû être appliquée pour la normalisation des masses mesurées

pour les différentes macromolécules. Ceci aurait probablement permis de diminuer davantage l'incertitude associée aux fractions macromoléculaires calculées. De plus, il aurait été possible de réaliser l'ensemble des quantifications à partir d'un seul et même lot de cultures de *M. florum* de départ, ce qui aurait permis de normaliser la totalité des données obtenues par exactement les mêmes données de dénombrement des cellules. Quoi qu'il en soit, les résultats présentés au Chapitre 4 semblent tout de même être une très bonne estimation des fractions macromoléculaires retrouvées chez *M. florum*, d'une part puisque ceux-ci concordent avec les fractions observées chez d'autres espèces bactériennes, mais également puisque la somme des fractions totalise presque la valeur mesurée pour la masse cellulaire sèche totale.

Récemment, notre laboratoire a proposé différentes stratégies de réduction du génome de *M. florum* à partir de données de conservation et d'essentialité des gènes obtenues par génomique comparative et par mutagenèse par transposons (Baby et al., 2018). Quoique très intéressantes, ces scénarios ne disposaient pas des données adéquates pour évaluer convenablement l'organisation transcriptionnelle des gènes de *M. florum*, conservant ainsi la totalité des régions intergéniques dans les génomes réduits proposés. L'objectif de ces prédictions n'était donc pas de proposer des organisations génomiques réduites fonctionnelles, mais bien d'explorer la composition génétique et la nature des fonctions retrouvées dans un génome minimal hypothétique basé sur celui de *M. florum*, et de comparer ces aspects avec la bactérie minimale JCVI-syn3.0. Par conséquent, les génomes réduits proposés n'étaient fort probablement pas viables, bien que ceci ne fut pas démontré (en utilisant, par exemple, des approches de génomique synthétique et de transplantation de génomes complets). Par contre, les UT présentées au Chapitre 4 offrent une opportunité particulièrement intéressante de raffiner ces prédictions et ainsi proposer des organisations génomiques potentiellement viables pour la cellule. Bien que basées en partie sur des prédictions de séquences terminatrices, ces UT couvrent plus de 90 % des gènes annotés de *M. florum*, et leur structure est grandement appuyée par les résultats de RNA-seq. De façon intéressante, les gènes non inclus dans les UT montrent un niveau de transcription très faible comparativement au reste des gènes, ce qui soulève des questions par rapport à leur importance chez *M. florum.* Ces gènes sont-ils exprimés à un très

bas niveau pour des raisons fonctionnelles, c'est-à-dire pour n'avoir qu'un niveau limité de protéines associées ? Ou peut-être ces gènes sont-ils exprimés seulement dans certaines conditions spécifiques rencontrées dans l'habitat naturel de *M. florum* ? Quoi qu'il en soit, étant exclus des UT, la technique de 5' -RACE utilisée pour identifier les sites d'initiation de la transcription n'a pas permis d'identifier de promoteur pour ces gènes. Certains de ces gènes pourraient en fait être exprimés par le biais de terminateurs partiellement inefficaces (Lalanne et al., 2018), ce qui pourrait expliquer leur faible taux de transcription. Toutefois, puisque l'identité des terminateurs inclus dans les UT de *M. florum* ne repose actuellement que sur des prédictions, l'évaluation de cette hypothèse s'avère plutôt difficile. Les prédictions de séquences terminatrices pourraient éventuellement être remplacées par des données obtenues par des méthodes comme le Rend-seq (Lalanne et al., 2018). Ceci permettrait non seulement de valider ces prédictions, mais également d'évaluer l'importance du mécanisme de fuite de tige terminatrice chez *M. florum*.

En plus d'avoir permis la reconstruction des UT de *M. florum*, les résultats de 5' -RACE générés au cours de mon doctorat ont pu mettre en évidence la séquence du promoteur consensus de cette bactérie. Cette séquence, retrouvée à plus de 400 sites dans le chromosome de *M. florum*, est typique des séquences reconnues par les facteurs σ^{70} bactériens. Toutefois, l'absence de conservation au niveau de la boîte -35 laisse supposer que celle-ci ne serait peut-être pas essentielle pour l'initiation de la transcription chez *M. florum*. En fait, l'absence d'une boîte -35 pourrait être compensée par la présence d'une boîte -10 étendue, comme observée chez d'autres espèces de Mollicutes (Fisunov et al., 2016; Lloréns-Rico et al., 2015). L'importance de ces éléments pour l'initiation de la transcription chez *M. florum* pourrait être validée en utilisant des approches de gènes rapporteurs couplés à des librairies de promoteurs aléatoires. Ceci permettrait également de comparer la force des promoteurs testés aux intensités des signaux observés par 5' -RACE. Dans ce contexte, des gènes codants pour des protéines fluorescentes pourraient très bien agir à titre de gènes rapporteurs chez *M. florum*, d'autant plus que nous avons récemment réussi à exprimer ce type de protéines à l'aide d'un transposon synthétique basé sur le transposon Tn*4001* (Chamberland et al., résultats non publiés).

Curieusement, un bon nombre (~200) de sites d'initiation de la transcription (TSS) observés chez *M. florum* n'étaient pas associés à une séquence promotrice ressemblant au promoteur consensus identifié. De plus, l'intensité de ces TSS était généralement beaucoup plus faible que les TSS associés au promoteur consensus, et ces TSS étaient majoritairement retrouvés à l'intérieur des séquences codantes prédites du génome. Pour l'instant, l'hypothèse la plus probable pour expliquer la présence de ces TSS est qu'ils représentent soit des artefacts expérimentaux (issus d'une dégradation partielle des ARNm, par exemple) ou soit une forme de bruit transcriptionnel causé par le pourcentage A-T particulièrement élevé (~73 %) du génome de *M. florum* (Lloréns-Rico et al., 2016). L'hypothèse que ces TSS soient le résultat de la reconnaissance de séquences promotrices par un second facteur σ s'avère peu probable considérant qu'un seul facteur σ (σ^{70}) est prédit chez *M. florum*, et qu'aucun motif n'a pu être obtenu suite à l'analyse des séquences placées en amont de ces sites. De manière plus intéressante, une partie (~100) des TSS associés au promoteur consensus de *M. florum* se sont également avérés positionnés à l'intérieur de séquences codantes prédites, soit dans le même sens ou dans le sens inverse à celles-ci. Alors que les caractéristiques de ces TSS sont dans ce cas-ci très similaires aux TSS identifiés dans les régions intergéniques du génome (mis à part leur intensité en moyenne plus basse), leur importance chez cette bactérie demeure pour l'instant inconnue. Il est possible que ces TSS soient impliqués dans la transcription de gènes ou autres éléments génétiques (des ARNnc, par exemple) positionnés en aval de ceux-ci. Toutefois, les gènes correctement orientés pour qu'un ou plusieurs TSS intragéniques puissent contribuer à leur expression se trouvent, en grande majorité, également associés à un TSS intergénique d'intensité supérieure. Similairement aux TSS non associés au promoteur consensus de *M. florum*, il se peut que ces TSS soient tout simplement le fruit de bruit transcriptionnel causé par le biais G-C du génome. Étant donné que ces TSS sont généralement faibles, leur impact sur la transcription normale des gènes est probablement négligeable. Il demeure néanmoins possible que ceux-ci aient une ou plusieurs fonctions associées selon leur contexte et environnement génomique immédiat. Ces TSS intragéniques pourraient, par exemple, être impliqués dans la transcription de formes raccourcies ou antisens d'ARNm utilisées pour produire des isoformes alternatives de protéines (Miravet-Verde et al., 2019; Mouilleron et al., 2016; Vanderperre et

al., 2012, 2013). Alors que l'existence de ces formes alternatives de protéines est de plus en plus reconnue chez les eucaryotes, celles-ci demeurent largement inexplorées chez les procaryotes. Théoriquement, les expériences de MS/MS possèdent la capacité de détecter ces formes alternatives de protéines. Toutefois, si leur expression est faible, ou si elles sont exprimées à partir du même cadre de lecture que la séquence codante classique qu'elles chevauchent, leur identification peut être particulièrement difficile. De plus, l'identification de ces formes alternatives de protéines par MS/MS requiert généralement l'utilisation de bases de données traduites à partir des six cadres de lecture possibles du génome (*6-frame database*). Ce type de base de données augmente de manière fulgurante le nombre de possibilités d'identification des peptides associés aux spectres et réduit par le fait même la confiance envers les protéines détectées (Deutsch et al., 2008; Veltri, 2008). Cet aspect peut toutefois être atténué par une étape de préfiltration des cadres de lecture ouverts prédits, en utilisant différents critères de sélection comme la longueur ou la localisation de ceux-ci. Jusqu'à présent, les efforts d'identification de protéines alternatives chez *M. florum* en utilisant des données de MS/MS se sont avérés infructueux. Toutefois, différentes approches de filtration des données (et bases de données) sont présentement en cours afin d'explorer cet aspect. La réalisation d'expériences de profilage ribosomique (Ribo-seq) pourrait également contribuer à la détection de ces protéines alternatives si, bien entendu, protéines alternatives il y a chez *M. florum*.

5.5 Base expérimentale pour le développement d'un GEM

Les GEMs constituent des outils particulièrement efficaces pour regrouper l'ensemble de l'information génomique disponible pour un organisme donné et explorer ses capacités métaboliques (Durot et al., 2009; King et al., 2016). En représentant l'ensemble du réseau métabolique sous forme mathématique, les GEMs permettent d'effectuer différents types de prédictions phénotypiques utiles pour la compréhension des systèmes biologiques et pour la planification d'expériences plus coûteuses en temps et en ressources (O'Brien et al., 2015). L'impact de délétions génétiques sur le temps de doublement, par exemple, peut être évalué de manière *in silico* à l'aide des GEMs. De plus, un nombre croissant d'algorithmes sont

développés afin de faciliter les efforts effectués en contexte de la génomique synthétique. Ces algorithmes permettent, par exemple, de prédire l'impact de réductions génomiques sur le comportement cellulaire, ou d'aider à la conception de génomes significativement réduits ou modifiés (Chalkley et al., 2019; Rees et al., 2018; Wang and Maranas, 2018). L'utilisation des GEMs constituerait donc un atout pour planifier la réduction du génome de *M. florum* à partir des données d'essentialité, de conservation et d'organisation des gènes (UT). Ces modèles pourraient également être utilisés afin de prédire l'impact de réorganisations génomiques importantes avant d'entreprendre des efforts de synthèse et d'assemblage, ce qui permettrait de réduire le nombre de possibilités à tester avec des approches de génomique synthétique.

La fiabilité des prédictions générées par les GEMs dépend grandement du niveau d'exactitude de plusieurs contraintes définies dans le modèle telles que le taux de croissance, la composition macromoléculaire de la cellule et la masse sèche de la cellule (Feist and Palsson, 2010; Lachance et al., 2019b). Les données de temps de doublement et de composition de la biomasse présentées au Chapitre 4 seraient donc particulièrement adaptées pour définir plusieurs de ces contraintes et permettre le développement d'un GEM de haute qualité pour *M. florum*, projet actuellement mené par Jean-Christophe Lachance au laboratoire. De plus, les données d'expression obtenues par RNA-seq et MS/MS et les données d'essentialité des gènes précédemment publiées (Baby et al., 2018) pourraient être utilisées afin de valider les flux métaboliques prédits par le modèle et d'identifier les cas où celui-ci de s'accorde pas avec les résultats expérimentaux. Cela permettrait de corriger le modèle au besoin, et potentiellement de raffiner les connaissances du réseau métabolique de *M. florum*. Alors que les données d'essentialité précédemment obtenues se basaient sur des expériences de mutagenèse par transposons possédant une résolution relativement faible (une insertion à toutes les ~280 pb), notre laboratoire travaille actuellement à développer une méthode de mutagenèse à haute densité basée sur le transposon Tn*4001*. Certains résultats préliminaires semblent d'ailleurs indiquer qu'il serait possible d'obtenir plus de 10 000 mutants d'insertion par transformation à l'aide de ce transposon.

La composition du milieu de culture doit également être définie dans le modèle afin de pouvoir générer des prédictions phénotypiques. Pour l'instant, cette composition est majoritairement hypothétique puisque le milieu de culture utilisé (ATCC 1161) contient plusieurs éléments non définis comme du sérum et de l'extrait de levure. Nous travaillons actuellement à développer un milieu de culture entièrement défini où il sera facile de faire varier les composantes afin de tester leur impact sur la croissance et le métabolisme de *M. florum*. Ce milieu de culture pourra entre autres être utilisé afin de tester la capacité de *M. florum* à cataboliser et croître sur différents sucres, et ensuite comparer ces résultats aux prédictions faites par le modèle. Dans ce contexte, il serait également très intéressant de mesurer les taux de consommation des différents sucres et déterminer l'impact de la variation de ces taux sur le temps de doublement de la cellule. Alternativement, il serait aussi possible de quantifier les taux de production des produits de fermentation prédits (acétate, lactate) en fonction de la concentration en sucre du milieu et mesurer les temps de doublement associés. Ceci pourrait être accompli à l'aide d'expériences de suivi de croissance dans lesquelles les concentrations cellulaires seraient mesurées par cytométrie en flux et les concentrations des métabolites par des méthodes de chromatographie en phase liquide à haute performance (HPLC). Les résultats obtenus permettraient de définir certains paramètres énergétiques importants dans le modèle, comme le *growth-associated maintenance* (GAM) et le *non-growth associated maintenance* (NGAM) associés à chacun des sucres. Les valeurs observées pour ces paramètres permettraient possiblement d'expliquer pourquoi *M. florum* possède un temps de doublement aussi faible comparativement à plusieurs autres Mollicutes. Une valeur de NGAM très faible, par exemple, signifierait que la plupart de l'énergie produite par *M. florum* est investie dans la croissance et non pas à des processus de maintenance. À titre d'exemple, *M. pneumoniae* possède un NGAM très élevé, ce qui explique (en partie) sa croissance extrêmement lente en laboratoire (Wodke et al., 2013).

Alors que les GEMs étaient initialement restreints à la modélisation des réseaux métaboliques des cellules, ceux-ci permettent désormais la modélisation de processus cellulaires au-delà du métabolisme, comme les mécanismes responsables de la synthèse de la machinerie d'expression des gènes (*ME-model*) (Lerman et al., 2012; Lloyd et al., 2018; Thiele et al., 2012). Ce type de

modèle permet non seulement de prédire les flux métaboliques optimaux pour la cellule selon les conditions fournies, mais également de prédire la composition optimale en protéines afin de soutenir ces flux métaboliques. Par conséquent, le développement de ce type de modèle serait une excellente façon d'intégrer les données d'abondance des protéines obtenues par MS/MS en contexte de la physiologie et du métabolisme de *M. florum*. De plus, puisque ce type de modèle permet d'investiguer l'impact sur la croissance de différentes contraintes comme la température et le volume cellulaire, il serait possible de comparer les valeurs mesurées (voir Chapitre 4) aux valeurs prédites comme étant optimales par le modèle (Chen et al., 2017; Liu et al., 2014). La réalisation d'expériences de profilage ribosomique pourrait également s'avérer particulièrement intéressante à effectuer. En effet, l'intégration de ce type de données dans le modèle permettrait de mettre en relation les taux de synthèse des protéines et leur abondance dans la cellule (Brar and Weissman, 2015; Ingolia et al., 2012). Advenant que les processus de transcription soient éventuellement modélisés dans les *ME-model*, l'abondance des transcrits d'ARN estimée par RNA-seq pourrait également y être intégrée (voir Chapitre 4), tout comme les taux de dégradation des ARN. Toutefois, les protocoles à haut débit utilisés pour mesurer la demi-vie des transcrits dépendent de la sensibilité de l'ARN polymérase à la rifampicine afin d'inhiber le processus d'initiation de la transcription (Chen et al., 2015; Selinger et al., 2003). Or, il s'avère que l'ARN polymérase des Mollicutes est insensible à la rifampicine en raison d'une substitution dans la séquence en acides aminés de la protéine (Gaurivaud et al., 1996). Pour être en mesure d'effectuer ce type d'expérience, il faudrait tout d'abord vérifier que la réversion de cette mutation restaure le phénotype de sensibilité à la rifampicine chez *M. florum*.

Au fur et à mesure que la diversité des processus cellulaires modélisés par les GEMs s'accroît, ces derniers sont appelés à jouer un rôle important dans de nouvelles sphères d'application, par exemple pour la reprogrammation de génomes entiers. Ceux-ci devront toutefois continuer de reposer sur des bases expérimentales solides, afin que les prédictions générées soient le plus près possible de la réalité. En intégrant un nombre particulièrement élevé de données expérimentales dans le contexte d'un GEM développé pour l'organisme quasi minimal *M. florum*, nous pensons qu'il sera possible d'atteindre une compréhension globale de la cellule

jusqu'à présent inégalée. Ceci permettra de mieux prévoir l'impact de réorganisations génomiques complexes sur la cellule, et ainsi d'accélérer le développement de génomes créés sur mesure.

5.6 Recodage du génome de *M. florum*

Compte tenu des limitations de la méthode de clonage et de transplantation de génomes complets à partir de la levure, de nouvelles approches doivent être développées afin d'explorer efficacement les organisations et compositions possibles du génome de *M. florum*. Une des approches possibles serait d'utiliser les propriétés uniques des sérine intégrases afin d'échanger des fragments génomiques avec des fragments synthétiques assemblés dans la bactérie hôte *E. coli* (Figure 5.1). En effet, différentes sérine intégrases provenant de phages ont été démontrées comme capables de catalyser la recombinaison entre des sites précis (*attB* et *attP*) de manière hautement efficace et spécifique (Brown et al., 2011; Stark, 2017). Des résultats préliminaires obtenus par une autre étudiante au laboratoire indiquent que ces intégrases seraient fonctionnelles chez *M. florum* (Chamberland et al., résultats non publiés). De plus, la compatibilité entre les sites reconnus par ces intégrases peut être contrôlée à l'aide de dinucléotides centraux, ce qui offre plusieurs possibilités de recombinaisons orthologues par intégrase. Il serait donc possible, à partir du génome de *M. florum* cloné dans la levure, d'ajouter des sites de recombinaison *attB* orthologues à tous les ~40 kb du chromosome, puis de procéder à une première étape de transplantation afin de générer une souche de *M. florum* portant ce chromosome modifié. Parallèlement, des fragments synthétiques flanqués de sites *attP* pourraient être assemblés dans la bactérie *E coli* en utilisant différentes méthodes comme l'assemblage de type Gibson (Gibson, 2011) ou le GoldenGate (Engler et al., 2008). Une fois assemblés, ces fragments synthétiques pourraient être transférés à *M. florum* par conjugaison, où l'expression d'une serine intégrase stimulerait l'échange de fragments. L'ajout d'un gène de sélection (*tetM* ou *aadA1*, par exemple) aux fragments synthétiques permettrait de sélectionner seulement les cellules ayant réalisé l'échange de cassettes. Cette approche modulaire pourrait être accomplie en parallèle pour l'ensemble des sections génomiques de 40 kb flanqués par les

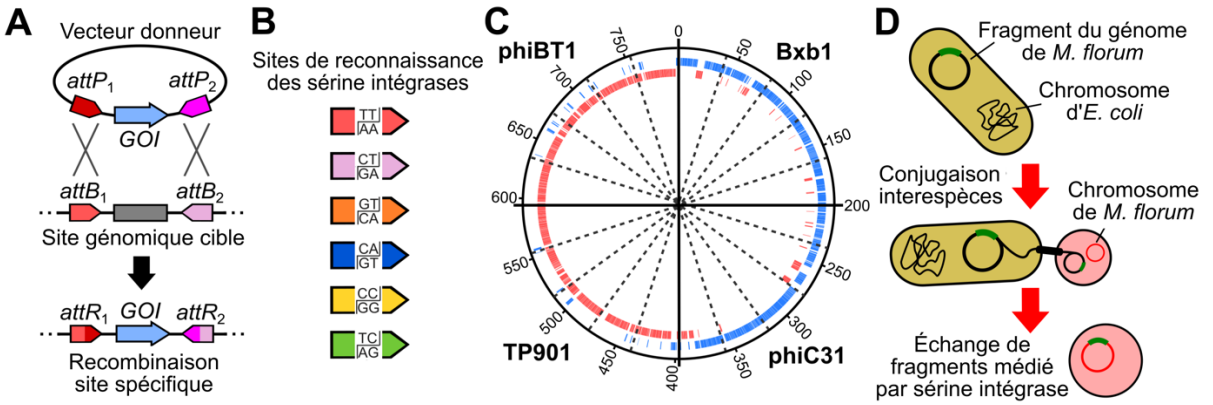

Figure 5.1. Approche modulaire d'ingénierie du génome de *M. florum*. A) Représentation schématique d'un évènement d'échange de cassette médié par l'action de sérine intégrases. L'échange s'effectue grâce à la recombinaison entre les sites *attP* et *attB*. *GOI* : gène d'intérêt. B) Le dinucléotide central des sites *attP* et *attB* assure l'orthogonalité des recombinaisons. C) Le chromosome de *M. florum* peut être divisé en quatre quadrants principaux, eux-mêmes subdivisés en sections d'environ 40 kb, pour un total de 20 sections chromosomiques. Chaque quadrant principal sera modifié à l'aide d'une sérine intégrase différente (Bxb1, phiC31, TP901, phiBT1), et les subdivisions seront flanquées de sites *attB* uniques assurant l'orthogonalité des recombinaisons. D) Les sections synthétiques de 40 kb peuvent être assemblées dans la bactérie *E. coli*, puis transférées dans *M. florum* afin de remplacer la séquence native.

sites *attB*. Advenant que les génomes modifiés ne soient pas viables, la composition ou l'organisation des fragments synthétiques pourrait être réévaluée (à l'aide des GEMs ou d'expériences de mutagenèse par transposons, par exemple) et rapidement modifiée dans *E. coli*, puis la procédure pourrait être répétée jusqu'à en comprendre la cause de l'échec. Les modifications pourraient ensuite entre combinées de manière hiérarchique à l'intérieur de souches de *M. florum*, pour finalement être rassemblées à l'intérieur d'une seule et même cellule.

L'approche d'ingénierie du génome de *M. florum* proposée dans cette section vise à faciliter et accélérer la boucle de conception, construction et test utilisée en génomique synthétique afin de faire émerger les règles de programmation des génomes. Cette approche pourrait, dans un premier temps, être utilisée afin de minimiser le génome de *M. florum*. Dans ce contexte, les fragments synthétiques assemblés dans la bactérie *E. coli* seraient conçus en fonction des données d'essentialité, de conservation et d'organisation des gènes (UT) précédemment

obtenues, mais également à l'aide des prédictions provenant d'un GEM spécifiquement reconstruit pour *M. florum*. Évidemment, cette approche ne se limite pas qu'aux efforts de minimisation; différents aspects et caractéristiques propres au génome de *M. florum* pourraient être explorés. Les fragments synthétiques à échanger pourraient, par exemple, être recodés afin d'évaluer l'impact d'un changement important au niveau du biais en G-C, ou bien éliminer l'utilisation de certains codons, comme précédemment démontré avec la bactérie *E. coli* (Hecht et al., 2017). L'importance de certains éléments comme les TSS intragéniques ou les ARNnc pourrait également être investiguée. Ultimement, cette approche pourrait être utilisée afin de réorganiser l'ensemble du génome de sorte à regrouper les gènes de fonctions similaires, éliminer les gènes non importants, reconcevoir les unités transcriptionnelles et adapter les réseaux de régulation en fonction des conditions retrouvées en laboratoire. Dans ce cas, on assisterait peut-être à la naissance de la toute première cellule artificielle réellement adaptée aux conditions de laboratoire, et par conséquent, à la première vraie plateforme de prototypage de génomes.

5.7 Conclusion et perspectives

Dans toute discipline d'ingénierie, le succès du processus de création dépend de trois grands facteurs : les ressources, les outils et les connaissances. Même si vous avez à votre disposition les meilleurs outils pour la fabrication d'un avion, sans les connaissances requises, les chances que celui-ci vole sont très minces… Puisque la biologie synthétique n'a pas encore atteint la maturité des disciplines comme le génie mécanique ou le génie électrique, celle-ci ne dispose pour l'instant que des ressources ; les connaissances du fonctionnement global des cellules ne sont encore que fragmentaires, et certains des outils les plus importants pour la programmation des génomes sont toujours manquants. Par conséquent, la biologie synthétique est encore à ce que la fin du 19^{e} siècle fût pour l'aviation ; la plupart des « avions » développés en biologie synthétique n'arrivent tout simplement pas à décoller, ou bien s'écrasent de manière inattendue. Dans les meilleurs des cas, ceux-ci réussissent à planer, tout au plus. Bien que très simpliste, cette analogie résume assez bien l'état actuel de nos capacités dans ce domaine.

Afin que la biologie synthétique puisse prendre son véritable envol, il va falloir d'une part acquérir les connaissances et les outils nécessaires, mais également être en mesure de construire et évaluer les prototypes proposés de manière efficace. Malheureusement pour nous (ou heureusement), les cellules ont perfectionné leur art pendant des millions d'années ; à l'heure actuelle, tenter de comprendre et reprogrammer l'entièreté d'une cellule eucaryote serait l'équivalent d'envoyer un avion à réaction dans les années 1800 et demander à ce qu'on la démonte et reprogramme pour en faire une locomotive à vapeur… ce serait complètement futile. Par conséquent, aussi bien débuter les efforts de reprogrammation cellulaire avec des cellules simples, même très simples. Les bactéries appartenant à la classe des Mollicutes, par exemple, représentent un excellent point de départ pour cette tâche. Par leur simplicité remarquable, celles-ci offrent une opportunité sans pareil de décortiquer le fonctionnement intégral d'une cellule vivante. Les travaux effectués dans le cadre de mon doctorat allaient justement en ce sens, c'est-à-dire accumuler une quantité très importante de données dans le but de mieux comprendre et définir une bactérie extrêmement simple, *M. florum*. De plus, ces travaux visaient également le développement d'outils spécifiquement conçus pour modifier le génome de cette bactérie. Ces outils serviront de base pour le développement d'approches plus sophistiquées afin d'explorer différentes compositions et organisations du génome de *M. florum*. Avec suffisamment de connaissances, ces approches pourront être utilisées afin de développer un châssis cellulaire simplifié dédié au prototypage de génomes, un outil au potentiel immense mais présentement inexistant en biologie synthétique.

Lorsque Tom Knight – un des pionniers dans le projet *M. florum* et cofondateur de Ginkgo Bioworks – présentait ces travaux portant sur le génome de *M. florum*, celui-ci décrivait souvent cet organisme comme « *an organism which absolutely no one cares about* ». En effet, cette bactérie était que très peu connue à ce moment, principalement parce que celle-ci ne causait aucune maladie ou perte économique importante. Pour résumer les travaux présentés dans cette thèse, il est possible d'affirmer que ceux-ci visaient à acquérir les connaissances et les outils moléculaires requis pour que, dans quelques années, on réfère plutôt *M. florum* comme « *an organism which absolutely everybody cares about* ». Le développement d'un châssis cellulaire

où la programmation de génome se fera de manière prévisible, efficace et reproductible aura des impacts très importants sur nos vies. En fait, lorsque Alan Turing travaillait sur la *Turing machine* vers le milieu des années 1930, celui-ci ne se doutait probablement pas que les concepts et les principes à la base de sa machine allaient un jour se retrouver au centre de nos ordinateurs personnels, téléphones, automobiles, caisses enregistreuses, partout en fait (Hopcroft, 2012)! Avec une telle plateforme, il est possible d'imaginer la création d'organismes complètement nouveaux et capables d'effectuer une variété de tâches pratiquement infinie. À l'aube d'une ère où les maladies infectieuses menacent de redevenir la principale cause mondiale de décès en raison de la multirésistance aux antibiotiques (Mobarki et al., 2019), ces organismes synthétiques pourraient jouer un rôle prédominant dans la lutte à ces infections. Malheureusement, comme toute technologie peut être utilisée à bon ou mauvais escient, il faudra rester vigilant par rapport à l'utilisation malveillante de ces technologies. Un univers de possibilités nous attend. Comme l'a si bien dit Einstein :

> *Imagination is more important than knowledge. For knowledge is limited, whereas imagination embraces the entire world, stimulating progress, giving birth to evolution.*
> *-Albert Einstein*

ANNEXE I
AUTRES PUBLICATIONS PERTINENTES

1. Precise Identification of Genome-Wide Transcription Start Sites in Bacteria by 5′-Rapid Amplification of cDNA Ends (5′-RACE)

Abstract : Transcription start sites are commonly used to locate promoter elements in bacterial genomes. TSS were previously studied one gene at a time, often through 5′-rapid amplification of cDNA ends (5′-RACE). This technique has now been adapted for high-throughput sequencing and can be used to precisely identify TSS in a genome-wide fashion for practically any bacterium, which greatly contributes to our understanding of gene regulatory networks in microorganisms.

Référence bibliographique : Matteau, D., and Rodrigue, S. (2015). Precise Identification of Genome-Wide Transcription Start Sites in Bacteria by 5′-Rapid Amplification of cDNA Ends (5′-RACE). In DNA-Protein Interactions, B.P. Leblanc, and S. Rodrigue, eds. (Springer New York), pp. 143–159.

Lien URL : https://link.springer.com/protocol/10.1007%2F978-1-4939-2877-4_9

2. Precise Identification of DNA-Binding Proteins Genomic Location by Exonuclease Coupled Chromatin Immunoprecipitation (ChIP-exo)

Abstract : DNA-binding proteins play a crucial role in all living organisms by interacting with various DNA sequences across the genome. While several methods have been used to study the interaction between DNA and proteins in vitro, chromatin immunoprecipitation followed by sequencing (ChIP-seq) has become the standard technique for identifying the genome-wide location of DNA-binding proteins in vivo. However, the resolution of standard ChIP-seq

methodology is limited by the DNA fragmentation process and presence of contaminating DNA. A significant improvement of the ChIP-seq technique results from the addition of an exonuclease treatment during the immunoprecipitation step (ChIP-exo) that lowers back-ground noise and more importantly increases the identification of binding sites to a level near to single-base resolution by effectively footprinting DNA-bound proteins. By doing so, ChIP-exo offers new opportuni- ties for a better characterization of the complex and fascinating architecture that resides in DNA-proteins interactions and provides new insights for the comprehension of important molecular mechanisms.

Référence bibliographique : Matteau, D., and Rodrigue, S. (2015). Precise Identification of DNA-Binding Proteins Genomic Location by Exonuclease Coupled Chromatin Immunoprecipitation (ChIP-exo). In DNA-Protein Interactions, B.P. Leblanc, and S. Rodrigue, eds. (Springer New York), pp. 173–193.

Lien URL : https://link.springer.com/protocol/10.1007%2F978-1-4939-2877-4_11

3. The Master Activator of IncA/C Conjugative Plasmids Stimulates Genomic Islands and Multidrug Resistance Dissemination

Abstract : Dissemination of antibiotic resistance genes occurs mostly by conjugation, which mediates DNA transfer between cells in direct contact. Conjugative plasmids of the IncA/C incompatibility group have become a substantial threat due to their broad host-range, the extended spectrum of antimicrobial resistance they confer, their prevalence in enteric bacteria and their very efficient spread by conjugation. However, their biology remains largely unexplored. Using the IncA/C conjugative plasmid pVCR94ΔX as a prototype, we have investigated the regulatory circuitry that governs IncA/C plasmids dissemination and found that the transcriptional activator complex AcaCD is essential for the expression of plasmid transfer genes. Using chromatin immunoprecipitation coupled with exonuclease digestion (ChIP-exo) and RNA sequencing (RNA-seq) approaches, we have identified the sequences recognized by

AcaCD and characterized the AcaCD regulon. Data mining using the DNA motif recognized by AcaCD revealed potential AcaCD-binding sites upstream of genes involved in the intracellular mobility functions (recombination directionality factor and mobilization genes) in two widespread classes of genomic islands (GIs) phylogenetically unrelated to IncA/C plasmids. The first class, SGI1, confers and propagates multidrug resistance in *Salmonella enterica* and *Proteus mirabilis*, whereas MGI*Vmi*1 in *Vibrio mimicus* belongs to a previously uncharacterized class of GIs. We have demonstrated that through expression of AcaCD, IncA/C plasmids specifically trigger the excision and mobilization of the GIs at high frequencies. This study provides new evidence of the considerable impact of IncA/C plasmids on bacterial genome plasticity through their own mobility and the mobilization of genomic islands.

Référence bibliographique : Carraro, N.*, Matteau, D.*, Luo, P., Burrus, V., and Rodrigue, S. (2014). The Master Activator of IncA/C Conjugative Plasmids Stimulates Genomic Islands and Multidrug Resistance Dissemination. PLoS Genet. *10*, e1004714.

* These authors contributed equally to this work.

Lien URL : https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004714

4. Unraveling the regulatory network of IncA/C plasmid mobilization: When genomic islands hijack conjugative elements

Abstract : Conjugative plasmids of the A/C incompatibility group (IncA/C) have become substantial players in the dissemination of multidrug resistance. These large conjugative plasmids are characterized by their broad host-range, extended spectrum of antimicrobials resistance, and prevalence in enteric bacteria recovered from both environmental and clinical settings. Until recently, relatively little was known about the basic biology of IncA/C plasmids, mostly because of the hindrance of multidrug resistance for molecular biology experiments. To circumvent this issue, we previously developed pVCR94ΔX, a convenient prototype that codes for a reduced set of antibiotic resistances. Using pVCR94ΔX, we then characterized the

regulatory pathway governing IncA/C plasmid dissemination. We found that the expression of roughly 2 thirds of the genes encoded by this plasmid, including large operons involved in the conjugation process, depends on an FlhCD-like master activator called AcaCD. Beyond the mobility of IncA/C plasmids, AcaCD was also shown to play a key role in the mobilization of different classes of genomic islands (GIs) identified in various pathogenic bacteria. By doing so, IncA/C plasmids can have a considerable impact on bacterial genomes plasticity and evolution.

Référence bibliographique : Carraro, N.*, Matteau, D.*, Burrus, V., and Rodrigue, S. (2015). Unraveling the regulatory network of IncA/C plasmid mobilization: When genomic islands hijack conjugative elements. Mob. Genet. Elements 34–38.

* These authors contributed equally to this work.

Lien URL : https://www.tandfonline.com/doi/full/10.1080/2159256X.2015.1045116

5. Transfer activation of SXT/R391 integrative and conjugative elements: unraveling the SetCD regulon

Abstract : Integrative and conjugative elements (ICEs) of the SXT/R391 family have been recognized as key drivers of antibiotic resistance dissemination in the seventh-pandemic lineage of *Vibrio cholerae*. SXT/R391 ICEs propagate by conjugation and integrate site-specifically into the chromosome of a wide range of environmental and clinical *Gammaproteobacteria*. SXT/R391 ICEs bear *setC* and *setD*, two conserved genes coding for a transcriptional activator complex that is essential for activation of conjugative transfer. We used chromatin immunoprecipitation coupled with exonuclease digestion (ChIP-exo) and RNA sequencing (RNA-seq) to characterize the SetCD regulon of three representative members of the SXT/R391 family. We also identified the DNA sequences bound by SetCD in MGI*Vfl*Ind1, a mobilizable genomic island phylogenetically unrelated to SXT/R391 ICEs that hijacks the conjugative machinery of these ICEs to drive its own transfer. SetCD was found to bind a 19-bp sequence

that is consistently located near the promoter −35 element of SetCD-activated genes, a position typical of class II transcriptional activators. Furthermore, we refined our understanding of the regulation of excision from and integration into the chromosome for SXT/R391 ICEs and demonstrated that *de novo* expression of SetCD is crucial to allow integration of the incoming ICE DNA into a naive host following conjugative transfer.

Référence bibliographique : Poulin-Laprade, D.*, Matteau, D.*, Jacques, P.-É., Rodrigue, S., and Burrus, V. (2015). Transfer activation of SXT/R391 integrative and conjugative elements: unraveling the SetCD regulon. Nucleic Acids Res. *43*, 2045–2056.

* These authors contributed equally to this work.

Lien URL : https://academic.oup.com/nar/article/43/4/2045/2411527

BIBLIOGRAPHIE

Abdel-Ghany, S.E., Hamilton, M., Jacobi, J.L., Ngam, P., Devitt, N., Schilkey, F., Ben-Hur, A., and Reddy, A.S.N. (2016). A survey of the sorghum transcriptome using single-molecule long reads. Nat. Commun. *7*, 11706.

Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Ech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. *46*, W537–W544.

Agarwal, K.L., Büchi, H., Caruthers, M.H., Gupta, N.K., Khorana, H.G., Klbppe, K., Kumar, A., Ohtsuka, E., RajBhandary, U.L., van de Sande, J.H., et al. (1970). Total synthesis of the structural gene for an alanine transfer ribonucleic acid from yeast. Nature *227*, 27–34.

Alper, H., Cirino, P., Nevoigt, E., and Sriram, G. (2010). Applications of synthetic biology in microbial biotechnology. J. Biomed. Biotechnol. *2010*, 1–2.

Altelaar, A.F.M., Munoz, J., and Heck, A.J.R. (2013). Next-generation proteomics: Towards an integrative view of proteome dynamics. Nat. Rev. Genet. *14*, 35–48.

Alwine, J.C., Kemp, D.J., and Stark, G.R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc. Natl. Acad. Sci. U. S. A. *74*, 5350–5354.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Andersen, J.S., and Mann, M. (2000). Functional genomics by mass spectrometry. FEBS Lett. *480*, 25–31.

Andrews, S. (2018). FastQC.

Andrews, S.J., and Rothnagel, J.A. (2014). Emerging evidence for functional peptides encoded by short open reading frames. Nat. Rev. Genet. *15*, 193–204.

Andrianantoandro, E., Basu, S., Karig, D.K., and Weiss, R. (2006). Synthetic biology: new engineering rules for an emerging discipline. Mol. Syst. Biol. *2*, 2006.0028.

Annaluru, N., Muller, H., and Mitchell, L. (2014). Total Synthesis of a functional designer Eukaryotic chromosome. Science (80-.). *344*, 55–58.

Arkin, A. (2008). Setting the standard in synthetic biology. Nat. Biotechnol. *26*, 771–774.

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. Nat. Genet. *25*, 25–29.

Aslam, B., Basit, M., Nisar, M.A., Khurshid, M., and Rasool, M.H. (2017). Proteomics: Technologies and their applications. J. Chromatogr. Sci. *55*, 182–196.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. a, Tomita, M., Wanner, B.L., and Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol. *2*, 1–11.

Babarinde, I.A., Li, Y., and Hutchins, A.P. (2019). Computational Methods for Mapping, Assembly and Quantification for Coding and Non-coding Transcripts. Comput. Struct. Biotechnol. J. *17*, 628–637.

Baby, V., Labroussaa, F., Brodeur, J., Matteau, D., Gourgues, G., Lartigue, C., and Rodrigue, S. (2017). Cloning and transplantation of the Mesoplasma florum genome. ACS Synth. Biol.

Baby, V., Lachance, J.-C., Gagnon, J., Lucier, J.-F., Matteau, D., Knight, T.F., and Rodrigue, S. (2018). Inferring the Minimal Genome of Mesoplasma florum by Comparative Genomics and Transposon Mutagenesis. MSystems *3*, e00198-17.

Baby, V., Labroussaa, F., Lartigue, C., and Rodrigue, S. (2019). Chromosomes synthétiques - Réécrire le code de la vie. Médecine/Sciences *35*, 753–760.

Bailey, T.L. (2011). DREME: Motif discovery in transcription factor ChIP-seq data. Bioinformatics *27*, 1653–1659.

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., and Zhang, J. (2013). Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. PLoS Comput. Biol. *9*, 5–12.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME Suite: Tools for motif discovery and searching. Nucleic Acids Res. *37*, 202–208.

Bainbridge, M.N., Warren, R.L., Hirst, M., Romanuik, T., Zeng, T., Go, A., Delaney, A., Griffith, M., Hickenbotham, M., Magrini, V., et al. (2006). Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics *7*, 1–11.

Bajantri, B., Venkatram, S., and Diaz-Fuentes, G. (2018). Mycoplasma pneumoniae : A Potentially Severe Infection . J. Clin. Med. Res. *10*, 535–544.

Baker, M. (2011). The next step for the synthetic genome. Nature *473*, 403–408.

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., and Kuster, B. (2007). Quantitative mass spectrometry in proteomics: a critical review. Anal. Bioanal. Chem. *389*, 1017–1031.

Barillot, E., Hupé, P., Calzone, L., Vert, J.-P., and Zinovyev, A. (2012). Experimental High-throughput Technologies for Cancer Research. In Computational Systems Biology of Cancer, Chapman, and Hall, eds. (CRC Press), p.

Basu, S., Gerchman, Y., Collins, C., Arnold, F., and Weiss, R. (2005). A synthetic multicellular system for programmed pattern formation. Nature *434*.

Bayer, T.S., and Smolke, C.D. (2005). Programmable ligand-controlled riboregulators of eukaryotic gene expression. Nat. Biotechnol. *23*, 337–343.

Becker, M.M., Graham, R.L., Donaldson, E.F., Rockx, B., Sims, A.C., Sheahan, T., Pickles, R.J., Corti, D., Johnston, R.E., Baric, R.S., et al. (2008). Synthetic recombinant bat SARS-like coronavirus is infectious in cultured cells and in mice. Proc. Natl. Acad. Sci. U. S. A. *105*, 19944–19949.

Becker, S.A., Feist, A.M., Mo, M.L., Hannum, G., Palsson, B.Ø., and Herrgard, M.J. (2007). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nat. Protoc. *2*, 727–738.

Benders, G.A., Noskov, V.N., Denisova, E.A., Lartigue, C., Gibson, D.G., Assad-Garcia, N., Chuang, R.-Y., Carrera, W., Moodie, M., Algire, M.A., et al. (2010). Cloning whole bacterial genomes in yeast. Nucleic Acids Res. *38*, 2558–2569.

Benner, C. (2019). HOMER; Finding Enriched Peaks, Regions, and Transcripts.

Bertin, C., Pau-Roblot, C., Courtois, J., Manso-Silván, L., Thiaucourt, F., Tardy, F., Le Grand, D., Poumarat, F., and Gaurivaud, P. (2013). Characterization of Free Exopolysaccharides Secreted by Mycoplasma mycoides Subsp. mycoides. PLoS One *8*.

Bertin, C., Pau-Roblot, C., Courtois, J., Manso-Silván, L., Tardy, F., Poumarat, F., Citti, C., Sirand-Pugnet, P., Gaurivaud, P., and Thiaucourt, F. (2015). Highly dynamic genomic loci drive the synthesis of two types of capsular or secreted polysaccharides within the Mycoplasma mycoides cluster. Appl. Environ. Microbiol. *81*, 676–687.

Bianco, S., Rodrigue, S., Murphy, B.D., and Gévry, N. (2015). Global Mapping of Open Chromatin Regulatory Elements by Formaldehyde-Assisted Isolation of Regulatory Elements Followed by Sequencing (FAIRE-seq). In DNA-Protein Interactions, pp. 261–272.

Bionumbers (2015). What is the macromolecular composition of the cell.

Blattner, F.R., Plunkett, G., Bloch, C.A.C.C.A., Perna, N.T.N., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., et al. (1997). The complete genome sequence of Escherichia coli K-12. Science (80-.). *277*, 1453–1462.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bolisetty, M.T., Rajadinakaran, G., and Graveley, B.R. (2015). Determining exon connectivity in complex mRNAs by nanopore sequencing. Genome Biol. *16*, 204.

Bordbar, A., Monk, J.M., King, Z.A., and Palsson, B.O. (2014). Constraint-based models predict metabolic and associated cellular functions. Nat. Rev. Genet. *15*, 107–120.

Borodina, T., Adjaye, J., and Sultan, M. (2011). A strand-specific library preparation protocol for RNA sequencing (Elsevier Inc.).

Brar, G.A., and Weissman, J.S. (2015). Ribosome profiling reveals the what, when, where and how of protein synthesis. Nat. Rev. Mol. Cell Biol. *16*, 651–664.

Breitling, R. (2010). What is systems biology? Front. Physiol. *1*, 1–5.

Breton, M., Tardy, F., Dordet-Frisoni, E., Sagne, E., Mick, V., Renaudin, J., Sirand-Pugnet, P., Citti, C., and Blanchard, A. (2012). Distribution and diversity of mycoplasma plasmids: lessons from cryptic genetic elements. BMC Microbiol. *12*, 257.

Breuer, M., Earnest, T.M., Merryman, C., Wise, K.S., Sun, L., Lynott, M.R., Hutchison, C.A., Smith, H.O., Lapek, J.D., Gonzalez, D.J., et al. (2019). Essential metabolism for a minimal cell. Elife *8*, 1–77.

Brown, W.R.A., Lee, N.C.O., Xu, Z., and Smith, M.C.M. (2011). Serine recombinases as tools for genome engineering. Methods *53*, 372–379.

Bull, A.T. (2010). The renaissance of continuous culture in the post-genomics age. J. Ind. Microbiol. Biotechnol. *37*, 993–1021.

Bumgarner, R. (2013). Overview of dna microarrays: Types, applications, and their future. Curr. Protoc. Mol. Biol. *6137*, 1–17.

Bunnik, E.M., and Le Roch, K.G. (2013). An Introduction to Functional Genomics and Systems Biology. Adv. Wound Care *2*, 490–498.

Bushmanova, E., Antipov, D., Lapidus, A., and Prjibelski, A.D. (2019). RnaSPAdes: A de novo transcriptome assembler and its application to RNA-Seq data. Gigascience *8*, 1–13.

Bustin, S.A. (2000). Absolute quantification of mrna using real-time reverse transcription polymerase chain reaction assays. J. Mol. Endocrinol. *25*, 169–193.

Cambray, G., Mutalik, V.K., and Arkin, A.P. (2011). Toward rational design of bacterial genomes. Curr. Opin. Microbiol. 1–7.

Cameron, D.E., Bashor, C.J., and Collins, J.J. (2014). A brief history of synthetic biology. Nat. Rev. Microbiol. *12*, 381–390.

Carraro, N., and Burrus, V. (2014). Biology of Three ICE Families : SXT/R391, ICEBs1, and ICESt1/ICESt3. Microbiol. Spectr. *2*, MDNA3-0008–2014.

Carraro, N., Matteau, D., Luo, P., Burrus, V., and Rodrigue, S. (2014). The Master Activator of IncA/C Conjugative Plasmids Stimulates Genomic Islands and Multidrug Resistance Dissemination. PLoS Genet. *10*, e1004714.

Carraro, N., Matteau, D., Burrus, V., and Rodrigue, S. (2015). Unraveling the regulatory network of IncA/C plasmid mobilization: When genomic islands hijack conjugative elements. Mob. Genet. Elements 34–38.

Cello, J., Paul, A.V.A., and Wimmer, E. (2002). Chemical synthesis of poliovirus cDNA: Generation of infectious virus in the absence of natural template. Science (80-.). *297*, 1016–1018.

Chalkley, O., Purcell, O., Grierson, C., and Marucci, L. (2019). The genome design suite: enabling massive in-silico experiments to design genomes. BioRxiv 681270.

Chao, M.C., Abel, S., Davis, B.M., and Waldor, M.K. (2016). The design and analysis of transposon insertion sequencing experiments. Nat. Rev. Microbiol. *14*, 119–128.

Chen, H., Shiroguchi, K., Ge, H., and Xie, X.S. (2015). Genome-wide study of mRNA degradation and transcript elongation in Escherichia coli . Mol. Syst. Biol. *11*, 808.

Chen, K., Gao, Y., Mih, N., O'Brien, E.J., Yang, L., and Palsson, B.O. (2017). Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation. Proc. Natl. Acad. Sci. U. S. A. *114*, 11548–11553.

Chen, W.H., Van Noort, V., Lluch-Senar, M., Hennrich, M.L., Wodke, J.A.H., Yus, E., Alibés, A., Roma, G., Mende, D.R., Pesavento, C., et al. (2016). Integration of multi-omics data of a genome-reduced bacterium: Prevalence of post-transcriptional regulation and its correlation with protein abundances. Nucleic Acids Res. *44*, 1192–1202.

Cheng, A.A., and Lu, T.K. (2012). Synthetic Biology : An Emerging Engineering Discipline. Annu. Rev. Biomed. Eng. *14*, 155–178.

Chopra-Dewasthaly, R., Marenda, M., Rosengarten, R., Jechlinger, W., and Citti, C. (2005). Construction of the first shuttle vectors for gene cloning and homologous recombination in Mycoplasma agalactiae. FEMS Microbiol. Lett. *253*, 89–94.

Chuang, H.-Y.H., Hofree, M., and Ideker, T. (2010). A decade of systems biology. … Cell Dev. Biol. *26*, 721–744.

Commichau, F.M., Pietack, N., and Stülke, J. (2013). Essential genes in Bacillus subtilis: a re-evaluation after ten years. Mol. Biosyst. *9*, 1068–1075.

Cooper, B. (2014). Proof by synthesis of Tobacco mosaic virus. Genome Biol. *15*, R67.

Cordova, C.M.M., Lartigue, C., Sirand-Pugnet, P., Renaudin, J., Cunha, R.A.F., and Blanchard, A. (2002). Identification of the origin of replication of the Mycoplasma pulmonis chromosome and its use in oriC replicative plasmids. J. Bacteriol. *184*, 5426–5435.

de Crécy, E., Metzgar, D., Allen, C., Pénicaud, M., Lyons, B., Hansen, C.J., and de Crécy-Lagard, V. (2007). Development of a novel continuous culture device for experimental evolution of bacterial populations. Appl. Microbiol. Biotechnol. *77*, 489–496.

Creecy, J.P., and Conway, T. (2015). Quantitative bacterial transcriptomics with RNA-seq. Curr. Opin. Microbiol. *23*, 133–140.

Croucher, N.J., and Thomson, N.R. (2010). Studying bacterial transcriptomes using RNA-seq. Curr. Opin. Microbiol. *13*, 619–624.

Danino, T., Mondragón-Palomino, O., Tsimring, L., and Hasty, J. (2010). A synchronized quorum of genetic clocks. Nature *463*, 326–330.

Datsenko, K. a, and Wanner, B.L. (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc. Natl. Acad. Sci. U. S. A. *97*, 6640–6645.

Daubenspeck, J.M., Jordan, D.S., and Dybvig, K. (2014). The Glycocalyx of Mollicutes. In Mollicutes: Molecular Biology and Pathogenesis, G. Browning, and C. Citti, eds. (Norfolk: Caister Academic Press), pp. 131–147.

Delahunty, C., and Yates, J.R. (2005). Protein identification using 2D-LC-MS/MS. Methods *35*, 248–255.

Deutsch, E.W., Lam, H., and Aebersold, R. (2008). Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. Physiol. Genomics *33*, 18–25.

Dey, B., Thukral, S., Krishnan, S., Chakrobarty, M., Gupta, S., Manghani, C., and Rani, V. (2012). DNA-protein interactions: Methods for detection and analysis. Mol. Cell. Biochem. *365*, 279–299.

van Dijk, E.L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. Trends Genet. *30*.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dow, S., Lucau-danila, A., Anderson, K., Arkin, A.P., Astromoff, A., Bakkoury, M. El, Bangham, R., Benito, R., Brachat, S., Andre, B., et al. (2002). Functional profiling of the Saccharomyces cerevisiae genome. Nature 387–391.

Dragosits, M., and Mattanovich, D. (2013). Adaptive laboratory evolution - principles and applications for biotechnology. Microb. Cell Fact. *12*, 1–17.

Durot, M., Bourguignon, P.Y., and Schachter, V. (2009). Genome-scale models of bacterial metabolism: Reconstruction and applications. FEMS Microbiol. Rev. *33*, 164–190.

Dybvig, K., and Voelker, L.L. (1996). Molecular biology of Mycoplasmas. Annu. Rev. Microbiol. 25–57.

Dymond, J.S., Richardson, S.M., Coombes, C.E., Babatz, T., Muller, H., Annaluru, N., Blake, W.J., Schwerzmann, J.W., Dai, J., Lindstrom, D.L., et al. (2011). Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. Nature *477*, 471–476.

Ebrahim, A., Lerman, J.A., Palsson, B.O., and Hyduke, D.R. (2013). COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Syst. Biol. *7*.

Ebrahim, A., Brunk, E., Tan, J., O'Brien, E.J., Kim, D., Szubin, R., Lerman, J.A., Lechner, A., Sastry, A., Bordbar, A., et al. (2016). Multi-omic data integration enables discovery of hidden biological regularities. Nat. Commun. *7*, 13091.

Edman, P. (1949). A method for the determination of amino acid sequence in peptides. Arch. Biochem. Biophys. *22*, 475.

Edwards, J.S., and Palsson, B.O. (1999). Systems properties of the Haemophilus influenzae Rd metabolic genotype. J. Biol. Chem. *274*, 17410–17416.

Egwu, G., Nicholas, R., Ameh, J.A., and Bashiruddin, J. (1996). Contagious bovine pleuropneumonia: An update. Vet. Bull. *66*, 875–888.

Elowitz, M., and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. Nature 335–338.

Elowitz, M., and Lim, W. a (2010). Build life to understand it. Nature *468*, 889–890.

Endy, D. (2005). Foundations for engineering biology. Nature *438*, 449–453.

Engler, C., Kandzia, R., and Marillonnet, S. (2008). A one pot, one step, precision cloning method with high throughput capability. PLoS One *3*.

Esvelt, K.M., and Wang, H.H. (2013). Genome-scale engineering for systems and synthetic biology. Mol. Syst. Biol. *9*, 1–17.

Esvelt, K.M., Carlson, J.C., and Liu, D.R. (2011). A system for the continuous directed evolution of biomolecules. Nature *472*, 499–503.

Feist, A.M., and Palsson, B.O. (2010). The biomass objective function. Curr. Opin. Microbiol.

13, 344–349.

Fejes, A.P., Robertson, G., Bilenky, M., Varhol, R., Bainbridge, M., and Jones, S.J.M. (2008). FindPeaks 3.1: A tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics *24*, 1729–1730.

Feng, J., Liu, T., and Zhang, Y. (2011). Using MACS to identify peaks from ChiP-seq data. Curr. Protoc. Bioinforma. 1–14.

Fisunov, G.Y., Garanina, I.A., Evsyutina, D. V., Semashko, T.A., Nikitina, A.S., and Govorun, V.M. (2016). Reconstruction of transcription control networks in mollicutes by high-throughput identification of promoters. Front. Microbiol. *7*, 1–15.

Fleischmann, R., Adams, M., and White, O. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science (80-.).

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R. a, Fleischmann, R.D., Bult, C.J., Kerlavage, a R., Sutton, G., Kelley, J.M., et al. (1995). The minimal gene complement of Mycoplasma genitalium. Science *270*, 397–403.

Fredens, J., Wang, K., de la Torre, D., Funke, L.F.H., Robertson, W.E., Christova, Y., Chia, T., Schmied, W.H., Dunkelmann, D.L., Beránek, V., et al. (2019). Total synthesis of Escherichia coli with a recoded genome. Nature.

Fritz, B.R., Timmerman, L.E., Daringer, N.M., Leonard, J.N., and Jewett, M.C. (2010). Biology by design: from top to bottom and back. J. Biomed. Biotechnol. *2010*, 232016.

Fung, E., Wong, W., Suen, J., Bulter, T., Lee, S., and Liao, J. (2005). A synthetic gene–metabolic oscillator. Nature *435*, 118–122.

Furey, T.S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat. Rev. Genet. *13*, 840–852.

Gardner, T.S. (2013). Synthetic biology: from hype to impact. Trends Biotechnol. *31*, 123–125.

Gardner, T., Cantor, C., and Collins, J. (2000). Construction of a genetic toggle switch in Escherichia coli. Nature 339–342.

Gasperskaja, E., and Kučinskas, V. (2017). The most common technologies and tools for functional genome analysis. Acta Medica Litu. *24*, 1–11.

Gaurivaud, P., Laigret, F., and Bové, J.M. (1996). Insusceptibility of members of the class Mollicutes to rifampin: Studies of the Spiroplasma citri RNA polymerase B-subunit gene. Antimicrob. Agents Chemother. *40*, 858–862.

Gaurivaud, P., Lakhdar, L., Le Grand, D., Poumarat, F., and Tardy, F. (2014). Comparison of

in vivo and in vitro properties of capsulated and noncapsulated variants of Mycoplasma mycoides subsp. Mycoides strain Afadé: A potential new insight into the biology of contagious bovine pleuropneumonia. FEMS Microbiol. Lett. *359*, 42–49.

Georgiadis, M.M., Singh, I., Kellett, W.F., Hoshika, S., Benner, S.A., and Richards, N.G.J. (2015). Structural basis for a six nucleotide genetic alphabet. J. Am. Chem. Soc. *137*, 6947–6955.

Gibson, D.G. (2011). Enzymatic assembly of overlapping DNA fragments. Methods Enzymol. *498*, 349–361.

Gibson, D., and Benders, G. (2008). Complete Chemical Synthesis, Assembly, and Cloning of a Mycoplasma genitalium Genome. Science (80-.). *319*, 1215–1220.

Gibson, D.G., Glass, J.I., Lartigue, C., Noskov, V.N., Chuang, R.-Y., Algire, M.A., Benders, G.A., Montague, M.G., Ma, L., Moodie, M.M., et al. (2010a). Creation of a bacterial cell controlled by a chemically synthesized genome. Science *329*, 52–56.

Gibson, D.G., Smith, H.O., Hutchison, C. a, Venter, J.C., and Merryman, C. (2010b). Chemical synthesis of the mouse mitochondrial genome. Nat. Methods *7*, 901–903.

Girolamo, F., Lante, I., Muraca, M., and Putignani, L. (2013). The Role of Mass Spectrometry in the "Omics" Era. Curr. Org. Chem. *17*, 2891–2905.

Glass, J. (2006). Essential genes of a minimal bacterium. Proc. … *103*, 425–430.

Glass, J.I., Merryman, C., Wise, K.S., Iii, C.A.H., and Smith, H.O. (2017). Minimal Cells — Real and Imagined. Cold Spring Harb Perspect Biol. *1*, 1–12.

Goodall, E.C.A., Robinson, A., Johnston, I.G., Jabbari, S., Turner, K.A., Cunningham, A.F., Lund, P.A., Cole, J.A., and Henderson, I.R. (2018). The essential genome of Escherichia coli K-12. MBio *9*, 1–18.

Gopalakrishnan, V., Krishnan, N.P., McClure, E., Pelesko, J., Crozier, D., Williamson, D.F.K., Webster, N., Ecker, D., Nichol, D., and Scott, J.G. (2019). A low-cost, open source, self-contained bacterial EVolutionary biorEactor (EVE). BioRxiv.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

Graves, P.R., and Haystead, T.A.J. (2002). Molecular Biologist's Guide to Proteomics. Microbiol. Mol. Biol. Rev. *66*, 39–63.

Gresham, D., and Dunham, M.J. (2014). The enduring utility of continuous culturing in experimental evolution. Genomics *104*, 399–405.

Griffiths, J. (2008). A brief history of mass spectrometry. Anal. Chem. *80*, 5678–5683.

Gu, C., Kim, G.B., Kim, W.J., Kim, H.U., and Lee, S.Y. (2019). Current status and applications of genome-scale metabolic models. Genome Biol. *20*, 1–18.

Guarino, A., Shannon, B., Marucci, L., Grierson, C., Savery, N., and Bernardo, M. di (2019). A low cost, open source Turbidostat design for in-vivo control experiments in Synthetic Biology. BioRxiv.

Guet, C.C., Elowitz, M.B., Hsing, W., and Leibler, S. (2002). Combinatorial synthesis of genetic networks. Science *296*, 1466–1470.

Gupta, R.S., Son, J., and Oren, A. (2019). A phylogenomic and molecular markers based taxonomic framework for members of the order Entomoplasmatales: proposal for an emended order Mycoplasmatales containing the family Spiroplasmataceae and emended family Mycoplasmataceae comprised of six genera. Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol. *112*, 561–588.

Hafner, M., Landgraf, P., Ludwig, J., Rice, A., Ojo, T., Lin, C., Holoch, D., Lim, C., and Tuschl, T. (2008). Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. Methods *44*, 3–12.

Ham, T.S., Lee, S.K., Keasling, J.D., and Arkin, A.P. (2008). Design and construction of a double inversion recombination switch for heritable sequential genetic memory. PLoS One *3*, e2815.

Han, X., Aslanian, A., and Yates III, J.R. (2008). Mass spectrometry for Proteomics. Curr. Opin. Chem. Biol. *12*, 483–490.

Hashim, F.A., Mabrouk, M.S., and Al-Atabany, W. (2019). Review of Different Sequence Motif Finding Algorithms. Avicenna J. Med. Biotechnol. *11*, 130–148.

Hashimoto, M., Ichimura, T., Mizoguchi, H., Tanaka, K., Fujimitsu, K., Keyamura, K., Ote, T., Yamakawa, T., Yamazaki, Y., Mori, H., et al. (2005). Cell size and nucleoid organization of engineered Escherichia coli cells with a reduced genome. Mol. Microbiol. *55*, 137–149.

Hecht, A., Glasgow, J., Jaschke, P.R., Bawazer, L.A., Munson, M.S., Cochran, J.R., Endy, D., and Salit, M. (2017). Measurements of translation initiation from all 64 codons in E. coli. Nucleic Acids Res. *45*, 3615–3626.

Hillmer, R.A. (2015). Systems Biology for Biologists. PLoS Pathog. *11*, 1–6.

Hirokawa, Y., Kawano, H., Tanaka-Masuda, K., Nakamura, N., Nakagawa, A., Ito, M., Mori, H., Oshima, T., and Ogasawara, N. (2013). Genetic manipulations restored the growth fitness of reduced-genome Escherichia coli. J. Biosci. Bioeng. *116*, 52–58.

Hoffmann, S.A., Wohltat, C., Müller, K.M., and Arndt, K.M. (2017). A user-friendly, low-cost turbidostat with versatile growth rate estimation based on an extended Kalman filter. PLoS One *12*, 1–15.

Hooshangi, S., Thiberge, S., and Weiss, R. (2005). Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. Proc. Natl. Acad. Sci. U. S. A. *102*, 3581–3586.

Hopcroft, J. (2012). On the Impact of Turing Machines - Theory and Applications of Models of Computation. M. Agrawal, S.B. Cooper, and A. Li, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 1–2.

Hoshika, S., Leal, N.A., Kim, M.J., Kim, M.S., Karalkar, N.B., Kim, H.J., Bates, A.M., Watkins, N.E., SantaLucia, H.A., Meyer, A.J., et al. (2019). Hachimoji DNA and RNA: A genetic system with eight building blocks. Science (80-.). *363*, 884–887.

Hoskisson, P. a, and Hobbs, G. (2005). Continuous culture--making a comeback? Microbiology *151*, 3153–3159.

Hughes, R.A., and Ellington, A.D. (2017). Synthetic DNA synthesis and assembly: Putting the synthetic in synthetic biology. Cold Spring Harb. Perspect. Biol. *9*.

Hutchison, C., Peterson, S., Gill, S., and Cline, R. (1999). Global transposon mutagenesis and a minimal Mycoplasma genome. Science (80-.). *286*, 2165–2170.

Hutchison, C.A., Chuang, R.Y., Noskov, V.N., Assad-Garcia, N., Deerinck, T.J., Ellisman, M.H., Gill, J., Kannan, K., Karas, B.J., Ma, L., et al. (2016). Design and synthesis of a minimal bacterial genome. Science (80-.). *351*, aad6253-1-aad6253-11.

Iacobucci, C., Götze, M., Ihling, C.H., Piotrowski, C., Arlt, C., Schäfer, M., Hage, C., Schmidt, R., and Sinz, A. (2018). A cross-linking/mass spectrometry workflow based on MS-cleavable cross-linkers and the MeroX software for studying protein structures and protein–protein interactions. Nat. Protoc. *13*, 2864–2889.

Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M., and Weissman, J.S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat. Protoc. *7*, 1534–1550.

Iriarte, A., Baraibar, J.D., Romero, H., and Musto, H. (2011). Selected codon usage bias in members of the class Mollicutes. Gene *473*, 110–118.

Isaacs, F.J., Hasty, J., Cantor, C.R., and Collins, J.J. (2003). Prediction and measurement of an autoregulatory genetic module. Proc. Natl. Acad. Sci. U. S. A. *100*, 7714–7719.

Iwadate, Y., Honda, H., Sato, H., Hashimoto, M., and Kato, J.I. (2011). Oxidative stress sensitivity of engineered Escherichia coli cells with a reduced genome. FEMS Microbiol. Lett. *322*, 25–33.

Janis, C., Lartigue, C., Frey, J., Wróblewski, H., Thiaucourt, F., Blanchard, A., and Sirand-Pugnet, P. (2005). Versatile use of oriC plasmids for functional genomics of Mycoplasma capricolum subsp. capricolum. Appl. Environ. Microbiol. *71*, 2888–2893.

Jewett, M., and Forster, A. (2010). Update on designing and building minimal cells. Curr. Opin. Biotechnol. *21*, 697–703.

Joyce, A.R., and Palsson, B.Ø. (2006). The model organism as a system: integrating "omics" data sets. Nat. Rev. Mol. Cell Biol. *7*, 198–210.

Juhas, M., Eberl, L., and Glass, J.I. (2011). Essence of life: essential genes of minimal genomes. Trends Cell Biol. *21*, 562–568.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. *44*, D457–D462.

Karas, B.J., Jablanovic, J., Sun, L., Ma, L., Goldgof, G.M., Stam, J., Ramon, A., Manary, M.J., Winzeler, E. a, Venter, J.C., et al. (2013). Direct transfer of whole genomes from bacteria to yeast. Nat. Methods *10*, 410–412.

Karlsen, E., Schulz, C., and Almaas, E. (2018). Automated generation of genome-scale metabolic draft reconstructions based on KEGG. BMC Bioinformatics *19*, 1–11.

Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M. V, Jacobs, J.M., Bolival, B., Assad-Garcia, N., Glass, J.I., and Covert, M.W. (2012). A whole-cell computational model predicts phenotype from genotype. Cell *150*, 389–401.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

Kessner, D., Chambers, M., Burke, R., Agus, D., and Mallick, P. (2008). ProteoWizard: Open source software for rapid proteomics tools development. Bioinformatics *24*, 2534–2536.

Khalil, A.S., and Collins, J.J. (2010). Synthetic biology: applications come of age. Nat. Rev. Genet. *11*, 367–379.

Kim, T.H., and Ren, B. (2006). Genome-Wide Analysis of Protein-DNA Interactions. Annu. Rev. Genomics Hum. Genet. *7*, 81–102.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

Kim, M., Rai, N., Zorraquino, V., and Tagkopoulos, I. (2016). Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli. Nat. Commun. *7*, 1–12.

King, K.W., and Dybvig, K. (1991). Plasmid transformation of Mycoplasma mycoides subspecies mycoides is promoted by high concentrations of polyethylene glycol. Plasmid *26*, 108–115.

King, K.W., and Dybvig, K. (1994). Mycoplasmal cloning vectors derived from plasmid pKMK1. Plasmid *31*, 49–59.

King, Z.A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J.A., Ebrahim, A., Palsson, B.O., and Lewis, N.E. (2016). BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. *44*, D515–D522.

Kitano, H. (2002). Systems biology: a brief overview. Science *295*, 1662–1664.

Kito, K., and Ito, T. (2008). Mass Spectrometry-Based Approaches Toward Absolute Quantitative Proteomics. Curr. Genomics *9*, 263–274.

Klumpp, S., Zhang, Z., and Hwa, T. (2009). Growth Rate-Dependent Global Effects on Gene Expression in Bacteria. Cell *139*, 1366–1375.

Knight, T.F. (2005). Engineering novel life. Mol. Syst. Biol. *1*, 2005.0020.

Kobayashi, H., and Kærn, M. (2004). Programmable cells: interfacing natural and engineered gene networks. Proc. Natl. Acad. Sci. *101*, 8414–8419.

Kobayashi, K., Ehrlich, S.D., Albertini, A., Amati, G., Andersen, K.K., Arnaud, M., Asai, K., Ashikaga, S., Aymerich, S., Bessieres, P., et al. (2003). Essential Bacillus subtilis genes. Proc. Natl. Acad. Sci. U. S. A. *100*, 4678–4683.

Kohl, P., Crampin, E.J., Quinn, T. a, and Noble, D. (2010). Systems biology: an approach. Clin. Pharmacol. Ther. *88*, 25–33.

Kono, N., and Arakawa, K. (2019). Nanopore sequencing: Review of potential applications in functional genomics. Dev. Growth Differ. *61*, 316–326.

Koohy, H., Down, T.A., Spivakov, M., and Hubbard, T. (2014). A comparison of peak callers used for DNase-Seq data. PLoS One *9*.

Koonin, E. V (2003). Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat. Rev. Microbiol. *1*, 127–136.

Krishnakumar, R., Assad-Garcia, N., Benders, G. a, Phan, Q., Montague, M.G., and Glass, J.I. (2010). Targeted chromosomal knockouts in Mycoplasma pneumoniae. Appl. Environ. Microbiol. *76*, 5297–5299.

Kühner, S., Van Noort, V., Betts, M.J., Leo-Madas, A., Batisse, C., Rode, M., Yamada, T., Maier, T., Bader, S., Beltran-Alvarez, P., et al. (2009). Proteome organization in a genome-

reduced bacterium. Science (80-.). *326*, 1235–1240.

Kukurba, K.R., and Montgomery, S.B. (2015). RNA sequencing and Analysis. Cold Spring Harb. Protoc. *11*, 951–969.

Kwok, R. (2010). Five hard truths for synthetic biology. Nature *463*, 288–290.

Labroussaa, F., Lebaudy, A., Baby, V., Gourgues, G., Matteau, D., Vashee, S., Sirand-Pugnet, P., Rodrigue, S., and Lartigue, C. (2016). Impact of donor-recipient phylogenetic distance on bacterial genome transplantation. Nucleic Acids Res. *44*, 8501–8511.

Labroussaa, F., Baby, V., Rodrigue, S., and Lartigue, C. (2019). La transplantation de génomes - Redonner vie à des génomes bactériens naturels ou synthétiques. Médecine/Sciences *35*, 761–770.

Lachance, J.-C., Rodrigue, S., and Palsson, B.O. (2019a). Synthetic Biology: Minimal cells, maximal knowledge. Elife *8*, 1–4.

Lachance, J.C., Lloyd, C.J., Monk, J.M., Yang, L., Sastry, A. V., Seif, Y., Palsson, B.O., Rodrigue, S., Feist, A.M., King, Z.A., et al. (2019b). BOFdat: Generating biomass objective functions for genome-scale metabolic models from experimental data. PLoS Comput. Biol. *15*, e1006971.

Lalanne, J.B., Taggart, J.C., Guo, M.S., Herzel, L., Schieler, A., and Li, G.W. (2018). Evolutionary Convergence of Pathway-Specific Enzyme Expression Stoichiometry. Cell *173*, 749-761.e38.

Lander, E., Linton, L., Birren, B., and Nusbaum, C. (2001). Initial sequencing and analysis of the human genome. Nature *409*.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lartigue, C., Duret, S., Garnier, M., and Renaudin, J. (2002). New plasmid vectors for specific gene targeting in spiroplasma citri. Plasmid *48*, 149–159.

Lartigue, C., Blanchard, A., Renaudin, J., Thiaucourt, F., and Sirand-Pugnet, P. (2003). Host specificity of mollicutes oriC plasmids: Functional analysis of replication origin. Nucleic Acids Res. *31*, 6610–6618.

Lartigue, C., Glass, J.I., Alperovich, N., Pieper, R., Parmar, P.P., Hutchison, C.A., Smith, H.O., and Venter, J.C. (2007). Genome transplantation in bacteria: Changing one species to another. Science (80-.). *317*, 632–638.

Lartigue, C., Vashee, S., Algire, M.A., Chuang, R.-Y., Benders, G.A., Ma, L., Noskov, V.N., Denisova, E.A., Gibson, D.G., Assad-Garcia, N., et al. (2009). Creating bacterial strains from

genomes that have been cloned and engineered in yeast. Science *325*, 1693–1696.

Lassmann, T., Hayashizaki, Y., and Daub, C.O. (2011). SAMStat: monitoring biases in next generation sequencing data. Bioinformatics *27*, 130–131.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for Computing and Annotating Genomic Ranges. PLoS Comput. Biol. *9*, 1–10.

Lee, K.S., Boccazzi, P., Sinskey, A.J., and Ram, R.J. (2011). Microfluidic chemostat and turbidostat with flow rate, oxygen, and temperature control for dynamic continuous culture. Lab Chip *11*, 1730–1739.

Lee, S.-W., Browning, G.F., and Markham, P.F. (2008). Development of a replicable oriC plasmid for Mycoplasma gallisepticum and Mycoplasma imitans, and gene disruption through homologous recombination in M. gallisepticum. Microbiology *154*, 2571–2580.

Lerman, J.A., Hyduke, D.R., Latif, H., Portnoy, V.A., Lewis, N.E., Orth, J.D., Schrimpe-Rutledge, A.C., Smith, R.D., Adkins, J.N., Zengler, K., et al. (2012). In silico method for modelling metabolism and gene product expression at genome scale. Nat. Commun. *3*.

Levin, J.Z.J.J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A., and Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat. Methods *7*, 709–715.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liao, Y., Smyth, G.K., and Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

Liu, F., Zheng, K., Chen, H.-C., and Liu, Z.-F. (2018). Capping-RACE: a simple, accurate, and sensitive 5′ RACE method for use in prokaryotes. Nucleic Acids Res. *46*.

Liu, J.K., O'Brien, E.J., Lerman, J.A., Zengler, K., Palsson, B.O., and Feist, A.M. (2014). Reconstruction and modeling protein translocation and compartmentalization in Escherichia coli at the genome-scale. BMC Syst. Biol. *8*, 1–15.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012a). Comparison of next-generation sequencing systems. J. Biomed. Biotechnol. *2012*, 251364.

Liu, Y., Han, Y., Huang, W., Duan, Y., Mou, L., Jiang, Z., Fa, P., Xie, J., Diao, R., Chen, Y., et

al. (2012b). Whole-genome synthesis and characterization of viable S13-like bacteriophages. PLoS One *7*, 1–7.

Lloréns-Rico, V., Lluch-Senar, M., and Serrano, L. (2015). Distinguishing between productive and abortive promoters using a random forest classifier in Mycoplasma pneumoniae. Nucleic Acids Res. *43*, 3442–3453.

Lloréns-Rico, V., Cano, J., Kamminga, T., Gil, R., Latorre, A., Chen, W.-H.H., Bork, P., Glass, J.I., Serrano, L., and Lluch-Senar, M. (2016). Bacterial antisense RNAs are mainly the product of transcriptional noise. Sci. Adv. *2*, 1–10.

Lloyd, C.J., Ebrahim, A., Yang, L., King, Z.A., Catoiu, E., O'Brien, E.J., Liu, J.K., and Palsson, B.O. (2018). COBRAme: A computational framework for genome-scale models of metabolism and gene expression. PLoS Comput. Biol. *14*, 1–14.

Lluch-Senar, M., Delgado, J., Chen, W., Lloréns-Rico, V., O'Reilly, F.J., Wodke, J.A., Unal, E.B., Yus, E., Martínez, S., Nichols, R.J., et al. (2015). Defining a minimal cell: essentiality of small ORF s and nc RNA s in a genome-reduced bacterium . Mol. Syst. Biol. *11*, 780.

Lokody, I. (2013). Synthetic biology: Recoding bacterial genomes. Nat. Rev. Genet. *14*, 822.

Long, Z., Nugent, E., Javer, A., Cicuta, P., Sclavi, B., Cosentino Lagomarsino, M., and Dorfman, K.D. (2013). Microfluidic chemostat for measuring single cell dynamics in bacteria. Lab Chip *13*, 947–954.

Lovato, A., Faoro, F., Gambino, G., Maffi, D., Bracale, M., Polverari, A., and Santi, L. (2014). Construction of a synthetic infectious cDNA clone of Grapevine Algerian latent virus (GALV-Nf) and its biological activity in Nicotiana benthamiana and grapevine plants. Virol. J. *11*, 1–15.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 1–21.

Lu, T.K., Khalil, A.S., and Collins, J.J. (2009). Next-generation synthetic gene networks. Nat. Biotechnol. *27*, 1139–1150.

Machado, D., Andrejev, S., Tramontano, M., and Patil, K.R. (2018). Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Res. *46*, 7542–7553.

Maes, D., Sibila, M., Kuhnert, P., Segalés, J., Haesebrouck, F., and Pieters, M. (2018). Update on Mycoplasma hyopneumoniae infections in pigs: Knowledge gaps for improved disease control. Transbound. Emerg. Dis. *65*, 110–124.

Maglennon, G.A., Cook, B.S., Matthews, D., Deeney, A.S., Bossé, J.T., Langford, P.R., Maskell, D.J., Tucker, A.W., Wren, B.W., and Rycroft, A.N. (2013). Development of a self-

replicating plasmid system for Mycoplasma hyopneumoniae. Vet. Res. *44*, 1–10.

Magoc, T., Wood, D., and Salzberg, S.L. (2013). EDGE-pro: Estimated Degree of Gene Expression in Prokaryotic Genomes. Evol. Bioinform. Online *9*, 127–136.

Mahmoudabadi, G., and Phillips, R. (2018). A comprehensive and quantitative exploration of thousands of viral genomes. Elife *7*, 1–26.

Mann, M., Hendrickson, R.C., and Pandey, A. (2001). Analysis of Proteins and Proteomes By Mass Spectrometry. Annu. Rev. Biochem. *70*, 437–473.

Mardis, E. (2007). ChIP-seq : welcome to the new frontier. Nat. Methods *4*, 613–614.

Marin-Sanguino, A., Vera, J., and Alves, R. (2018). Editorial: Foundations of theoretical approaches in systems biology. Front. Genet. *9*.

Markx, G.H., Davey, C.L., and Kell, D.B. (1991). The permittistat: a novel type of turbidostat. J. Gen. Microbiol. *137*, 735–743.

Matteau, D., and Rodrigue, S. (2015a). Precise Identification of Genome-Wide Transcription Start Sites in Bacteria by 5′-Rapid Amplification of cDNA Ends (5′-RACE). In DNA-Protein Interactions SE - 9, B.P. Leblanc, and S. Rodrigue, eds. (Springer New York), pp. 143–159.

Matteau, D., and Rodrigue, S. (2015b). Precise Identification of DNA-Binding Proteins Genomic Location by Exonuclease Coupled Chromatin Immunoprecipitation (ChIP-exo). In DNA-Protein Interactions SE - 11, B.P. Leblanc, and S. Rodrigue, eds. (Springer New York), pp. 173–193.

Matteau, D., Baby, V., Pelletier, S., and Rodrigue, S. (2015). A Small-Volume, Low-Cost, and Versatile Continuous Culture Device. PLoS One *10*, e0133384.

Matteau, D., Pepin, M., Baby, V., Gauthier, S., Arango Giraldo, M., Knight, T.F., and Rodrigue, S. (2017). Development of oriC -based plasmids for Mesoplasma florum. Appl. Environ. Microbiol. *83*, 1–16.

McCoy, R.E., Basham, H.G., Tully, J.G., Rose, D.L., Carle, P., and Bové, J.M. (1984). Acholeplasma florum, a New Species Isolated from Plants. Int. J. Syst. Bacteriol. *34*, 11–15.

McGeachy, A.M., Meacham, Z.A., and Ingolia, N.T. (2019). An Accessible Continuous-Culture Turbidostat for Pooled Analysis of Complex Libraries. ACS Synth. Biol. *8*, 844–856.

McGowin, C.L., and Anderson-Smits, C. (2011). Mycoplasma genitalium: An Emerging Cause of Sexually Transmitted Disease in Women. PLoS Pathog. *7*.

McIlwain, S., Mathews, M., Bereman, M.S., Rubel, E.W., MacCoss, M.J., and Noble, W.S. (2012). Estimating relative abundances of proteins from shotgun proteomics data. BMC

Bioinformatics *13*, 308.

Messer, W. (2002). The bacterial replication initiator DnaA. DnaA and oriC, the bacterial mode to initiate DNA replication. FEMS Microbiol. Rev. *26*, 355–374.

Miller, A.W., Befort, C., Kerr, E.O., and Dunham, M.J. (2013). Design and use of multiplexed chemostat arrays. J. Vis. Exp. e50262.

Minato, Y., Gohl, D.M., Thiede, J.M., Maruyama, F., Baughn, A.D., and Harcombe, W.R. (2019). Genomewide Assessment of Mycobacterium tuberculosis Conditionally Essential Metabolic Pathways. MSystems *4*, e00070-19.

Miravet-Verde, S., Ferrar, T., Espadas-García, G., Mazzolini, R., Gharrab, A., Sabido, E., Serrano, L., and Lluch-Senar, M. (2019). Unraveling the hidden universe of small proteins in bacterial genomes. Mol. Syst. Biol. *15*, 1–17.

Mitchell, L.A., Wang, A., Stracquadanio, G., Kuang, Z., Wang, X., Yang, K., Richardson, S., Martin, J.A., Zhao, Y., Walker, R., et al. (2017). Synthesis, debugging, and effects of synthetic chromosome consolidation: synVI and beyond. Science (80-.). *355*, eaaf4831.

Mobarki, N., Almerabi, B., and Hattan, A. (2019). Antibiotic Resistance Crisis. Int. J. Med. Dev. Ctries. *40*, 561–564.

Moffitt, J.R., Lee, J.B., and Cluzel, P. (2012). The single-cell chemostat: an agarose-based, microfluidic device for high-throughput, single-cell studies of bacteria and bacterial communities. Lab Chip *12*, 1487–1494.

Monod, J. (1950). La technique de culture continue. Théorie et applications. Ann. Inst. Pasteur (Paris). *79*, 390–410.

Montague, M.G., Lartigue, C., and Vashee, S. (2012). Synthetic genomics: Potential and limitations. Curr. Opin. Biotechnol. *23*, 659–665.

Morimoto, T., Kadoya, R., and Endo, K. (2008). Enhanced recombinant protein productivity by genome reduction in Bacillus subtilis. DNA … 73–81.

Morowitz, H.J. (1984). The completeness of molecular biology. Isr. J. Med. Sci. *20*, 750–753.

Mouilleron, H., Delcourt, V., and Roucou, X. (2016). Death of a dogma: Eukaryotic mRNAs can code for more than one protein. Nucleic Acids Res. *44*, 14–23.

Munshaw, S., Bailey, J.R., Liu, L., Osburn, W.O., Burke, K.P., Cox, A.L., and Ray, S.C. (2012). Computational Reconstruction of Bole1a, a Representative Synthetic Hepatitis C Virus Subtype 1a Genome. J. Virol. *86*, 5915–5921.

Mushegian, A. (1999). The minimal genome concept. Curr. Opin. Genet. Dev. 709–714.

Mushegian, A., and Koonin, E. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc. Natl. Acad. Sci. *93*, 10268–10273.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science (80-.). *320*, 1344–1349.

Nauwelaers, D., van Houtte, M., Winters, B., Steegen, K., van Baelen, K., Chi, E., Zhou, M., Steiner, D., Bonesteel, R., Aston, C., et al. (2011). A synthetic HIV-1 subtype C backbone generates comparable PR and RT resistance profiles to a subtype B backbone in a recombinant virus assay. PLoS One *6*.

Navas-Castillo, J., Laigret, F., Tully, J., and Bové, J.M. (1992). Mollicute Acholeplasma florum possesses a gene of phosphoenolpyruvate sugar phosphotransferase system and it uses UGA as tryptophan codon. C R Acad Sci III. *315*, 43–48.

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., et al. (2012). BEDOPS: High-performance genomic feature operations. Bioinformatics *28*, 1919–1920.

Noble, J.E., and Bailey, M.J.A. (2009). Quantitation of Protein. In Methods in Enzymology, (Elsevier Inc.), pp. 73–95.

Novick, A., and Szilard, L. (1950). Description of the Chemostat. Science (80-.). *112*, 715–716.

Noyce, R.S., Lederman, S., and Evans, D.H. (2018). Construction of an infectious horsepox virus vaccine from chemically synthesized DNA fragments. PLoS One *13*, 1–16.

O'Brien, E.J., Monk, J.M., and Palsson, B.O. (2015). Using genome-scale models to predict biological capabilities. Cell *161*, 971–987.

Oberhardt, M.A., Palsson, B., and Papin, J.A. (2009). Applications of genome-scale metabolic reconstructions. Mol. Syst. Biol. *5*, 1–15.

Orth, Jeffrey D., Ines Thiele, B.Ø.P. (2010). What is flux balance analysis? Nat. Biotechnol. *28*, 245–248.

Park, P.J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. *10*, 669–680.

Peebo, K., and Neubauer, P. (2018). Application of Continuous Culture Methods to Recombinant Protein Production in Microorganisms. Microorganisms *6*, 56.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295.

Peterson, S.N., and Fraser, C.M. (2001). The complexity of simplicity. Genome Biol. *2*, comment2002.1-2002.8.

Pettersson, B., and Johansson, K.-E. (2002). Taxonomy of Mollicutes. In Molecular Biology and Pathogenicity of Mycoplasmas, S. Razin, and R. Herrmann, eds. (New York (USA): Springer), pp. 1–30.

Pilizota, T., and Yang, Y.-T. (2018). "Do It Yourself" Microbial Cultivation Techniques for Synthetic and Systems Biology: Cheap, Fun, and Flexible. Front. Microbiol. *9*, 1–9.

Pollack, J.D., Williams, M. V., and McElhaney, R.N. (1997). The comparative metabolism of the mollicutes (Mycoplasmas): The utility for taxonomic classification and the relationship of putative gene annotation and phylogeny to enzymatic function in the smallest free-living cells. Crit. Rev. Microbiol. *23*, 269–354.

Pósfai, G., Plunkett, G., Fehér, T., Frisch, D., Keil, G.M., Umenhoffer, K., Kolisnychenko, V., Stahl, B., Sharma, S.S., de Arruda, M., et al. (2006). Emergent properties of reduced-genome Escherichia coli. Science *312*, 1044–1046.

Poulin-Laprade, D., Matteau, D., Jacques, P.-É., Rodrigue, S., and Burrus, V. (2015). Transfer activation of SXT/R391 integrative and conjugative elements: unraveling the SetCD regulon. Nucleic Acids Res. *43*, 2045–2056.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Radmila, H., Toloue, M., and Tian, B. (2017). RNA-seq methods for transcriptome analysis. Wiley Interdiscip. Rev. RNA *8*, 1–24.

Rashid, N.U., Giresi, P.G., Ibrahim, J.G., Sun, W., and Lieb, J.D. (2011). ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. Genome Biol. *12*.

Ravikumar, V., Nalpas, N.C., Anselm, V., Krug, K., Lenuzzi, M., Šestak, M.S., Domazet-Lošo, T., Mijakovic, I., and Macek, B. (2018). In-depth analysis of Bacillus subtilis proteome identifies new ORFs and traces the evolutionary history of modified proteins. Sci. Rep. *8*, 17246.

Rees, J., Chalkley, O., Landon, S., Purcell, O., Marucci, L., and Grierson, C. (2018). Designing Minimal Genomes Using Whole-Cell Models. BioRxiv 344564.

Renaudin, J., Marais, A., Verdin, E., Duret, S., Foissac, X., Laigret, F., and Bové, J.M. (1995). Integrative and free Spiroplasma citri oriC plasmids: Expression of the Spiroplasma phoeniceum spiralin in Spiroplasma citri. J. Bacteriol. *177*, 2870–2877.

ReuB, D.R., Altenbuchner, J., Mäder, U., Rath, H., Ischebeck, T., Sappa, P.K., Thürmer, A., Guérin, C., Nicolas, P., Steil, L., et al. (2017). Large-scale reduction of the Bacillus subtilis

genome: Consequences for the transcriptional network, resource allocation, and metabolism. Genome Res. *27*, 289–299.

Rhee, H.S., and Pugh, B.F. (2011). Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. Cell *147*, 1408–1419.

Rhee, H.S., and Pugh, B.F. (2012). ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. In Current Protocols in Molecular Biology, (John Wiley & Sons, Inc.), p. Unit21.24.

Rice, L.B. (1998). Tn916 family conjugative transposons and dissemination of antimicrobial resistance determinants. Antimicrob. Agents Chemother. *42*, 1871–1877.

Richardson, S.M., Mitchell, L.A., Stracquadanio, G., Yang, K., Dymond, J.S., DiCarlo, J.E., Lee, D., Huang, C.L.V., Chandrasegaran, S., Cai, Y., et al. (2017). Design of a synthetic yeast genome. Science (80-.). *355*, 1040–1044.

Rideau, F., Le Roy, C., Descamps, E.C.T., Renaudin, H., Lartigue, C., and Bébéar, C. (2017). Cloning, Stability, and Modification of Mycoplasma hominis Genome in Yeast. ACS Synth. Biol. *6*, 891–901.

Rinaudo, K., Bleris, L., Maddamsetti, R., Subramanian, S., Weiss, R., and Benenson, Y. (2007). A universal RNAi-based logic evaluator that operates in mammalian cells. Nat. Biotechnol. *25*, 795–801.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Rossi, M.J., Lai, W.K.M., and Pugh, B.F. (2018). Simplified ChIP-exo assays. Nat. Commun. *9*, 1–13.

Röst, H.L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F., Andreotti, S., Ehrlich, H.C., Gutenbrunner, P., Kenar, E., et al. (2016). OpenMS: A flexible open-source software platform for mass spectrometry data analysis. Nat. Methods *13*, 741–748.

Sabidó, E., Selevsek, N., and Aebersold, R. (2012). Mass spectrometry-based proteomics for systems biology. Curr. Opin. Biotechnol. *23*, 591–597.

Saiki, R., Scharf, S., Faloona, F., Mullis, K., Horn, G., Erlich, H., and Arnheim, N. (1985). Enzymatic Amplification of β-Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia. Science (80-.). *230*, 1350–1354.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison, C.A., Slocombe, P.M., and Smith, M. (1977). Nucleotide sequence of bacteriophage phiX174 DNA. Nature *265*, 687–695.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science (80-.). *270*, 467–470.

Schindler, D., Dai, J., and Cai, Y. (2018). Synthetic genomics: a new venture to dissect genome fundamentals and engineer new functions. Curr. Opin. Chem. Biol. *46*, 56–62.

Schloss, J. (2008). How to get genomes at one ten-thousandth the cost. Nat. Biotechnol. *26*, 1113–1115.

Scholte, F.E.M., Tas, A., Martina, B.E.E., Cordioli, P., Narayanan, K., Makino, S., Snijder, E.J., and van Hemert, M.J. (2013). Characterization of Synthetic Chikungunya Viruses Based on the Consensus Sequence of Recent E1-226V Isolates. PLoS One *8*.

Selinger, D.W., Saxena, R.M., Cheung, K.J., Church, G.M., and Rosenow, C. (2003). Global RNA half-life analysis in Escherichia coli reveals positional patterns of transcript degradation. Genome Res. *13*, 216–223.

Serandour, A. a, Brown, G.D., Cohen, J.D., and Carroll, J.S. (2013). Development of an Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties. Genome Biol. *14*, R147.

Shahid, M.A., Marenda, M.S., Markham, P.F., and Noormohammadi, A.H. (2014). Development of an oriC vector for use in Mycoplasma synoviae. J. Microbiol. Methods *103*, 70–76.

Shang, Y., Wang, M., Xiao, G., Wang, X., Hou, D., Pan, K., Liu, S., Li, J., Wang, J., Arif, B.M., et al. (2017). Construction and Rescue of a Functional Synthetic Baculovirus. ACS Synth. Biol. *6*, 1393–1402.

Shen, Y., Wang, Y., Chen, T., Gao, F., Gong, J., Abramczyk, D., Walker, R., Zhao, H., Chen, S., Liu, W., et al. (2017). Deep functional analysis of synII, a 770-kilobase synthetic yeast chromosome. Science (80-.). *355*.

Shinhara, A., Matsui, M., Hiraoka, K., Nomura, W., Hirano, R., Nakahigashi, K., Tomita, M., Mori, H., and Kanai, A. (2011). Deep sequencing reveals as-yet-undiscovered small RNAs in Escherichia coli. BMC Genomics *12*, 428.

Simon, J., Giresi, P., Davis, I., and Lieb, J. (2012). Using FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) to isolate active regulatory DNA. Nat. Protoc. *7*, 256–267.

Sirand-Pugnet, P., Citti, C., Barré, A., and Blanchard, A. (2007a). Evolution of mollicutes: down a bumpy road with twists and turns. Res. Microbiol. *158*, 754–766.

Sirand-Pugnet, P., Lartigue, C., Marenda, M., Jacob, D., Barré, A., Barbe, V., Schenowitz, C., Mangenot, S., Couloux, A., Segurens, B., et al. (2007b). Being pathogenic, plastic, and sexual while living with a nearly minimal bacterial genome. PLoS Genet. *3*, 744–758.

Skelding, D., Hart, S.F.M., Vidyasagar, T., Pozhitkov, A.E., and Shou, W. (2018). Developing a low-cost milliliter-scale chemostat array for precise control of cellular growth. Quant. Biol. *6*, 129–141.

Sleator, R.D. (2010). The story of Mycoplasma mycoides JCVI-syn1.0. Bioeng. Bugs *1*, 231–232.

van der Sloot, A., and Tyers, M. (2017). Synthetic Genomics: Rewriting the Genome Chromosome by Chromosome. Mol. Cell *66*, 441–443.

Smith, H.O., Iii, C.A.H., Pfannkoch, C., and Venter, J.C. (2003). Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. Proc. Natl. Acad. Sci. *100*, 15440–15445.

Smith, R., Mathis, A.D., Ventura, D., and Prince, J.T. (2014). Proteomics, lipidomics, metabolomics: A mass spectrometry tutorial from a computer scientist's point of view. BMC Bioinformatics *15*.

Sohka, T., Heins, R. a, Phelan, R.M., Greisler, J.M., Townsend, C. a, and Ostermeier, M. (2009). An externally tunable bacterial band-pass filter. Proc. Natl. Acad. Sci. U. S. A. *106*, 10135–10140.

Song, L., Zhang, Z., and Grasfeder, L. (2011). Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome … *21*, 1757–1767.

Spencer, N., and Fernandes, W. (2010). Human genome at ten: the sequence explosion. Nature *464*, 670–671.

Stark, W.M. (2017). Making serine integrases work for us. Curr. Opin. Microbiol. *38*, 130–136.

Stricker, G., Engelhardt, A., Schulz, D., Schmid, M., Tresch, A., and Gagneur, J. (2017). GenoGAM: Genome-wide generalized additive models for ChIP-Seq analysis. Bioinformatics *33*, 2258–2265.

Stricker, J., Cookson, S., Bennett, M.R., Mather, W.H., Tsimring, L.S., and Hasty, J. (2008). A fast, robust and tunable synthetic gene oscillator. Nature *456*, 516–519.

Sundaram, A.Y.M., Hughes, T., Biondi, S., Bolduc, N., Bowman, S.K., Camilli, A., Chew, Y.C., Couture, C., Farmer, A., Jerome, J.P., et al. (2016). A comparative study of ChIP-seq sequencing library preparation methods. BMC Genomics *17*, 1–12.

Takahashi, C.N., Miller, A.W., Ekness, F., Dunham, M.J., and Klavins, E. (2015). A Low Cost, Customizable Turbidostat for Use in Synthetic Circuit Characterization. ACS Synth. Biol. *4*, 32–38.

Tanita Casci (2001). ChIP on chips. Nat. Rev. Genet. *2*, 2309.

TerMaat, J.R., Pienaar, E., Whitney, S.E., Mamedov, T.G., and Subramanian, A. (2009). Gene synthesis by integrated polymerase chain assembly and PCR amplification using a high-speed thermocycler. J. Microbiol. Methods *79*, 295–300.

Thiele, I., Fleming, R.M.T., Que, R., Bordbar, A., Diep, D., and Palsson, B.O. (2012). Multiscale Modeling of Metabolism and Macromolecular Synthesis in E. coli and Its Application to the Evolution of Codon Usage. PLoS One *7*.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.

Tigges, M., Marquez-Lago, T.T., Stelling, J., and Fussenegger, M. (2009). A tunable synthetic mammalian oscillator. Nature *457*, 309–312.

Toprak, E., Veres, A., Yildiz, S., Pedraza, J.M., Chait, R., Paulsson, J., and Kishony, R. (2013). Building a morbidostat: an automated continuous-culture device for studying bacterial drug resistance under dynamically sustained drug inhibition. Nat. Protoc. *8*, 555–567.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Tsarmpopoulos, I., Gourgues, G., Blanchard, A., Vashee, S., Jores, J., Lartigue, C., and Sirand-Pugnet, P. (2016). In-Yeast Engineering of a Bacterial Genome Using CRISPR/Cas9. ACS Synth. Biol. *5*, 104–109.

Tumpey, T.M., Basler, C.F., Aguilar, P. V., Zeng, H., Solórzano, A., Swayne, D.E., Cox, N.J., Katz, J.M., Taubenger, J.K., Pales, P., et al. (2005). Characterization of the reconstructed 1918 Spanish influenza pandemic virus. Science (80-.). *310*, 77–80.

Valouev, A., Johnson, D.S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R.M., and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat. Methods *5*, 829–834.

Vanderperre, B., Lucier, J.-F., and Roucou, X. (2012). HAltORF: a database of predicted out-of-frame alternative open reading frames in human. Database (Oxford). *2012*, 1–5.

Vanderperre, B., Lucier, J.-F., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S., Wisztorski, M., Salzet, M., Boisvert, F.-M., and Roucou, X. (2013). Direct detection of

alternative open reading frames translation products in human significantly expands the proteome. PLoS One *8*, e70698.

Vaudel, M., Burkhart, J.M., Zahedi, R.P., Oveland, E., Berven, F.S., Sickmann, A., Martens, L., and Barsnes, H. (2015). PeptideShaker enables reanalysis of MS-derived proteomics data sets. Nat. Biotechnol. *33*, 22–24.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. (1995). Serial analysis of gene expression. Science (80-.). *270*, 484–487.

Veltri, P. (2008). Algorithms and tools for analysis and management of mass spectrometry data. Brief. Bioinform. *9*, 144–155.

Venetz, J.E., Medico, L. Del, Wölfle, A., Schächle, P., Bucher, Y., Appert, D., Tschan, F., Flores-Tinoco, C.E., Van Kooten, M., Guennoun, R., et al. (2019). Chemical synthesis rewriting of a bacterial genome to achieve design flexibility and biological functionality. Proc. Natl. Acad. Sci. U. S. A. *116*, 8070–8079.

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 *291*, 1304–1351.

van Vliet, A.H.M. (2010). Next generation sequencing of microbial transcriptomes: challenges and opportunities. FEMS Microbiol. Lett. *302*, 1–7.

Vu, H.L.X., Ma, F., Laegreid, W.W., Pattnaik, A.K., Steffen, D., Doster, A.R., and Osorio, F.A. (2015). A Synthetic Porcine Reproductive and Respiratory Syndrome Virus Strain Confers Unprecedented Levels of Heterologous Protection. J. Virol. *89*, 12070–12083.

Waites, K.B., and Talkington, D.F. (2004). Mycoplasma pneumoniae and its role as a human pathogen. Clin. Microbiol. Rev. *17*, 697–728.

Wang, H.H. (2010). Synthetic genomes for synthetic biology. J. Mol. Cell Biol. *2*, 178–179.

Wang, L., and Maranas, C.D. (2018). MinGenome: An in Silico Top-Down Approach for the Synthesis of Minimized Genomes. ACS Synth. Biol. *7*, 462–473.

Wang, B., Tseng, E., Regulski, M., Clark, T.A., Hon, T., Jiao, Y., Lu, Z., Olson, A., Stein, J.C., and Ware, D. (2016). Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nat. Commun. *7*, 11708.

Watson, J., and Jordan, E. (1989). The Human Genome Program at the National of Health. Genomics *5*, 654–656.

Welte, T., Torres, A., and Nathwani, D. (2012). Clinical and economic burden of community-acquired pneumonia among adults in Europe. Thorax *67*, 71–79.

Westermann, A.J., Gorski, S.A., and Vogel, J. (2012). Dual RNA-seq of pathogen and host. Nat. Rev. Microbiol. *10*, 618–630.

Wetterstrand, K. (2019). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).

Wilbanks, E.G., and Facciotti, M.T. (2010). Evaluation of algorithm performance in ChIP-seq peak detection. PLoS One *5*.

Win, M., and Smolke, C. (2008). Higher-order cellular information processing with synthetic RNA devices. Science (80-.).

Winder, C.L., and Lanthaler, K. (2011). The use of continuous culture in systems biology investigations. In Methods in Enzymology, (Elsevier Inc.), pp. 261–275.

Wodke, J.A.H., Pucha ka, J., Lluch-Senar, M., Marcos, J., Yus, E., Godinho, M., Gutierrez-Gallego, R., dos Santos, V.A.P.M., Serrano, L., Klipp, E., et al. (2013). Dissecting the energy metabolism in Mycoplasma pneumoniae through genome-scale metabolic modeling. Mol. Syst. Biol. *9*, 653–653.

Wodke, J.A.H., Alibés, A., Cozzuto, L., Hermoso, A., Yus, E., Lluch-Senar, M., Serrano, L., and Roma, G. (2015). MyMpn: A database for the systems biology model organism Mycoplasma pneumoniae. Nucleic Acids Res. *43*, D618–D623.

Wolf-Yadlin, A., Hu, A., and Noble, W.S. (2016). Technical advances in proteomics: New developments in data-independent acquisition. F1000Research *5*, 419.

Wu, Y., Li, B.Z., Zhao, M., Mitchell, L.A., Xie, Z.X., Lin, Q.H., Wang, X.X., Xiao, W.H., Wang, Y., Zhou, X., et al. (2017). Bug mapping and fitness testing of chemically synthesized chromosome X. Science (80-.). *355*.

Xavier, J.C., Patil, K.R., and Rocha, I. (2014). Systems Biology Perspectives on Minimal and Simpler Cells. Microbiol. Mol. Biol. Rev. *78*, 487–509.

Xie, Z.X., Li, B.Z., Mitchell, L.A., Wu, Y., Qi, X., Jin, Z., Jia, B., Wang, X.X., Zeng, B.X., Liu, H.M., et al. (2017). “Perfect” designer chromosome v and behavior of a ring derivative. Science (80-.). *355*.

Yan, C., Sun, H., and Zhao, H. (2016). Latest surveillance data on mycoplasma pneumoniae infections in children, suggesting a new epidemic occurring in Beijing. J. Clin. Microbiol. *54*, 1400–1401.

Yang, M., Yang, Y., Chen, Z., Zhang, J., Lin, Y., Wang, Y., Xiong, Q., Li, T., Ge, F., Bryant, D.A., et al. (2014). Proteogenomic analysis and global discovery of posttranslational modifications in prokaryotes. Proc. Natl. Acad. Sci. *111*, E5633–E5642.

Yang, R., Han, Y., Ye, Y., Liu, Y., Jiang, Z., Gui, Y., and Cai, Z. (2011). Chemical synthesis of bacteriophage G4. PLoS One *6*, 2–7.

Yi, H., Cho, Y.-J., Won, S., Lee, J.-E., Jin Yu, H., Kim, S., Schroth, G.P., Luo, S., and Chun, J. (2011). Duplex-specific nuclease efficiently removes rRNA for prokaryotic RNA-seq. Nucleic Acids Res. *39*, e140.

Young, N.L., and Bieniasz, P.D. (2007). Reconstitution of an infectious human endogenous retrovirus. PLoS Pathog. *3*, 0119–0130.

Yus, E., Maier, T., Michalodimitrakis, K., van Noort, V., Yamada, T., Chen, W.-H.W.-H., Wodke, J.A.H.J.A.H., Güell, M., Martínez, S., Bourgeois, R., et al. (2009). Impact of Genome Reduction on Bacterial Metabolism and Its Regulation. Science (80-.). *326*, 1263–1268.

Yus, E., Lloréns-Rico, V., Martínez, S., Gallo, C., Eilers, H., Blötz, C., Stülke, J., Lluch-Senar, M., and Serrano, L. (2019). Determination of the Gene Regulatory Network of a Genome-Reduced Bacterium Highlights Alternative Regulation Independent of Transcription Factors. Cell Syst. *9*, 143-158.e13.

Zhang, L.-Y., Chang, S.-H., and Wang, J. (2010). How to make a minimal genome for synthetic minimal cell. Protein Cell *1*, 427–434.

Zhang, W., Zhao, G., Luo, Z., Lin, Y., Wang, L., Guo, Y., Wang, A., Jiang, S., Jiang, Q., Gong, J., et al. (2017). Engineering the ribosomal DNA in a megabase synthetic chromosome. Science (80-.). *355*.

Zhang, Y., Liu, T., Meyer, C. a, Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhang, Z., Boccazzi, P., Choi, H.-G., Perozziello, G., Sinskey, A.J., Jensen, K.F., and Sinskey, J. (2006). Microchemostat-microbial continuous culture in a polymer-based, instrumented microbioreactor. Lab Chip *6*, 906–913.

Zhu, W., Smith, J.W., and Huang, C.M. (2010). Mass spectrometry-based label-free quantitative proteomics. J. Biomed. Biotechnol. *2010*.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol. Cell *65*, 631-643.e4.