UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Stanimir KAMBAREV

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Nantes
sous le sceau de l'Université Bretagne Loire

École doctorale : *Biologie-Sante*

Discipline : Microbiologie
Spécialité : *Biologie des organismes*
Unité de recherche : *UMR 892 Inserm – 6299 CNRS*

Soutenue le *08/11/2016*
Thèse N° :

GeXplore : développement d'une approche génomique pour l'étude des interactions hôte-pathogène.

JURY

Président du Jury :	**Guy GOROCHOV**, Professeur des universités – Praticien Hospitalier, Université Pierre et Marie Curie, Paris
Rapporteurs :	**Hervé BLOTTIERE**, Directeur de Recherche, Micalis Institute, Jouy en Josas **Vincent CATTOIR**, Professeur des universités – Praticien Hospitalier, Université de Rennes 1
Examinateurs :	**Guy GOROCHOV**, Professeur des universités – Praticien Hospitalier, Université Pierre et Marie Curie, Paris **Hervé BLOTTIERE**, Directeur de Recherche, Micalis Institute, Jouy en Josas **Jean-Jacques TOULME**, Professeur émérite, Inserm U1212, Université de Bordeaux **Vincent CATTOIR**, Professeur des universités – Praticien Hospitalier, Université de Rennes 1
Directeur de Thèse :	**Frédéric PECORARI**, Chargé de Recherche, CNRS, UMR Inserm 892 – CNRS 6299
Co-directeur de Thèse :	**Stéphane CORVEC**, Maître de Conférences des Universités –Praticien Hospitalier, CHU de Nantes et Faculté de Médecine de Nantes

Contents

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisors Frédéric Pecorari and Stéphane Corvec for their continuous support during my PhD study and related research, for their patience, motivation, and immense knowledge. Their guidance helped me in the most difficult moments of my training as well as in the writing of this thesis.

Besides my supervisors, I would like to thank the rest of my thesis committee: Dr Laurent Marsollier and Dr Emmanuel Coron for their insightful comments and encouragement during these three years and a half. My sincere thanks also goes to Prof. Guy Gorochov, Dr Hervé Blottière, Prof. Jean-Jacques Toulme and Prof. Vincent Cattoir who agreed on being my defense jury members.

I also thank my labmates Valentina, Barbara, Axelle, Ben and Petar for their stimulating discussions and for being such good colleagues and supportive friends. In particular, I am grateful to Ghislaine Behar-Gigi for sharing her enormous experience with me and for her constant and patient support during my evolution as a young scientist.

I would like to extend my gratitude to the ARMINA consortium for funding my PhD thesis and to Prof. Michel Cherel and Team 13 of UMR 892 for extending my training so I can finish my research. I would also like to thank Audrey Donnart from the genomic platform of Nantes and Erich Charpentier from the BiRD platform for being such kind colleagues and for their technical support related to NGS performance and analysis.

No words can express my gratitude towards my best friends Hristo and Vasil who keep being constant support in my life for over a decade now. I would like to thank also Swapnil, Neha,

Kubat, David, Mattieu, Iyanar, Simon, Marine, Benoit and Johann for being such friends and family during my great time in Nantes.

Last but not the least, I would like to thank my beloved family: my girlfriend Elitsa, my sister Ina and my parents for supporting me spiritually throughout every endeavor in my life and for always being there for me.

Abbreviations

BHI	Brain-Hearth Infusion
MALDI-TOF	Matrix-Assisted Laser Desorption/Ionization- Time Of Flight
DNA	Deoxyribonucleic acid
EDTA	Ethylene-diamine-tetraacetic acid
TE	Tris-EDTA
SDS	Sodium Dodecyl Sulfate
RT	Room Temperature
TAE	Tris-Acetate-EDTA
PFGE	Pulsed-Field Gel Electrophoresis
NCBI	National Center for Biological Information
NGS	Next Generation Sequencing
WGS	Whole-genome sequencing
MLST	Multi-Locus Sequence Typing
OrthoANI	Orthologous Average Nucleotide Identity
BLAST	Basic Local Alignment Search Tool
BRIG	BLAST Ring Image Generator
ORF	Open Reading Frame
PBS	Phosphate-Buffered Saline
TBS	Tris-Buffered Saline
WB	Washing Buffer
BSA	Bovine Serum Albumin

Synopsis

Chapitre 1 : Introduction

Dans ce chapitre, nous présentons des approches classiques mais aussi plus récentes pour l'étude des interactions hôte-pathogène. Ces dernières décennies ont vu l'émergence d'une trentaine de nouvelles maladies infectieuses (Schlipköter & Flahault 2010) et la résurgence d'anciennes maladies d'origine bactérienne, virale ou parasitaire affectant un grand nombre de personnes. Dans le même temps, l'utilisation généralisée d'antibiotiques a conduit au développement de résistances aux antibiotiques et à l'affaiblissement inévitable de l'arsenal thérapeutique disponible. Au total, les maladies infectieuses sont responsables de plus d'un quart des décès dans le monde.

La recherche sur les interactions hôte-pathogène est un domaine en constante évolution. Ces interactions sont réalisées par l'intermédiaire d'une multitude de différents mécanismes. Dans ce contexte, il est essentiel de comprendre les mécanismes de pathogenèse de divers micro-organismes et de mieux définir leurs interactions avec l'hôte au niveau cellulaire et moléculaire, puisque cette connaissance est essentielle pour le développement de nouvelles approches préventives (vaccins), thérapeutiques (identification de cibles) et diagnostiques.

Il est couramment admis qu'environ 80% des protéines dans la cellule fonctionnent exclusivement sous forme de complexes protéiques, qu'ils soient transitoires ou permanents (Berggård et al. 2007). En ce qui concerne l'interaction hôte-pathogène, il est donc essentiel de caractériser les interactions protéine-ligand, afin de mieux comprendre la biochimie de la cellule ainsi que sa physiologie pour identifier les interactions cruciales dans l'interactome hôte-pathogène. En effet, les interactions interspécifiques protéine-ligand révèlent des stratégies et des aspects de la capacité du pathogène à coloniser des niches spécifiques (par exemple pour échapper à la réponse immunitaire de l'hôte).

De nombreuses études ont ainsi porté sur des interactions protéine-ligand *in vitro* et *in vivo* et ont discuté de la culture des agents pathogènes, des systèmes modèles hôtes et des outils bioinformatiques pour corréler les grands ensembles de données obtenues et les résultats cliniques (Horvatić et al. 2016).

Une multitude de technologies d'exposition (« display » en Anglais) à base de banques combinatoires a été développée et optimisée au cours des 30 dernières années ou plus (pour une revue, se référer à Galan et al. 2016). En principe, ces technologies sont utiles pour l'identification des interactions moléculaires et ont été largement employées pour l'évolution dirigée de diverses protéines telles que des enzymes et des anticorps. Plus précisément, ces approches ont été utilisées pour le développement à des fins diagnostiques et thérapeutiques dans les domaines du cancer, et des maladies auto-immunes et infectieuses. En fonction de leur mode d'action, les technologies d'exposition peuvent être divisées en catégories. Par exemple, *in vivo* et *in vitro*, ou cellulaire et acellulaire. Cependant, quelle que soit leur performance spécifique, toutes ces techniques fonctionnent selon le principe suivant: une banque diversifiée de molécules souhaitées est créée (protéine, anticorps, peptide, etc.). La banque obtenue peut être diversifiée artificiellement en utilisant diverses méthodes de randomisation ou peut être dérivée de sources naturelles (ADN génomique). La banque est ensuite criblée sur plusieurs tours contre un ou plusieurs ligands d'intérêt qui conduit à l'enrichissement de séquences spécifiques codant pour des peptides / protéines avec une certaine affinité anti-ligand. Enfin, la sortie de sélection est analysée pour déterminer le phénotype souhaité. Un groupe particulier de technologies d'exposition à base de banque, que nous appelons «génome entier», ont été mises au point et utilisées avec succès dans le domaine de l'interaction hôte-pathogène. Ces techniques comprennent la technologie ANTIGENome, l'exposition par phage Shotgun (« Shotgun phage display ») et l'exposition par

ribosome (« ribosome display »). Elles seront discutées plus en détail dans cette partie car elles ont servi de base pour le développement de la méthode présentée ici, GeXplore.

Chapter 2: Séquençages de génomes entiers et analyses.

L'ADN est la source ultime de l'information dans un organisme vivant. La quête pour le séquençage de l'ADN a commencée peu de temps après que sa structure chimique a été élucidé en 1953 (Watson & Crick 1953). Malheureusement, il a fallu attendre deux décennies pour l'avènement du séquençage classique de Sanger et encore 20 ans avant que les premiers séquençages de génomes bactériens soient réalisés (Sanger et al. 1977; Fleischmann et al. 1995; Fraser et al. 1995; Hutchison 2007). Toutefois, le séquençage de génomes entiers (« Whole Genome Sequencing » ou WGS) est resté un luxe pour les petites et moyennes installations de recherche pendant un certain temps en raison de son coût encore élevé (Barbosa et al. 2014).

Cet inconvénient a limité sa large diffusion jusqu'au le développement de la «deuxième» ou «nouvelle génération» de séquençage (« Next Generation Sequencing » ou NGS) en 2005 qui a rendu le WGS abordable pour de nombreux laboratoires révolutionnant le domaine de la génomique bactérienne (Shendure & Ji 2008; Zhang et al. 2011). Les séquenceurs 454 GS FLX (Roche), Illumina (Genome Analyzer) et SOLiD (Life Technologies) sont les plates-formes les plus répandues sur le marché, chacune ayant ses propres avantages et inconvénients (Metzker 2010; Van Dijk, Auger, et al. 2014). Par conséquent, nous avons assisté à une explosion de la disponibilité de l'information génomique pour les procaryotes (au moment de la rédaction en août 2016, il y avait plus de 72 000 séquences génomiques procaryotes publiquement disponibles (NCBI, 2016) et leur nombre ne cesse de croître de façon exponentielle. Comme mentionné dans le chapitre 1, le principal objectif de notre travail était le développement d'une approche d'exposition de génomes entiers entièrement *in vitro,* GeXplore, pour étudier les interactions hôte-pathogène. L'une des

étapes les plus importantes de notre approche est la cartographie d'une sortie de sélection potentiellement enrichie par rapport à une séquence du génome de référence. Logiquement, si un gène codant pour une certaine protéine sélectionnée n'est pas présente dans la séquence de référence, elle sera pas mise en correspondance sur le génome de référence et sera donc omise lors de l'étape d'analyse. Lorsque notre projet a été lancé en février 2013, le nombre de génomes accessibles au public pour deux des espèces utilisées dans nos expériences, *M. ulcerans* et *S. gallolyticus* était extrêmement limitée (un et quatre génomes, respectivement). En outre, la plasticité du génome est un trait connu parmi les bactéries qui pourraient atteindre des dimensions méga-bases dans certaines espèces (Land et al. 2015). Ainsi, il apparaît comme un prérequis de séquencer les génomes entiers des souches pour qu'elles soient étudiées avec succès par des approches de type « génomes entiers » telles que GeXplore.

Dans cette partie du manuscrit, nous décrivons le WGS de *S. aureus* FP_SA_ST25, de *M. ulcerans* S4018 et *S. gallolyticus* - NTS31106099 en utilisant la technologie Illumina. Leurs assemblages *de novo* ont été utilisés comme génomes de référence à différents stades du développement et de l'application de notre méthode. Les séquences ont également été partiellement caractérisées par génomique comparative. Ainsi, un transposon conjugatif de type Tn916, que nous avons nommé Tn6263, qui confère un groupe de gènes de résistance aux antibiotiques, a été découvert dans *S. gallolyticus* NTS31106099. Sa prévalence dans d'autres isolats cliniques a été étudiée. Enfin, les génomes de deux autres isolats cliniques de *S. gallolyticus* (NTS31301958 et NTS31307655) ont été séquencés afin de caractériser un autre élément, Tn6331, qui a été identifié dans leurs génomes par criblage par PCR.

Chapitre 3: Préparation et validation de l'expression *in vitro* de banques génomiques combinatoires.

L'exposition par ribosome est un outil puissant pour étudier et /ou concevoir des interactions entre un peptide ou une protéine avec une autre molécule d'intérêt. Comme les autres technologies d'exposition, elle offre la possibilité de coupler le phénotype au génotype en liant un peptide ou une protéine à l'acide nucléique le codant. Son principal avantage est sa capacité à fonctionner dans un cadre entièrement *in vitro* pour la traduction et la transcription ce qui permet de sonder des banques de diversités gigantesques, jusqu'à 10^{12} molécules (Pluckthun 2012).

Pour être étudié par exposition par ribosome, le génome d'un agent pathogène doit être fragmenté de façon aléatoire et sous-cloné dans un vecteur d'expression contenant l'ensemble des régions fonctionnelles requises pour la transcription / traduction *in vitro* et la sélection. Des amplifications par PCR permettent ensuite de convertir chaque construction créée en une séquence linéaire fonctionnelle pour son exposition sur ribosome. Enfin, la banque est criblée contre un ou plusieurs ligands d'intérêt.

Malheureusement, en dépit de ses avantages mentionnés, la relative complexité de la technique d'exposition sur ribosome a entravé sa large utilisation dans le domaine des maladies infectieuses. À notre connaissance, il n'y a que deux groupes qui ont publié son utilisation, dans les deux cas pour l'identification de candidats vaccins (Weichhart et al. 2003; Lei et al. 2008). Même si les deux études ont été réalisées à l'aide des banques combinatoires de génomes entiers, seul la premiere décrit une tentative pour caractériser la banque (Henics et al. 2003). Cependant, aussi pertinente qu'elle a pu apparaître en son temps, l'analyse de la banque a été effectuée sur un échantillon de seulement 500-1000 clones, un nombre presque négligeable par rapport à la taille de la banque d'origine et au potentiel des plates-formes NGS actuellement disponibles. En outre, les deux groupes ont produit les banques *in vivo* en transformant *E. coli* avec les constructions de la banque,

qui est l'étape demandant le plus de travail et qui réduit de manière significative la diversité de la banque (Mullen et al. 2006).

Dans cette partie de notre travail, nous présentons le développement et l'optimisation d'une procédure de préparation entièrement *in vitro* de banques qui ne nécessite donc pas l'étape de transformation. Des banques génomiques combinatoires de *S. aureus* FP_SA_ST25, *S. gallolyticus* NTS 31106099 et *M. ulcerans* S4018 ont été préparées en utilisant une stratégie alternative de clonage assistée par G/C. Nous avons utilisé cette stratégie de clonage pour augmenter le niveau de l'insertion du fragment aléatoire et pour réduire la « recircularisation » des molécules de vecteur vide. La qualité des banques obtenues a été caractérisée par NGS pour leur diversité et leur contenu. Nous avons observé une réduction significative de la couverture du génome pour les banques de *S. aureus* analysées en raison d'un biais de l'amplification par PCR associé aux régions des génomes ayant un contenu G + C en dessous de 29% qui sont sous-représentées dans les banques. Cependant, nous avons constaté que la banque de *M. ulcerans* S4018 couvre entièrement sa séquence de référence qui a une taille de 5,4 Mb et un contenu en G + C de 64%.

Chapitre 4: Optimisation et validation de GeXplore.

Les premières études par exposition sur phage « shotgun » ont montré que jusqu'à 40% des clones enrichis après deux tours de sélection codent pour un peptide affin du ligand lors de l'utilisation de banques de fusion au gène III de phage. Cependant, l'augmentation du nombre de tours de sélection n'a pas entraîné un enrichissement plus élevé pour les clones désirés (Jacobsson & Frykberg 1995). La technologie ANTIGENome est une approche alternative qui a été mis au point pour identifier des candidats vaccins potentiels à partir de banques génomiques pathogènes spécifiques en utilisant des anticorps provenant de patients (conditions multi-ligands). De plus, une seule tentative d'utilisation d'une variante de cette technique de l'ANTIGENome basée sur l'exposition sur

ribosome a été réalisée en utilisant une banque génomique de *S. aureus* préparée *in vivo*. Cette étude a montré que la méthode de sélection comportant une étape cellulaire a certaines limitations d'expression (de nombreux peptides qui ont été sélectionnés par la version basée sur l'exposition par ribosome sont omis en utilisant l'approche à base de cellules). Néanmoins, pour les deux approches de nombreux clones analysés codent des peptides appartenant à des protéines exposées à la surface, ce qui souligne leur efficacité. Cependant, la caractérisation quantitative détaillée des sorties de sélection n'a pas été possible jusqu'à l'avènement de la plate-forme NGS.

Dans ce travail, nous avons cherché à développer une approche de type « génome entier » se déroulant entièrement *in vitro* qui pourrait être utilisée pour les deux types de sélections mono et multi-ligands. Dans une telle approche, les banques initiales et de sorties de sélection sont obtenues exclusivement par amplification PCR. Cependant, la PCR, comme déjà mentionné est la principale source de biais dans les échantillons génomiques tels que les banques NGS. Un autre problème critique est la recombinaison qui se produit entre des fragments non complémentaires lors de PCR multi-matrices qui peuvent altérer de manière significative la qualité d'entrée/sortie par l'accumulation de produits hétéroduplexes dégénérées. Ce problème peut éventuellement être évité par l'utilisation de la PCR en émulsion (ePCR) (Diehl et al. 2006). Cette technique tire parti de la compartimentation de la matrice PCR très hétérogène à la fois par la taille et par la séquence dans des microgouttelettes aqueuses en émulsion dans phase huileuse (émulsions eau-dans-huile).

D'autres points doivent être pris en considération lorsque l'on développe un système d'exposition comme le nombre de lavages à l'étape de sélection et le choix de la matrice de sélection utilisée. Notre expérience des sélections par exposition sur ribosome nous a appris que les deux pourraient avoir un impact crucial sur le succès de la sélection puisque la première aide à différencier les interactions non spécifiques des interactions désirées et la deuxième pourrait être responsable du niveau élevé de ce bruit de fond. Le nombre optimal de lavages nécessaires peut être spécifique à

chaque cas liant/ligand et nécessite une détermination empirique. Le niveau de bruit de fond sélectionné par la matrice peut être contrôlé par l'alternance des matériaux des différentes matrices utilisées (plaques ELISA, billes) tout au long de la sélection. Nos priorités se sont portées sur de grandes efficacités et spécificités de sélection, car elles sont cruciales, en particulier cette dernière. Dans cette partie du travail, nous expliquons l'optimisation étape par étape de notre méthode, avec un accent particulier sur la qualité des entrées et sorties, les étapes de lavage optimales et le choix optimal de la matrice de sélection. Enfin, nous avons pu valider l'efficacité de la sélection et de la spécificité de GeXplore en utilisant un système modèle, utilisé précédemment par d'autres équipes pour validation de l'exposition sur phages « shotgun ».

Chapitre 5: Application de GeXplore

Les pathogènes utilisent une multitude de molécules diverses pour s'attacher à l'hôte, se soustraire à son système immunitaire et diffuser dans ses tissus, menant éventuellement à une maladie infectieuse. De son côté, le système immunitaire de l'hôte neutralise les microbes intrus au niveau cellulaire ainsi que moléculaire. Par conséquent, la réponse immunitaire humorale développée pourrait être considérée comme une « empreinte immunologique» qui peut être utilisée pour la recherche dans les génomes entiers des protéines antigéniques / immunogènes correspondantes (Weichhart et al. 2003). Des anticorps spécifiques de pathogènes tels que les IgG et IgA, provenant de sérums de patients ont déjà été utilisés dans les technologies d'exposition comme ANTIGENome pour l'identification des immunoprotéomes des agents pathogènes humains (Meinke et al. 2005). Ces résultats soulignent le potentiel des technologies d'exposition pour être utilisées comme un test d'immuno-capture pour la découverte de facteurs de virulence inconnus à partir de banques génomiques d'agents pathogènes.

Streptococcus gallolyticus ssp. *gallolyticus,* anciennement connu sous le nom de *Streptococcus bovis* biotype I est une bactérie à Gram positif qui se trouve dans le tract gastro-intestinal des animaux et des humains (Schlegel et al. 2003). Cependant, il a également été reconnu comme un pathogène opportuniste en raison de sa capacité à provoquer une septicémie chez les oiseaux, la mastite du bétail, ainsi que des bactériémies et/ou endocardites infectieuses chez les humains. Son association étroite avec l'incidence du cancer colorectal s'est multipliée au cours des quatre dernières années et deux rapports récents ont examiné de façon critique les informations disponibles (Abdulamir et al. 2011; Boleij et al. 2011). Cependant, malgré les recherches approfondies sur le sujet, la raison de cette association reste inconnue.

Mycobacterium ulcerans est l'agent causal de la maladie tropicale négligée de l'ulcère de Buruli. C'est la troisième mycobactériose la plus répandue dans le monde après la tuberculose et la lèpre et elle a été diagnostiquée dans plus de 30 pays dans le monde entier (Merritt et al. 2010). Le principal facteur de virulence de cette bactérie est une exotoxine macrolide appelée mycolactone conduisant à une ulcération étendue de la peau en raison de sa cytotoxicité. Les méthodes de diagnostic actuelles impliquent la microscopie, la culture, l'histopathologie et la PCR spécifique d'IS2404. Cependant, toutes ces méthodes de laboratoire nécessitent un personnel qualifié et des équipements techniques, qui sont tous deux difficiles à fournir dans les régions rurales éloignées et pauvres (Sakyi et al. 2016). Par conséquent, un outil de diagnostic bon marché et adapté au terrain de l'environnement est nécessaire. En outre, aucun vaccin n'est disponible à jour.

Par conséquent, nous avons décidé d'utiliser notre approche optimisée et validée, GeXplore, pour l'identification des protéines immunitaires pertinentes de *S. gallolyticus* et *M. ulcerans*. Ces protéomes peuvent contenir des facteurs de virulence inconnus et des protéines immunodominantes. Des ensembles représentatifs de sérums de patients infectés par *S. gallolyticus* (n = 15) et *M. ulcerans* (n = 24) ont été collectés et utilisés pour la purification des anticorps de

types IgG et IgA. Les banques génomiques de *S. gallolyticus* NTS31106099 et *M. ulcerans* S4018 ont été sélectionnées contre les pools d'IgG/IgA purifiés. Après trois tours de sélection, les sorties obtenues ont été séquencées avec la plateforme Illumina NGS et analysées. Un ensemble représentatif des régions enrichies a été sélectionnés, sur la base de la couverture de séquençage après le dernier tour de sélection.

List of tables

List of figures

List of primers

Name	Sequence (5'-3')	Length	T_{anneal}	5'-Phospate
Tn6263_Rec_F	GCATACAACTGAAAGCATATTTCC	24	60.7	No
Tn6263_Rec_R	GTGAAAGAAGTAGAAGTAATCAAAGC	26	60.5	No
Tn6263_1	CGTCGTTAACTCCTCATTCTACGACAGC	35	67	No
Tn6263_2	GATGGTGCTAAATTTAAACCACAAAGAAAAAATGC	28	66.9	No
Tn6263_3	CGGATTTTATCACCCTCACTTGTAAACACG	30	66.7	No
Tn6263_4	CCATGGCGCGTGACATCAAGC	21	67.5	No
link_F	GAAGCTTTATATGGCCTCGGGGGCCGAATTC	31	71.5	Yes
RDV1_R	GGATCCGTGATGGTGATGGTGATGCGATCCTCTC	34	73.2	Yes
FseI_F	GGCCTATATGTTAACCTCAAGCTTTATATGGCCTCGGG	38	71.8	Yes
FseI_R	GGCCTAATATCTCGGATCCGTGATGGTGATGG	32	71.1	Yes
MAG_R	CCAGCCACGGATATATCTCCTTCTTAAAGTTAAACAAAATTATTTCTAGAGGG	53	71.1	Yes
T7C	ATACGAAATTAATACGACTCACTATAGGGAGACCACAACGGTTTCCCTC	49	73.5	No
TolAext	CGCACACCAGTAAGGTGTGCGGTTTCAGTTGCCGCTTTCTTTCTTGCTTCAGCTGCAGCTGCTT	66	79.9	No
int_F	GGAGATATATCCGTGGCTGG	20	58.8	Yes
int_R	CCCGAGGCCATATAAAGC	18	59.2	Yes
ST25_1_F	GCAGAAAATCTTCTCACAGG	20	58.4	No
ST25_1_R	CTCTCTTTTAAATATTTTTCTATTTTGC	28	58	No
ST25_2_F	CTTGTGGTAATAAAGAAAAAGAGG	24	58.3	No
ST25_2_R	CAAATGTATATTATCCATGACATGTTG	27	59.6	No
ST25_3_F	CAATTCTATTACTACTTTGTATATTGAGC	29	59.5	No
ST25_3_R	CCTATAAAGATAAATACCACACCC	24	58.6	No
ST25_4_F	CAGCATTAATTTGTTGCGTG	20	58.6	No
ST25_4_R	GGTTTAACATTAGATAGACAGCC	23	58.8	No
ST25_5_F	GCTTTTTTCACTACTTATATTAATTAAAAC	30	58.7	No
ST25_5_R	CTTCTTTTTTATATGGATAAATGAAAGG	28	58.7	No
ST25_6_F	GAATTTTTAGATATGGAACAAAATGC	26	58.9	No
ST25_6_R	CTAAAACTTCCATCAAACAGTTC	23	58.3	No
ST25_7_F	CACTATTTTGATGGTTTATAATTTTG	26	57	No
ST25_7_R	GATTATTTGTTGAAGGCTTTGATAC	25	59.3	No
ST25_8_F	GAAACTCTTGATTCTTAAGTTTCG	24	58.5	No
ST25_8_R	GATTATGAAAGGTCTCTTAGATGG	24	58.6	No
ST25_9_F	CATCATATAGAAAATTCTTTAATGACG	27	58.1	No
ST25_9_R	GAAAGAAGTTAATTTTAAGGTTGTC	25	57.8	No
ST25_10_F	GATAGCTAAAGCGATATTTGTATTAG	26	58.3	No
ST25_10_R	GTTCATTTTTTCATACTTTCCTTTTG	26	59.1	No
sbi_F	GCGTTGAACCACCTTGAATTAGTATAGTAAC	31	65.9	No
sbi_R	CTATTTGTAAGCGTTTTCGAATTAACTGTTCG	32	65.9	No
spA_F	GTTGTCACTTGGATTCAAATGACATTTTAAATC	33	65.4	No
spA_R	CTCTATTACGCAAGTGTGCTGTATTCTAAAG	31	65.7	No
T7s	ATA CGA AAT TAA TAC GAC TCA CTA TAG G	28	60.8	No
TolAs	CCG CAC ACC AGT AAG G	16	58.7	No

List of publications

Accepted:

- **Kambarev S**, Caté C, Corvec S, Pecorari F, Draft Genome Sequence of Erythromycin-Resistant *Streptococcus gallolyticus subsp . gallolyticus* NTS 31106099 Isolated from a Patient with Infective Endocarditis and Colorectal Cancer. *Genome Announc* 2015; **3**: 2–3.

In preparation:

- **Kambarev S**, Pecorari F, Corvec S, Draft Genomes of *S. gallolyticus* NTS31301958 and NTS31307655 containing Tn*6331*. *Genome Announc*
- **Kambarev S**, Corvec S, Chauty A, Marion E, Marsollier L, Pecorari F, Draft Genome Sequence of *Mycobacterium ulcerans* S4018. *Genome Announc*
- **Kambarev S**, Pecorari F, Corvec S, High Prevalence of Novel Tn*916*-like Elements in Erythromycin-Resistant *S. gallolyticus*. *JAC*
- **Kambarev S,** Corvec S, Pecorari F. GeXplore: a genome-wide approach to studying host-pathogen interactions.
- **Kambarev S,** Corvec S, Pecorari F. How to explore virulence factors by using genome based molecular techniques? Review

The student has also contributed to the preparation of:

- Aubin GG, Bémer P, **Kambarev S**, *et al. Propinibacterium namnetense* sp.nov., isolated from a human bone infection. *Int J Syst Evol Microbiol* 2016 (in press).

- Aubin GG, **Kambarev S**, Bémer P *et al*. Draft Genome Sequence of Highly Rifampin-Resistant *Propionibacterium namnetense* NTS 31307302 T Isolated from a Patient with a Bone Infection. *Genome Announc* 2016; **4**: 1–2.

Chapter 1

Introduction

The last decades have seen the emergence of some thirty new infectious diseases (Schlipköter & Flahault 2010) and the resurgence of old diseases of bacterial, viral or parasitic origin affecting a large number of people. At the same time, the widespread use of antimicrobials has led to the development of resistance and the inevitable weakening of the available therapeutic arsenal. In total, infectious diseases are responsible for over a quarter of deaths worldwide.

The research on host-pathogen interactions is a constantly evolving field. These interactions are accomplished via a multitude of various mechanisms (figure on page 25). To successfully settle in its host, a pathogenic organism must be able to colonize its tissues but also to efficiently escape its immune system by developing appropriate strategies. It is a fight of genes and proteins from both sides. Who wins depends on whether the host develops an infection or not. Moreover, a higher level of complexity arises when the pathogens evolve and become resistant to a host's defense mechanisms (Sen et al. 2016). Such pathogens pose serious challenges for treatment which place the entire human population in danger of such long-lasting persistent infections. Some of these infections even increase the rate of mortality. In this context, it is essential to understand the pathogenic mechanisms of various microorganisms and to better define their host interactions at cellular and molecular level since this knowledge is essential for development of new preventive (vaccines), therapeutic (target identification) and diagnostic approaches.

It is currently accepted that about 80% of the proteins in the cell operate exclusively in protein complexes, whether transient or constant ones (Berggård et al. 2007).

A)	Classification of host-pathogen interactions.

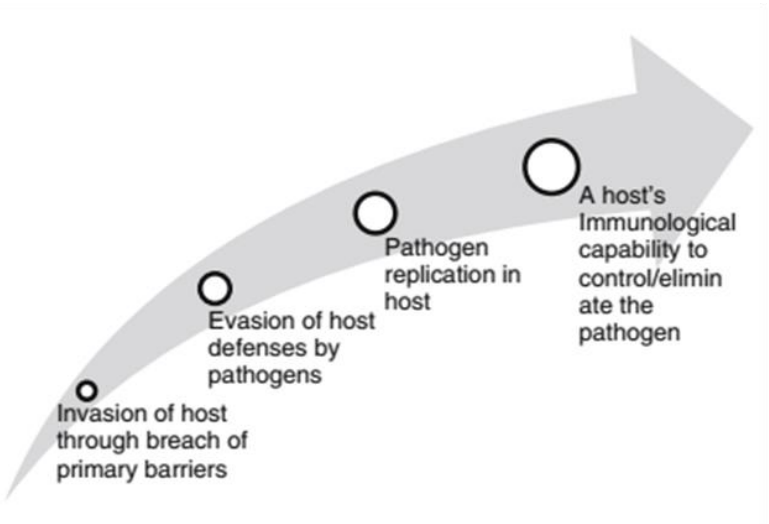


Figure 1.1. Taken from Sen *et al*, 2016.

With regards to host-pathogen interaction it is therefore essential to characterize protein-ligand interactions in order to better understand cell biochemistry and physiology and to define crucial hot-spots in the host-pathogen interactome. Indeed, interspecies protein-ligand interactions reveal pathogen infection strategies and aspects of the pathogen's ability to colonize specific niches (e.g. to evade host's immune response). Numerous studies have addressed protein-ligand interactions *in vitro* and *in vivo* and have discussed pathogen cultivation, model host systems and bioinformatic tools to correlate the obtained large data sets and clinical outcomes (Horvatić et al. 2016).

Today, there are a plethora of approaches available to investigate host-pathogen interactions at either cellular or molecular level, ranging from traditional biochemistry and genetics to sophisticated display systems for identification of protein-ligand interactions and/or discovery of potential virulence factors at a genome-wide scale.

1.1.Traditional biochemical and genetic approaches.

1.1.1. Biochemical approaches for purification and studying of virulence factors.

Historically, the discovery of protein-ligand interactions between host and pathogen has been initially performed on a small scale by studying one particular, most often surface-exposed or secreted, bacterial protein and the interacting host cells or proteins by classical biochemical approaches. Such methods include immuno-chemistry and molecular techniques like Enzyme-Linked Immunosorbent Assay (ELISA), Western blot, and Polymerase Chain reaction (PCR). For example, fibronectin-dependent bridging between *S. aureus* fibronectin-binding proteins (FnBPs) and host cell integrin α5β1 was show to be a conserved mechanism for *S. aureus* invasion of human cells (Sinha et al. 1999).

This conclusion was drawn using western blots and various *S. aureus* strains, each defective for a specific protein (mutations or deletions) and soluble FnBP to inhibit invasion. A more recent example is about Sbi, a multifunctional bacterial protein of *S. aureus*, which binds host complement components Factor H and C3 as well as IgG and beta(2)-glycoprotein I and interferes with innate immune recognition. By using several methods (adsorption experiments, recombinant cloning and purification, expression of factor H deletion mutant, antibody binding to Sbi fragments, protein binding assays, ELISA, inhibition of protein binding, etc…), it has been demonstrated that Sbi mediates innate and adaptive immune escape (Haupt et al. 2008).

1.1.2. Molecular genetic approaches for gene cloning from pathogenic into avirulent receptor strain.

Non-pathogenic bacteria could be used as recipients for studying a potential virulence factor. For example, a gene (*noxR1*) in *M. tuberculosis* has been associated to resistance to reactive

intermediates of nitrogen oxidation in macrophages. The authors have performed screenings of recombinant *M. smegmatis* or *E. coli* strains containing libraries of *M. tuberculosis* genes, followed by the selection of recombinants with enhanced survival and identification of a recombinant gene that conferred the observed phenotype (Ehrt et al. 1997).

1.1.3. Transposon mutagenesis generating a collection of mutants with low or high virulence.

Transposons are genetic elements that can move within or between genomes by either replicative or 'cut-and-paste' mechanisms mediated by an enzyme called transposase. When insertion of such element occurs in a certain gene it could lead to suppression of its expression. The most frequently used application of transposons has been insertional mutagenesis, in which a library of bacterial strains, each containing a single randomly located transposon, is constructed. Then, mutants with attenuated virulence are identified using animal host models. Finally, the disrupted gene responsible for the observed phenotype is identified by PCR. A good example for such technique is Signature-Tagged Mutagenesis (STM). For example, in one study a library of 1248 *S. aureus* mutants was screened using STM which resulted in identification of 50 mutants with attenuated virulence. Subsequent individual analysis of these mutants led to the conclusion that genes *fem*A and *fem*B, which are involved in the formation of cell wall peptidoglycan, could have virulence potential (Mei et al. 1997). Several methods derived from the initial transposon mutagenesis technique have been developed during last years, especially by combining it with massive parallel sequencing (van Opijnen & Camilli 2013).

1.2.Identification of virulence genes expressed *in vivo*.

1.2.1. *In vivo* expression technology (IVET) for detection of genes expressed inside the host.

In vivo expression technology (also known as IVET) is a method used to determine which bacterial genes are upregulated under particular environmental conditions *in vivo*, especially those, promoted during the course of infection (Mahan et al. 1993). In practice, the genomic DNA from the pathogen is partially fragmented to get multiple promoter-containing genomic fragments and then ligating those fragments into plasmids, upstream of a promoter-less reporter gene such as one for antibiotic resistance. The recombinant plasmids are transformed into a strain, deficient in the reporter gene. The foreign promoter-containing gene ligated within the plasmid permits homologous recombination between the plasmid and the pathogen's chromosome. Consequently, a series of pathogen recombinants are generated, where reporter genes are fused randomly to different promoter-containing genes from the pathogen. The recombinant pathogens are used to inoculate a host model (e.g. mouse). The promoters with upregulated activity during infection will contain active antibiotic-resistance genes thus allowing survival of the recombinants, while those with non-upregulated promoters will die in the presence of antibiotics. The survived recombinants are isolated, their plasmids are extracted and the respective promoter-containing fragment sequenced to identify the up-regulated gene. This approach has been used to study staphylococcal virulence (Lowe et al. 1998). In this study, 45 staphylococcal genes which were induced during infection in a murine renal abscess model were identified. Among them, six were previously known and 11 were uncharacterized. Eleven corresponding strain mutants were constructed and showed attenuated virulence compared with the wild-type parent, which suggest they may encode staphylococcal virulence factors.

1.3.“Omic” approaches to identifying virulence genes.

The complexity of host-pathogen interactions cannot be properly revealed by classical biochemical or genetic studies in a research-relevant timescale. As a solution to this problem, technological advances in more recent high-throughput technologies which deal with large-scale analysis of genes and proteins have enabled a comprehensive investigation of host–pathogen interactions at molecular level.

1.3.1. Genomic subtractive hybridization (GSH) for identification of genes, present in virulent strains only.

Subtractive hybridization is a technique for identifying and characterizing differences between two populations of nucleic acids. Comparison of DNA from virulent strains of bacterial pathogens with DNA from less virulent or avirulent close relatives allows the identification of genomic regions potentially involved in virulence. Such regions are often associated with pathogenicity islands and their characterisation can lead to a greater understanding of the pathogenesis of infectious disease. The method has been first reported in 1990 by Straus and Ausubel (Straus & Ausubel 1990). The initial step involved the removal of sequences from wild-type DNA (termed “driver”) which are present in both wild-type and mutant genomes (termed “tester”). DNA corresponding to the deleted region is enriched by allowing a mixture of denatured wild-type and biotinylated mutant DNA to re-associate. After re-association, avidin-coated beads are used to remove the biotinylated sequences. Following the ligation of short oligonucleotide sequences (adaptors), the recovered DNA is amplified using the adaptor sequence as a PCR primer. This amplified DNA was available for sequencing or used as a probe to screen strains and identify sequences in genomic libraries.

This approach was applied to *S. aureus* to identify two open reading-frames that might encode virulence factors (El Adhami & Stewart 1997; El-Adhami 1999)

1.3.2. Selective Capture of Transcribed Sequences (SCOTS) for analysis of mRNA from host-recovered bacteria.

Selective Capture of Transcribed Sequences (SCOTS) is a PCR-based RNA analysis. SCOTS was developed to directly identify bacterial genes rather than promoter regions. This method allows the selective capture of bacterial cDNAs from total cDNA libraries prepared from infected cells or tissues, using hybridization to biotinylated, bacterial genomic DNA. The obtained cDNA mixtures are then enriched for sequences which are transcribed preferentially during growth in the host, using additional hybridizations to bacterial genomic DNA in the presence of cDNA similarly prepared from bacteria grown *in vitro*. This method was developed by Graham in 1999 to identify genes expressed by *M. tuberculosis* in macrophages (Graham & Clark-Curtiss 1999). Several functional categories were identified related to DNA repair, nutrient acquisition, cell wall metabolism, virulence factors, membrane-associated protein, transcriptional activator, and phagocytosis response. The SCOTS approach has also been used with success in many Gram-negative bacteria.

1.3.3. *In Vivo*-Induced Antigen Technology (IVIAT).

In Vivo-Induced Antigen Technology (IVIAT) IVIAT accomplishes the same goals as IVET but does not rely on animal models. Instead, it identifies genes expressed during a human infection by using sera from patients to probe for genes specifically expressed *in vivo*. IVIAT was initially developed to study the oral pathogen *Actinobacillus actinomycetemcomitans* (Handfield et al. 2000). Sera from patients who have developed an infection caused by the pathogen under study are

pooled and absorbed with cells of the pathogen grown *in vitro*. Unbound antibodies are collected as they are expected to bind antigens that are expressed only *in vivo*. An expression library of the pathogen's DNA is generated similarly to IVET approach and clones are probed with the previously collected absorbed antibodies. Reactive clones, which are producing antigens that are expressed during a natural infection but not during *in vitro* cultivation, are purified and their cloned DNA sequenced. This approach was also applied to *Mycobacterium tuberculosis* and allowed identification of antigens specifically expressed or upregulated during infection and growth *in vivo*. Two enzymes appeared to be potential targets for drug development: DNA polymerase III and dihydrolipoamide dehydrogenase (Deb et al. 2002).

1.3.4. Microarray technology based on DNA oligonucleotides or expressed proteins corresponding to all genes in the pathogen's genome.

These methods allow one to analyse the expression of thousands of genes simultaneously using DNA microarrays, which typically contain ≈ 6000 spots of DNA on a 2×2 cm square. This is particularly useful to study host-pathogen interactions, as changes in bacterial gene expression also occur during infection or a particular growth condition. For example, a study has been conducted to investigate the gene expression of *S. aureus* under conditions for biofilm formation, in which bacteria are known to be more resistant to antibiotics and the immune system than their planktonic counterparts (Resch et al. 2005). In this study, mRNA was isolated from cells grown under both conditions and used for hybridization with DNA microarrays. Many genes involved in cell envelope were shown to be upregulated under biofilm formation conditions. These factors might contribute to survival, persistence and growth in a biofilm environment. Using physiological and biochemical tests, the authors have confirmed the up-regulation of urease, formate dehydrogenase, proteases and the synthesis of staphyloxanthin observed with the microarray.

Another type of microarrays displays hundreds to thousands of ordered recombinant proteins from a pathogen or its host. Recently, two host-pathogen interactions were studied by: 1) spotting 159 *S. aureus* recombinant proteins which were tested for binding to 75 human recombinant extracellular proteins, 2) spotting ≈2300 human recombinant proteins which were tested against *Neisseria meningitidis* adhesin (NadA), an important vaccine component against serogroup B meningococcus (Scietti et al. 2016). This study allowed the identification of the interaction between the *S. aureus* immune evasion protein FLIPr (formyl-peptide receptor like-1 inhibitory protein) and the human complement component C1q, key players in the offense-defense fighting. Also, an interaction between meningococcal NadA and human LOX-1 (low-density oxidized lipoprotein receptor), an endothelial receptor, was identified. These interactions were validated with functional tests.

1.3.5. Proteomic approaches.

Proteomic profiling characterizes the spectrum of proteins expressed in bacteria under varying growth conditions using two-dimensional gel electrophoresis and mass-spectroscopy. Although this approach has provided many insights on various host-pathogens interactions, this method is particularly sensitive to the sample preparation procedures. For example, membrane and cell wall proteins are usually extracted in low amounts due to their low abundance and solubility constraints that are caused by their hydrophobicity.

The work of Hempel *et al* (Hempel et al. 2011) illustrates the impact of sample preparation on the proteomic output. To get a comprehensive identification of *S. aureus* proteins which are surface-exposed or secreted under conditions of iron-limited conditions, the authors applied three different procedures for sample preparation: trypsin shaving of bacteria, biotinylation of bacteria, and precipitation of the supernatant. Each of the approaches led to identification of different subsets of

proteins detected by mass-spectrometry. A total of 1210 proteins were identified, and while these results were complementary, only 120 of them were detected by the different approaches. Finally, 29 proteins were found in altered amounts showing that surface-exposed proteins such as IsdA, IsdB, IsdC and IsdD were strongly induced.

1.3.6. Immunoproteomics.

An elegant means for elucidating the behaviour of a pathogen during encounters with its host is the interrogation of the host's immune system. This is based on the assumption that the adaptive immune system will only react to antigens to which the host has been exposed. However, it is also possible antigens expressed by the microorganism to be ignored by the immune system.

Numerous antigens from various pathogens have been purified or expressed in recombinant form, and antibodies from pathogen-exposed individuals have been tested for binding to them using techniques such as ELISA, peptide arrays and Western blots (for a review please refer to Holtfreter et al. 2010). This low-throughput method has been up-scaled to the so-called immunoproteomic or serum proteome approach - protein extracts of the microorganisms of interest are resolved onto two gels by two-dimensional gel electrophoresis and then, one of the gels is used for protein staining and identification. The other gel is processed in parallel which is then blotted onto membranes and incubated with serum antibodies. Following visualization of antibody binding, the immunoblots are matched with the protein gel and spots of interest are excised, digested and identified by mass spectrometry. In a pioneering work, Vytvytska *et al* (Vytvytska et al. 2002) used 100 sera from healthy individuals and patients suffering from *S. aureus* infections. They were screened for antibodies against staphylococcal lysates and recombinant proteins representing surface antigens. This led to the identification of 15 proteins including known and novel vaccine

candidates, including serine-aspartate repeat containing protein D, an immuno-dominant staphylococcal antigen.

1.4. Library-based display technologies.

A multitude of library-based display technologies has been developed and optimized over the last 30 years or so (for a review, please refer to Galan et al. 2016). In principle, such technologies are used for the identification of molecular interactions and have been widely used for directed evolution of various proteins such as enzymes and antibodies. More specifically, library-based approaches have been used for development of diagnostic and therapeutic purposes in cancer, autoimmune and infectious diseases. Depending on their mode of action, display technologies could be divided in categories. For example, *in vivo* and *in vitro* or cellular and acellular. However, no matter their specific performance all such technologies operate according to the following principle – first, a diverse library of desired molecules is created (protein, peptide, antibody, etc). The obtained library could be artificially diversified using various randomization approaches or could be derived from natural sources (genomic DNA). The library is then screened for several rounds against one or many ligands of interest which leads to enrichment of specific sequences encoding peptides/proteins with certain anti-ligand affinity. Finally, the selection output is analyzed for the desired phenotype. A particular group of library-based display technologies which we refer to as "genome-wide" have been developed and successfully used in the field of host-pathogen interaction. These techniques include ANTIGENome technology, Shotgun phage display and ribosome display and will be discussed in more details in the remaining part of this section as they have served as a basis for the development of the here presented method, GeXplore.

1.4.1. Shotgun phage display.

Development

In principle, phage display is a method for identification of peptide/protein-ligand interactions. It consists of fusing a diverse DNA library to a gene, encoding a phage coat protein and then screening the obtained phage library against a certain ligand of interest. Originally, the technology has been developed by Smith, showing that a foreign DNA sequence could be successfully inserted between the N- and C- terminal domains of the pIII minor coat protein, encoded by gene III of filamentous phage. Also, in this seminal work the author has shown that the foreign DNA insert does not interfere with the function of the coat protein, and more importantly, it is "immunologically" accessible when expressed on the phage particle (Smith 1985). Originally, the technique has been designed using filamentous phage and with its spread in the scientific community several strains of the Ff class, such as M13, fd and f1 have been adopted for its performance. Currently, the most established platforms are based on N-terminal fusion to the coat proteins pIII and pVIII, even though all five of them could potentially be used for display (Mullen et al. 2006). About 3-5 copies of the former and about 3000 of the latter are expressed per phage particle (Jacobsson et al. 2003). An improvement of the display on filamentous phage is the development of "phagemid" vectors which are artificial chimeras of phage and plasmid vectors. They contain two origins of replication (for phage and *E. coli*, respectively), gene III and/or gene VIII, multiple cloning site and antibiotic resistance gene (Russel et al. 2004). However, they are devoid of all other genes, necessary for complete phage assembly. Thus, they could be grown as plasmids in *E. coli* and then packed into phage particles using helper phage (Mullen et al. 2006). The main advantages of the phagemid expression system are two – first its gives the expression of only one copy of fusion protein per phage particle which allows the differentiation between weak- and strong-affinity phage particle and second, allows the expression of longer polypeptides since

the phage infection ability will be conferred by the helper phage (Jacobsson & Frykberg 1995). However, the filamentous phage-based display platform has been shown to have certain drawbacks, concerning the proper expression of the fused peptide/protein which has led to the development of alternative versions, using lytic phage platforms such as λ, P4, T4 and T7 (Castagnioli et al, 2001). Phage display has been used for several different applications such as selection of protein scaffolds, antibodies and peptides (Kügler et al. 2013). More specifically, in the field of infectious diseases, the techniques has been used for studying host-pathogen interactions, identification of potential vaccine candidates, epitope mapping and discovery of novel bacterial adhesins (Mullen et al. 2006). A particular genome-wide technology, shotgun phage display has been developed for the latter application (Jacobsson & Frykberg 1995). In this seminal work, the authors illustrated the potential of phage display for affinity selection of ligand-binding domains of bacterial receptors. The efficiency of the approach has been validated by choosing appropriate and well-characterized model system – a random genomic library of *S. aureus* with size of $9.2\ 10^6$ clones has been prepared and screened against human IgGs and fibronectin. After one-two rounds of selection, all five IgG-binding domains of protein A and the two fibronectin-binding domains of protein FnbA were successfully identified in the analyzed output. Furthermore, this study was the very experiment in which the second IgG binding protein of *S. aureus*, SpA, was discovered.

Principle and performance.

The general principle of phage display is illustrated on page 37 and shotgun phage display follows the same principle (Jacobsson et al. 2003). Logically, the method begins with extraction of a pathogen's genome which is then randomly sheared into fragments with a desired size using ultrasound. The use of restriction enzymes for fragmentation is not recommended as it could be associated with a sequence-dependent digestion pattern.

A)	Principle of phage display.

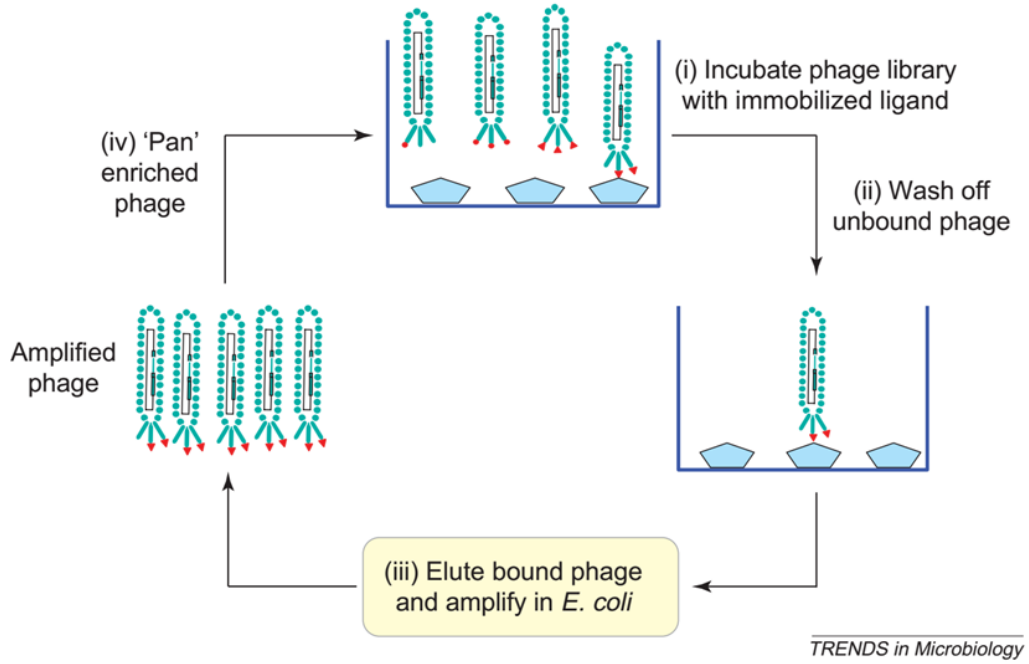


Figure 1.2. Taken from Mullen *et al*, 2006.

The fragment size range depends on the specific goal of the study – for example if ligand-binding genes are sought, the fragments need to be longer. In case of search for particular ligand-binding domains, the use of shorter fragments is advisable. In general, it has been shown that identification of clones possessing up to 1500 bp fragments using fusion to gene VIII of filamentous phage is possible (Jacobsson et al. 2003). After shearing, the ends of the genomic fragments are repaired with T_4 DNA polymerase and cloned into a blunt-ended phage or phagemid vector. The 3'-ends of the vector need to be dephosphorylated since blunt-end ligation are complicated by the high level of vector recircularization. Additionally, such kind of ligations require fine optimization since improper vector/insert ratio could lead to generation of undesired products like multiple empty vectors or fragment concatamers. Finally, the ligated library needs to be transformed into *E. coli* for enrichment of the ligated constructs. This step is considered as a drawback because

transformation is tedious procedure with low reproducibility. Furthermore, multiple transformation attempts might be required due to the high number of transformants needed. The number of clones required depends on the size of the genomic fragments and the size of the bacterial genome of interest. Some important things need to be considered when calculating the number of clones needed. First, theoretically, in such kind of random genomic libraries only one in 18 clones could be expected to be cloned in-frame with the vector in the proper orientation. Second, it is known that not all fragments are expressed with the same efficiency which would eventually result in biased library content. Obviously, a certain library size which sufficiently exceeds the minimum number of clones is recommended.

Once the phage library has been prepared, it is being screened against a ligand of interest at a step called panning. The remaining non-coated surface needs to be further covered using blocking reagent such as Tween 20 or Triton X100, or some kind of irrelevant protein. However, if the latter case is to be used, an additional step, pre-panning, needs to be included prior panning. This is to ensure that no phage particles with affinity for the blocking agent are being enriched during the panning procedure. Once the library has been let to interact with the ligand, the unbound particles need to be removed by performing certain number of washes. As it will be discussed in chapter 4 of the presented thesis, the washing step is crucial for display approaches since it determines the differentiation between specific peptide/protein-ligand interaction and unspecific background. Therefore, the number of washes might need to be empirically adjusted to the particular ligand(s) of interest. After washing, the bound phage particles are eluted at low pH, neutralized and used for infection of *E. coli*. The grown phage clones are collected for subsequent selection rounds (re-panning).

Applications.

Shotgun phage display has been used extensively for identification of bacterial adhesins (Jacobsson et al. 2003). For example, fibrinogen-binding protein Fbe of *Staphylococcus epidermidis* was discovered by panning shotgun phage display library against immobilized fibrinogen (Nilsson et al. 1998). Alternative version of the technology using the modified display vector of Crameri and Suter (Crameri & Suter 1993) has been also applied for epitope mapping using whole-genome libraries (Palzkill et al. 1998). In this study, a genomic library of *E. coli* MG1655 has been created and screened against polyclonal anti-RecA protein sera. Another study has reported the discovery of novel fibronectin-binding protein SFS from *Streptococcus equi*, which was identified by panning genomic library of *S. equi* Bd3221 against human fibronectin (Lindmark & Guss 1999). Following these early reports, the approach has been applied to multiple pathogens including *S. aureus* (Bjerketorp et al. 2002), *Streptococcus dysgalactiae* (Almeida et al. 2000), *Lactobacillus casei* (Muñoz-Provencio et al. 2011), *Staphylococcus hyicus* (Rosander et al. 2011) and *Leptospira interrogans* (Ching et al. 2012; Lima et al. 2013)

1.4.2. ANTIGENome technology.

Development

ANTIGENome technology is a cell-surface display technology which has been designed for identification of putative vaccine candidates out of genomic libraries. Originally, the platform has been developed by first choosing surface proteins of *E. coli* which could provide optimal display of random genomic libraries with various fragment size range which could be easily sorted using magnetic cell sorting (MACS) selection (Etz et al. 2002). The display abilities of four membrane proteins in *E. coli* – FhuA, BtuB, OmpA and LamB were analyzed (Hildegard et al. 2001). All these proteins serve as phage receptors. Protein Fhu is an outer membrane receptor protein which

facilitates the uptake of iron from the environment (Coulton et al. 1986). Protein BtuB is also an outer membrane protein, which is the cell receptor, responsible for binding and uptake of vitamin B_{12} (Heller & Kadner 1985). The third protein, OmpA is one of the most abundant and well-studied membrane proteins of *E. coli* which has been shown to be involved in the cell interaction with bacteriophages and colicins as well as in conjugation (Freudl 1989). Later on, it has been proposed that the protein play a crucial role for the structural integrity of the cell outer membrane (Wang 2002). The last of the chosen proteins, LamB is an outer membrane proteins with two functions – it participates in the uptake of maltose and maltodextrins and serves as phage receptor (Charbit et al. 1988). In their study, Etz *et al* have chosen to examine several outer membrane proteins for the following reasons. First, as already mentioned, the main goal of the study was the identification of optimal platforms for surface display of random genomic libraries. Therefore, the highly-diverse nature of such libraries might require the use of several alternative platforms which could allow for proper expression of wider range of foreign polypeptides since individual platforms have been shown to have limited display abilities. Second, the family of outer membrane proteins provides multiple platform candidates and several had already been successfully used for the display of short synthetic peptides. Finally, several foreign peptides could be expressed in fusion to the multiple extracellular loops of such proteins. The display ability of the four proteins has been characterized using well characterized epitopes T7tag and myc. Inserts of various size have been fused to various loops of the proteins and their display assessed by Western blot, fluorescence-activated cell sorting (FACS), MACS and sensitivity to bacteriophages and colicin. Accordingly, the study shows that the most potent protein is FhuA with efficient display of up to 250 amino acids, followed by BtuB with inserts of at least 86 amino acids. However, the other two proteins, OmpA and LamB, were found to be less efficient with up to 40 amino acids since insertion of longer fragments resulted in impaired transport or assembly to the outer membrane. Additionally, the expression of OmpA

fusions was found to be quite low probably due to competition with the abundant wild-type protein on the cell surface. Therefore, FhuA and BtuB were shown to be suitable platforms for expression of longer polypeptides, while LamB could be used for display of shorter ones. Together, the three proteins were proposed as a panel of platform proteins, suitable for cell surface display of highly diverse genomic libraries. Shortly after the optimization of the display platforms was released the first report for application of ANTIGENome (Etz et al. 2002). A random genomic libraries of the methicillin-resistant *S. aureus* COL strain with various fragment sizes have been prepared and screened against high-titer sera obtained from patients with various *S. aureus* infections as two of the mentioned platform proteins – LamB and FhuA have been used for display. As a result, a set of 60 proteins have been identified as most of them have been found to be surface exposed or secreted. This study contributes to the field of genome-wide display technologies by highlighting their potential for studying host-pathogen interactions. Furthermore, from technical point of view this study provides detailed information about the library preparation procedure which was further extended in a following paper focused on library quality and representativity (Henics et al. 2003).

Principle and performance.

With regards to host-pathogen interaction, ANTIGENome was developed as a method for identification of immune-relevant proteomes of human pathogens using patient-derived antibodies. As such, it is obvious that its proper performance depends on two main requirements – high-quality genomic libraries of the pathogen of interest as well as collection, characterization and selection of disease-relevant sera. The principle of the method is represented on page 42. It follows the general concept of a genome-wide display technology. The method starts with extractions of the pathogen's genome which is then fragmented into fragments with a certain length range.

A)	Principle of ANTIGENome technology.

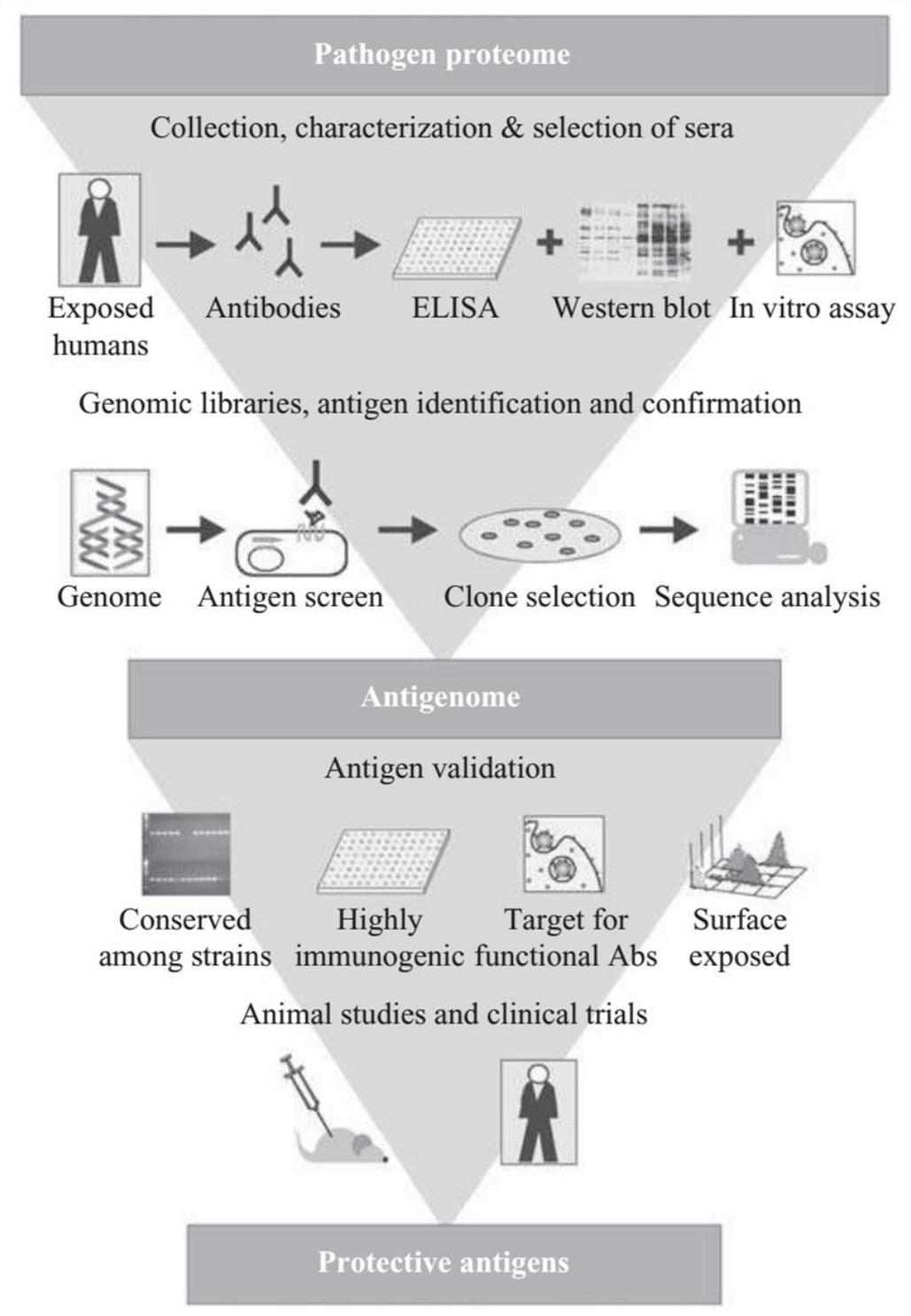


Figure 1.3. Taken from Nagy *et al*, 2004.

Since the method relies on antibody/antigen interaction, libraries of two different fragment lengths are prepared – 30-50 bp and 150-300 bp to facilitate selection of linear and conformational epitopes, respectively. Short DNA fragments are prepared using DNAse I shearing since ultrasound fragmentation is not efficient and reproducible below 100 bp (Henics et al. 2003). Longer DNA fragments are prepared using ultrasound shearing. Once the fragments have been prepared, their ends are carefully repaired with T_4 DNA polymerase, followed by ligation into a SmaI-digested vector. The next step is the enrichment of the ligated construct via transformation in *E. coli* which we consider as a main disadvantage of this technology. Due to the random nature of the genomic library, achieving a good representation of the pathogens genome would require obtaining a sufficient number of clones which would depend on the chromosome's size. Also, similarly to the phage display libraries, discussed in the previous section only one in 18 clones will be cloned in-frame with proper orientation.

Therefore, multiple transformation steps will be needed which are tedious and with low reproducibility. In order to increase the efficiency of this step, prior to ligation into surface display vector, the genomic fragments are first ligated into a frame-selection vector pMAL4.1, which is constructed on the backbone of pEH1 (Etz et al. 2002; Henics et al. 2003). The vector harbors an out-of-frame β-lactamase gene, which is located downstream of a cloning site containing SmaI restriction site for library insertion. The vector has been designed in such a way that a +1 frame shift should restore the proper expression of the β-lactamase gene, thus resulting in ampicillin-resistant clones. Once the library has been ligated in pMAL4.1, the obtained constructs are transformed into *E. coli* and grown on ampicillin-containing plates. The next step is an important contribution of ANTIGENome to the field of display technologies – even if the method has been developed before the development of the massively-parallel DNA sequencing, a library quality control was still included using classical Sanger sequencing.

After frame selection, about 500-1000 randomly picked clones are being sequenced and then mapped to a publicly available genome (if present) of the pathogen of interest. While this number of clones is insufficient to provide detailed information about the quality of the genomic library, it does give some useful information such as size range of the cloned genomic fragments as well as their distribution over the reference genome. After the genomic fragments have been frame-selected and their quality validated by sequencing, they are excised from the frame-selection vector and transferred into the surface display vectors, encoding the platform proteins FhuA and LamB. This step has two main disadvantages – first it requires directional cloning of the frame-selected fragments which could be achieved with restriction enzymes. However, due to the genomic nature of the library, the use of rare-cutting enzymes is needed which narrows the applicability of the method to genomes for which such kind of enzymes are available. And second, this sub-cloning step is performed *via* second transformation in *E. coli*, which reduces additionally the coverage of the original input library. Finally, once the library has been transferred into the display vectors, it could be screened against human sera. The reasons for using human sera in selection is the host-to-host variation – model animals could show different disease manifestation and mount different immune response than humans which could lead to identification of irrelevant proteins.

Applications.

ANTIGENome has been used essentially for identification of novel vaccine candidates. The first reports have been focused on *S. aureus* and *Streptococcus pneumoniae* (Etz et al. 2002; Giefing et al. 2008). Later on the approach has been applied to larger array of human pathogens such as *Staphylococcus epidermidis*, *Streptococcus agalactiae*, *Streptococcus pyogenes* (Meinke et al. 2005), *Helicobacter pylori* (Meinke et al. 2009), *Borrelia* (Poljak et al. 2012), *Moraxella catarrhalis* (Smidt et al. 2013) and *Klebsiella pneumoniae* (Lundberg et al. 2013).

1.4.3. Ribosome display.

Development.

In principle, ribosome display is an in vitro method for identification of protein-ligand interactions, allowing for fast screening of highly diverse DNA libraries. The main advantages of this method over phage or cell surface display techniques is the faster performance and superior size of input library. The idea about ribosome display has been inspired by the early works on immunoprecipitation of polysomes. In these works it has been shown that an mRNA transcript, encoding a certain protein could be identified from an mRNA library by immunoprecipitating nascent polypeptides on polysomes (Korman et al. 1982; Kraus & Rosenberg 1982). Later on, similar ideas have been proposed by others. For example, in their seminal paper on systematic evolution of ligands by exponential enrichment, also known as SELEX, Tuerk and Gold proposed the use of ribosome display as an *in vitro* evolution technique, due to the possibility of selecting ribosome-associated mRNAs against a desired target (Tuerk & Gold 1990). Also, similar approach for cell-free screening of novel genes and polypeptides has been proposed by Kawasaki in an early patent application (Kawasaki 1997). Eventually, the earliest attempts for ribosome display have been reported by Mattheakis *et al*, who used "polysome" display of short peptides for identification of peptide-ligand interactions (Mattheakis et al. 1994; Mattheakis 1996). A diversified library of decapeptides has been screened for four rounds against an immobilized monoclonal antibody in a coupled *in vitro* transcription/ translation *E. coli* S30 system and clones with binding affinities of 7-140 nM have been identified. These works represent one of the first attempts for *in vitro* screening of highly-diverse libraries of up to 10^{12} members, demonstrating their potential for identification of peptide-ligand interactions. However, it has been argued that this method is not efficient for studying the interaction of full-length proteins with ligands since it would depend on their native conformation which will be assumed only at the end of the mRNA transcript. Therefore,

just the very last polypeptides on the polysome could be expected to be functional (Hanes & Plückthun 1997). Eventually, the development of the method has been completed by Hanes and Plückthun, who reported the use of ribosome display for expression of full-length proteins (Hanes & Plückthun 1997). In this work, the authors reported the successful *in vitro* expression of a single-chain Fab fragment variable (scFv) on the surface of stalled ribosomes. The fragment has been properly folded and selected 10^8-fold for five antigen-affinity selection rounds.

From technical point of view, this study reports several crucial improvements of the Mattheakis' polysome concept which have significantly improved the method's performance and extended the range of its applications. Namely:

- To avoid the folding problem mentioned above, an M13-derived 88-amino acid spacer (also called tether) has been fused to the C-terminus of the scFv fragment, thus allowing it to protrude from the ribosome tunnel and to assume its native three-dimensional structure.
- The stop codon in the display construct has been removed, causing the so-called ribosome "stalling" which is the essence of the method – by being stalled, the ribosome links the genotype (the mRNA transcript) to its phenotype (the nascent polypeptide chain). The mRNA-ribosome-polypeptides assemblies are called ternary complexes, which after *in vitro* translation are being stabilized at low temperature on ice and 50 mM magnesium acetate.
- The efficiency of ribosome display has been found to be 2 orders of magnitude higher when *in vitro* transcription and translation were performed separately.
- 5' and 3' loops have been introduced in the display construct to protect the mRNA transcripts from exonuclease degradation. Collectively, the study reports a 15 times improvement of the selection efficiency after introduction of the C-terminal tether and terminal loops.

Principle and performance.

The mode of action of ribosome display is illustrated on page 48 (Pluckthun 2012). The method begins with a starting DNA library which encodes certain peptide(s)/proteins(s) of interest. This initial library must be then converted into the so-called ribosome display construct containing all functional and structural regions for efficient *in vitro* selection. This step is accomplished by ligating the starting (input) library into the multiple cloning site (MCS) of a ribosome display vector (pRDV). In general, directional or blunt-end cloning is used at this step, depending on the nature of the input library (synthetic, gene, genomic). However, in some cases alternative cloning strategies might be required (described in chapter 3). Upstream of the vector's MCS are positioned a T7 promotor, a 5'-loop region and a ribosome binding site (RBS). Downstream of the MCS are located a spacer which tethers the nascent polypeptide chain to the ribosome during translation and the 3'-loop region. Intentionally, the construct does not contain a stop codon. After the library fragments have been ligated into the RDV vector, the obtained constructs are used for further conversion of the library into a linear template for *in vitro* transcription. This step is accomplished by amplifying the ligation products with primers, specific for the ends of the ribosome display construct. Once the input DNA library has been prepared, it is *in vitro* transcribed and the obtained mRNA is used for *in vitro* translation. This is the step at which the genotype is coupled with the phenotype – ternary ribosomal complexes between the nascent polypeptide chain (library member), the ribosome and the mRNA are formed. The bulk of ribosomal complexes are then let to interact with the ligand (target) of interest which could be performed on a solid surface (e.g. microplate) or in solution followed by capture of the formed complexes on the surface of magnetic or other types of beads (agarose, etc). The choice of selection matrix might be crucial for successful experiment since it could determine the enrichment of desired molecules or level of unspecific output (background).

A)	Principle of ribosome display.

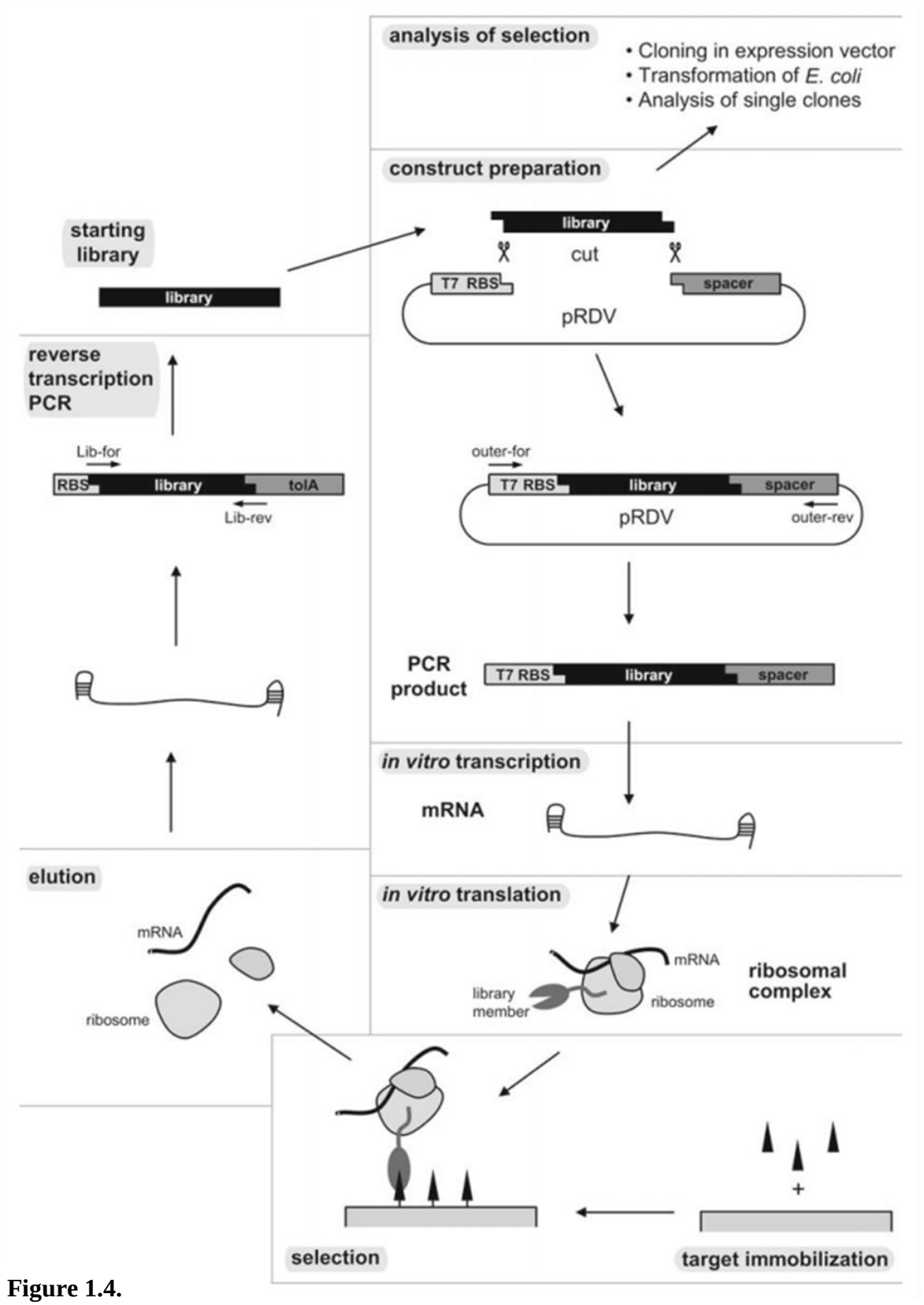


Figure 1.4.

Taken from Plückthun *et al*, 2012.

After certain number of washes, which may need to be empirically determined, the retained complexes are destabilized by the addition of chelating agent (EDTA), eluted and reverse-transcribed. This step could be performed using flanking or internal primers. Finally, the obtained cDNA is amplified by PCR and thus the output is ready to be used for another selection round.

Eventually, after several number of rounds (for example 3-5) the final output is sub-cloned in expression vector, transformed in *E. coli* and single clones are further analyzed for the desired phenotype.

Applications.

Since its advent in late 1990s, ribosome display has been extensively used for a wide range of applications which have been recently reviewed (Douthwaite & Jackson 2012). These include optimization of antibody affinity (Lewis & Lloyd 2012), evolution of protein stability (Buchanan 2012) and selection of Sac7d scaffolds (Mouratou et al. 2012) to mentioned but a few. Unfortunately, the use of the method in the field of infectious diseases has been limited. The first report for the use of ribosome display as a genome-wide approach was published by the developers of the previously discussed ANTIGENome technology (Weichhart et al. 2003). In this report a random genomic library of methicillin-resistant *Staphylococcus aureus* COL strain has been screened against high-titer sera, obtained from patients with *S. aureus* wound- and catheter-related infections (Etz et al. 2002). After four rounds of selection, most of the identified clones were found to map on 75 genes, predicted to encode mainly surface-exposed or secreted proteins.

With regards to the application of ribosome display to infectious diseases, this study reports on the following important findings:

- The ability of ribosome display to select for immunoreactive peptides has been validated using myc epitope and anti-myc antibody. At first round the myc mRNA has been

intentionally provided at 1:1x10^6 ratio with unrelated fragment, respectively. Importantly, it has been shown that after only three rounds of selection, the myc epitope was dominant in the selection output. This result correlates with the finding of Hanes and Plucthun and underscores the potential of ribosome display for identification of antigen/antibody interaction (Hanes & Plückthun 1997).

- After genome-wide screening, most of the identified clones were found to map over open-reading frames (ORF) assigned to cell envelope function (25%), transporter proteins (25%) or pathogenesis (9%). These findings confirm that ribosome display could be used to identify immune-relevant proteins by selection against infection-relevant antibodies.
- The theoretical advantage of ribosome display over cell surface system has been experimentally confirmed. Indeed, when ORFs identified *in vitro* were transferred to an FhuA-based platform in *E. coli* (Etz et al. 2002), about 55% of the clones could not be expressed.

Even though these findings highlight the potential of ribosome display for studying host-pathogen interactions and confirm its advantage over related technologies, one important drawback could be still pointed out – the random genomic library of *S. aureus* has been prepared according to the blunt-end cloning strategy described for ANTIGENome (Etz et al. 2002). This procedure relies on subsequent enrichment of the ligated genomic fragment *via* transformation in *E. coli*. However, this step has been pointed as a bottleneck since it reduces the coverage of the obtained library due to its low efficiency. Furthermore, the step is labor-intensive and demanding. Finally, as it will be described in chapter 3 of the presented thesis, such ligation strategy is not compatible with fully *in vitro* library preparation by PCR.

Unfortunately, to our knowledge, there has been only one other attempt to use ribosome display in a similar context (Lei et al. 2008). In this study, cDNA libraries of the swine pathogen

Actinobacillus pleuropneumoniae were prepared and screened for one round against serotype specific antibodies. Even though the paper reports on the discovery of potential vaccine candidates, it is not technically directed and, therefore, does not provide significant findings about the selection process itself. Additionally, since the experiments have been performed with cDNA, the content of the input libraries could be expected to be reduced compared to a whole-genome library because it would be dependent on culture growth conditions.

However, a book chapter written by the first author of this study, Lei (Lei 2012), provides very detailed protocols for every step of the selection process. Especially important for the development of our method was the described library preparation protocol. In contrast to the study of Weichhard *et al*, the protocol recommends single nucleotide overhang-assisted cloning strategy – the 3'-ends of the genomic fragments are dA-tailed *via* the non-template nucleotide addition ability of Taq polymerase and subsequently ligated into pGEM-T vector. This procedure is believed to have superior efficiency than the problematic blunt-end cloning. Furthermore, as it will be illustrated in chapter 3, even if we did not use this strategy, it inspired the development of the G/C-assisted cloning strategy which allowed us to avoid the tedious transformation in *E. coli*.

1.5. Objectives and overview of the presented thesis.

Here, we present the development of a completely *in vitro* genome-wide display approach to studying host-pathogen interactions, which we term **GeXplore**, standing for ***Ge***nome ***Explor***ation. A schematic of the approach is presented on page 53. Our method is based on ribosome display and takes advantage from its ability to couple phenotype with genotype as well as to work with highly-diverse libraries of up to 10^{12} molecules.

Through the course of the thesis, our main objectives were:

- Preparation of reference genome sequences in order to facilitate strain-specific analysis of input libraries and selection outputs.
- Development of a general procedure for *in vitro* library preparation, devoid of the tedious and low-reproducible transformation step.
- Characterization of the prepared genomic libraries using Illumina next-generation DNA sequencing (NGS).
- Optimization of the selection conditions in order to obtain optimal specificity.
- Pilot validation of the method's performance under single-ligand conditions by performing a complete selection cycle (3 consecutive rounds) based on well-characterized protein-protein interactions and characterization of the selection outputs by Illumina NGS.
- Challenging the method's performance under multi-ligand conditions by screening genomic libraries of *Streptococcus gallolyticus* and *Mycobacterium ulcerans* for three selection rounds against patient-derived antibodies.

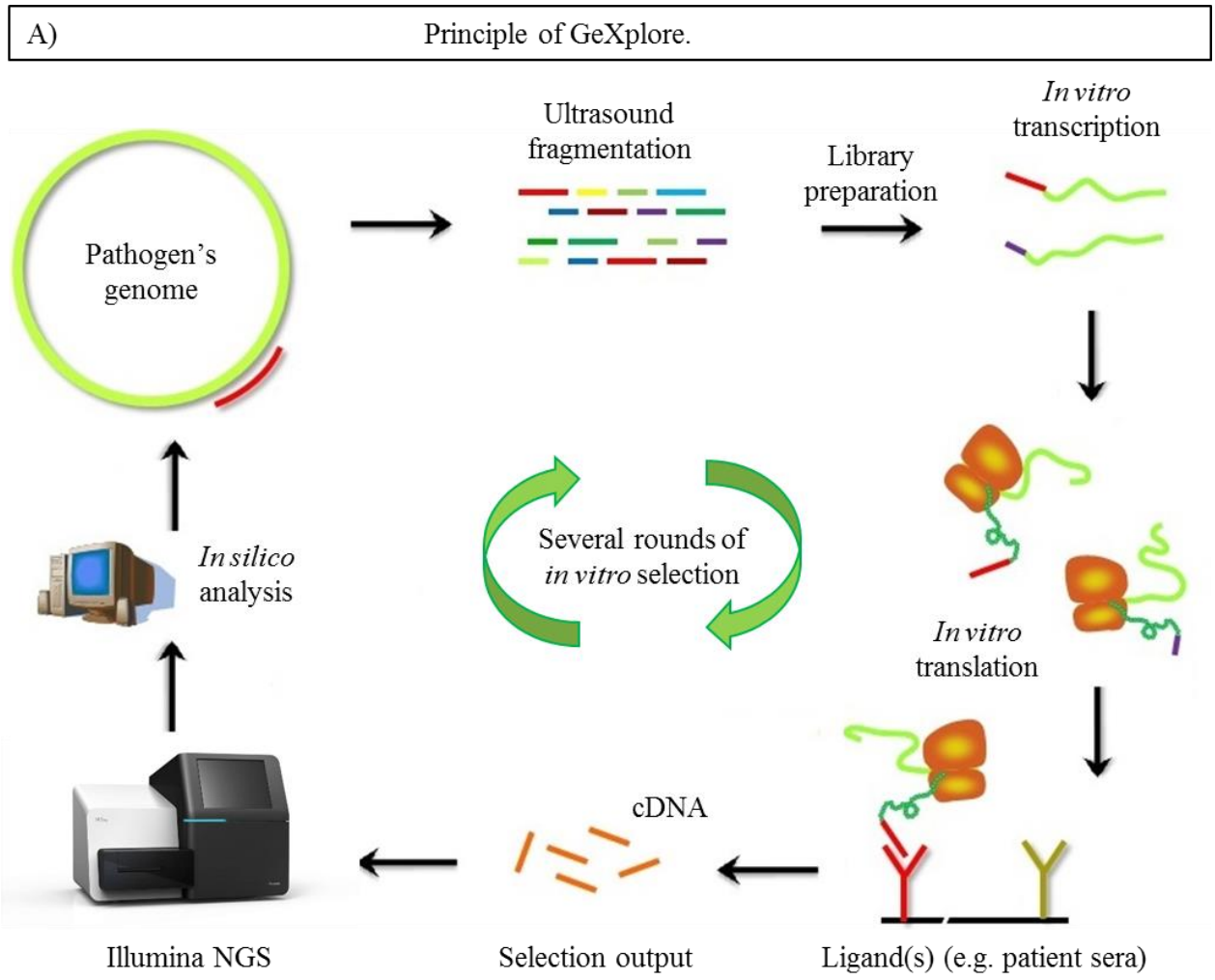


Figure 1.5.

Previous genome-wide approaches such as phage and cell surface display have been used for identification of multiple bacterial adhesins and putative vaccine candidates. However, certain limitations of these approaches, including library coverage and labor-intensive performance have been reported, which we believe could be avoided by the use of ribosome display. Unfortunately, only two such attempts have been performed earlier and in both of them library preparation was performed *via* transformation. And finally, the performance of these approaches have not yet been characterized by currently available NGS platforms. Therefore, we approached the development of our method by focusing on these particular points.

The first part of **chapter 1** addresses classical and more recent approaches to studying host-pathogen interaction and reflects the transition from low-throughput biochemical and genetic methods to more efficient proteomic and genomic techniques allowing for more detailed analysis of this complex relationship. The second part of the chapter introduces the three most-commonly used genome-wide approaches in the field of infectious diseases since they served as a basis for the development of our method.

In **chapter 2** we describe the preparation of reference genome sequences which were used for analyzing the quality of the input libraries and selection outputs. Draft genome sequences of *S. aureus, S. gallolyticus* and *M. ulcerans* were sequenced with Illumina platform, *de novo* assembled at a draft genome level and partially characterized with respect to their strain-specific mobiliomes.

In **chapter 3** we present the development of *in vitro* library preparation procedure based on PCR amplification right after fragment ligation. An alternative GC-assisted cloning strategy was developed and used for the preparation of random genomic libraries of *S. aureus*, *S. gallolyticus* and *M. ulcerans* which were then characterized by NGS. We report an important source of %G+C bias, leading to a coverage threshold at 29% G+C.

In **chapter 4** we explain the fine tuning of GeXplore's performance by optimization of various selection parameters and its pilot validation, based on *S. aureus*/ IgG interaction. We demonstrate that our method is highly-specific, at least under single-ligand conditions, and is able to identify ligand binding domains out of whole-genome libraries.

In **chapter 5** we challenge the performance of our method under multi-ligand conditions by screening genomic libraries of *S. gallolyticus* and *M. ulcerans* against patient-derived antibodies and characterizing the obtained outputs by NGS. We demonstrate that GeXplore is able to identify multiple regions out of whole-genome libraries, encoding for potentially antigenic/immunogenic proteins.

Chapter 2

Bacterial whole-genome sequencing and analysis

2.1. Introduction.

DNA is the ultimate source of information about a living organisms. The quest for DNA sequencing has begun soon after its chemical structure has been elucidated in 1953 (Watson & Crick 1953). However, it has taken two more decades for the classical Sanger biochemistry to arise (Sanger et al. 1977) and another 20 years before the first bacterial genomes have been completed (Fleischmann et al. 1995; Fraser et al. 1995; Hutchison 2007). These seminal reports have initiated a great deal of work which has led to a revolution in our understanding about the genetic basis and organization of microbial life. Unfortunately, Whole-Genomes Sequencing (WGS) has remained a luxury for small and middle research facilities for quite some time due to the still-high cost of performance per genome (Barbosa et al. 2014).

This disadvantage has hampered its wide spread until the development of the so-called "second" or "Next-Generation" Sequencing (NGS) technologies in 2005 which have made WGS affordable to many laboratories and have revolutionized the field of bacterial genomics (Shendure & Ji 2008; Zhang et al. 2011). The 454 GS FLX (Roche), Illumina (Genome Analyzer) and SOLiD (Life Technologies) are the most established platforms on the market, each having its own advantages and disadvantages (Metzker 2010; Van Dijk, Auger, et al. 2014) These technologies have dramatically reduced the cost of WGS and have made it possible to sequence a complete bacterial genome in a matter of hours.

Consequently, we have witnessed an explosion in the availability of prokaryotic genomic information - at the time of writing (August, 2016) there were more than 72 000 publicly-available prokaryotic genome sequences (NCBI, 2016) and their number continues to grow exponentially. Unfortunately, as mighty as they may be, NGS technologies have had some negative consequences on the public nucleotide databases such as reduced genome quality and biased genome availability – almost 90% of the bacterial genomes in GenBank are currently at draft (partial) level and while some species are represented by thousands of genomes, the available genomic information about others is scarce (Land et al. 2015). Additionally, the value of a newly sequenced bacterial genome has been recently discussed with an accent put on the limitations of draft genome use (Barbosa et al. 2014). Accordingly, special attention has been recommended when approaching several types of analyses which quality may be impaired by the lack of genome "completeness" such as total gene number determination, horizontal gene transfer and pan- or phylogenomics studies. However, a recent study involving more than 32000 genomes has shown that only about 10% of the draft genomes studied were of significantly reduced quality (Land et al. 2015). Furthermore, another study has compared draft *versus* complete versions of the same genomes using various sequencing technologies and the authors have concluded that Illumina technology is the cost-effective choice for microbial WGS, associated with minimal loss of information (Mavromatis et al. 2012).

As mentioned in chapter 1, the main goal of our work was development of fully-*in vitro*, whole-genome display technology, GeXplore, for studying host-pathogen interactions. One of the most important steps of our approach is the mapping of a potentially enriched selection output to a reference genome sequence. Logically, if a gene encoding a certain selected protein is not present in the reference sequence it will not be mapped on the reference genome and therefore will be omitted during the analysis step. When our project was started in February 2013, the number of publicly available genomes for two of the species used in our experiments, *M. ulcerans* and *S.*

gallolyticus was extremely limited (one and four genomes, respectively). Furthermore, genome plasticity is a known trait among bacteria which could reach mega-base dimensions in some species (Land et al. 2015). All things considered, it is obvious that the need for whole-genome sequencing of the very strains studied by genome-wide approaches such as GeXplore might be a pre-requisite for their successful application.

In this part of the manuscript, we describe the WGS of *S. aureus* FP_SA_ST25, *M. ulcerans* S4018 and *S. gallolyticus* – NTS31106099 using Illumina technology. Their *de novo* assemblies were used as reference genomes at various stages of the development and application of our method. The drafts were also partially characterized by means of comparative genomics. Accordingly, a novel Tn*916*-like conjugative transposon called Tn*6263* which confers an antibiotic resistance gene cluster was discovered in *S. gallolyticus* NTS31106099 and its prevalence among other clinical isolates was partially investigated. Finally, the draft genomes of two other *S. gallolyticus* clinical isolates – NTS31301958 and NTS31307655 were sequenced in order to characterize another novel element, Tn*6331*, identified in their genomes by PCR screening.

2.2. Materials and methods.

All manipulations with commercial kits were performed according to the manufacturer recommendations unless when stated otherwise. All primers used in the present work are enlisted in the table on page 21.

2.2.1. Strains and culture conditions.

Strain *Staphylococcus aureus* FP_SA_ST25 is an isolate available in our laboratory collection. When needed it was grown from glycerol stocks overnight at 37°C on BHI streak plates (Sigma) and then 20-40 mL liquid culture was prepared from a single colony in BHI broth (Sigma) under

identical conditions with 150 rpm agitation. A collection of 60 clinical isolates of *Streptococcus gallolyticus* ssp. *gallolyticus*, mostly recovered from bloodstream infection, isolated at Nantes University Hospital (Centre Hospitalier Universitaire de Nantes, CHU) during the period between 2007 and 2015 were kindly provided by Stephane Corvec. Identification of the bacterial strains has been carried out using VitekMS® MALDI-TOF technology (bioMérieux, Marcy l'Etoile, France). When needed, liquid cultures were prepared as follows. Columbia agar plates, supplemented with 5% horse blood (Oxoid, United Kingdom) were inoculated from stock cultures followed by overnight incubation at 37°C in an atmosphere of 5% CO_2. Liquid cultures were prepared from single colony in BHI broth (Oxoid, Dardilly, France) under identical conditions. Strain *Mycobacterium ulcerans* S4018 was kindly provided by Laurent Marsollier from University of Angers. It has been originally isolated from a cutaneous lesion of a patient with confirmed Buruli ulcer in Benin, Africa. Strain cultivation has been performed as follows. After inoculation onto Lowenstein-Jensen medium, growth has been monitored weekly for 5 months. The strain has then been grown on Middlebrook 7H9 agar enriched with Oleic Albumin Dextrose Catalase growth supplement.

2.2.2. Genomic DNA extraction.

Genomic DNA from *M. ulcerans* was extracted as follows - 20 µL of 50 mg/mL lysozyme (Sigma) and 5 mg/mL of RNAse A (Sigma) were added to every 50 mg of cells followed by overnight incubation at 37°C. The cells were then centrifuged for 5 min at 8000 rpm at room temperature (RT), re-suspended in 700 µL of 1 x TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0) containing 1% SDS and 25 µg/mL of Proteinase K (Sigma), and incubated for 1 h at 50°C. The pellet was then centrifuged for 5 min at 8000 rpm at RT, re-suspended in 800 µL of Lysis C solution (6% guanidine hydrochloride, 1% Tween 20, 1% Nonidet P-40) and incubated for another 1 h at 37°C.

The cell suspension was transferred into a 2-mL screw-cap tube containing ~250 µL of glass beads (Sigma) and 500 µL of chloroform, and agitated in bead-beater for 2 x 30 second pulses at speed 5. The obtained lysate was centrifuged for 10 min at 13000 rpm and the aqueous phase was then extracted twice into a fresh tube containing 500 µL of chloroform and 100 µL of 1 x TE buffer. After centrifugation for 15 min at 13 000 x rpm at RT, the aqueous phase was transferred into a fresh tube containing 600 µL of isopropanol and 60 µL of 3 M sodium acetate, and precipitated on ice for 30 min. Finally, the pellets were washed with 70% ethanol, dried and dissolved in 1 x TE buffer. Genomic DNAs from *S. aureus* FP_SA_ST25 was extracted according to the explained in-house protocol with minor modification – namely, lysis was performed with 5 units of lysostaphin (Sigma)/ 50 mg of cells for 15 min at 42°C instead of with 1 mg/mL lysozyme overnight at 37°C. The beat beater agitation was omitted. Genomic DNA from *S. gallolyticus* NTS31106099 was extracted either using the commercial DNeasy Blood and Tissue Kit (Qiagen, Germany) or the explained in-house protocol with minor modification – cells were lysed with 1 mg/mL lysozyme for 1 hour at 37°C instead of overnight at 37°C. The beat beater agitation was omitted again. Amplifiable total DNA from all other studied clinical isolates of *S. gallolyticus* was performed with the InstaGene matrix (Bio-Rad). Concentration of all DNA preparations was determined using Nanodrop 2000c instrument and their quality was assessed on 1% agarose 1 x TAE gels.

2.2.3. Whole-genome sequencing.

Whole-genome sequencing was performed exclusively using Illumina next-generation sequencing by synthesis. Draft genome of *S. gallolyticus* NTS31106099 was sequenced in collaboration with Bo Segerman from the Veterinary Institute of Umea, Sweden. A sequencing library has been prepared using Nextera XT Library Preparation Kit (Illumina, USA) and sequenced on MiSeq sequencer (Illumina, USA). *De novo* assembly has been performed with SPAdes 2.5.1. Short and

low-coverage contigs were filtered out using CLC Sequence Viewer 7.0 or later (CLC Bio, Qiagen). All other genomes were sequenced in collaboration with the Genomics and Bioinformatics Core Facility of Nantes (GenoBiRD). Sequencing libraries were prepared with NEBNext Ultra DNA Library Prep Kit for Illumina (NEB) and sequenced on MiSeq sequencer (Illumina, USA). *De novo* assembly was performed and optimized using Velvet 1.2.10 (Zerbino & Birney 2008) and VelvetOptimizer 2.2.5 (Zerbino 2011), respectively. All contig reordering was performed using Mauve 2.3.1 (Darling et al. 2004) or later. When necessary, the multi-contig drafts were merged into a single DNA string using Artemis 16.0.0 (Rutherford et al. 2000).

2.2.4. Genome analysis and comparative genomics.

All draft genomes were submitted to NCBI GenBank and annotated through the NCBI Prokaryotic Genome Automatic Annotation Pipeline (PGAAP) (Angiuoli et al. 2008). General sequence analysis was performed with CLC Sequence Viewer 7.0 or later (CLC Bio, Qiagen). Whole-genome comparisons were performed using Mauve 2.3.1 or later (Darling et al. 2004), NCBI BLAST (Altschul et al. n.d.) or WebACT (http://www.webact.org/WebACT/home) (Abbott et al. 2005). Comparisons were visualized using CLC Sequence Viewer, Mauve, BRIG 0.95 (Alikhan et al. 2011) or Easyfig 2.2.2 (Sullivan et al. 2011). Whole genome sequence relatedness was estimated and visualized on the basis of OrthoANI value calculation and comparison using Oat 0.93 (Lee et al. 2016). The following analyses were performed essentially through the server of Center for Genomic Epidemiology (CGE) (Aarestrup et al. 2012) at http://www.genomicepidemiology.org/. Strain typing was performed using MLST 1.8 or later (Larsen et al. 2012). Acquired antibiotic resistance genes were identified using ResFinder 2.1 or later (Zankari et al. 2012). Identification of virulence genes and spa-typing of *S. aureus* FP_SA_ST25 were performed with VirulenceFinder 1.5 (Joensen et al. 2014) and spaTyper 1.0 (Bartels et al. 2014), respectively.

2.2.5. Antibiotic susceptibility testing and determination of erythromycin resistance determinants.

Antibiotic susceptibilities of *S. gallolyticus* clinical isolates were determined using VITEK2 AST cards (bioMérieux) and interpreted according to the EUCAST recommendations (www.eucast.org). Determination of erythromycin resistance determinants was performed by PCR according to Leclercq *et al* (Leclercq et al. 2005).

2.2.6. Prevalence of Tn*6263* or related elements in clinical isolates of *S. gallolyticus*.

The prevalence of Tn*6263* or related elements among the 60 *S. gallolyticus* clinical isolates mentioned in point 2.2.1 was determined by PCR. The presence of three specific regions of Tn*6263* was assessed according to Brouwer *et al* (Brouwer et al. 2011). Accordingly, primers Tn*6263*_1 to Tn*6263*_4 were designed to anneal 100 bp up- and downstream of the transposon ends (Tn*6263*_1+Tn*6263*_2 for right end and Tn*6263*_3+Tn*6263*_4 for the left one), thus giving two amplicons of ~200 bp. Primers rec_F and rec_R were designed specific for the amplification of a 997-bp amplicon from the serine recombinase gene (UG96_07020 from the draft genome sequence of strain NTS31106099, acc. no. JYKU00000013). All screenings were performed in 25-µL PCRs with the following composition: 1 x HF buffer (Thermo Fisher Scientifc), 0.2 mM dNTPs (Thermo Fisher Scientifc), 0.5 µM primers, 4% DMSO, 0.5 units of Phusion HotStart II DNA Polymerase (Thermo Fisher Scientifc) and 10 ng of total DNA. Amplification was performed in MJ Mini thermocycler (Bio-Rad) using the general program [10 sec/98^{O}C, 30 x (10 sec/98^{O}C, 30 sec/T_{anneal}^{O}C, 20 sec/72^{O}C) and 5 min/72^{O}C] and monitored on 1.5 % agarose 1 x TAE gels.

2.2.7. Molecular typing.

Genetic relatedness among 17 isolates of *S. gallolyticus*, containing the recombinase of Tn*6263* was investigated by Pulsed-Field Gel Electrophoresis (PFGE) typing. Whole-cell DNA was digested with SmaI restrictase overnight at 25°C and migration was performed through 1% agarose/0.5 x TAE gel using CHEF-DR II instrument (Bio-Rad). PFGE-profiles were analyzed using BioNumerics software (Applied Maths, Saint-Martens-Latem, Belgium) and interpreted according to Tenover's criteria (Tenover et al. 1995).

2.2.8. Nucleotide accession numbers.

The draft genome sequences of *S. aureus* FP_SA_ST25, *S. gallolyticus* NTS31106099, NTS31301958, NTS31307655 and *M. ulcerans* S4018 studied in this work have been deposited at DDBJ/EMBL/GenBank under the accession no. LXFD00000000, JYKU00000000, MAMV00000000, LXFC0000000 and MDUB00000000, respectively.

2.3. Results.

2.3.1. Whole-genome sequencing, *de novo* assembly and annotation.

A comprehensive table summarizing each step of the whole-genome sequencing, *de novo* assembly and annotation is presented on page 63. The column "Feature" contains information and various metrics which may help the reader in understanding the performance and interpreting the whole-genome sequencing, *de nov*o assembly and annotation steps. The %Q30 parameter is a standard metrics for qualifying Illumina sequencing run, calculated by the FASTQC software (Brown, 2015). It represents the percentage of reads with sequence quality ≥ 30 on the classical Phred scale (Ewing et al. 1998), which indicates a base-call accuracy of 99.9%.

Table 2.1. Whole genome sequencing and analysis.

Step	Feature	Strains				
		S. aureus FP_SA_ST25	*M. ulcerans* S4018	*S. gallolyticus* NTS31106099	*S. gallolyticus* NTS31301958	*S. gallolyticus* NTS31307655
Next-generation sequencing (NGS)	NGS Library	NEBNext Ultra	NEBNext Ultra	Nextera	NEBNext Ultra	NEBNext Ultra
	NGS technology	Illumina MISeq	Illumina MISeq	Illumina MISeq	Illumina MISeq	Illumina MISeq
	Total number of reads	3078364	1807792	10190802	1499382	1807792
	%Q30	98.35	93.61	*	98.35	98.07
	%GC	33	64	39	37	37
De novo assembly	Assembler	Velvet 1.2.10	Velvet 1.2.10	SPAdes 2.5.1	Velvet 1.2.10	Velvet 1.2.10
	Optimization	Velvet Optimizer 2.2.5	Velvet Optimizer 2.2.5	*	Velvet Optimizer 2.2.5	Velvet Optimizer 2.2.5
	Number of iterations	7	8	*	7	7
	Optimal hash value (bp)	133	113	*	121	121
	Optimal coverage cutoff	0.76	4.3	*	0.85	0.71
	Total number of contigs	39	505	350	22	30
	Shortest contig (bp)	265	225	857	241	241
	Longest contig (bp)	1336980	160017	583716	1180450	600556
	Number of contigs > 1 kb	7	265	17	14	20
	N50 (bp)	1001174	37602	226282	1180450	272370
	Total bases in contigs	2782532	5402811	2475980	2330998	2332206
	Total bases in contigs > 1kb	2770670	5324987	2311566	2327965	2328247
Automated annotation	Annotation method	PGAAP/ GeneMarkS+	PGAAP/ GeneMarkS+	PGAAP/ GeneMarkS+	PGAAP/ GeneMarkS+	PGAAP/ GeneMarkS+
	Total number of genes	2874	**	2302	2320	2321
	CDS	2773	**	2198	2253	2254
	Pseudo genes	58	**	38	42	49
	CRISPR arrays	-	**	2	2	2
	rRNA genes	101	**	6	7	7
	tRNA genes	59	**	59	56	56
	ncRNA genes	4	**	1	4	4
	Accession #	LXFD00000000	MDUB00000000	JYKU00000000	MAMV00000000	LXFC00000000

** - these values are not available since we did not obtain the data from the collaborator.*

*** - these values are not available since the draft genome has not yet been released.*

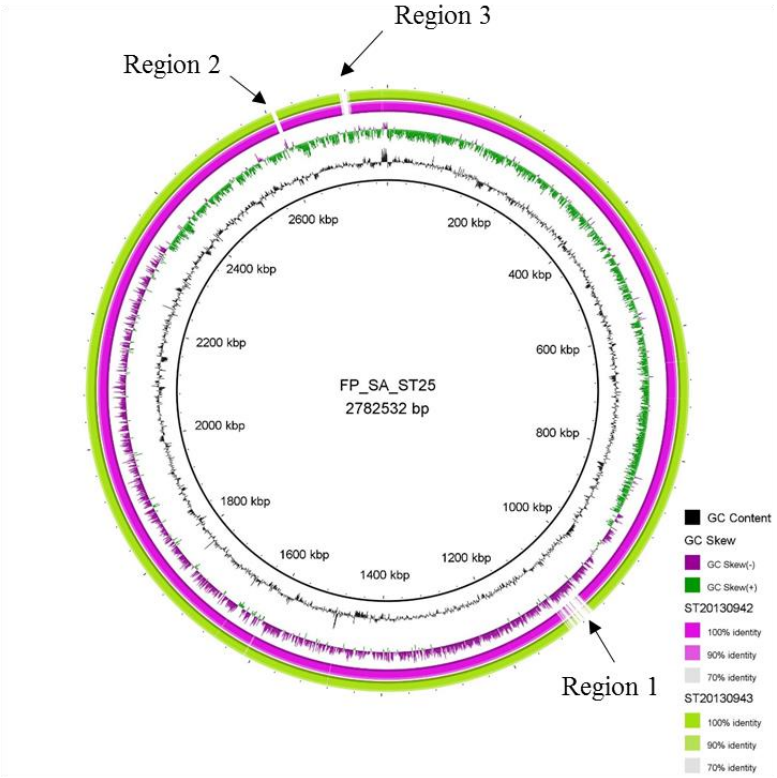
Another important parameter is the overall G+C content (%GC) of the sequenced DNA sample as it may indicate a bias in the sequencing library content if an expected value for it is already available (Brown, 2015). In our particular case, the %GC content of all 5 samples was found to be very close to the expected %GC values – 32.7% for *S. aureus* (Kuroda et al. 2001), 37.6% for *S. gallolyticus* (Rusniok et al. 2010) and 65% for *M. ulcerans* (Stinear et al. 2007; Röltgen et al. 2012). However, a good example about how this parameter may indicate a serious problem in the library quality will be discussed in chapter 3 of the manuscript. Hash length, also known as k-mer length, and coverage cutoff are parameters which values have direct impact on the quality of *de novo* assembly with Velvet assembler (Zerbino & Birney 2008) but they need to be empirically determined for each genome. In order to streamline this step, a special script called VelvetOptimizer has been developed to automatically optimize these parameters by iterative alterations (Zerbino 2011). The number of iterations and the optimal values of these two parameters are indicated for each assembly in the table. The N_{50} is one of the most commonly used quality metrics for genome assembly describing its contiguity (Yandell & Ence 2012). It is defined as the sequence length N for which 50% of the total number of bases in the assembly are contained in sequences with length $L<N$ (paraphrased definition of Broad Institute, https://www.broadinstitute.org/crd/wiki/index.php/N50). Generally speaking, the longer the N_{50}, the better the assembly. Note that the draft genome of *M. ulcerans* S4018 has the biggest total length (5.4 Mb) but the lowest N_{50} (38 kb) indicating quite low assembly quality. Regarding the annotation step, we think that discussion about the total gene content of the assemblies, other than mere number indication is inappropriate due to the draft level of the genomes.

2.3.2. Comparative genomics.

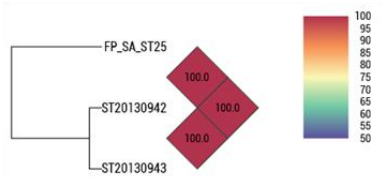
Staphylococcus aureus FP_SA_ST25

A good approach to properly order the contigs of a newly sequenced draft genomes is to use the most closely related available complete genome of the species (Edwards & Holt 2013). Having in mind the myriad of *S. aureus* complete genomes in GenBank we used BLAST search through the non-redundant database (nr/nt) to identify the best reference sequence. We identified two recently released complete genomes (Trouillet-Assant et al. 2016) which show more than 99% nucleotide identity to our strain. The ring diagram (panel A) on page 66 represents BLASTn whole-genome comparison of our draft genome to the complete genomes of the two best BLASTn matches – *S. aureus* isolates ST20130942 (acc. no. CP012976) and ST20130943 (acc. no. CP012974). These genomes share 100% and 99.97% OrthoANI value between each other and FP_SA_ST25, respectively (shown on panel B). All three isolates were confirmed to belong to ST-25 of the Enright MLST scheme (Enright & Day 2000). Currently, the PubMLST isolate database (http://pubmlst.org/) contains about 138 isolates (mainly invasive, methicillin-sensitive) under this sequence type with either human or animal origin. Identification of acquired antibiotic resistance genes with ResFinder indicated the presence of *nor*A and *bla*Z genes, encoding an efflux-mediated fluoroquinolone resistance and a beta-lactamase, respectively. The former gene is 91.59% identical to the wild type *nor*A gene of *S. aureus* SA-1119 (acc. no. M97169) (Kaatz et al. 1993) and does not seem to be associated to a particular mobile element. The latter gene shows 99.88% identity to the Tn*552*-related *bla*Z gene of *Staphylococcus haemolyticus* (acc. no. NVH97A) (Anthonisen et al. 2002). Strain FP_SA_ST25 was found to contain three additional regions compared to the reference genomes (indicated with black arrows as regions 1, 2 and 3 on panel A), accounting for 2.42% of its genome.

A) *S. aureus* FP_SA_ST25 against strain ST20130942 and ST20130943



B) Relatedness between the three genomes based on OrthoANI comparison.



C) Relatedness between the three regions and their reference elements.

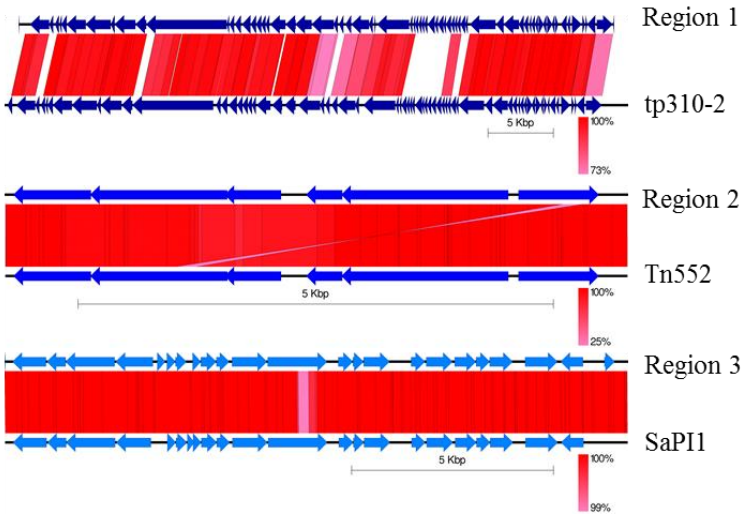


Figure 2.1. Comparative genomics of *S. aureus* FP_SA_ST25.

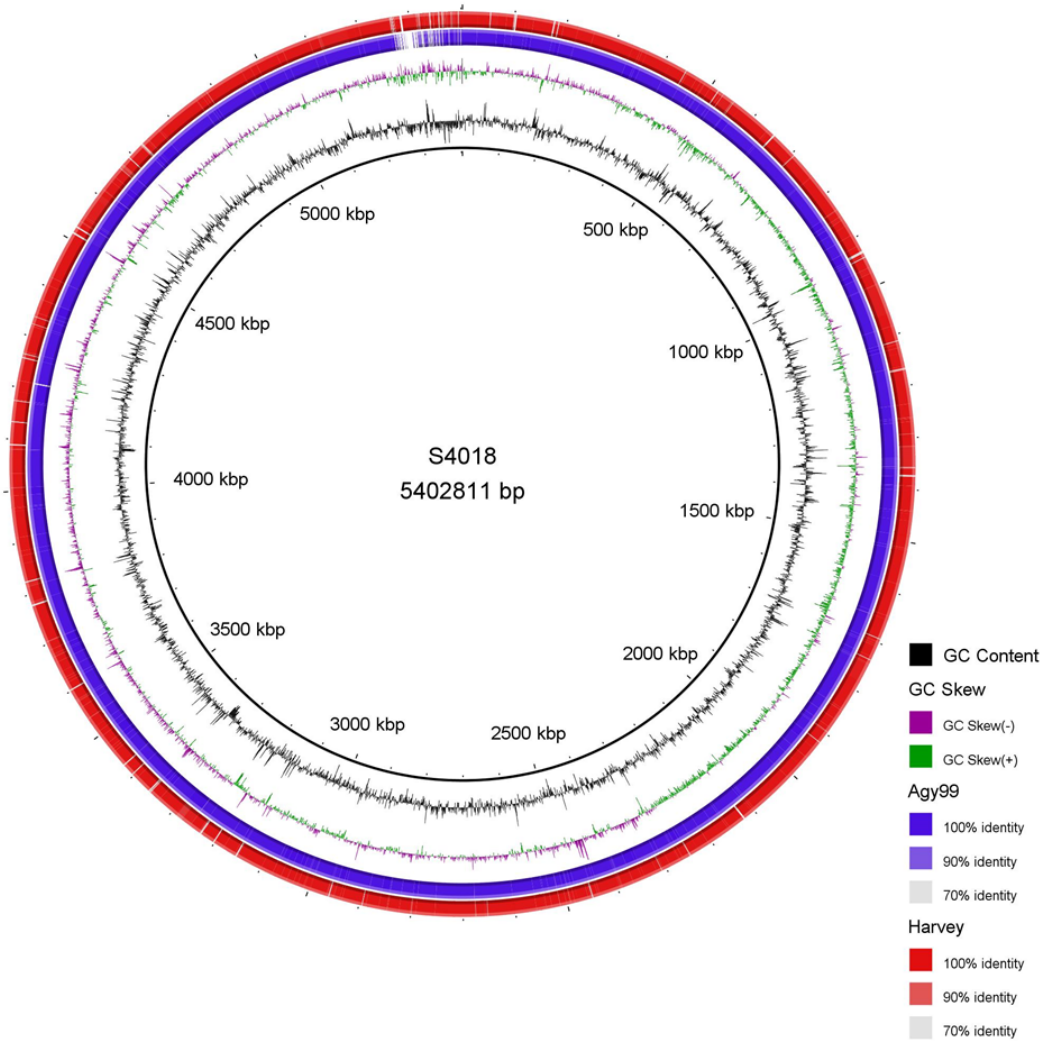
Region 1 (coordinates 970981..1016535 on contig LXFD01000039) is a 45.6-kb genomic island flanked by 34-bp direct repeats (DRs) which appears to have inserted into a gene encoding a hypothetical protein SAST42_01445 (ST20130942 genome numbering, acc. no. CP012976).The closest match to region 1 was found to be a staphylococcal phage tp-310-2 (93% coverage and 96% sequence identity) from *S. aureus* spa-310 (acc. no. EF462198). The second region (coordinates 835784..842343 on contig LXFD01000005) is a 6.5-kb element which is flanked by 8-bp DRs and seems to have inserted into an intergenic region between two genes encoding a hypothetical protein SAST42_03747 and a major facilitator protein SAST42_00201 (ST20130942 genome numbering, acc. no. CP012976). Eventually, this region was found to be the very genomic islet which harbors the mentioned *bla*Z gene and it was confirmed to be a Tn*552*-like element – it shows 98% sequence identity at 99% coverage to Tn*552* (X52734) described by Rowland and Dyke (Rowland & Dyke 1990). Region 3 (coordinates 939766..955084 on contig LXFD01000005) is a 15.3-kb genomic island flanked by 66-bp imperfect DRs which was found inserted into an intergenic region between the *rps*R and SAST42_03261 genes of strain ST20130942 (CP012976). This mobile genetic element is almost identical (100% coverage, 99% sequence identity) to the recently described *Staphylococcus aureus* pathogenicity island SaPIivm10 of *S. aureus* IVM10 (acc. no. AB716349) (Sato'o et al. 2013). Individual TBlastX comparisons of the three regions to their related elements are presented on panel C.

***Mycobacterium ulcerans* S4018**

Before we sequenced the draft genome of strain S4018, they were only two publicly available genome sequences of *M. ulcerans* – the first and only complete genome of this species from strain Agy99 (acc.no CP000325) which has been published back in 2007 (Stinear et al. 2007) and another non-described draft genome sequence of strain Harvey (JAOL00000000) released in 2014.

However, despite the lack of publicly available genome information about this species, WGS and comparative genomic studies have been performed since the sequencing of strain Agy99. Unfortunately, these sequences have never been deposited to any public databases. Even though we did not succeed to obtain assembly of good quality, as indicated by the large number of contigs and the low N_{50}, we were still able to compare the obtained draft sequence to the two other genomes available. The comparison is presented on panel A of page 69. Surprisingly, this analysis revealed a possible explanation about the low quality of our assembly. Initially, we supposed that the large number of contigs (about 505) was due to the high %GC content of *M. ulcerans* – regions with extremes in %GC content have been shown to be underrepresented in NGS libraries which leads to reduction in the assembly contiguity during the *de novo* assembly step and the presence of gaps in the final sequence (Chen et al. 2013). However, when we aligned our draft to the complete genome of strain Agy99 and analyzed the contig boundaries, we observed that almost every contig was flanked by a very short gap, containing one or very few genes. When we analyzed the genes present in the gaps we found that at least 213 of the gaps contain identical gene encoding the transposase IS*2404* and at least 71 - the transposase of IS*2606*. The accumulation of these two elements is characteristic for *M. ulcerans* (Stinear et al. 1999) and sequencing of strain Agy99 has revealed that it contains 213 copies of IS*2404* and 91 copies of IS*2606* (Stinear et al. 2007). The estimated numbers of IS*2404* and IS*2606* in strain Agy99 is in good agreement with the number of gaps in our draft which we associated with these elements – 213 *vs* 213 and 91 *vs* 71. This correlation could explain the low quality of our draft as it has been shown that except %GC extremes another common reason for gaps in genome assemblies with short reads (<1000 bp) is associated with the presence of mobile element genes such as insertion sequences, integrases and transposases (Barbosa et al. 2014).

A) *M. ulcerans* S4018 against strains Agy99 and Harvey.



B) Relatedness between the three genomes based on OrthoANI comparison.

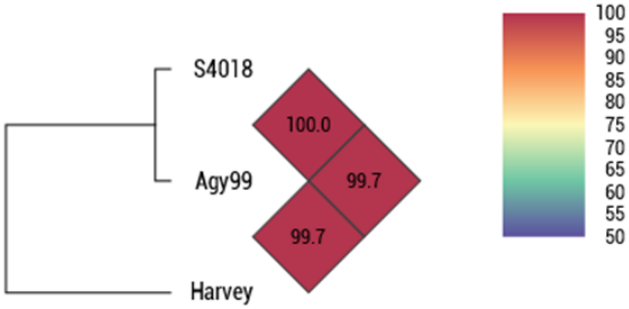


Figure 2.2. Comparative genomics of *M. ulcerans* S4018.

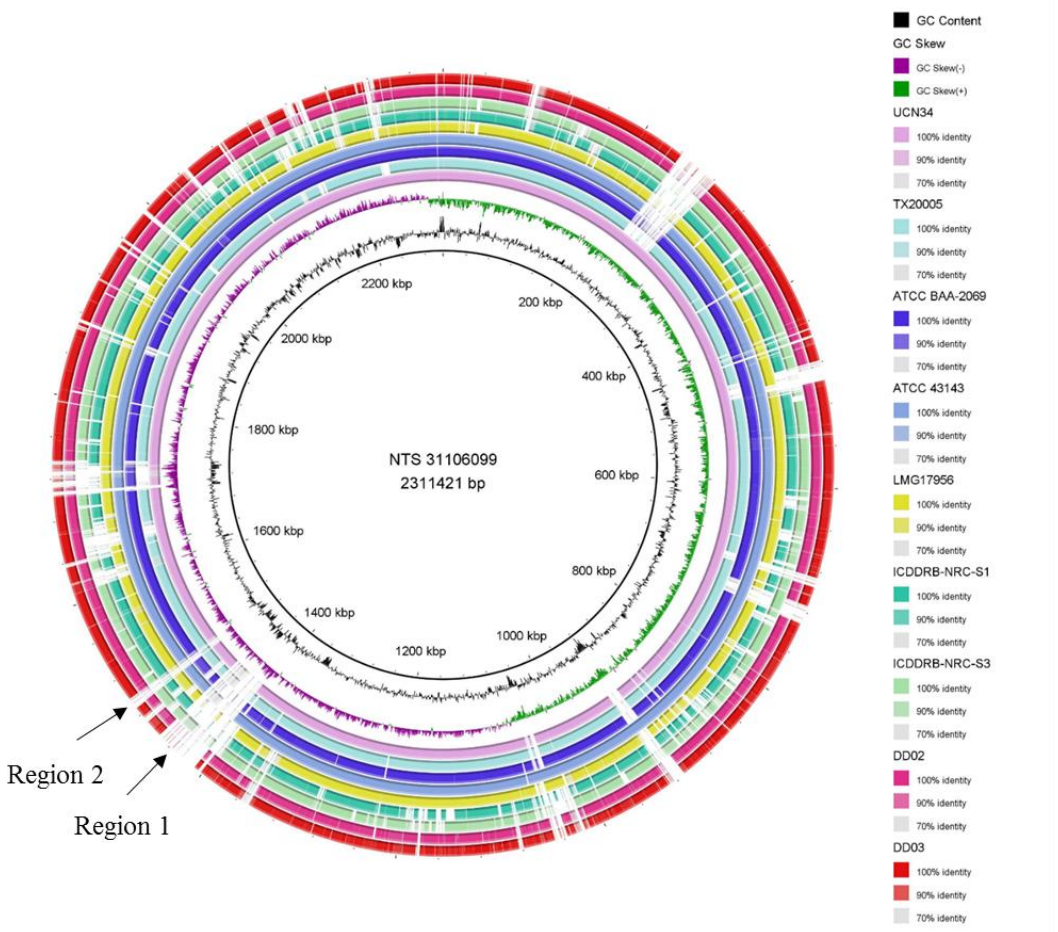
In one study between 18 and 77% of these genes have been found to be non-reconstructible during *de novo* assembly (Kingsford et al. 2010). Another study, in which isolates of *M. ulcerans* have been sequenced using Illumina platform and assembled with Velvet/VelvetOptimizer, has reported almost identical findings – the authors have reported even lower N_{50} value than ours (18399 *vs* 37602 bp) and have found that as much as 51% of the contig boundaries terminated at the mentioned mobile elements (Doig et al. 2012).

***Streptococcus gallolyticus* NTS31106099**

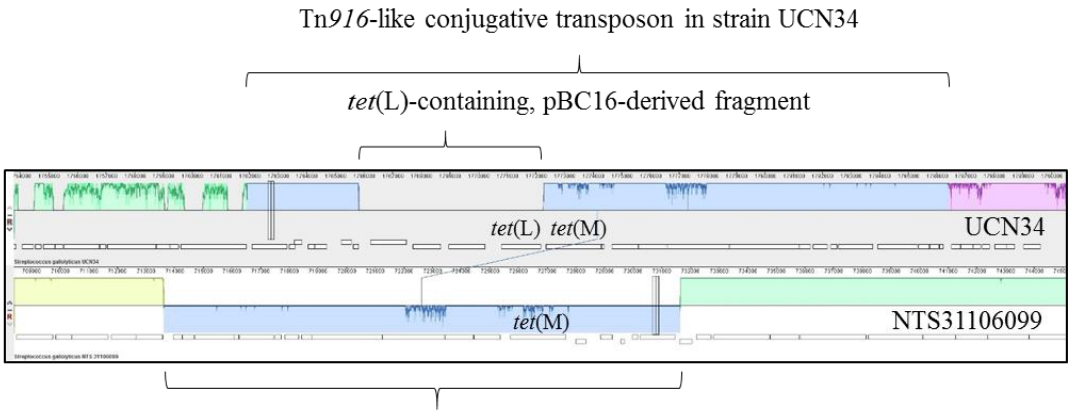
Multilocus sequence typing (MLST) of strain NTS31106099 according to the recently developed scheme by Dumke et al (Dumke et al. 2014) revealed a novel allelic combination (aroE-4, glgB-4, nifS-9, p20-5, tkt-5, trpD-4 and uvr-4) which was submitted to the *S. gallolyticus* database at PubMLST (http://pubmLst.org/sgallolyticus/) and assigned a sequence type number ST-91 (isolate id=281). Identification of acquired antibiotic resistance genes indicated the presence of tetracycline [*tet*(M), coordinates 22903..24822 on JYKU01000009], aminoglycoside [*aph*(3')-III and *ant*(6)-Ia, coordinates 198225..199022 and 199654..200562 on JYKU01000013, respectively) and macrolide [*erm*(B), coordinates 202938..203675 on JYKU01000013] determinants. Note that the *tet*(M) gene was found in different location than the other three genes, which are clustered together, suggesting the lack of genetic linkage between them. Panel A on page 72 illustrates BLASTn comparison between the draft genome of strain NTS31106099 and all other publicly available genomes of *S. gallolyticus* – TX20005 (acc.no. AEEM00000000) (Sillanpää et al. 2009), UCN34 (acc.no. NC_013798) (Rusniok et al. 2010), BAA-2069 (acc.no. NC_015215) (Hinse et al. 2011), ATCC 43143 (acc.no. NC_017576) (Lin et al. 2011), LMG 17956 (acc.no. CCBC00000000) (Romero-Hernández et al. 2015), ICDDRB-NRC-S1 and ICDDRB-NRC-S3 (acc.no. NZ_CP013688 and LPVQ00000000, respectively) (Sarker et al. 2015), DD02 and DD03 (acc.no.

LQOF00000000 and LQXV00000000, respectively) (Denapaite et al. 2015). Interestingly, two genomic regions appeared to be specific for strain NTS31106099 (labeled as regions 1 and 2 on panel A). Region 1 has a total length of 44 640 bp and was found to be inserted in a putative RNA methyltransferase gene corresponding to GALLO_1429, encoding a putative methyl transferase in the UCN34 reference genome. No DRs could be distinguished at the ends of the region. The G+C content of this genomic island is slightly higher than the rest of the genome (38 *vs* 37.5%). Interestingly, we were not able to identify this element in any other sequence present in the nt/nr or wgs databases of GenBank. Annotation revealed about 50 CDS predicted to be involved in conjugal transfer and clinically-relevant accessory functions such as antibiotic resistance [the *aph*(3')-III → *ant*(6)-Ia → *erm*(B) cluster mentioned above] and cell adhesion (putative collagen-binding protein UG96_07295). Further annotation analysis indicated that as many as 35 (70%) of the predicted genes have received functional annotation based on their highest similarity to homologs belonging to representatives of class Clostridia. Based on these lines of evidence, we concluded that the element is a putative conjugative transposon with clostridial origin, which was registered to the transposon number registry (http://www.ucl.ac.uk/eastman/research/departments/microbial-diseases/tn) under the number Tn*6263* in accordance with the revised nomenclature for transposable genetic elements (Roberts et al. 2008). Region 2 has a total length of 10 614 bp and was found to be inserted in an 18 bp target sequence TGATTATTTTTTAAGGTT, which includes the final 15 bp of an enolase gene homologous to the *eno* gene of strain UCN34 (position 1533609..1533707). The insertion of this element has resulted in duplication of the target sequence creating two 18-bp perfect DRs. The G+C content of the sequence is lower than the overall value for the whole assembly – 34.1% *vs* 37.5%.

A) *S. gallolyticus* NTS31106099 against all other publicly available genomes of *S. gallolyticus* .



B) Rearrangement of the Tn*916*-like element of isolate UCN34 in isolate NTS31106099.



Inverted Tn*916*-like conjugative transposon in strain NTS31106099, lacking the *tet*(L)-containing, pBC16-derived fragment.

Figure 2.3. Comparative genomics of *S. gallolyticus* NTS31106099.

This element was also not identified in the public nucleotide databases. About 18 genes were annotated (UG96_07440..UG96_07530) as 50% of them are of unknown function (labeled as hypothetical proteins). Among the other 9 genes, annotation revealed an integrase and another phage-associated protein. However, we were not able to further characterize the element, due to the lack of related but described elements in the public nucleotide databases. The *tet*(M) gene, identified with ResFinder was traced to a known Tn*916*-like conjugative transposon, originally described and inserted at identical position in the genome of strain UCN34 (Rusniok et al. 2010). However, while in strain UCN34 this element contains a fragment of plasmid pBC16 of *Bacillus cereus* (Palva et al. 1990) carrying *tet*(L) determinant, in strain NTS31106099 the transposon seems to have lost its plasmid-derived sequence and to have inverted its position (panel C on page 72).

2.3.3. Characterization of Tn*6263*.

BLASTn search through the nt/nr database of GenBank for previously characterized Tn*6263*-identical or related elements indicated that conjugative transposon CTn*7* of *Clostridium difficile* 630 described by Sebaihia *et al* (Sebaihia et al. 2006) is the closest such element. It belongs to the Tn*916*/*1545* family of conjugative transposons and has similar organization to our element with as much as 85% sequence identity in the homologous regions. Conjugative transposons from this family have a similar clustered organization, consisting of recombination, conjugation and regulation modules (Ciric et al. 2000). Detailed comparison of Tn*6263* to Tn*916* and CTn*7* revealed its modular structure and a vast array of accessory genes (panel A on page 74). Regarding its recombination module (labeled in pink on the panel A), Tn*6263* contains a single large serine recombinase which shares 86% sequence identity with the serine recombinase of CTn*7*. Both elements seem to target putative RNA methylase genes.

A) Tn*6263* agaist Tn*916* of *E. faecalis* DS16 and CTn*7* of *C. difficile* 630.

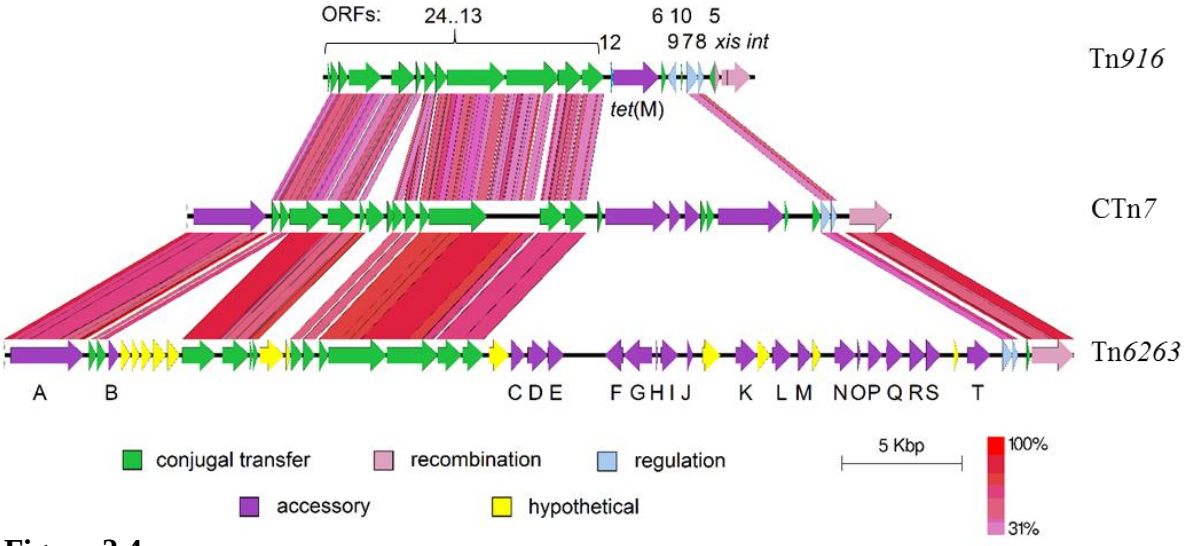


Figure 2.4.

B) Accessory genes in Tn*6263*.

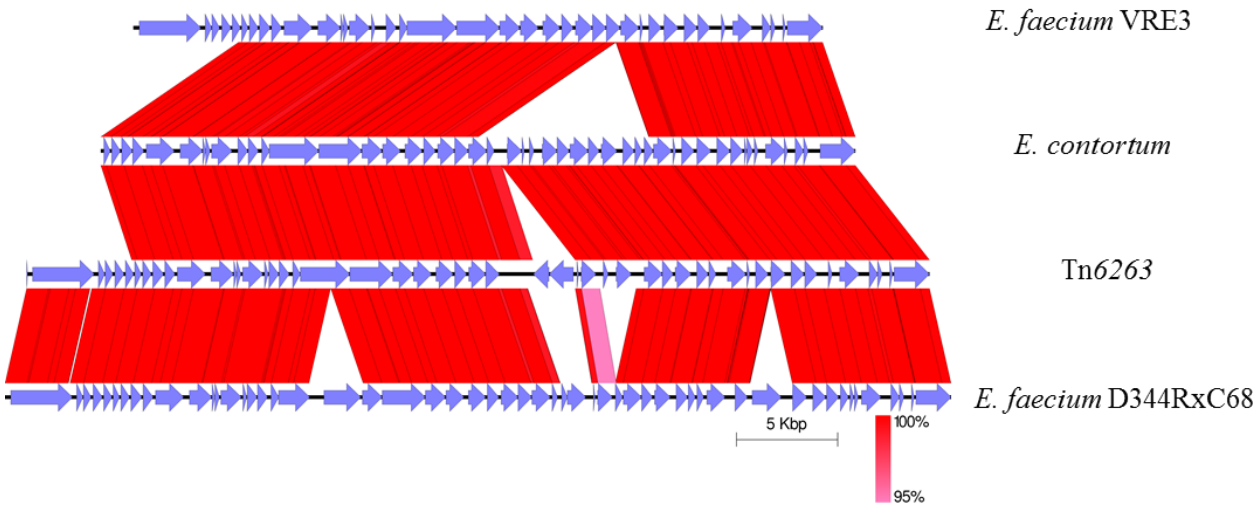
Label	CDS	Product	Best hit	Database
A	UG96_07295	collagen-binding protein	WP_009757279	RefSeq
B	UG96_07280	glyoxylase	WP_003437214	RefSeq
C	UG96_07180	tetR transcriptional regulator	WP_009902354	RefSeq
D	UG96_07175	ATP-binding protein	WP_009892238	RefSeq
E	UG96_07170	ABC transporter permease	WP_016295773	RefSeq
F	UG96_07150	transposase	WP_005585552	RefSeq
G	UG96_07145	integrase	WP_014062983	RefSeq
H	UG96_07140	ermB leader peptide	P21233	SwissProt
I	UG96_07135	ermB	P06573	SwissProt
J	UG96_07130	pirin	WP_002380754	RefSeq
K	UG96_07115	aminoglycoside adenylyltransferase/ant(6)-Ia	WP_002349868	RefSeq
L	UG96_07105	aminoglycoside phosphotransferase/aph(3')-III	WP_001637127	RefSeq
M	UG96_07100	transcriptional regulator	WP_001079940	RefSeq
N	UG96_07085	cation diffusion facilitator family protein	WP_009892244	RefSeq
O	UG96_07080	membrane protein	WP_016295776	RefSeq
P	UG96_07075	tetR family transcripttional regulator	WP_003434340	RefSeq
Q	UG96_07070	methyltransferase	WP_003434342	RefSeq
R	UG96_07065	phosphoesterase PA-phosphatase-like protein	WP_005929954	RefSeq
S	UG96_07060	XRE transcriptional regulator	WP_013976756	RefSeq
T	UG96_07040	AraC family transcriptional regulator	WP_018597285	RefSeq

Table 2.2.

However, while the excision of CTn*7* has been recently confirmed experimentally (Brouwer et al. 2011) and its target and coupling sequences have been determined, further work is needed to explain the mobility of Tn*6263* since no DRs were found to flank the ends of the element. Two genes of our element (UG96_07030 and UG96_07035, labeled with light blue on panel A) are homologous to ORFs 7 and 8 in Tn*916* which have been proposed to be parts of its regulation module (Su et al. 1992). In terms of its conjugation module, Tn*626*3 contains the complete conjugation module of Tn*916* (labeled with green on panel A) except the last open reading frame (ORF) 24 which is also missing in CTn*7*. Interestingly, at this position both elements, CTn*7* and Tn*6263*, contain a gene (UG96_07295 in Tn*6263*, labeled with purple and letter A) encoding a LPxTG-motif protein which has been annotated as putative collagen-binding protein. Such proteins are present at identical position on another transposon of *C. difficile* 630 – CTn*1* (Sebaihia et al. 2006) as well as on Tn*5386* of *Enterococcus faecium* D344R (Rice et al. 2005) and Tn*6079* from an infant metagenome sample which has been shown be carried on a *S. gallolyticus* genome (de Vries et al. 2011). Recently, the functional assignment of these genes has been questioned (Brouwer et al. 2011) since they were annotated as "collagen-binding" proteins due to the presence of a domain, which is homologous to domain B of the collagen-binding protein Cna of *S. aureus* (Patti et al. 1992). However, domain B of adhesin Cna has been shown not to be involved in the collagen binding. Rather, it has been proposed to serve as a molecular "stalk" which projects the collagen-binding domain A from the cell surface (Rich et al. 1998). Therefore, the exact function of these proteins remains uncertain. It has been proposed that their presence on Tn*916*-like elements might be due to erroneous or variable excision reactions (Roberts & Mullany 2011). Since the three mentioned modules in Tn*6263* are similar to the corresponding regions of Tn*916*, which serves as a scaffold for Tn*916*/Tn*1545* family of conjugative transposons, we concluded that our element belongs to this family (Roberts & Mullany 2009). Apart from the modules facilitating its

lateral mobility, Tn*6263* contains about 11 accessory genes with unknown function (annotated as hypothetical proteins, labeled with yellow on the panel A) and a set of 20 accessory genes with various functions (labeled with purple and capital letters from A to T on the panel A). A more detailed information about these functionally-assigned accessory genes is shown on panel B. Additional BLASTn searches for identical or related elements through the WGS database of GenBank revealed three non-described, highly-similar (99-100% sequence identity) elements in the genomes of *E. faecium* VRE3 (acc.no. JSET00000000) (Khan et al. 2015) and D344RxC68 (acc.no. LRHK01000000) and one strain of [*Eubacterium*] *contortum* 2789STDY5834876 (acc.no. CYZU01000000). Some insights on the putative origin of the antibiotic resistance determinants identified by ResFinder on Tn*6263* are given on page 77. A 9.5-kb DNA cassette (coordinates 197 207..206761 on contig JYKU01000013) containing the aminoglycoside/macrolide resistance cluster mentioned above was found to be similar in organization to the 4.2 kb macrolide-aminoglycoside-streptothricin (MAS) element originally discovered on Tn*1545*/Tn*6003* of *Streptococcus pneumonie* Ar4 (acc.no. AM410044) (Cochetti et al. 2007) with as much as 99% sequence identity in the corresponding homologous regions. Both elements, the cassette of Tn*6263* and the MAS element of Tn*1545*/Tn*6003*, harbor the aminoglycoside/streptothricin gene cluster described in *E. faecium* (acc.no. AF330699) (Werner et al. 2001) – gene UG96_07110 on Tn*6263*, annotated as a hypothetical protein was actually found to be the streptothricin resistance determinant *sat*4 thus giving the organization [*aph*(3')-III→*sat*4→*ant*(6)-Ia→*erm*(B)]. Additonally, the MAS element has been proposed to be a rearrangement of the 50-kb conjugative multiresistance plasmid pRE25 of *E. faecalis* RE25 (acc.no. X92945) (Schwarz et al. 2001; Cochetti et al. 2007). This seems to be also the case for the cassette on our transposon as the complete 9.5-kb sequence accounted for 16% of the whole plasmid with 99% sequence identity. The sequence relatedness between the three elements is evident from panel B on page 77.

A) Uncharacterized elements from NCBI GenBank, highly-related to Tn*6263*.



B) Tn*6263* against Tn*6003* of *S. pneumonie* Ar4 and pRE25 of *E. faecalis* RE25.

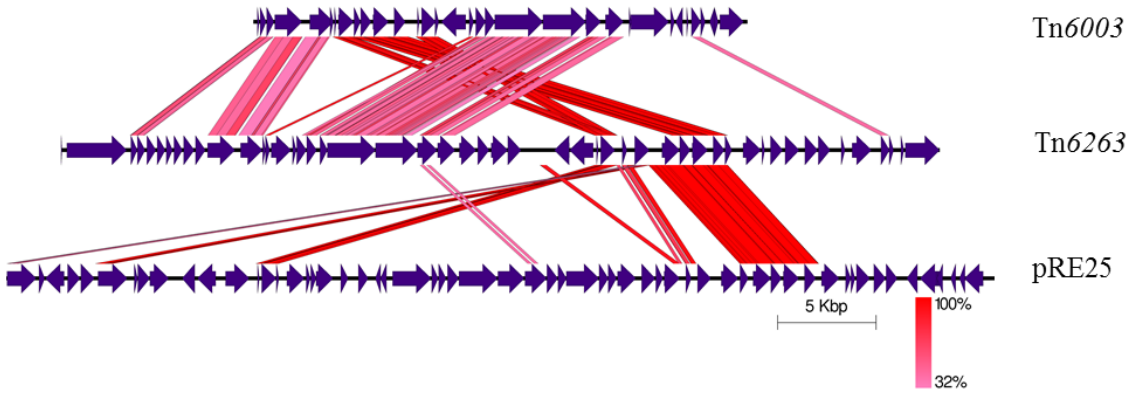


Figure 2.5. Tn6263 against other non-characterized and characterized elements.

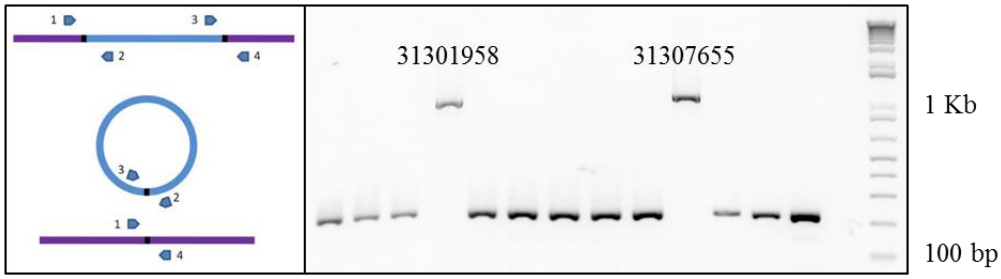
2.3.4. Prevalence of Tn*6263* among clinical isolates of *S. gallolyticus*.

We studied the prevalence of Tn*6263* or related elements in a collection of 60 clinical isolates of *S. gallolyticus*, isolated at Nantes University Hospital during the period between 2007 and 2015. The obtained results are shown on page 79. Antibiotic susceptibility testing indicated that 67% (n=40) and 59% (n=35) of all strains are highly resistant to tetracycline and erythromycin, respectively. PCR screening for *erm* resistance determinants A, B and C revealed that *erm*B is the dominant resistance gene which was found in 94% (n=33) of the highly-resistant isolates. No possession of *erm*A or *erm*C was observed. Surprisingly, PCR screening showed that about 69% (n=24) of all *erm*B-positive isolates tested were positive for the recombinase gene of Tn*6263* UG96_07020, suggesting the presence of identical or related element in their genomes. Additionally, amplicons of primer couples (1+2) and (3+4) (see Materials and methods and the scheme on page 79) were obtained for all isolates that tested positive for the recombinase gene, suggesting the insertion of Tn*6263* or related element at identical target site (the putative RNA methyltransferase gene) in all isolates examined. Identical (1+2) amplicons of 192 bp were obtained from all recombinase-positive isolates (n=24). However, identical (3+4) amplicons of 222 bp were observed in about 92% (n=22) of them - two of the isolates, NTS31301958 and NTS31307655, showed identical but longer amplicons of ~1.1 kb (panel B on page 79). Isolates, which give (1+2) and/or (3+4) amplicons only, or recombinase amplicons only were not identified. We then attempted to assess the genetic relatedness of the recombinase-positive isolates by means of molecular typing. About 71% (n=17) of them were typeable by Pulsed-Field Gel Electrophoretic (PFGE) analysis. The obtained results are presented on panel C on page 79. Among all 17 isolates, only one cluster comprised of two isolates with almost indistinguishable SmaI-profiles was observed. Interestingly, these were found to be the same isolates which gave the abnormal (3+4) amplicons – NTS31301958 and NTS31307655. All other isolates presented with diverse profiles.

A) Prevalence of Tn*6263* among clinical isolates of *S. gallolyticus*.

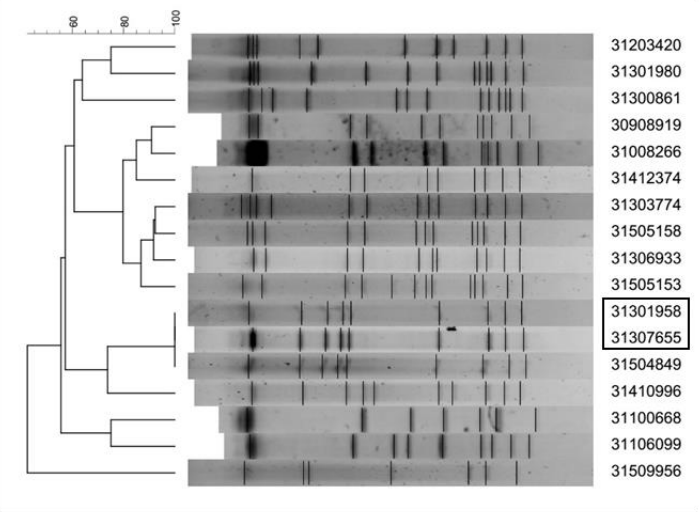
Total	Resistant to tetracycline	Highly-resistant to erythromycin	*erm*B gene present	Tn*6263*-recombinase gene present	Left-end 200-bp amplicon	Right-end 200-bp amplicon
60	40	35	33	24	24	22
	67%	59%	55%	40%	40%	37%

B) Variation in the length of the right-end amplicon (primers 3+4) in two of the tested isolates.



Taken from Brouwer et al, 2011

C) Relatednes between clinical isolates of *S. gallolyticus*, containing Tn*6263* recombinase.



D) Relatedness between Tn*6263* and Tn*6331*.

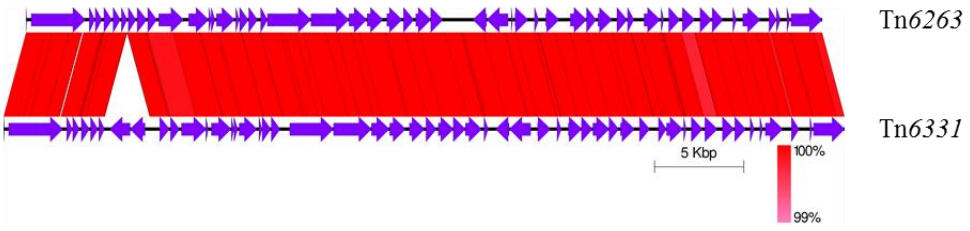


Figure 2.6. Prevalence of Tn*6263* in clinical isolates of *S. gallolyticus*.

Taken together, these lines of evidence suggest significant prevalence of Tn*6263* or closely-related elements among the tested highly erythromycin-resistant clinical isolates of *S. gallolyticus*. These elements seem to insert into an identical target site in the chromosomes of genetically-diverse isolates.

2.3.5. Draft genome sequencing of isolates NTS31301958 and NTS31307655.

The draft genomes of isolates NTS31301958 and NTS31307655 were sequenced in an attempt to better understand and explain the observed difference in the insertion site of their harbored elements, in comparison to Tn*6263*. Whole-genome comparison correlated with the results obtained from the PFGE typing – the two sequences share 100% OrthoANI. This suggests that they might represent isolates of the same or very close strain of *S. gallolyticus*. Both isolates were found to contain an identical element which is less than 100% identical to Tn*6263* in terms of its nucleotide and gene contents – this transposon shares 99% sequence identity with Tn*6263* and contain two more genes in its conjugation module. Therefore, the element was submitted to the Tn number registry and assigned transposon number Tn*6331* in accordance with the recommendations of Roberts *et al* (Roberts et al. 2008). Comparison between Tn*6263* and Tn*6331* is presented on panel D of page 79. The reason for the longer (1+3) amplicons obtained from these isolates at the PCR screening step was found to be an additional 883-bp sequence which seems to have inserted at position 1332 in the Tn*6263*-target methyltransferase gene GALO_1429 (UCN34 numbering, data not shown). The same insertion, but without Tn*6331* or any other element inserted, was found in one of the publicly available genomes of *S. gallolyticus* – ATCC BAA-2069. It introduces a single acetyltransferase gene before the target site (c_14530 in the complete genome of strain ATCC BAA-2069, accession no. FR824043).

2.4. Discussion.

The main goals of this part of the work was to sequence, *de novo* assemble and annotate the genomes of the three strains of interest – *S. aureus* FP_SA_ST25, *S. gallolyticus* NTS31106099 and *M. ulcerans* S4018 at a draft level by means of Illumina NGS technology in order to use them as reference sequences for selection output analysis during the development, validation and pilot application of our *in vitro* selection method, GeXplore. We were able to obtain good assemblies for two of the species – *S. aureus* and *S. gallolytcus* having between 17 and 39 contigs using Illumina platform, Velvet assembler and VelvetOptimizer. We concluded that our assemblies are of good quality since they contain much lower number of contigs than the average number of contigs for draft genomes in GenBank, which is 190 (Land et al. 2015). However, we obtained an assembly of lesser quality for *M. ulcerans* S4018 due to the intrinsic accumulation of two mobile elements in its genome – IS*2404* and IS*2606* as explained earlier. This could be explained by the fact that *de novo* assembly of genomes, containing highly repetitive regions is complicated by the short read length of the current NGS technologies, including Illumina (Miller et al. 2010). Therefore, a more recent "third-generation" platforms, performing at longer read lengths such as PacBio (Rhoads & Au 2015) could be a better choice for sequencing *M. ulcerans* genomes. All our assemblies were annotated through the Prokaryotic Genome Automated Annotation Pipeline (PGAP) of NCBI (Angiuoli et al. 2008). We chose this annotation method as it has been improved compared to its original release and it is thought to offer higher performance than alternative pipelines due to its novel pan-genome approach to protein annotation (Tatusova et al. 2016).

As mentioned earlier we envisioned to use the obtained drafts as reference sequences for output analysis at various level of development and application of our method, GeXplore. Having in mind that mobile genetic elements may harbor virulence determinants, implicated in host-pathogen interactions, we also attempted to partially characterize our assemblies by comparative genomics

with respect to their strain-specific mobiliomes. Thus, we identified numerous genomic island which seem to have been acquired by horizontal gene transfer (HGT) in all of our assemblies, except in *M. ulcerans* S4018. Importantly, we identified a *S. aureus* pathogenicity island in our strain, *S. aureus* FP_SA_ST25, which has been recently described (Sato'o et al. 2013), together with two other known elements. We also identified a novel Tn*916*-like conjugative transposon Tn*6263* in strain *S. gallolyticus* NTS31106099, which carries a 9.5 kb DNA cassette, related to the MAS element of Tn*6003*. We concluded that this element has a clostridial origin since most of its genes were annotated on the basis of highest similarity to clostridial genes. It seems quite related to Ctn7 of *C. difficile* 630 and highly similar to non-described elements which we identified in WGS of *E. faecium* and *E. contortum*. Even though the available genomic information about *S. gallolyticus* is scarce, few comparative genomics studies could be found in the literature (Rusniok et al. 2010; Hinse et al. 2011; Lin et al. 2011). Early insights into the genome of this species has suggested its active involvement in HGT with other *Firmicutes* into the gut, mainly *Enterococci*, *Bacilli* and *Clostridia*. Additionally, these studies have reported the presence of several putative Tn*916*-like elements, carrying tetracycline resistance genes. Another element, Tn*6079*, containing tetracycline and erythromycin resistance determinants has also been suggested to be borne by *S. gallolyticus*. However, our element shows only limited similarity to those elements.

Chapter 3

Preparation and validation of random genomic libraries for *in vitro* expression.

3.1. Introduction.

Ribosome display is a powerful tool for studying and/or designing the interaction of peptide or protein to another molecule of interest. Like other display technologies it has the ability to couple phenotype and genotype by linking a certain peptide/protein to its encoding nucleic acid. Its main advantage is the ability to operate in a setting of completely *in vitro* transcription/ translation process allowing for probing libraries of enormous diversities - up to 10^{12} molecules (Pluckthun, 2012).

Considering this advantage, the molecular nature of host-pathogen interactions and the highly-diverse content of prokaryotic genomes, it is no wonder that there have been attempts to use ribosome display in the field of infectious diseases (Weichhart et al. 2003; Lei et al. 2009). Similarly to the earlier shotgun phage display (Jacobsson et al. 2003; Mullen et al. 2006) and ANTIGENome technology (Meinke et al. 2005), ribosome display can be potentially used for genome-wide search of unknown virulence factors, unique diagnostic targets or immunodominant epitopes independently of culture conditions and/or regulation of gene expression.

To be studied by ribosome display, a pathogen's genome needs to be randomly fragmented and sub-cloned into an expression vector containing all functional regions required for successful *in vitro* transcription/translation and selection. Subsequent amplification by PCR converts each

created construct into a linear and functional ribosome display library. Finally, the library is screened against single or multiple ligands of interest. Therefore, it is clear that generation of high-quality and representative genomic library is a pre-requisite for successful application of the method.

Unfortunately, despite its mentioned advantages, the relative complexity of ribosome display has hampered its wide use in the field of infectious diseases. To our knowledge, there are only two groups that have reported its use, both for identification of vaccine candidates (Weichhart et al. 2003; Lei et al. 2008). Even though both studies have been performed using random whole-genome libraries, only the first one is associated with an attempt for comprehensive library characterization (Henics et al. 2003). However, as relevant as it might have been for its time, the library analysis have been performed by sampling about 500-1000 clones only – a number almost negligible compared to the original library size and the potential of the currently available NGS platforms. Additionally, both groups have performed *in vivo* library enrichment by transforming *E. coli* with the library constructs, which is labor-intensive step known to significantly reduce the library coverage (Mullen et al. 2006).

In this part of our work we present the development and optimization of a fully *in vitro* library preparation procedure which does not require transformation step. Random genomic libraries of *S. aureus* FP_SA_ST25, *S. gallolyticus* NTS 31106099 and *M. ulcerans* S4018 were prepared using an alternative G/C-assisted cloning strategy which we used to increase the level of fragment insertion and reduce the recircularization of empty vector molecules. The quality of the obtained libraries was characterized by NGS with regards to their coverage and content. We observed significantly reduced genome coverage in the analyzed libraries of *S. aureus* which was caused by strong PCR amplification bias associated with the genome G+C content – genomic regions with G+C content below 29% were underrepresented in the libraries. However, we found that the library

of *M. ulcerans* S4018 covers its complete reference sequence which has size of 5.4 Mb and G+C content of 64%.

3.2. Materials and methods.

All procedures accomplished with the use of commercial kits were performed according to the manufacturer recommendation unless otherwise stated. All primers used in the mentioned experiments could be found on page 21.

3.2.1. Description of ribosome display vector pFP-RDV1 and its modification – pSK-GeX1 and pSK-GeX2.

Vector pFP-RDV1 is a 2787-bp basic ribosome display vector created by our group on the basis of pUC18 plasmid (ATCC 37253) (unpublished vector). It contains a β-lactamase gene (ampicillin resistance) as a selection marker (1173..2587) and a 484-bp region (152..635) possessing all functional regions required for successful cloning, *in vitro* transcription, translation and selection of a potential DNA sequence. A BamHI/HindIII cloning site is positioned at location 271..302 (271..276 and 279..302, respectively) as four stop codons (277..279, 280..282, 290..292 and 294..296) are located between the two restriction sites. Upstream of the BamHI restriction site are located an MRGS(His)6 tag (241..270), a ribosome binding site (RBS, 227..232), a stem-loop (178..198) and a T7 promoter (162..177). Immediately after the HindIII restriction site are located two linker sequences (297..326 and 339..359) which are followed by a 276-bp "tether" region, TolA (360..635) ending with a hairpin loop region (613..635). Linearized pFP-RDV1, free of any closed-circular form was prepared by PCR using primers link-F and RDV1_R. Vector pSK-GeX1 is a variant of pFP-RDV1 containing two modifications. First, an additional 139-bp fragment was inserted between the EcoRI and AgeI restriction sites downstream of the cloning site (321..326 and

339..344, respectively, pFP-RDV1 numbering). This sequence encodes a chitin-binding domain which can be used in a contra-selection step, if needed. Second, the 20-bp fragment between the BamHI and HindIII restriction sites was modified with primers FseI_F and FseI_R in order to insert an FseI restriction site (GGCCGGCC) which can be used for forced cloning, if necessary. Vector pSK-GeX2 is essentially pFP-RDV1, except that the MRGS(His)6 tag has been removed by PCR using primers MAG and link-F.

3.2.2. Early attempts for random genomic library preparation.

Library assembly using the classical megaprimer method (Sarkar and Sommer, 1990), chain-reaction cloning (Pachuk et al. 2000) and blunt-end approach (Henics et al. 2003) were applied initially in an attempt to prepare *in vitro* expression random genomic libraries. As no success was achieved and no meaningful results were obtained with any of the mentioned approaches, this part of the work was dropped and will neither be presented nor discussed in more detail throughout the manuscript.

3.2.3. Pilot attempt for T/A- assisted library preparation.

A pilot T/A -assisted library preparation was attempted according to Lei (Lei, 2012).

Preparation of high-quality ribosome display vector.

Linearized pSK-GeX1, free of any closed-circular form was prepared in 50-µL PCRs using phosphorylated primers FseI_F and FseI_R under the following conditions: 1 x HF buffer (Thermo Fisher Scientifc), 0.2 mM dNTPs, 0.5 µM primers, 4% DMSO, 0.5 ng of template (BamHI/HindIII-digested pFP-RDV1), 0.5 units of Phusion HotStart II DNA polymerase (Thermo Fisher Scientifc). Amplification was performed in MJ Mini thermocycler (Bio-Rad) as follows: initial denaturation

for 30 sec at 98°C, 30 cycles of (10 sec at 98°C/ 30 sec at 65°C/ 20 sec at 72 °C) and final extension for 5 min at 72°C. Products were analysed on 1.5% agarose/ 1 x TAE gels. Primer removal was performed by adding 100 units of Exonuclease I (Thermo Fisher Scientific) directly to the PCR mixture followed by incubation for 2 h at 37°C and clean-up using Wizard SV Gel and PCR Clean-Up System (Promega).

Preparation of simplified model fragment library.

Simplified model of genomic library was prepared by digesting about 4 µg of pUC19 with 10 units of Hpy99I (NEB) in 50-µL reaction using 1 x NEBuffer 4 buffering system (NEB) and 1 x NEBSA bovine serum albumin (NEB). After incubation for 2 hours at 37 °C, the digestion was terminated by heating for 20 min at 65 °C, the products were purified using Wizard SV Gel and PCR Clean-Up System (Promega). Digestion completion was monitored on 1.5% agarose/ 1 x TAE gel.

dT- tailing of ribosome display vector.

Single-nucleotide tailing of pSK-GeX1 was performed with Taq polymerase according to Zhou and Gomez-Sanchez (Zhou & Gomez-Sanchez 2000). Briefly, about 2 µg of PCR-prepared, blunt-ended vector were tailed with single dT 3'-overhang using 5 units of GoTaq Flexi DNA polymerase (Promega) in a 100-µL reaction having the following compositions: 1 x GoTaq Flexi buffer, 2.5 mM $MgCl_2$ and 1 mM dTTP. After tailing for 2 hours at 72 °C the products were purified using Qiaquick PCR Purification Kit (Qiagen).

End-repair and dA- tailing of the model fragment library.

End-repair of the model fragment library was performed using the NEBNext End Repair Module (NEB). A 100-µL reaction containing 1.7 µg of purified model library was prepared and incubated

for 30 min at 20°C followed by reaction clean-up with Qiaquick PCR Purification Kit (Qiagen). After being purified, the fragments were dA-tailed in a 50-µL reaction using the NEBNext dA-tailing module (NEB). After tailing, the products were purified using Qiaquick PCR Purification Kit (Qiagen).

Fragment cloning.

Ligation of dA-tailed model library into dT-tailed pSK-GeX1 was performed in 10-µL reactions using T_4 DNA ligase (Thermo Fischer Scientific). Two different ligations were set – one containing only 60 fmoles (around 114 ng) of dT-tailed vector and another one, containing 60 fmoles of both vector (around 114 ng) and fragments (around 15 ng). Each reaction mixture had the following composition – 1 x T_4 DNA ligase buffer (Thermo Fisher Scientifc), 3.5% PEG8000 (Sigma) and 2.5 Weiss units of T_4 DNA ligase (Thermo Fisher Scientifc). All tubes were incubated overnight at 20°C followed by ligase inactivation for 10 min at 80°C and FseI-digestion. Ligation efficiency was assessed using 1.0 µL of ligation mixture in 25-µL PCR with the following composition: 1 x HF buffer (Thermo Fisher Scientifc), 0.2 mM dNTPs, 0.5 µM primers T7C and TolAext, 4% DMSO, 0.5 units of Phusion HotStart II DNA polymerase (Thermo Fisher Scientifc). Two-step amplification was performed in MJ Mini thermocycler (Bio-Rad) as follows: initial denaturation for 30 sec at 98°C, 30 cycles of (10 sec at 98°C/ 20 sec at 72°C) and final extension for 5 min at 72°C. Products were analysed on 1.5% agarose/ 1 x TAE gels.

3.2.4. Development of GC-based ligation strategy.

***Single-nucleotide tailing of ribosome display vector*.**

Single-nucleotide tailing of pSK-GeX1 using all four deoxynucleotide triphosphates (dNTPs) was performed in separate reactions according to point 3.2.3 with the following modifications. Six 100-

µL reactions (one for each dNTP + two negative controls) were prepared with the following compositions: 1 x GoTaq Flexi buffer, 2.5 mM $MgCl_2$ and 1 mM corresponding dNTP. After tailing for 2 hours at 72 °C, the products were purified using Qiaquick PCR Purification Kit (Qiagen) and monitored on 1.5% agarose/ 1 x TAE gel.

End-repair and single-nucleotide tailing of the model fragment library.

End-repair of the model fragment library with all four dNTPs was performed in separate reactions using the NEBNext End Repair Module (NEB). Four 100-µL reactions (one for each dNTP) were prepared, each containing about 1.7 µg of purified model library, followed by incubation for 30 min at 20°C and clean-up with Qiaquick PCR Purification Kit (Qiagen). Once being purified, the fragments were tailed in four 50-µL reactions with the NEBNext dA-tailing module (NEB) according to the manufacturer recommendation except that lab-made dA-free tailing buffer (100 mM Tris-HCl, 500 mM NaCl, 10 mM DTT, pH 7.5-8.0 at 10 x) was used and a single type of dNTP was added to each reaction. After tailing, the products were purified using Qiaquick PCR Purification Kit (Qiagen) and monitored on 1.5% agarose/ 1 x TAE gel.

Cloning optimization.

Ligation of tailed fragments into ribosome display vector, tailed with the complementary single-nucleotide overhang was performed in 10-µL reactions using T_4 DNA ligase (Thermo Fischer Scientific). About ten ligation reactions were set, divided into 4 sample series - A+T, T+A, G+C and C+G combinations of vector + fragments, respectively, and two controls. Each sample series contained two tubes: one tube containing only 60 fmoles (around 114 ng) of vector and another tube, containing 60 fmoles of each – vector (around 114 ng) and fragments (around 15 ng). The controls contained G-tailed vector with and without ligase. Each reaction mixture had the following

composition – 1 x T_4 DNA ligase buffer (Thermo Fisher Scientifc), 3.5% PEG8000 (Sigma) and 2.5 Weiss units of T4 DNA ligase (Thermo Fisher Scientifc). All tubes were incubated overnight at 20°C followed by ligase inactivation for 10 min at 80°C. Ligation efficiency was assessed using 1.0 µL of ligation mixture in 25-µL PCR with the following composition: 1 x HF buffer (Thermo Fisher Scientifc), 0.2 mM dNTPs, 0.5 µM primers T7C and TolAext, 4% DMSO, 0.5 units of Phusion HotStart II DNA polymerase (Thermo Fisher Scientifc). Two-step amplification was performed in MJ Mini thermocycler (Bio-Rad) as follows: initial denaturation for 30 sec at 98°C, 30 cycles of (10 sec at 98°C/ 20 sec at 72°C) and final extension for 5 min at 72°C. Products were analysed on 1.5% agarose/ 1 x TAE gels.

3.2.5. Fragmentation of genomic DNA.

For details about the DNA extraction procedure, you can refer to chapter 2. Digestion of genomic DNA with DNAse I was performed essentially according to Henics et al (Henics et al. 2003). Ultrasound fragmentation of genomic DNA was performed as follows – about 2 µg of genomic DNA were dissolved in 100 µL of 1 x TE buffer (10 mM Tris, 1 mM EDTA, pH7.5) and randomly fragmented into 100-300 or 200-1500 bp fragments using Bioruptor Standard (Diagenode) under the following settings: high intensity level, 30 sec on/off cycling with total duration of 15 min for 200-2000 bp and 60 min for 100-300 bp at 4 °C. Libraries with two different fragment size ranges were used in order to allow display of several ligand-binding domains or complete conformational epitopes. Fragmentation profiles were analysed either with TapeStation instrument (Agilent Technologies, USA) or on 1.5% agarose/ 1 x TAE gels.

3.2.6. Preparation of random genomic libraries using the developed G/C cloning strategy.

Short (100-300) and long (200-1500) fragments of *S. gallolyticus* NTS31106099 DNA were C-tailed and cloned into G-tailed pSK-GeX1 according to point 3.2.4. Additional cloning control was included in the ligation samples in order to assess the size range of the ligated fragments – about 1.7 µg of 1.0 kb+ DNA ladder (Invitrogen) were processed and cloned accordingly. The obtained amplicons were analysed on 1.5% agarose/ 1 x TAE gels.

3.2.7. Preparation of pilot *in vitro* expression random genomic libraries of *Staphylococcus aureus* FP_SA_ST25.

Linearized pSK-GeX2 was prepared from pFP-RDV1 according to point 3.2.3 using phosphorylated primers link_F and MAG_R. Short (100-300) and long (200-1500) fragments of *S. aureus* FP_SA_ST25 DNA were C-tailed and cloned into G-tailed pSK-GeX2 according to point 3.2.4. Once efficient ligation was validated for both libraries by PCR, the whole ligation mixtures were used as templates for 2 x 250 µL preparatory PCRs performed exclusively according to point 3.2.4. The amplicons were analysed on 1.5% agarose/ 1 x TAE gels and primer removal was performed by adding 100 units of exonuclease I (Thermo Fisher Scientifc) directly to the PCR mixture followed by incubation for 2 h at 37°C. The PCR products were then purified using Wizard SV Gel and PCR Clean-Up System (Promega). Finally, the purified amplicons were concentrated down to 200 ng/ µL in a speedvac instrument and analysed on 1.5% agarose/ 1 x TAE gel. The short-fragment library of *S. aureus* FP_SA_ST25 was designated ***Sa*si_1** standing for ***S***. ***a****ureus* ***s***hort ***i***nput ***1*** in order to be easily differentiated from any following modified library. This library was further characterized by Illumina NGS sequencing.

3.2.8. Next-Generation Sequencing of the pilot *Sa*si_1 library.

Removal of the two constant regions, originating from the ribosome display vector was performed by PCR. About 10 ng of *Sa*si_1 was used as a template for 100-µL PCR performed according to point 3.2.4 except that primers T7C and TolAext were replaced by primers int_F and int_R, which flank the vector cloning site. Amplicon analysis and subsequent processing (primer removal and clean-up) were according to point 3.2.4. After the constant regions were removed, about 1 µg of shortened *Sa*si_1 was used as an input for preparation of paired-end sequencing library using NEBNext Ultra DNA Library Prep Kit for Illumina (NEB) according to the kit manual except that size selection and PCR clean-ups were performed using NucleoMag NGS Clean-up and Size Select (Macherey-Nagel) instead of Agencourt AMPure XP (Beckman Coulter, Inc). After preparation, all libraries were submitted to the GenoBiRD Core Facility of Nantes, where they were controlled on TapeStation instrument (Agilent Technologies, USA), quantified by qPCR and sequenced on MiSeq instrument (Illumina, USA).

3.2.9. Analysis of the NGS output.

The ends of the raw datasets were trimmed with cutadapt 1.8.1 in order to remove the constant regions originating from primers int_F/int_R and the Illumina sequencing primers. Read quality control was performed with FASTQC. The reads were then mapped using bwa 0.7.10-r789 or through the GALAXY server (https://galaxyproject.org/) to the draft genome of *S. aureus* FP_SA_ST25, discussed in chapter 2. Mapping coverage was visually assessed in IGV 2.3.72 and then visualized using BRIG 0.95.

3.2.10. Probing *Sa*si_1 for underrepresented genomic regions by PCR.

Ten primer couples (ST_25_1...10) were designed for ten evenly distributed regions in the draft genome of *S. aureus* FP_SA_ST25. The primers were designed using OligoAnalyzer 3.1 (https://eu.idtdna.com/calc/analyzer) under the following parameter setting: 0.5 µM primers, no Na^{+}, 1.5 mM Mg^{++} and 0.2 mM dNTPs. Primer couples 1, 2, 3, 5, 7, 8, 9 and 10 were designed to be specific for 8 evenly distributed regions of the draft genome, which were found to be underrepresented in *Sa*si_1 library. Primer couples 4 and 6 were designed to be specific for two well-represented regions and served as positive PCR controls. Three different states of the *S. aureus* FP_SA_ST25 genome were tested – non-fragmented genomic DNA, genomic fragments and library *Sa*si_1. Probing was performed with 10 ng of template in 50-µL PCRs with the following composition: 1 x HF buffer (Thermo Fisher Scientifc), 0.2 mM dNTPs, 0.5 µM primers, 4% DMSO, 0.5 units of Phusion HotStart II DNA polymerase (Thermo Fisher Scientifc). Amplification was performed in MJ Mini thermocycler (Bio-Rad) as follows: initial denaturation for 30 sec at 98°C, 30 cycles of (10 sec at 98°C/ 30 sec at 57°C and 5 sec at 72°C) and final extension for 5 min at 72°C. Products were analysed on 1.5% agarose/ 1 x TAE gels.

3.2.11. Correction of the library preparation protocol.

After several cloning optimizations, forced and up-scaled ligation of an optimized short-fragment *S. aureus* FP_SA_ST25 library *Sa*si_2 (standing for ***S***. ***a****ureus* ***s***hort ***i***nput ***2)*** was performed as follows: about 150 ng of C-tailed fragments were ligated into 500 ng of G-tailed pSK-GeX2 in 50-µL reactions with the following composition: 1 x T_4 DNA ligase buffer (Thermo Fisher Scientific), 3.5% PEG8000 and 25 Weiss units of T_4 DNA ligase (Thermo Fisher Scientific).

Therefore, the same conditions were used also for the preparation of *S. gallolyticus* NTS31106099 short-fragment genomic library (*Sg*si) and *M. ulcerans* S4018 short-fragment genomic library

(*Mu*si). The ligation efficiency of the three libraries (*Sa*si_2, *Sg*si and *Mu*si) was assessed according to point 3.2.3, but then the ligation mixtures were used for 250-µL preparatory PCRs with the following modified composition: 1 x Q5 buffer (NEB), 0.2 mM dNTPs, 0.5 µM oligos RDV1_F and TolAkurz, 5 units of Q5 DNA Polymerase (NEB). Two-step amplification was performed in MJ Mini thermocycler (Bio-Rad) as follows: initial denaturation for 30 sec at 98°C, 10 cycles of (10 sec at 98°C/ 75 sec at 65°C) and final extension for 5 min at 65°C. The amplicons of the three libraries were identically processed (primer removal and clean-up) according to point 3.2.7. After concentration to 200 ng/µL *Sa*si_2, *Sg*si and *Mu*si were processed for NGS as follows: 200 ng of library concentrate were used as a template for a 100-µL internal PCRs as explained in point 2.8, in order to remove the vector-derived constant regions. Illumina library preparation, sequencing and analysis were performed as explained for *Sa*si_1 in points 2.8-2.9. *Sa*si_2 was mapped on the genome of *S. aureus* FP_SA_ST25, *Sg*si on *S. gallolyticus* NTS31106099 and *Mu*si on the genome of *M. ulcerans* S4018.

3.3. Results.

3.3.1. Early attempts for library preparation.

As mentioned in the previous section, we did not succeed to use the megaprimer method (1st method tested) (Sarkar and Sommer, 1990) and chain reaction cloning (Pachuk et al. 2000) for library preparation. Unfortunately, we were not able to obtain any product, compatible with ribosome display with any of the methods. The former method appeared to be unable of assembling the desired product when fragments with various length are used as a template in a setting of multi-template PCR and the latter failed to generate any closed circular constructs (data not shown). Therefore, these approaches were abandoned.

Cloning using vector with phosphorylated blunt ends (2nd method tested), according to Henics et al, (Henics et al. 2003) resulted in almost its complete recircularization and no inserted fragments were observed after PCR amplification of the ligated constructs. A slight increase in the level of fragment insertion was observed when the vector was dephosporylated (data not shown). Unfortunately, such ligation strategy was incompatible with our goal for "*in vitro* library preparation" as the repair of the single-strand nicks resulting from end dephosphorylation would eventually require passage through a living system. This approach was also abandoned.

3.3.2. Initial attempt for T/A-cloning.

Even if fragment ligation in dephosphorylated vector failed to generate sufficient amount of inserted fragments, it suggested that preventing the vector re-circularization by some means will increase the level of fragment insertion. One possible approach is the single-nucleotide tailing of the ligation parties, also known as T/A-cloning. Such an approach has already been used for generation of random genomic libraries for ribosome display. However, the authors have still used transformation in *E. coli* to enrich the ligated constructs (Lei, 2012). Therefore, we tried to increase the fragment insertion by using this technique (3rd method tested). The results from this experiment are shown on page 96. Panels A and B show the preparation of the two DNA parties which are about to be ligated. A linearized, blunt-ended and phosphorylated pSK-GeX1 was synthetized *in vitro* by PCR using primers FseI_F and FseI_R and the product was validated free of any closed-circular form by PCR with T7C and TolAkurz (data not shown).

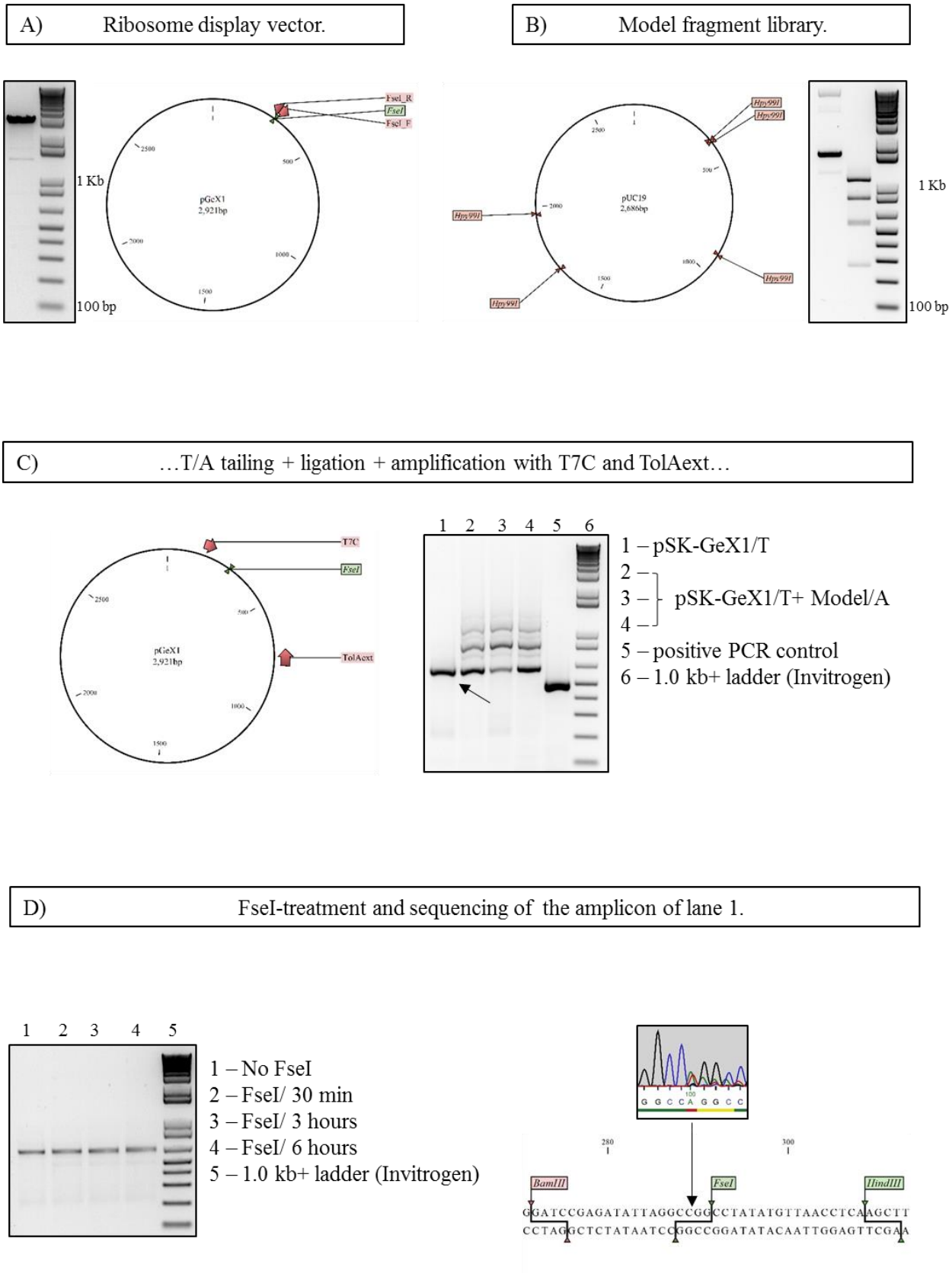


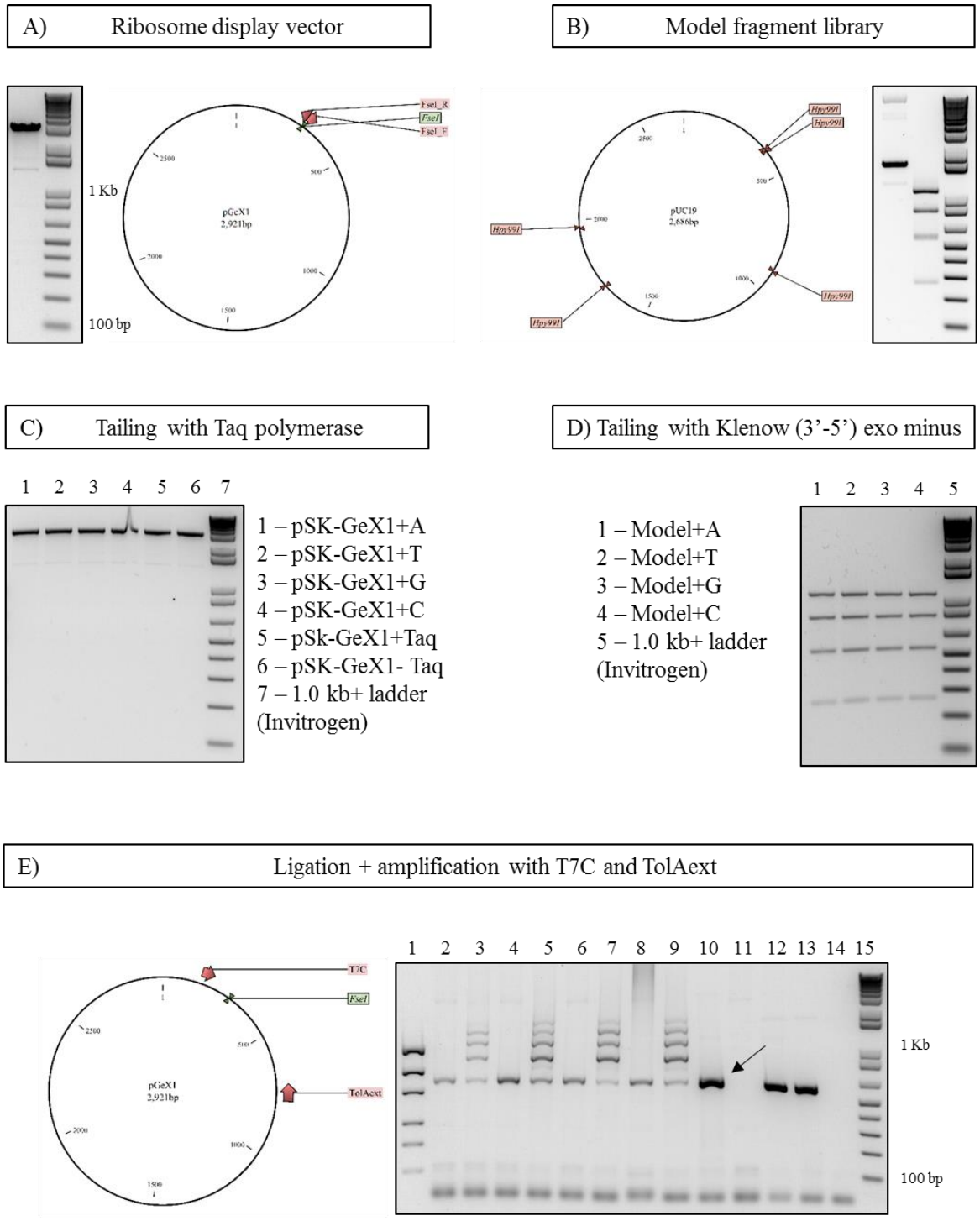
Figure 3.1. Initial attempt for T/A cloning.

A simplified fragment library, consisting of just few fragments was derived from pUC19. The vector was treated with Hpy99I restriction enzyme, resulting in a set of 5 fragments with different size – 17, 259, 522, 794 and 1094 bp. Since the length of the shortest fragment is 17 bp only, its presence was negligible and it was not observed during the subsequent applications, thus giving the model library 4 fragments with range of 259...1094 bp. Panel C shows the results of ligating dA-tailed model library into dT-tailed pSK-GeX1 and amplifying the resulting constructs with primers T7C and TolAext in order to obtain linear ribosome display templates. Lane 1 shows the level of vector recircularization when only dT-tailed pSK-GeX1 is present in the ligation mixture. Lanes 2, 3 and 4 show the level of vector recircularization and fragment insertion when dA-tailed model library is ligated into dT-tailed pSK-GeX1. The band, corresponding to re-circularized vector is indicated with black arrow. As evident from the figure, the level of recircularization was quite high even when dT-overhangs are attached to the 3'-ends of the vector. However, the level of fragment insertion was improved – two bands can be clearly distinguished higher of the empty amplicon, indicating that fragment insertion has occurred. In comparison, we did not observe any inserted fragments with the blunt-end approach.

As we expected some level of recircularization, we intentionally introduced FseI restriction site in pSK-GeX1 with primers FseI_F and FseI_R which is reconstituted when the vector recircularize. Panel D shows the results of digesting the amplicon of lane 1 with FseI restrictase. Surprisingly, the treatment resulted in almost no digestion. Interestingly, subsequent sequencing of the amplicon indicated that a single T/A bp is inserted in the middle of the restriction site, thus making the amplicon resistant to digestion. So far, we have not been able either to identify the reason for or to explain this insertion.

3.3.3. Development of G/C-cloning procedure.

Even though T/A cloning improved the fragment insertion level, we still observed quite high amount of re-circularized vector in the library preparations. In order to improve the efficiency of the single nucleotide-assisted cloning in both reducing the vector recircularization and increasing the level of fragment insertion we had to develop an alternative cloning strategy. Therefore, we decided to explore the other three possible nucleotide combinations as single-nucleotide overhangs. The obtained results are indicated on page 99 (4^{th} method tested). Identical to page 96, panels A and B show the preparation of two DNA parties which are about to be ligated. Panels C and D show the pSK-GeX1 and the model library, respectively, after being split into separate batches and tailed with different single-nucleotide 3'-overhangs (A, T, G or C). Panel E shows the results after the tailed fragments have been ligated into complementary-tailed vector and the corresponding vector/fragment combinations have been amplified with primers T7C and TolAext. As can be seen on the electrophoregram, the band intensity of lane 10, which corresponds to blunt-end pSK-GeX1 + ligase is comparable to the intensity of the positive PCR controls (lanes 12 and 14) which contains an identical amount of closed-circular template. This result suggests that significant part of the blunt-ended vector has re-circularized (indicated again with black arrow). Lanes 2, 4, 6 and 8 indicate the reduction of the vector re-circularization caused by the addition of different single-nucleotide 3'-overhang (A, T, G and C, respectively) at the vector ends (compare to blunt-ends, lane 10). As evident from the gel, lanes 2, 4, 6, 8 and 10 can be ranked according to their band intensity in the following decreasing order: 10>>4>8>6>2 corresponding to blunt ends>>T>C>G>A. Lanes 3, 5, 7 and 9 represent the difference in the ligation efficiency of each complementary vector/fragment combination. Concerning the intensity of their empty amplicons, the lanes follow the same order as lanes 2, 4, 6, 8 and 10 – 10>>5>9>7>3 corresponding again to blunt ends>>T>C>G>A.



1 – Mass Ruler DNA ladder
2 – pSK-GeX1/A
3 – pSK-GeX1/A + Model/T
4 – pSK-GeX1/T
5 – pSK-GeX1/T + Model/A
6 – pSK-GeX1/G
7 – pSK-GeX1/G + Model/C
8 – pSK-GeX1/C
9 – pSK-GeX1/C + Model/G
10 – pSK-GeX1/ blunt-end + ligase
11 – pSK-GeX1/ blunt-end – ligase
12 and 13 – positive PCR controls
14 – negative PCR control
15 – 1.0 kb+ ladder (Invitrogen)

Figure 3.2. Development of G/C cloning procedure.

However, while the overall intensities of the bands corresponding to the ligated model fragments are almost indistinguishable in lanes 5, 7 and 9, the fragments in lane 3 are significantly paler than the rest. This suggests that the insertion of T-tailed fragments in A-tailed pSK-GeX1 is lower than the other three overhang combinations even if A-tailed vector showed the lowest level of recircularization. Therefore, we selected the next most optimal combination for preparation of our libraries, which is ligation of C-tailed fragments into G-tailed vector.

3.3.4. Fragmentation of genomic DNA.

As mentioned earlier, our approach could potentially identify peptide/protein/ligand interactions at a genome-wide scale. However, the information about the limitation of ribosome display for the size of the expressed fragment is limited especially when multi-length genomic libraries are to be screened. Additionally, the peptide/protein/ligand interaction itself might be dependent on the size of the expressed fragment (a conformational epitope and its cognate antibody, for example). Having these considerations in mind, we envisioned the use of libraries with various fragment lengths between 100 and 1500 bp. Therefore, we needed a reproducible and easy-to-control method for fragmentation of genomic DNA.

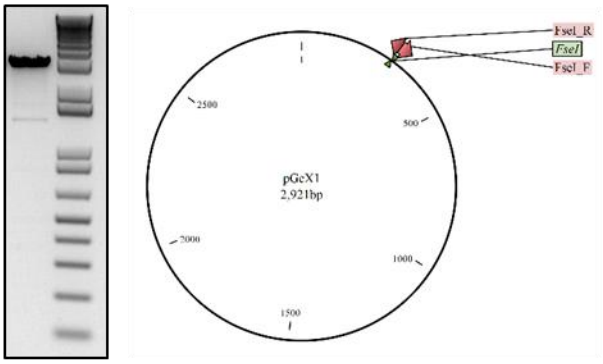
The three most common methods for shearing genomic DNA are random enzymatic digestion, nebulization and sonication, which have been recently compared in the light of the current NGS technologies and shown to be of equal overall performance (Knierim et al. 2011). Previous display technologies have made successful use of both, sonication and/or DNAse I digestion for preparation of genomic fragments (Jacobsson & Frykberg 1995; Henics et al. 2003). Therefore, we decided to try both methods. DNAse I digestion was performed essentially according to Henics et al (Henics et al. 2003) but as we did not succeed to obtain satisfying and reproducible results with this approach, we dropped it (data not shown).

Ultrasound sonication was performed with Bioruptor Standrad instrument (Diagenode) which allowed us to fragment our DNA in reproducible and easily controllable manner – we were able to prepare genomic fragments in two different ranges – 100-300 bp and 200-1500 bp, which suited perfectly our needs (page 102, panel B).

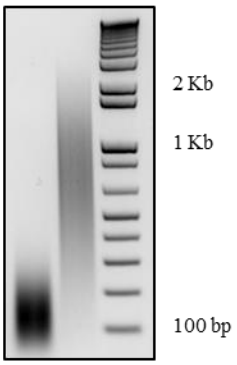
3.3.5. Preparation of pilot *in vitro* expression random genomic libraries using the optimized G/C-assisted cloning strategy.

Having the genomic DNA sheared into fragments with desired length and the G/C-cloning strategy optimized, we attempted to create a pilot genomic library of *S. gallolyticus* NTS31106099 using the developed protocol described here. The obtained results are presented on page 102. Again, panels A and B show the two parties which are to be ligated – pSK-GeX1 and fragmented genomic DNA. Panel B shows the short (100-300 bp) and long (200-2000 bp) genomic fragments and the 1.0 kb+ DNA ladder which was also cloned in a vector as a fragment size control. Panel C shows the pilot library preparation experiment. Lanes 2 and 3 show the impact of G-tailing to the recircularization of pSK-GeX1. Clearly, lane 2 (blunt-ended and phosphorylated pSK-GeX1) has much higher intensity than lane 3 (G-tailed pSK-GeX1). Lane 4 shows the results of blunt-ended model fragment library being ligated into blunt-ended, phosphorylated pSK-GeX1 – clearly, no inserted fragment can be distinguished. Lane 5 represents the ligation of C-tailed 1.0 kb+ DNA ladder (Invitrogen) which has been ligated to a G-tailed pSK-GeX1. As evident from the figure, 6 different bands can be distinguished corresponding to inserted fragments of 100, 200, 300, 400, 500 and 650 bp. Additionally, just a pale band corresponding to recircularized vector is visible. Lanes 6 and 7 depict the two pilot libraries of *S. gallolyticus* NTS31106099 – one built using 100-300 bp fragments and one using 200-2000 bp fragments, respectively. Again, the band corresponding to empty amplicon in lanes 6 and 7 is almost negligible.

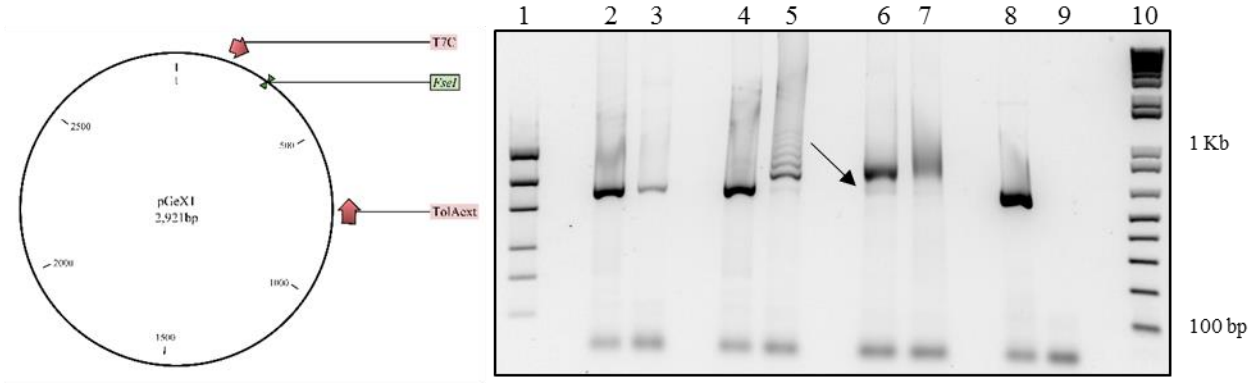
A) Ribosome display vector



B) Fragmented DNA and 1.0 kb+ ladder



C) …G/C tailing + ligation + amplification with T7C and TolAext…



1 – Mass Ruler DNA ladder
2 – pSK-GeX1
3 – pSK-GeX1/G
4 – pSK-GeX1 + Model
5 – pSK-GeX1/G + 1.0 kb+ ladder/C
6 – pSK-GeX1/G + short fragments/C
7 – pSK-GeX1/G + long fragments /C
8 – positive PCR controls
9 – negative PCR control
10 – 1.0 kb+ DNA ladder (Invitrogen)

Figure 3.3. Preparation of pilot in vitro expression libraries using G/C cloning

This result clearly indicates that the developed G/C cloning strategy allows for efficient and completely *in vitro* cloning of genomic fragments into a circular vector and their subsequent conversion into a linear ribosome display constructs by PCR with negligible presence of empty constructs.

3.3.6. Characterization of short-fragment library *Sa*si_1 of *S. aureus* FP_SA_ST25 by NGS.

After additional optimization, related to the ribosome display vector which will be discussed later in the manuscript (chapter 4), we proceeded with detailed characterization of our *in vitro* expression genomic libraries by Illumina NGS sequencing. Initially, we prepared a short- (100-300 bp) and long-fragment (200-1500 bp) libraries of *S. aureus* FP_SA_ST25 using pSK-GeX2 according to our optimized protocol. The short-fragment library was designated *Sa*si_1, in order to be easily differentiated from any following optimized or modified library. This library was sequenced by Illumina technology and its quality was compared to the Illumina NEBNext library which we used for sequencing the genome of strain FP_SA_ST25. The latter served as a library "gold standard". About 3.1 million reads of Illumina NEBNext and 1.4 million reads of *Sa*si_1 were mapped on the draft genome of strain FP_SA_ST25 using Burrows-Wheeler transform (Li & Durbin 2009). The obtained results are shown on page 104. Panel A illustrates a comparison between the coverage of the two libraries. The first graph (the second innermost ring in black) indicates the fluctuations in the %GC content over the reference genome sequence with a window size of 50 bp. The next two graphs outwards (colored in red) visualize the overall coverage of the two libraries and represent a direct visualization of the paired-end reads aligned to their corresponding positions in the reference genome. The outermost graphs belongs to our *in vitro* expression random genomic library *Sa*si_1 and the second outermost – to the Illumina NEBNext library

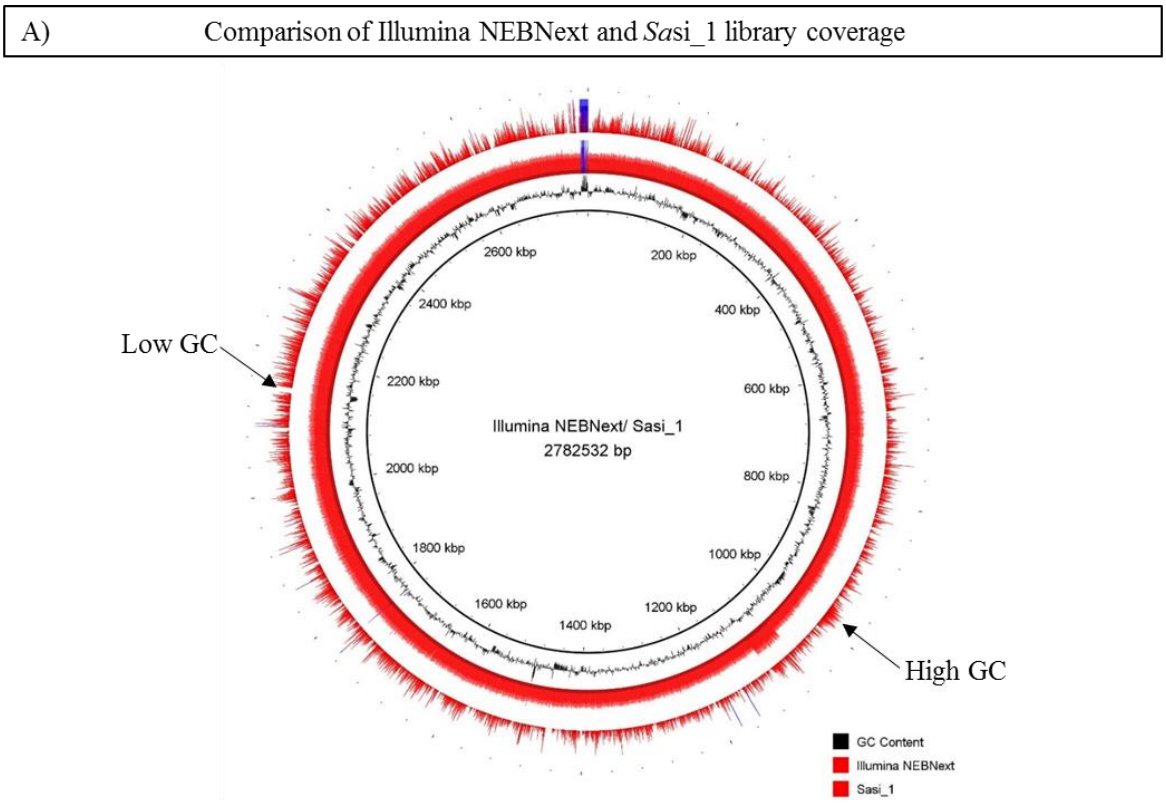


Figure 3.4. NGS analysis of pilot library *Sa*si_1

As evident from the picture, the coverage of the latter (Illumina NEBNExt) is uniform and the reads cover completely the reference genome in an evenly manner. However, the former one (*Sa*si_1) has uneven read distribution indicating a biased library content – some regions are over-represented while others are missing. Note the correlation between the under-represented regions (gaps) and their low GC content. Panel B represents a more detailed visualization of the read mapping in two representative genomic regions with IGV software – one with high (left) and one with low (right) %GC content, pointed with black arrows on panel A. The same trend could be observed – the %GC-low region is not present in our library. This interesting observation pointed towards a plausible explanation of the observed bias – as mentioned in chapter 2, the overall %GC content of a NGS sample could be informative about its quality. Taking into account the common origin of the two samples (the genome of *S. aureus* FP_SA_ST25) one could expect that they would also have an equal %GC content. However, when we compared the two values we found that our library has significantly higher %GC content than the Illumina NEBNext library and the genome of *S. aureus* FP_SA_ST25, 41% *vs* 33%, respectively. Panel C shows the per-read GC count of the two libraries as calculated by the quality control software FASTQC. While Illumina NEBNext library has an overall %GC of 33% and a 13%-57% per-read GC range, elevated values were observed for *Sa*si_1, 41% and 29%-61%, respectively.

Taken together, these result indicated that our pilot library had incomplete genome coverage and that the coverage loss seemed to be due to a %GC bias.

3.3.7. An attempt for improvement of the library coverage.

Eventually, NGS of *Sa*si1 revealed a %GC-associated bias, resulting in increased abundance of fragments originating from GC-rich genomic regions and complete lack of fragments having G+C content below 29%.

In an attempt to improve the library coverage, we first sought a way to identify the source of the GC-bias and then we modified several parameters of the established library preparation protocol. Finally, as we observed under-representation of GC-low regions and the genome of *S. aureus* FP_SA_ST25 has the lowest %GC content among our three species of interest (33% *vs* 38% and 64% for *S. gallolyticus* NTS31106099 and *M. ulcerans* S4018, respectively), we hypothesized that libraries of the latter two genomes should have better coverage than *Sa*si_1. The results obtained from the performed experiments are presented on page 107. Panel A illustrates the identification of the bias source. We designed 10 primer couples specific for 10 evenly distributed regions over the genome of *S. aureus* FP_SA_ST25 as 8 of them were chosen from the under-represented genomic regions and 2 of them from the well-presented in order to serve as internal positive controls. We probed intact DNA, fragmented DNA and *Sa*si_1 with the designed primers. The source of bias was traced to the amplification step during which the ligated fragments are converted into a linear and functional ribosome display templates. Note the reduction in the intensity of the bands corresponding to GC%-low regions (the lowest gel on panel A). This result correlates with previous studies, focused on identifying biases in NGS libraries, which have shown that PCR is the main source of bias in such libraries and that it could depend on the %GC content of the template, the type of polymerase used and the number of amplification cycles performed (Oyola et al. 2012; Dabney & Meyer 2012; Van Dijk, Jaszczyszyn, et al. 2014). Therefore, we compared the bias in identical libraries amplified with Phusion II Hot Start DNA polymerase and Q5 DNA polymerase as well as Taq, Pfu and Kappa DNA polymerases using the mentioned PCR-based assay. Unfortunately, no polymerase showed reduction in the library bias (data not shown). The next parameters modified were the overall vector/fragment amount used for ligation as well as their ratio. Accordingly, the vector input amount was increased from 114 to 500 ng per ligation and the fragment amount – from 15 to 150 ng.

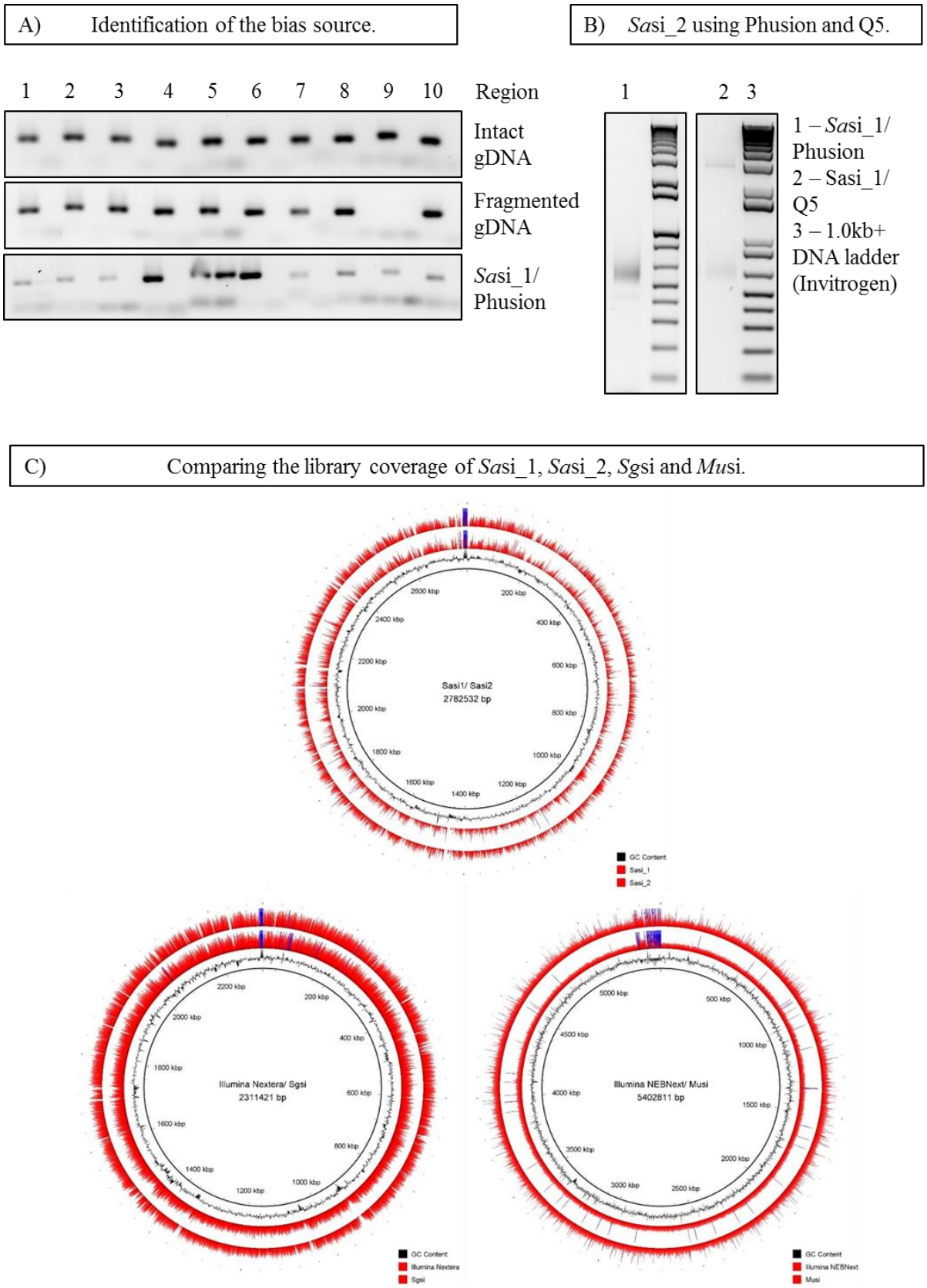


Figure 3.5. Modification of library preparation protocol.

Finally, the amplification conditions were also modified – the ligated constructs were amplified according to the NEBNext Ultra Library Prep Kit for Illumina. Namely, the number of cycles were reduced from 30 to 10, Q5 DNA Polymerase (NEB) was used instead of Phusion II HotStart (Thermo Fisher Scientific) and the extension step was modified from 20 sec at 72 °C to 75 seconds at 65 °C. We chose to replace Phusion with Q5 polymerase for further use as a pre-caution since the former enzyme has been particularly shown to generate biased NGS libraries, depleted of AT-rich loci (Aird et al. 2010).

The *S. aureus* library amplified following the modified protocol was termed *Sa*si_2. Panel B shows a comparison between *Sa*si_1 and *Sa*si_2 after the PCR amplification step, no shift in the library fragment range was observed. Expectedly, the reduced number of cycling resulted in significantly lower amount of PCR product, which however, was enough as a template for performing the subsequent application – *in vitro* transcription.

Short-fragment libraries of *S. gallolyticus* NTS31106099 and *M. ulcerans* S4018 (*Sg*si and *Mu*si, respectively) were prepared according to the modified protocol used for the preparation of *Sa*si_2. Their amplification profiles after the preparative PCR step were identical to the one shown for Sa*si*_2 on panel B. All three libraries were sequenced by NGS. Panel C illustrates comparison of their genome coverage. The upper ring compares the coverage of *Sa*si_1 and the modified *Sa*si_2. As evident form the picture, no coverage improvement was achieved by altering the mentioned parameters. The ring on lower left shows comparison between our library *Sg*si_1 and the Nextera NGS library used for sequencing the draft genome of *S. gallolyticus* NTS31106099 explained in chapter 2. Interestingly, even though the Nextera library covered completely the reference sequence, similar fluctuations in the read mapping depth were observed in both libraries. Additionally, the library coverage of *Sg*si_1 was found to be much higher than the one of *Sa*si_1 and *Sa*si_2.

The ring on the lower right shows the coverage of our library *Mu*si_1, compared to the coverage of the NEBNext NGS library used for the sequencing of the draft genomes of *M. ulcerans* S4018. Strikingly, a complete coverage and smaller fluctuations in the read depth of *Mu*si_1 were observed. Note that the size of *M. ulcerans* genome is almost twice the size of *S. aureus* – 5.4 *vs* 2.78 Mb and still our *in vitro* expression library had complete coverage over the reference sequence. Finally, we analysed the %GC range of each library by comparing it to the %GC range of its corresponding genome. Accordingly, the three libraries had %GC range of (29-63, 29-63 and 45-79) while their corresponding genomes had range of (13-63, 19-63 and 45-85). Evidently, the %GC range of the two "compromised" libraries, *Sa*si_1 and *Sg*si, has been shifted as both of them have lower limit of 30%. Since the lower limit in their corresponding genomes is different (13% and 19%) but both of them has the same lower limit of 30%, we think that this could be the %GC threshold of our method – genomic regions having %GC lower than 30% are not represented in our libraries, according to our developed protocol.

3.4. Discussion.

Intuitively, routine directional cloning procedures based on restriction enzyme-derived cohesive ends are not applicable for construction of random genomic libraries, exactly because of their "random" nature. Such libraries have been mainly prepared by the use of the classical blunt-end cloning approach. However, this technique has been proposed to be about 10-100 times less efficient than sticky-end cloning (Sambrook and Russel, 2000). It proceeds according to a double-hit intermolecular mechanism – the incoming genomic fragment must first be ligated to one of the vector ends and then to the other one. Additionally, in order to occur, this reaction must outcompete the more probable, single-hit intramolecular vector re-circularization. In order to favor fragment insertion, the vector ends are usually being dephosphorylated, which reduces its re-circularization.

The obtained constructs are then transformed in *E. coli* as the single-strand nicks resulting from the end dephosphorylation are repaired in the host cell. However, transformation has limited efficiency which results in reduced coverage of the genomic library. Additionally the procedure has low reproducibility and is labor intensive for the production of large libraries (with more than 10^9-10^{10} independent members). As we wanted our method, GeXplore, operates entirely *in vitro*, we attempted to develop a completely *in vitro* procedure for preparation of random genomic libraries, which is devoid of the problematic transformation step.

Several existing approaches were initially used without any success, despite a lot of time spent to optimize them. Then, an alternative G/C-assisted cloning strategy was developed, inspired by the classical T/A cloning. The latter approach did not perform well in reduction the level of vector recircularization - we observed a single T/A bp inserted at the vector junction. We could propose only one plausible explanation for this event to occur, even if it is a bit far-fetched – insertion of T/A base pair would require a vector with single dT-overhang and a complementary dATP or ATP present in the ligation mixture. Most probably, molecules with single dT-overhang are present in the dT-tailed vector batch because it is unlikely that the tailing reaction has proceeded with 100% efficiency. Even if it had, it is possible that some dT-overhang have been attacked by contaminating exonucleases. Residual dATP is also possibly present in the tailed vector and fragment batches since the clean-up procedure with commercial kit is not able to remove 100% of the unconsumed dNTPs after the PCR and the end-repair steps. Additionally, the T_4 ligase buffer contains significant amounts of ATP. Therefore, if a residual dATP or ATP molecule anneals to the single dT-overhang of the vector and is subsequently ligated by the T_4 polymerase, this could create a T/A blunt end, which could ligate to the opposite blunt end of the vector and could explain the sequenced amplicon. However, no support about this hypothesis was found in the literature.

In an attempt to avoid the observed drawback of T/A cloning we compared all other possible combinations of single-nucleotide overhangs. Regarding the vector tailing, we observed the following order of reduction in the level of empty vector – blunt-ends>>T>C>G>A. This result correlates with the reported preference of Taq polymerase for non-template single-nucleotide addition (Clark 1988). However, we observed a strange trend in the fragment insertion level of the four nucleotide combinations – while A, C and G overhangs showed almost comparable levels, T-tailed fragments showed significantly lower insertion level than the rest. Taking into account that A-tailed vector showed the lowest level of recircularization and that the 3'-5' exonuclease-deficient Klenow fragment is known to have equal preference for all four dNTPs when provided individually (Clark et al. 1987), one could expect the insertion of T-tailed fragments into A-tailed vector would be the most efficient combination among the four. Yet, we observed exactly the contrary. So far we have not found an explanation for this result.

After we developed a successful cloning strategy, we used it to prepare random genomic libraries of *S. aureus* FP_SA_ST25, *S. gallolyticus* NTS31106099 and *M. ulcerans* S4018 which are compatible with ribosome display. Since there is no information about the genome coverage of such libraries, we decided to characterize them using Illumina NGS technology with regards to their genome coverage. After sequencing, we mapped the obtained reads to the reference genomes, as explained in chapter 2. The library coverage of the three libraries, *Sa*si_1, *Sg*si and *Mu*si, varied between quite reduced, intermediate and complete, respectively. Strikingly, this trend was not associated with the size of the corresponding genomes, which is 2.8, 2.3 and 5.4 Mb, respectively. Contra intuitively, the library made of the largest genome had complete coverage.

We then analyzed the %GC range of the three libraries and compared it with the range of their corresponding genomes and we found out that genomic regions with %GC content below 30% are not present in our libraries. This trend did not change even when we modified the library

preparation protocol by (i) increasing the input amount, (ii) adding an annealing step, (iii) reducing the number of amplification cycles and (iv) changing the DNA polymerase used. We have not been able to solve this issue so far. However, a recent study might provide with a possible explanation (Pan et al. 2014). The authors have hypothesized that DNA polymerases have a biased preference for different oligonucleotides during initiation of DNA polymerization and have used NGS to test their hypothesis. In an elegant experiment, a synthetic library of random sequences has been amplified using several commercial polymerases and the diversity of the obtained amplicons has been examined using high throughput sequencing. Eventually, the authors have observed a positive amplification bias for all tested polymerases towards an increasing GC content of the six base pairs involved in the primer-template interaction. The report also tries to explain the bias by relating it to the transition of DNA from B-to-A from during interaction with the DNA polymerase since this transition has been shown to be important for many protein-DNA interactions. Furthermore, earlier studies have reported that GC-rich DNA undergoes B-to-A transition easier than GC-poor DNA (Tolstorukov et al. 2001). Therefore, these authors have proposed that GC-rich motifs are preferentially amplified by the DNA polymerase since its binding to such motifs is facilitated by their easier B-to-A form transition. Even though our libraries are somehow different than the synthetic library used in this study, it seems possible that similar scenario drives the GC bias observed in our experiments. In our libraries, all amplicons contain a 92-bp constant region, originating from the ribosome display vector. Therefore, the polymerization priming for all amplicons could be expected to proceed in an even manner. However, after the end of this region, the upcoming template will be of different GC content due to the random nature of the genomic library. Thus, it might be speculated that after the end of the constant region the amplicons originating from genomic fragments with different GC content will also have different tendency to

undergo B-to-A transition, which could result in the observed GC bias. Of course, this hypothesis requires further experimental confirmation.

As of why this GC bias is not so profound in NGS libraries, we think that the explanation is the starting amount of template. Our NGS libraries were prepared using at least 1.0 µg of DNA and because of this, we performed only 4 PCR cycles at the amplification step in the protocol. However, we have not been able to optimize our GeXplore library preparation protocol for such big amount of genomic fragments. Additionally, similar GC bias was observed for the Nextera library used to sequence the genome of *S. gallolyticus* NTS31106099 (the lower left ring diagram on page 107). Even if the library covered the complete reference sequence, the most underrepresented genomic regions coincided with the gaps in our GeXplore library. Nextera technology differs from the NEBNext protocol in that fragmentation and tagging steps are performed simultaneously (thus, "tag-mentation"). However, we are not certain if this difference could be the reason for the observed bias in Nextera library. It is possible that the Nextera sample just provides better view on the input library due to the much bigger number of reads sequenced (10.2 million *vs* 2 million for Nextera and NEBNext respectively). However, it is also possible that such a GC bias occurs as a general phenomenon when random genomic libraries are extensively amplified. Unfortunately, this observation was not described in pioneering works on ribosome display used as a genome-wide technique because NGS was not available at that time. We could propose the following explanation for the more pronounced GC bias in our GeXplore libraries – we have observed that a sufficient number of amplification cycles (30 for the initial and 10 for the modified protocols) are needed to obtain sufficient amount of linear library product for *in vitro* transcription. This extensive amplification might be the reason for the biased library content. In comparison, NEBNext libraries are amplified for four cycles only. Of course, this hypothesis needs to be experimentally confirmed.

Chapter 4

Optimization and validation of GeXplore.

4.1. Introduction.

The relatively limited use of display technologies like phage or cell surface display to explore host-pathogen interactions on a genome-wide scale has resulted in a limited knowledge about many aspects of their performance such as library coverage, per-round enrichment, selection-induced biases (non-specific interactions), etc.

Early studies with shotgun phage display have shown that up to 40% of the clones, enriched after two selection rounds encoded a ligand-binding peptide when using gene III fusion libraries. However, subsequent increase in the number of selection rounds has not resulted higher enrichment of desired clones (Jacobsson & Frykberg 1995). Modification of the technology to use gene VIII instead of gene III fusion libraries has resulted in better enrichment, with 75-100% of the selected clones which have been found to encode for the desired ligand-binding peptide (Jacobsson & Frykberg 1996). However, these studies report on the selection efficiency of shotgun phage display in the setting of single-ligand selections, since the genomic libraries were being screened against one molecule of interest only.

ANTIGENome technology is an alternative approach which has been developed to identify potential vaccine candidates from pathogen-specific genomic libraries using patient-derived antibodies (multi-ligand conditions). Indeed, this technology has proven its high-throughput potential, as it has been used for the identification of the immune-relevant proteomes (antigenomes) of several important human pathogens of the genera *Staphylococcus* and *Streptococcus*, consisting

of more than 100 proteins per species (Meinke et al. 2005). Additionally, a single attempt for ribosome display-based ANTIGENome technology, performed using *in vivo* prepared genomic library of *S. aureus* has shown that the cell-based display approach has certain expression limitations – many peptides that were selected by the ribosome-display based version were omitted using the cell-based approach. Nevertheless, for both approaches many of the analyzed clones have been found to encode peptides, belonging to surface-exposed proteins, which underscores their efficiency. However, detailed quantitative characterization of the selection outputs has not been possible until the advent of the NGS platforms.

In the present work, we aimed at the development of a completely *in vitro* genome-wide approach which could be used for both single- and multi-ligand selections. In such an approach, input libraries and selection outputs are obtained exclusively by PCR amplification. However, as already mentioned, PCR has been shown to be the main source of bias in genomic samples such as NGS libraries. Another critical issue is the recombination between non-complementary fragments inherent in multi-template PCR which may significantly impair the input/output quality by accumulation of degenerated heteroduplex products. If present, this problem could be avoided by the use of the so-called emulsion PCR (ePCR) (Diehl et al. 2006). This technique takes advantage of the compartmentalization of the highly heterogeneous (by both, size and sequence) PCR template into an aqueous microdroplets, emulsified in oil phase (water-in-oil emulsions).

Other points which need to be considered when one performs a display system are the number of washes at the panning step and the choice of selection matrix used. Our experience in ribosome display has taught us that both could have a crucial impact on the selection success since the former helps to differentiate unspecific background from desired interaction and the latter could easily be responsible for an elevated level of this very background. The number of washes needed could be in many cases binder/ligand specific and requires empirical determination. The level of selection

matrix-driven background could be handled by alternating matrices of different material (plates, beads) throughout the selection.

High selection efficiency and specificities were our priorities, since they are crucial, especially in the latter case. In this part of the work we explain the step-by-step optimization of our method, with a special accent on input and output quality, optimal washing step and optimal choice of selection matrix. Finally, we validated the selection efficiency and specificity of GeXplore using a model system, previously used for the validation of shotgun phage display.

4.2. Materials and methods.

All manipulations with commercial kits were performed according to the manufacturer recommendations unless when stated otherwise. All primers used in the present work are enlisted in the table on page 21.

4.2.1. Emulsion PCR.

Emulsion stability was controlled according to Williams *et al* (Williams et al. 2006). Briefly, 2 x 100-µL PCRs (reaction A and B) were prepared. Reaction one contained 1 x HF (Thermo Fisher Scientific), 0.2 mM dNTPS, 0.5 µM primers T7C and TolAkurz, 0.05% BSA (Thermo Fisher Scientific), 1 unit of Phusion Hot Start II DNA polymerase (Thermo Fisher Scientifc) and about 10^9 template molecules (around 4 ng of pSK-GeX1, containing 157-bp insert, encoding anti-Fc Affitin C3). Reaction A served as a positive control. Reaction B was prepared with identical composition except no primers were added and the pSK-GeX1+C3 template was replaced by pSK-GeX1 (around 3.3 ng for 10^9 molecules). After the reactions were prepared, each of them was split in 2 x 50-µL aliquots, thus giving A1, A2, B1 and B2. Reactions A2 and B2 served as standard (open) PCR controls. Aliquots A1 and B1 were emulsified using the protocol of Shutze *et al*

(Schütze et al. 2011). Briefly, each 50-µL aliquot was added to about 300 µL of pre-cooled surfactant, made of 73% Tegosoft (Evonik, Germany), 20% mineral oil (Sigma, Germany) and 7% ABIL WE (Evonik). Emulsification was performed at top speed of standard benchtop vortex for 5 min at 4°C (in a cold room) as the PCR mixture was added in 10-µL volumes at 1 min intervals. Once the A1 and B1 aliquots were emulsified, the resulting emulsions were mixed, split in thin-walled PCR tubes (100-µL emulsion/tube) and amplified in a MJ Mini thermos cycler (Bio-Rad) together with the "open" controls under the following conditions - initial denaturation for 30 sec at 95°C, 40 cycles of (10 sec at 95°C/ 20 sec at 72 °C) and final extension for 5 min at 72°C. After amplification, the tubes of each reaction were pooled in a 2-mL tube and broken by adding 1 mL of isobutanol, followed by centrifugation for 1 min at 16 000 x g/RT. After phase separation, the aqueous phases were purified using Qiaquick PCR Purification Kit and analysed on 1.5% agarose/1 x TAE gel. The same protocol was used for the amplification of *S. aureus* FP_SA_ST25 genomic libraries.

4.2.2. Assessment of library degeneration during standard PCR.

Degeneration of the library profiles during PCR was assessed according to Williams *et al* (Williams et al. 2006). Briefly, a short – and a long – fragment libraries of *S. aureus* FP_SA_ST25 were prepared according to the protocol developed in chapter 3. About 2 µL of the ligations were then used as a template for 2 x 50-µL PCRs with the following composition - 1 x HF (Thermo Fisher Scientific), 0.2 mM dNTPS, 0.5 µM primers T7C and TolAext, 4% DMSO (Thermo Fisher Scientifc), 0.5 units of Phusion Hot Start II DNA polymerase (Thermo Fisher Scientifc). The two reactions were then split into 5 x 10-µL aliquots and amplified as follows - initial denaturation for 30 sec at 98°C, 27 x (10 sec at 98°C/ 20 sec at 72°C) and final extension for 5 min at 72°C. One tube was removed from the cycler at every 3rd cycle starting from 15th until 27th cycles and kept on ice

until the final extension step was reached. All reactions were then completed together. The products were analysed on 1.5% agarose/1 x TAE gel. The same experiment was repeated with primers T7s and TolAs including annealing step for 30 sec at 60˚C.

4.2.3. *In vitro* transcription/translation.

About 1.2 µg of library concentrate (about 6 µL, for concentrate preparation please refer to chapter 3) was used as a template for 20 µL *in vitro* transcription using TranscriptAid T7 High Yield Transcription Kit (Thermo Fisher Scientifc). After 4 h of incubation at 37˚C, template DNA was removed by adding 25 units of DNAse I (Thermo Fisher Scientifc) directly to the transcription reaction, followed by incubation for 15 min at 37˚C and inactivation for 10 min at 70˚C. The RNA was then purified using standard LiCl precipitation. Residual template DNA was removed using NucleoSpin RNA kit (Macherey-Nagel) and controlled by PCR using primers T7C and TolAext. *In vitro* translations were performed as follows: about 5 µg of purified RNA of genomic library were translated in 30-µL reactions with the following composition: 1.61 µL of H_2O, 0.5 µL of 200 mM Methionine, 12.59 µL of Premix [5.3 mM adenosine triphosphate (ATP), 13.4 mM guanosine triphosphate (GTP), 2.7 mM cyclic adenosine monophosphate (cAMP), 80 mM acetyl phosphate, 1.33 mg/mL *E. coli* tRNA, 133 mM Tris– acetate pH 7.4 at 4°C, 6.7% (w/v) PEG 8000, 0.93 mM of all natural amino acids except methionine, 53 mg/mL folinic acid, 534 mM L-glutamic acid monopotassium, 15 mM magnesium acetate, 9.7 mM a ssrA], 12.50 µL of S30 extract, 0.3 µL of PDI (22 µM, Sigma) and 2.5 µL of RNA (2.0 µg/µL). After incubation for 7 min at 37˚C the translation reactions were stabilized with 121 µL of stop solution [1 x Washing buffer (WB) [50 mM Tris-HCl, 150 mM NaCl, 50 mM Mg $(CH3COO)_2$, pH 7.5], 0.5% BSA, 2.5 mg/mL heparin, 0.1% Tween 20) and centrifuged for 5 min at 20 000 x g/ 4˚C to pellet any precipitates present.

4.2.4. Experimental design for validation of GeXplore.

The experimental design for validating the efficiency of our method was performed according to the classical work for validation of shotgun phage display technology, conducted by Jacobsson and Frykberg (Jacobsson & Frykberg 1995). Briefly, a random genomic library of *S. aureus* is prepared and selected against human IgGs using the display technology to be tested in order to identify the two IgG-binding proteins of this pathogen (or their Fc-binding domains). We essentially used this system during all following optimizations. Library *Sa*si_1 of *S. aureus* FP_SA_ST25 was used as a test library. Purified human Fc fragment (Bethyl Laboratories, Inc.) was used as a ligand instead of whole IgGs to avoid possible cross-reactivity with the Fab, which was biotinylated *in vitro* as follows – the ligand was diluted in PBS to 10 µM and treated with 3.3 mg/mL of Sulfo-NHS-LC-LC-biotin (Pierce) in 50-µL reaction for 1 h at 4°C. Non-reacted biotin was removed by overnight dialysis against at least 1000 volumes of TBS using Slide-A-Lyzer Mini Dialysis Columns (Pierce). Target biotinylation was quantified by performing standard HABA assay and validated by ELISA. Dual-gene short-fragment library consisting of *sbi* and *spa* genes only was constructed similarly to Zhang *et al* (Zhang et al. 1998). Briefly, primers sbi_F, sbi_R, spa_F and spa_R were used to amplify 2.5-kb (2012541..5015092) and 2.2-kb (2410671..2412907) regions from the *S. aureus* FP_SA_ST25 chromosome, containing the *sbi* and *spa* genes, respectively. Amplicon fragmentation and library preparation was performed essentially according to the developed protocol described in chapter 3.

4.2.5. Initial washing optimization.

In order to identify the optimal number of washes for the *Sa*si_1/Fc system, an optimization experiment was performed comparing the selection outputs after 10, 20 and 30 washes. Anti-Fc Affitin C3, previously developed in our laboratory was used as a positive selection control (Behar

et al. 2013). Affitins are artificial affinity proteins derived from extremophilic scaffold proteins found in various Archaea (Mouratou et al. 2007; Correa et al. 2014). One selection round was performed on flat-bottom 96-well polystyrene plate (MaxiSorp, NUNC). Prior target immobilization, a series of four wells (one positive control and 3 test wells for 10, 20 and 30 washes) were pre-coated with 100 µL of NeutrAvidin (Sigma), then loaded with 100 µL of 150 nM biotinylated Fc, washed three times with PBS and twice with WB, and finally kept at 4°C until needed. Pre-panning and panning were performed for 1 h at 4°C with mild agitation using about 145 µL of translation mixture per well/ tube. The wells were then washed 10, 20 or 30 times for 1 min (first half of the washes with 1 x WB/0.5% Tween 20/ 0.5% BSA and second half with 1 x WB/0.5% Tween 20 only). Selected RNA was eluted with elution buffer (50 mM Tris-HCl, 1.5 mM NaCl, 20 mM EDTA, pH 7.5) containing 50 ng/mL RNA carrier (*S. cerevisiae* RNA). Eluates were purified using RNeasy Mini Kit (Qiagen) and reverse transcribed in 20-µL reaction mixture with RevertAid Reverse Transcriptase (Thermo Fisher Scientifc) using TolAkurz primer. Finally, the obtained cDNA was amplified in 250-µL PCR with the following composition - 1 x HF (Thermo Fisher Scientific), 0.2 mM dNTPS, 0.5 µM primers T7C and TolAext, 4% DMSO (Thermo Fisher Scientifc), 0.5 units of Phusion Hot Start II DNA polymerase (Thermo Fisher Scientifc) using the program [initial denaturation for 30 sec at 98°C, 30 x (10 sec at 98°C/ 20 sec at 72°C) and final extension for 5 min at 72°C]. The reaction products were analysed on 1.5% agarose/1 x TAE gel.

4.2.6. Removal of MRGS-$(His)_6$-tag.

The MRGS-$(His)_6$-tag was removed from pFP-RDV1 by PCR, resulting in pSK-GeX2. Briefly, the vector was prepared using primer MAG in combination with primer link-F under conditions

identical to the one explained in chapter 3 for pSK-GeX1. This primer modifies the vector in a way that the translated fragment begins with the amino acids M-A-G instead of M-R-G-S-$(H)_6$.

4.2.7. Washing optimization of pSK-GeX2-derived libraries.

A second washing experiment was performed, identical to the one explained in point 4.2.5 except that the short-fragment library of *S. aureus* was prepared using pSK-GeX2 vector and the test series was 5, 10 and 15 washes instead of 10, 20 and 30.

4.2.8. Selection inhibition with erythromycin.

A single selection round was performed according to point 4.2.5. Six samples were prepared as a series of: 2 x pFP-RDV1-C3 as a positive selection control and a duplicate of pSK-GeX2-derived short-fragment library of *S. aureus*. Prior addition of the RNA template to the translation mixture, erythromycin was added to each duplicate at 1 ng/µL final concentration and the samples were incubated for 5 min at 4°C (on ice).

4.2.9. Performance of a complete selection cycle (3 consecutive selection rounds).

Three rounds of selection were performed using two (short- and long-fragment) pSK-GeX2-derived libraries of *S. aureus* FP_SA_ST25 prepared using the basic protocol explained in chapter 3. In order to increase the amount of selection output, four wells were processed per library at first selection round and the obtained outputs were pooled before proceeding with round 2. The other two rounds were completed using one well per library only. Once a round was completed, the selection output was processed as explained in point 4.2.5 and used as a template for the next one. Ten washes were performed at each round.

4.2.10. Sub-cloning and analyzing the selection output.

The selection output from the third round of selection was migrated in 1.5% agarose gel and the two differentiated areas of the output band were excised, separated in tubes containing 100 µL of 1 x TE buffer and designated U (upper region) and L (lower region). The amplicons were let to diffuse out of the gel slices overnight at 4°C and about 5 µL of the obtained eluates were used as a template for PCR using internal primers. Two 50-µL PCRs were prepared with the following composition - 1 x HF (Thermo Fisher Scientific), 0.2 mM dNTPS, 0.5 µM primers T7s and int_R, 4% DMSO (Thermo Fisher Scientifc), 0.5 units of Phusion Hot Start II DNA polymerase (Thermo Fisher Scientifc) using the program [initial denaturation for 30 sec at 98°C, 30 x (10 sec at 98°C/ 20 sec at 72°C) and final extension for 5 min at 72°C]. The products were analysed on 1.5% agarose/1 x TAE gel, treated with Exonuclease I for 2 h at 37°C to remove residual primers and purified using Wizard SV Gel and PCR Clean-Up System (Promega). Finally the obtained amplicons were digested with XbaI and HindIII restriction enzymes, subcloned into XbaI/HindIII-digested pFP-RDV1 and transformed in *E.coli*. The constructs within the grown clones were extracted with Wizard SV Minipreps DNA Purification System (Promega) and sequenced by standard automated Sanger sequencing. The obtained sequences were translated using CLC Sequence Viewer. The topology of the obtained peptides was assessed using TMpred server at http://www.ch.embnet.org/software/TMPRED_form.htmL (Hoffman and Stoffel, 1993).

4.2.11. Improvement of the selection specificity.

The complete selection cycle explained in point 2.10 was reproduced with the following modifications – the selection was proceeded using the output obtained from the first round explained in point 2.10. Round two and three were performed identically, except that 50 µL avidin-

agarose (Sigma) and pre-coated magnetic beads (BioAdem, StreptAvidin Plus, Ademtech) were used instead of polystyrene plates as selection matrices, respectively.

4.2.12. NGS analysis of the output from the improved selection.

Analysis of the selection outputs by NGS was performed essentially as explained in chapter 3.

4.3. Results.

4.3.1. Emulsion PCR.

Recombination is a well-known complication during multi-template PCR such as amplification of random genomic libraries. Therefore, we attempted to apply the protocol described by Shutze *et al* (Schütze et al. 2011) for recombination-free amplification of our expression libraries in emulsion PCR (ePCR). The obtained results are presented on page 124.

The most important parameter of ePCR is the stability of the generated water-in-oil emulsion. In order to verify that the emulsion is stable during the thermal cycling, a simple experiment needs to be performed – two separate PCRs are prepared with identical composition except that the primers are omitted in one of them. A template with known length is then added to the reaction with primers and another template with shorter length is added to the reaction without primers. After being emulsified separately, the two reactions are pooled and cycled. In case of stable emulsion no primer exchange will occur during the cycling resulting in only one amplicon, originating from the longer template. Logically, if the emulsion is unstable two amplicons will be observed resulting from the fusion of micro-droplets from the two emulsions. The result from this experiment is shown on panel A. Evidently, only one band was observed (lanes 1 and 3) in the emulsified reactions, indicating that the emulsion has been stable during the 40 thermal cycles.

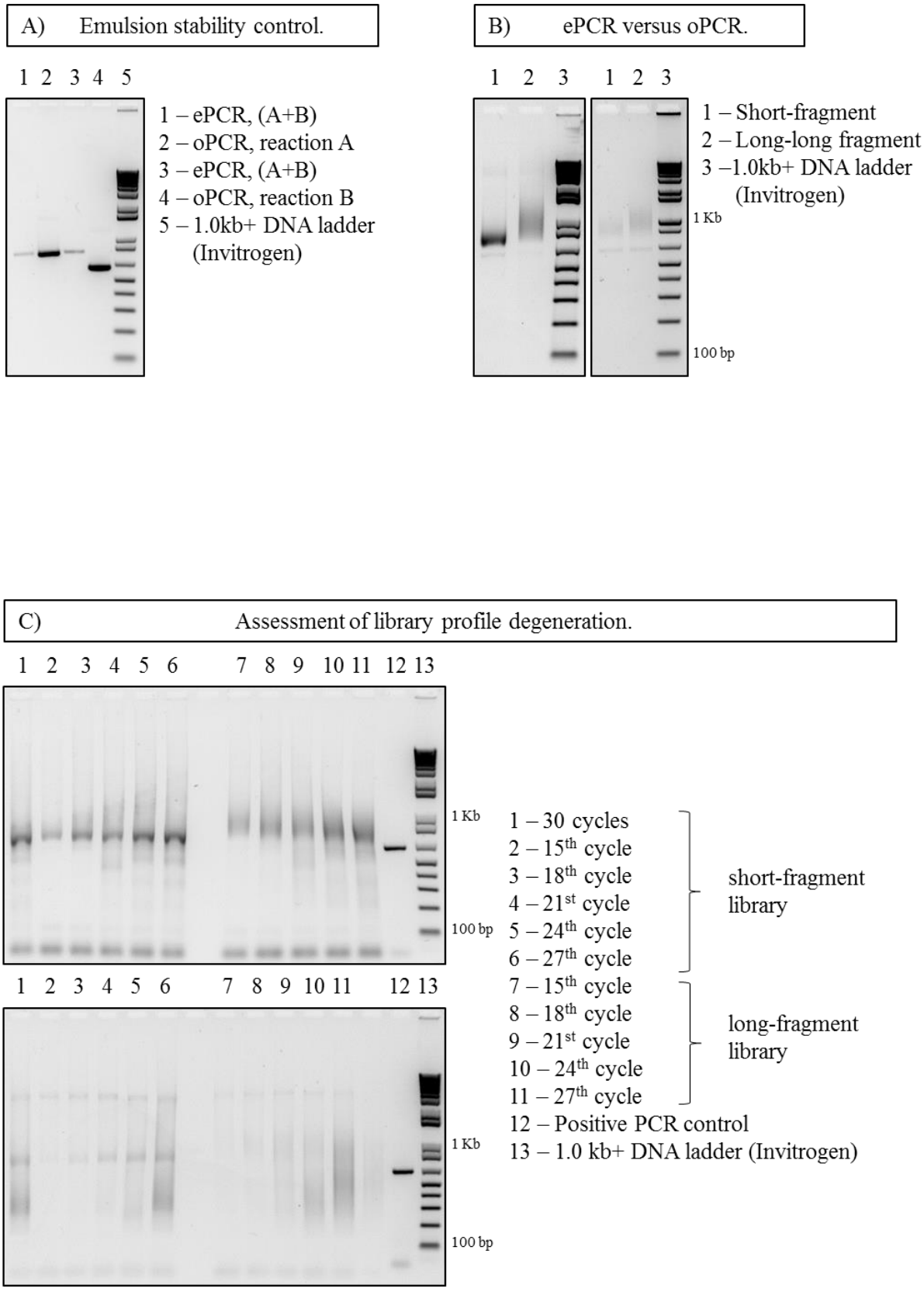


Figure 4.1. Assessment of library degeneration.

Having validated the emulsion stability, we proceeded with amplification of short- and long-fragment libraries of *S. aureus* FP_SA_ST25 using this optimized protocol for ePCR. Panel B shows a comparison between libraries amplified under standard and emulsion conditions. As can be seen on the gel, the obtained profiles were quite similar. However, the amount of linear library obtained after ePCR was extremely low to be sufficient for the subsequent step in the method – *in vitro* transcription which requires at least 1 µg of DNA template for satisfying RNA yields.

4.3.2. Assessment of the library degeneration.

Since we did not succeed to increase the yield of linear library with ePCR, we decided to assess the level of recombination during standard PCR which can be approximated by another simple experiment, reported by Williams *et al*, (Williams et al. 2006). In such an experiment, a multi-template library (genomic or cDNA) is being amplified under standard conditions as prior amplification the reaction mixture is split in several aliquots which are subsequently being removed with a known step from the machine during the cycling progress. The tubes are stored on ice until the final extension step is reached and then all samples are extended altogether. By analyzing the collected samples on an agarose gel, the gradual degeneration (if any) in the library profile could be observed. We performed this experiment by amplifying short- and long- fragment libraries of *S. aureus* FP_SA_ST25 with primers T7C and TolAext (the same couple we use for the linear library generation). Panel C shows the result from this experiment. Even though a minimal shift in the library profiles was observed during the late amplification cycles, the profiles of our libraries remained remarkably stable since this result is quite different than the one reported by Williams et al (Williams et al. 2006).

By comparing our conditions to the reported ones, we figured that the reason for this difference could be the two-step PCR used in our experiment. As mentioned earlier, ribosome display relies

on the presence of 5'- and 3'- stem-loops for protecting the ends of the translated constructs from being degraded by exoribonucleases present in the reaction mixture (Hanes & Plückthun 1997). However, this event cannot be fully avoided and ends degradation, which might compromise the full recovery of enriched fragments, does occur to a certain extend. A way to deal with this issue is the use of longer primers (50-70 bp) which anneal to longer parts of the fragment termini. Due to their length, these primers allow the performance of two-step PCR, which does not include annealing step at lower than the extension temperature (72°C).

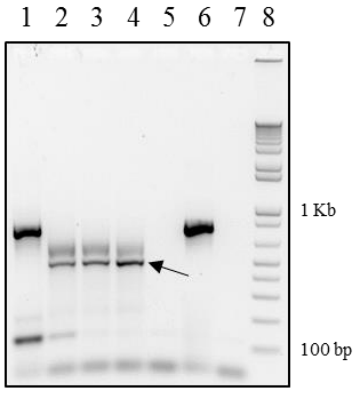
Therefore, we hypothesized that the two-step conditions might have prevented the library degeneration in our experiment by reducing the annealing between non-complementary templates (so-called heteroduplexes). These products might be favored at lower temperatures and have been proposed to drive PCR-recombination (Meyerhans et al. 1990; Thompson et al. 2002; Kanagawa 2003).

In order to test our hypothesis we reproduced the mentioned experiment using shorter primers [T7s (28 bp) and TolAs (16 bp)] and including an annealing step for 30 sec at 60°C. The results from this experiment (panel C) clearly indicate that adding a step with lower temperature in the cycling reduces the library quality by increasing the generation of multiple shorter amplicons with various size. Taking these results into consideration, we concluded that our libraries might be not significantly affected by PCR-recombination under the two-step conditions we used. Therefore, since we did not succeed to increase the yield of ePCR, we abandoned this approach and proceeded with the use of standard "open" PCR under two-step conditions.

4.3.3. Washing optimization.

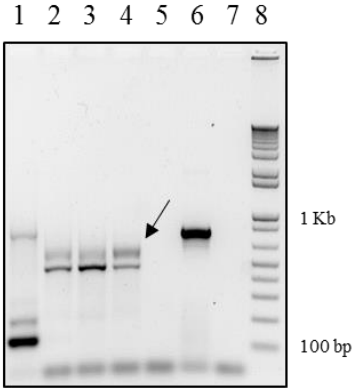
Washing is another important parameter of the selection process as the identification of desired product depends on it. Therefore we attempted a series of experiments to optimize this parameter. The results could be found on page 128. Initially, we performed a single round of selection using the explained *S. aureus*/ Fc system in order to optimize the number of washes. A short-fragment library of *S. aureus* FP_SA_ST25 was panned against biotinylated human Fc fragment in three wells, which were washed 10, 20 and 30 times for 1 min, respectively (panel A). Unexpectedly, while the level of selected products was reducing with increase in the washing number (lanes 2, 3 and 4), we observed exactly the opposite trend for the empty vector in the same samples– as could be seen on the gel, the intensity of the corresponding band increases with increase in the number of washes (indicated with black arrow). There were not many options for this to happen except if the empty construct has been translated during the selection and some part of it interacts strongly with a component in the system such as selection matrix (plate), blocking agent (BSA) or another molecule. These options were plausible due to the following reasons. First, if an empty, non-modified pFP-RDV1 was used, it would be prevented from being translated by the presence of 4 stop codons between the BamHI and HindIII restriction sites. However, as we modified this vector into pSK-GeX1 (for details, please refer to chapter 3), this region was not present anymore. Second, if translated, the empty pSK-GeX1 vector would contain MRGS-$(His)_6$ at its N-terminus and single cysteine at position 46. Surveying the literature about these interactions suggested that both of them could have caused the selection of empty vector construct. On one hand, a recent study (Holmberg et al. 2012) has reported that his-tagged proteins show increased adsorption onto tissue culture polystyrene (TCPS). TCPS, compared to the completely hydrophobic polystyrene, has been "hydrophylized" by being derivatized with variety of functional groups in order to increase its protein adsorption ability.

A) Initial washing optimization.



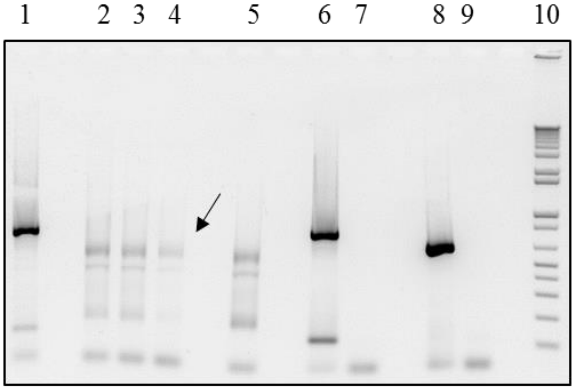
1 – C3 Affitin + Fc
2 – Short-fragment library + Fc/ 10 washes
3 – Short-fragment library + Fc/ 20 washes
4 – Short-fragment library + Fc/ 30 washes
5 – Negative RT control
6 – Positive PCR control
7 – Negative PCR control
8 – 1.0kb+ DNA ladder (Invitrogen)

B) Selection in the presence of iodoacetamide and imidazole.



1 – C3 Affitin + Fc
2 – Short-fragment library + Fc/ 10 washes
3 – Short-fragment library + Fc + iodoacetamide/ 10 washes
4 – Short-fragment library + Fc + imidazole/ 10 washes
5 – Negative RT control
6 – Positive PCR control
7 – Negative PCR control
8 – 1.0kb+ DNA ladder (Invitrogen)

C) Final washing optimization.



1 – C3 Affitin + Fc
2 – Short-fragment library + Fc/5 washes
3 – Short-fragment library + Fc/10 washes
4 – Short-fragment library + Fc/15 washes
5 – Short-fragment library + IgG/5 washes
6 – Positive RT control
7 – Negative RT control
8 – Positive PCR
9 – Negative PCR
10 – 1.0kb+ DNA ladder (Invitrogen)

Figure 4.2. Washing optimization.

The authors have proposed that there is a direct interaction between the six histidines in the His-tag and the carboxyl groups on the plate surface since addition of imidazole or EDTA reduced the adsorption of the tagged proteins. On the other hand, the only free cysteine in BSA, Cys34, is known to be able to react with free cysteines in another proteins. This interaction could be prevented by alkylating the free cysteine with iodoacetamide. Therefore, as derivatized polystyrene MaxiSorp plates similar to TCPS were used in our selections and BSA was used as a blocking agent, we decided to assess their impact on the selection of empty vector.

We designed a single-round selection experiment in which short-fragment library of *S. aureus* was selected against human Fc in the presence of imidazole and iodoacetamide. The results are presented on panel B. Unexpectedly, the addition of iodoacetamide had positive effect on the vector enrichment (lane 3). However, the addition of imidazole caused lower enrichment of empty vector and increased the selection of library fragments (indicated with black arrow). This result suggested that the MRGS-His tag was indeed involved in the undesired selection of empty vector. Therefore, we removed it from the vector using primers link_F and MAG_R, and reproduced the washing experiment presented on panel A identically, except that we used 5, 10 and 15 washes. The results are depicted on panel C. As evident from the picture, the intensity of the empty vector reduces with increase in the number of washes.

We chose 10 washes for our further experiments since 15 washes reduced quite much the intensity of the library fragments (indicated with black arrow).

4.3.4. Validation of GeXplore.

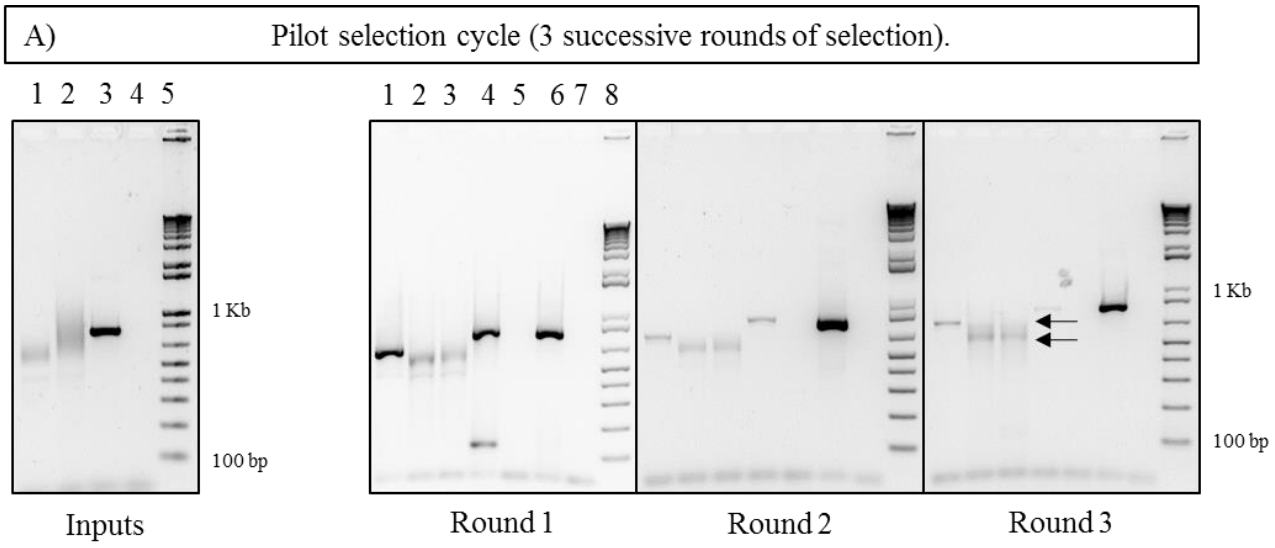
Having optimized the washing parameter we proceeded with validation of our method. In order to test the specificity and efficiency of GeXplore, we performed three rounds of selection using *S. aureus* short- and long-fragment libraries against human Fc fragment, expecting to identify mainly genomic fragments, encoding the Fc-binding domains of the two proteins, Sbi and SpA, in the selection output. Having in mind the compromised coverage of our *S. aureus* libraries, prior selection we verified that the two proteins were present in library *Sa*si_1 which was analyzed by NGS (for details, please refer to chapter 3).

The obtained results are presented on page 131. Panel A represents the two input libraries as well as the output evolution throughout the complete selection cycle (1st, 2nd and 3rd rounds). The electrophoregrams are ordered from left to right as follows – input libraries, first, second and third rounds, respectively.

Lane 1 corresponds to a positive selection control (pFP-RDV1+ anti-Fc Affitin C3). Lanes 2 and 3 correspond to short- and long-fragment library, respectively. All other lanes represent as follows – lanes 4 and 5 correspond to positive and negative reverse transcription (RT) controls, respectively; lanes 6 and 7 correspond to positive and negative PCR controls, respectively; lane 8 corresponds to 1.0 kb+ DNA ladder (Invitrogen).

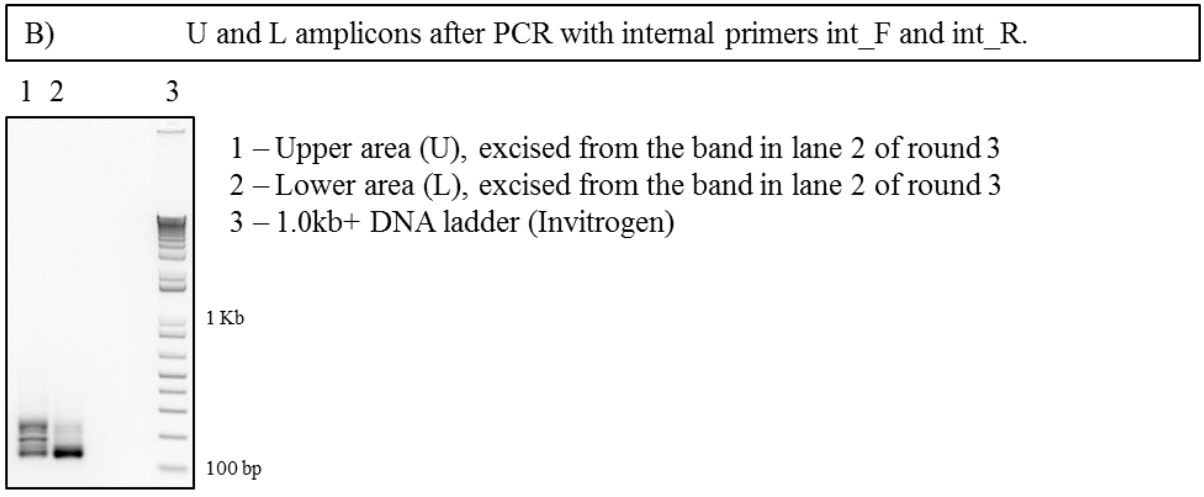
As evident from the picture, the range of the library profiles for both libraries has significantly narrowed at as early as 1st rounds and then has kept its range identical until 3rd round. The observed outputs correspond to genomic fragments with size 100-300 bp. Interestingly, we observed two different areas in the output profiles as the lower area appeared to be more presented.

In order to gain some preliminary insights on the obtained output, we excised the two areas separately and amplified them with internal primers int_F and int_R (Panel B). The obtained amplicons were then sub-cloned and transformed in *E. coli*.



1 – Short-fragment
2 – Long-long fragment
3 – Positive PCR control
4 – Negative PCR control
5 – 1.0kb+ DNA ladder (Invitrogen)

1 – C3 Affitin + Fc
2 – Short-fragment library + Fc/ 10 washes
3 – Short-fragment library + Fc/ 10 washes
4 – Positive RT control
5 – Negative RT control
6 – Positive PCR control
7 – Negative PCR control
8 – 1.0kb+ DNA ladder (Invitrogen)



1 – Upper area (U), excised from the band in lane 2 of round 3
2 – Lower area (L), excised from the band in lane 2 of round 3
3 – 1.0kb+ DNA ladder (Invitrogen)

Figure 4.3. Pilot selection cycle.

The constructs of 7 clones from the upper area and 8 from the lower were extracted and 6 clones per series were sequenced. The obtained sequences were mapped to the draft genomes of *S. aureus* FP_SA_ST25. The results are presented on page 133.

As evident from panel A, all sequences originated from two genomic regions – contig 5 and 39. Two clones (L3 and L7) were presented in duplicate and one (U5) in triplicate. No clone was found to originate from the *spa* gene. However, two clones (U1 and U3) were mapped to the *sbi* gene.

All other clones from the selection output mapped to non-IgG-binding proteins. An interesting observation was that all clones except L2 and L3 were found to originate from genes, encoding putative membrane-associated proteins. Since we performed the pilot selection exclusively on polystyrene plates, which are coated with proteins due to hydrophobic interactions, we hypothesized that this might have led to enrichment of fragments, encoding transmembrane regions of various membrane-associated proteins.

Therefore, we analyzed the peptides encoded by all non-IgG-binding clones using the TMpred server which predicts protein membrane-spanning regions and their orientation. Indeed, all analyzed clones except L2 and L3 were predicted to encode one or more transmembrane helices (data not shown). Panel B presents sequence alignment of translated clones U1 and U3 to the original Ig4 clone isolated from Jacobsson and Frykberg (Jacobsson & Frykberg 1995) the minimal IgG binding polypeptide proposed by Zhang *et al* (Zhang et al. 1998) and the domain I, proposed by Burman *et al* (Burman et al. 2008). The small size of clone U3 is due to the presence of HindIII-restriction site in the *sbi* gene, rather than being a result of functional selection. We were able to select a fragment, quite similar in size to the ones reported in previous studies. However, both clones from *sbi* were found only in the upper area from the migrated selection output.

A) Analysis of the sequenced U and L clones from the pilot selection output against Fc.

Clone	Contig	Position	Size (bp)	CDS	Product	TM
U1	5	242410..242438	29	A3650_10585	Immunoglubolin-binding protein Sbi	No
U2	5	843237..843330	94	A3650_13385	Quinolone resistance protein	Yes
U3	5	242416..242590	175	A3650_10585	Immunoglubolin-binding protein Sbi	No
U5	39	69506..69593	88	A3650_00780	Haloacid dehalogenase	Yes
U6	5	843237..843330	94	A3650_13385	Quinolone resistance protein	Yes
U7	5	734630..734711	82	A3650_12890	PTS glucose transporter subunit IICBA	Yes
L1	5	631731..631803	73	A3650_12445	Na/Pi contransporter	Yes
L2	39	802705..802835	131	A3650_04490	Na/Ala symporter	Yes
L3	5	896169..896253	85	A3650_13640	NADH-dependent flavin oxidoreductase	No
L4	5	896169..896253	85	A3650_13640	NADH-dependent flavin oxidoreductase	No
L7	39	69506..69593	88	A3650_00780	Haloacid dehalogenase	Yes
L8	39	69506..69593	89	A3650_00781	Haloacid dehalogenase	Yes

Table 4.1.

B) Alignment of clones U1 and U3 to Sbi peptides from selected studies.

```
1  TKHQTTQNNYVTDQQKAFYQVLHLKGITEEQRNQYIKTLREHPERAQEVFSESLKDSKNPDRRVAQQNADYNVLKNDNLTEQEK
2     QTTQNNYVTDQQKAFYQVLHLKGITEEQRNQYIKTLREHPERAQEVFSESLK
3                 QQKAFYQVLHLKGITEEQRNQYIKTLREHPERAQEVFSESLKDS
4        QNNYVTDQQKA
5           NYVTDQQKAFYQVLHLKGITEKQRNQYIKTLREHPERAQEVFSESLKDSKNPDRRVAQP
```

1 – Original clone Ig4, isolated by Jacobsson and Frykberg (Jacobsson and Frykberg, 1995)
2 – Minimal IgG-binding polypeptide, proposed by Zhang *et al* (Zhnag et al, 1998)
3 – Domain I of protein Sbi, proposed by Burman *et al* (Burman et al, 2008)
4 – Clone U1, isolated in this study
5 – Clone U3, isolated in this study

Figure 4.4.

Most of the clones from the lower area were found to originate from non-IgG-binding proteins, suggesting that this part of the output might be enriched due to unspecific interaction. Therefore we concluded that there is a problem with the selection specificity of our method.

4.3.5. Optimization of the selection specificity.

Taking into account the lines of evidence mentioned in the previous point, we hypothesized that the reason for the low specificity of our method is the non-specific hydrophobic interaction between non-blocked hydrophobic patches on the plate surface and membrane-spanning regions of membrane-associated proteins. In order to test out hypothesis, we modified the selection conditions by alternating the selection matrix at each round. We repeated the selection cycle, starting with the output from first selection round performed on plate and completed the other two rounds on avidin-agarose and magnetic beads, conjugated with streptavidin.

Comparison between the outputs of initial and modified round 2 and 3 is presented on page 135. As evident from panel A, the alternation of the selection matrix has significantly reduced the intensity of lower area of the output profile, which was found to contain non-IgG-binding proteins. Panel B shows the evolution of the library profiles after each round of selection. Before being migrated, all samples were amplified with internal primers int_F and int_R in order to remove the constant regions, originating from the ribosome display vector. The two electrophoregrams represent the samples series for the short- (upper picture) and long-fragment (lower picture) libraries. In each series, lane 1 corresponds to the input library and lane 2 – to round 1, performed on plate. The next lanes compare the output profiles after selection on plate (lanes 3 and 5) or agarose/magnetic beads (lanes 4 and 6, respectively) for second and third rounds.

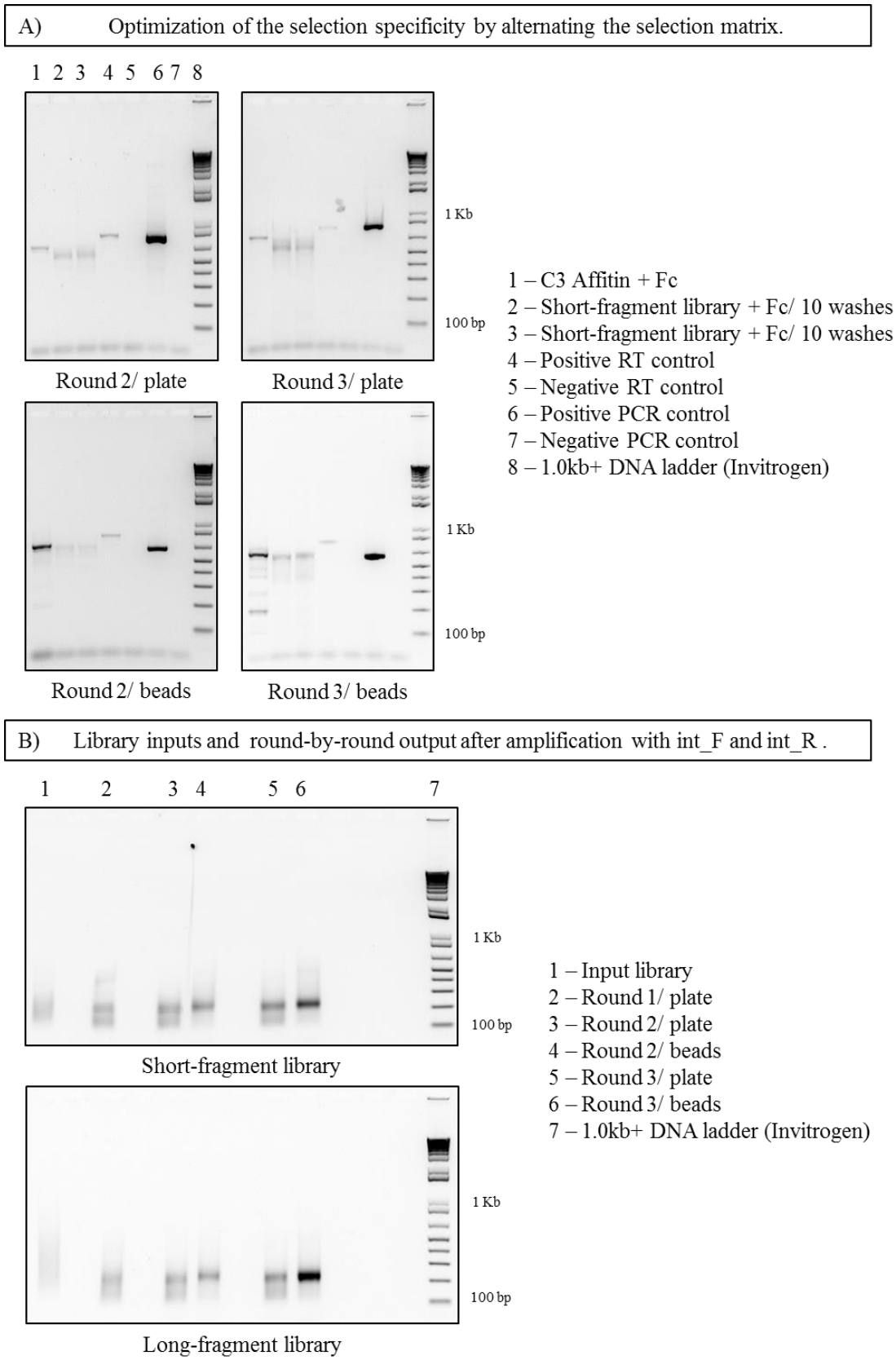


Figure 4.5. Optimized selection cycle.

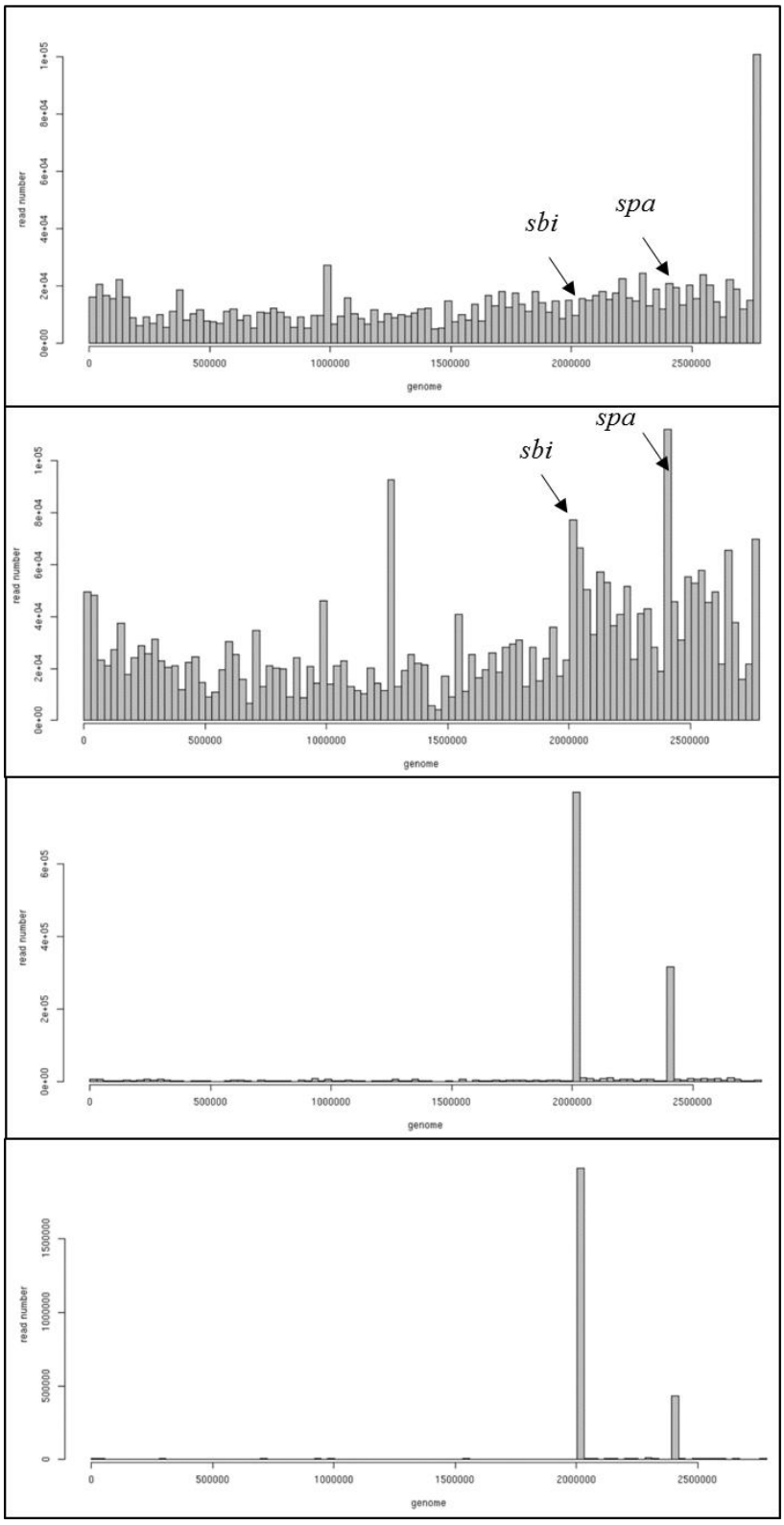
The reduction in the unspecific enrichment (lower area in the output profiles) is even more noticeable after the PCR with internal primers – for both libraries, the output of round three has a narrow homogeneous profile with size range of 200-250 bp. Considering the size of the *sbi* and *spa* IgG-binding domains (130 and 140 bp, respectively), it seems that the outputs of both libraries consist of fragments encoding one or maximum two domains.

We can propose two possible explanations of this result – the first one is that only fragments with maximum length of 250-bp could be translated using our *in vitro* expression system. The second one is that there could be a selection advantage of single rather than multi-domain fragments under our conditions. However, the former one is unlikely since the size limit of the cell-free *E. coli* expression system is known to be about 70 KDa, corresponding to more than 600 amino acids (Schaffitzel et al, 2006). Therefore, we think that minimal fragments containing at least a single binding domain, are preferentially selected over longer ones under our conditions.

4.3.6. Analysis of the selection outputs by NGS.

Having optimized the selection specificity we proceeded with NGS analysis of the selection outputs. In order to get a per-round insight on the performance of GeXplore, we sequenced all outputs obtained from the optimized plate/beads selection performed with the short-fragment library and mapped the reads to the genomes of *S. aureus* FP_SA_ST25. A genome-level representation of the obtained results is shown on page 137. The four plots correspond to the input library + 1st, 2nd and 3rd rounds from the top to the bottom, respectively. The plots were created by mapping the reads to the reference genome using bwa software, then splitting the genome into 100 consecutive windows of 27825-bp and calculating the mapping coverage for each window using the bedtools program.

A)	Round-by-round NGS analysis of the short-fragment library output.



Input short-fragment library

Round 1 on plate

Round 2 on agarose beads

Round 3 on magnetic beads

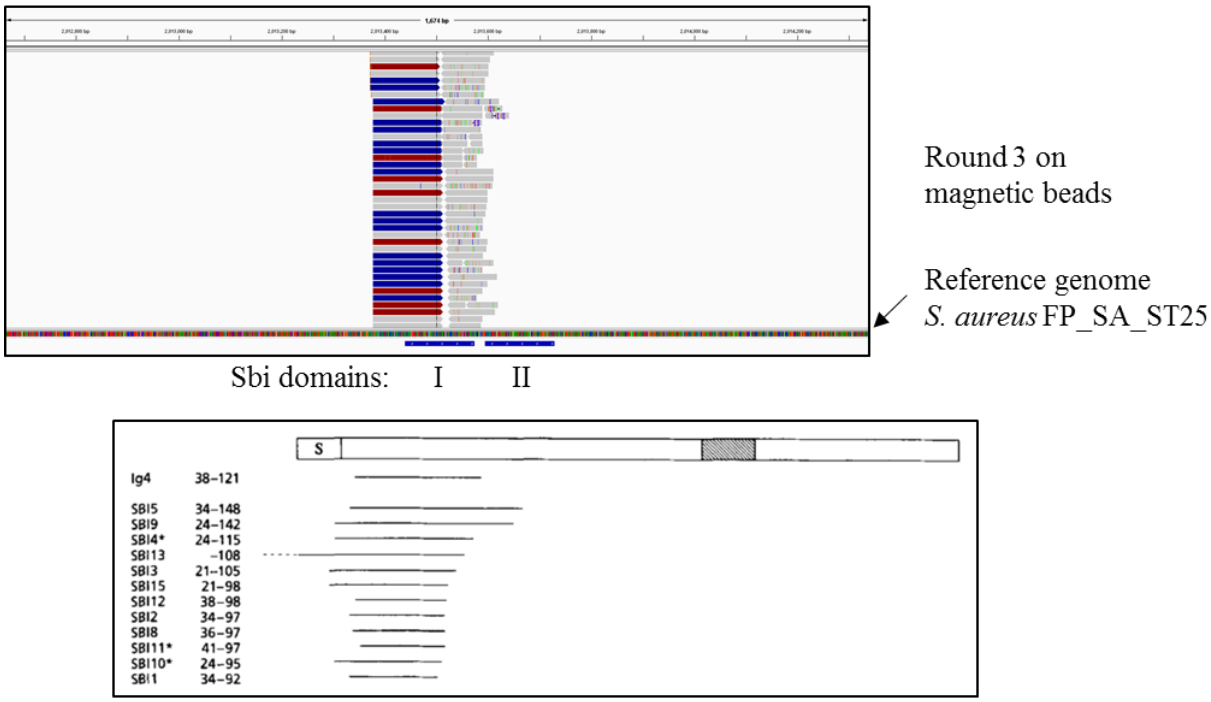
Figure 4.6. Validation of GeXplore (genome-wide view)

The results were plotted using R platform. The coordinates of the target genes *sbi* [2013292..2014605] and *spa* [2410988..2412490] are indicated with arrows on the first two plots. As can be seen on panel A, the number reads mapped to the target loci gradually start to increase at the first round together with many other genomic regions. At second round, the number of reads mapped to the target loci continues to grow and becomes dominant in the sample, while the rest of the genome is significantly underrepresented.

At third round, almost all reads in the sample mapped to the target loci and reads, originating from the rest of the genome were either missing or reduced to minimum. This result clearly indicates the selection power of GeXplore since we were able to identify two target proteins from a genome-wide library. Additionally, fragments originating from the target loci became dominant in the selection output at as early as second round of selection.

A domain-level representation of the read mapping for 3rd round only is presented on page 139. Panels A and B show zoom-ins of the read mapping focused on the target loci (*spi* and *spa*, respectively), visualized using IGV software. Regarding the *sbi* locus, we were able to select fragments encoding mainly the IgG-binding domain I of protein Sbi. This result correlates to the report Zhang *et al*, in which a gene library of *sbi* was made and screened against human IgG using shotgun phage display in order to identify the IgG-binding domain of the protein (Zhang et al. 1998). Accordingly, all fragments examined were found to originate from a region of the *sbi* gene identical to the one we observed in our experiment, corresponding to IgG-binding domain I. In order to confirm our results we also created dual-gene expression library consisting of only *sbi* and *spa* and performed two rounds of selection essentially as the one discussed in point 4.3.5. Yet, no fragments which encode domain II were identified (data not shown).

A) Zoom-in of the *sbi* region genomic region in *S. aureus* FP_SA_ST25.



The results from the domain mapping experiment of Zhang *et al* (picture taken from Zhang et al, 1998)

B) Zoom-in of the *spa* genomic region in *S. aureus* FP_SA_ST25.

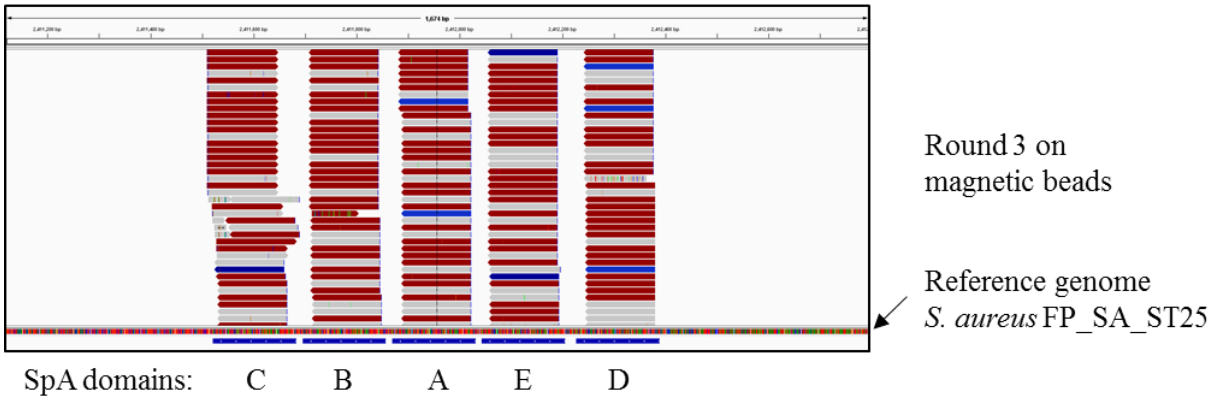


Figure 4.7. Validation of GeXplore (domain-level view)

Even though a second IgG-binding domain II has been predicted since the discovery of the protein (Zhang et al. 1998) and its IgG-binding has been later experimentally confirmed (Atkins et al. 2008), it seems that both, shotgun phage display and GeXplore fail to identify fragments, encoding this region. So far we have not been able to explain this result.

With regards to the *spa* gene, we were able to select fragments, encoding all five IgG-binding domains of protein Spa. The enrichment of fragments from all domains is obvious from the read mapping to the target locus, visualized on panel B.

4.4. Discussion.

With this part of our work, we completed the development of GeXplore. We have developed a genome wide, completely *in vitro* method for single- or multi-ligand selection. We launched our method using short- and long-fragment *in vitro*-expression random genomic libraries of *S. aureus* FP_SA_ST25, which we describe in chapter 3. The libraries were screened against human biotinylated human Fc fragment as a step-by-step parameter optimization was performed throughout the selection process. Having performed numerous optimizations, we attempted an experiment aiming at validating the selection efficiency and specificity of our method using a model *S. aureus*/ Fc system, previously used to validate shotgun phage display (Jacobsson & Frykberg 1995). Finally, the short-fragment input library as well as its corresponding 1^{st}-, 2^{nd}- and 3^{rd}- round outputs were sequenced by NGS and mapped to genome of *S. aureus* FP_SA_ST25, described in chapter 2.

During the optimization, we focused our attention on several points which could be potentially troublesome. Since our completely *in vitro* approach relies essentially on PCR to obtain input libraries and selection outputs, the first parameter which we assessed was the PCR-induced template degeneration. Recombination between non-complementary fragments during multi-

template PCR has first been reported by Meyerhans *et al* (Meyerhans et al. 1990). The study shows that up to 5.4% of the obtained amplicons could be recombinant molecules when two different HIV1 *tat* sequences are simultaneously used as a template. Importantly, even if the authors had not been able to prove the factors driving the recombination process, they have suggested that this event is a general phenomenon and should be expected when using "heterogeneous genetic material". Later studies have uncovered different aspects of the biased nature of multi-template PCR (for a review, please refer to Kanagawa 2003). Currently, it is well accepted that the performance of such PCR is complicated due to three main factors – (i) the heteroduplex formation and subsequent recombination mentioned above, (ii) short fragments are preferentially amplified over longer molecules and (iii) highly-diverse templates fail to re-anneal properly after denaturation step (Schütze et al. 2011).

In order to avoid such complications, an elegant approach have been developed termed emulsion PCR – the heterogeneous templates are separated and individually amplified in a water-in-oil emulsion (Diehl et al. 2006; Griffiths & Tawfik 2006). Importantly, such a strategy has been adopted by one of the pioneering NGS technology, Roche's 454, underscoring its efficiency. Even though creating a water-in-oil emulsion, stable enough to resist the thermal cycling during amplification is tricky and might require special equipment, streamlined protocols could be found in the literature (Schütze et al. 2011; Williams et al. 2006). As we anticipated some complications with our libraries we attempted to apply the ePCR developed by Schutze *et al*, and additionally assessed the degeneration of our libraries according to Williams *et al*. Thus, we succeeded to create a stable water-in-oil emulsion, able to withstand more than 40 thermal cycles. However, we failed to obtain sufficient amount of library amplicons to supply the following *in vitro* transcription step. Therefore, we dropped this approach and focused on determining the impact of standard multi-template PCR on our libraries. Surprisingly, we observed minimal shift in the library profiles,

which we attributed to the two-step PCR we used to amplify the libraries. As the library degeneration originates from the non-specific interaction between non-complementary fragments, we think that the lack of annealing step in our amplification program reduces this interaction. We confirmed our hypothesis by amplifying our libraries using annealing step and observed degeneration of the library profile. Here, it is important to mention the reason for using shorter versions of primers T7C and TolAext instead of just adding the annealing step to the cycling program. Primers T7C and TolAext anneal to the hairpin regions of the ribosome display construct. At temperatures below 72°C, both primers are in the form of almost perfect hairpins which severely impairs their priming efficiency. Therefore, to test our hypothesis we had to change two parameters – the primer length and the cycling conditions.

Having checked this parameter, we proceeded with optimization of the selection conditions – we performed a series of optimization experiments which revealed the following problems. First, our initial washing optimization trial showed significant enrichment of empty ribosome display construct, originating from the recircularization of the empty ribosome display vector. Considering the almost negligible level of empty vector construct in the input library, achieved by our G/C cloning strategy, this results could be explained only if the empty construct was translated in-frame during the selection and the product interacted with some component of the selection system. Due to the presence of $(His)_6$-tag at the N-terminus of the construct, we decided to test our hypothesis by performing another selection experiment in the presence of imidazole. Eventually, the addition of imidazole to the translation mixture influenced the selection output – the level of empty construct was reduced and the level of recovered genomic fragments was increased. This result suggested that the enrichment of empty construct was indeed related to the presence of $(His)_6$-tag. Importantly, our result correlated to the findings of an earlier study which has reported an increased adsorption of tagged proteins to derivatized polystyrene due to the interaction between $(His)_6$-tag

and the carboxyl groups on the plate surface (Holmberg et al. 2012). Therefore, we removed the tag from our ribosome display vector and repeated the washing experiment. Based on the obtained results, we chose 10 washes for our selection conditions.

Having resolved this issue, we proceeded our work with performing a complete selection cycle of 3 rounds on polystyrene plates using the mentioned *S. aureus*/ Fc model system. Such kind of system is particularly suitable for validation of our approach due to the following reasons – (i) it has already been used as a model system for the validation of related display technology, shotgun phage display, (ii) it challenges the ability of the tested technology to specifically identify few ligand-binding domains out of thousands of potential peptides/ proteins encoded by the whole-genome library and (iii) it is based on known and well-characterized protein-ligand interactions. The bacterium *S. aureus* possesses many proteins which facilitate its immune evasion and host interaction. Surface protein A (SpA) is a multi-domain protein of 42 kDa which is attached to the cell surface through a typical LPxTG motif. The protein is able to bind immunoglobulins G (IgGs) with 4 or 5 repeat domains (E, D, A, B, C) exhibiting IgG binding properties. Currently, it is well-accepted that the protein "cloacks" the bacterial cell with IgGs, thus preventing its recognition by Fc receptors on phagocytes (Atkins et al. 2008) and facilitating the immune evasion of the pathogen. *S. aureus* possesses another multi-domain protein, Sbi, which also has IgG binding activity. Importantly, this protein has been discovered during the very same experiment, used for validation of shotgun phage display (Jacobsson & Frykberg 1995). The protein contains 4 domains (I – IV) as domains I and II have been shown to carry the IgG-binding activity. More recently, the other two domains have been shown to interfere the complement pathways by interacting with complement factor C3 (Atkins et al. 2008).

Therefore, by performing three rounds of selection, we expected to identify all 7 Fc binding domains (two for Sbi and five for SpA) enriched in the selection output and all other genomic regions depleted, if not completely missing.

Generally, presentation of target ligand during selection could be performed in two different ways – immobilized on plate or in solution, followed by biotin/avidin capture. Studies using phage display have reported better recovery yields when performing first selection round on plate, probably due to a greater capture efficiency (Steiner et al. 2008). However, since the target coating is driven mainly by hydrophobic interactions, these could lead to its partial or complete denaturation. Additionally, again due to the hydrophobic interactions, there is a risk of high unspecific binding. Therefore, polystyrene plates must be used with caution and an optimization should be developed and performed for every binder/ligand system. To avoid target denaturation during our experiment, we biotinylated the Fc fragment and loaded it on Neutravidin pre-coated plates prior selection. After three rounds of selection on plate, we observed differentiation of two areas in the selection output band. Having considered the risk of high background on plate, we performed preliminary output analysis by excising the two output areas (upper and lower) from the gel, re-amplifying the obtained fragments and sub-cloning them separately. We then sequenced few of the obtained clones. While we did not identify any fragment belonging to the *spa* gene, encoding protein SpA, we identified two clones, U1 and U3 from the upper output area, which encoded the domain I of protein Sbi. As promising as these results may be, they indicated significant problem with our selection conditions – all other clones were found to belong to non-IgG binding proteins, suggesting high unspecific recovery. Also, we found that all clones except two encoded membrane-spanning regions of membrane-associated proteins, as predicted by the TMPred server. Thus, we suspected that the reason for the high background was that all three rounds were performed on plate which might have favored unspecific hydrophobic interactions.

Since it has been shown that alternating selection on plate with selection in solution reduces the level of unspecific recovery (Steiner et al. 2008) we reproduced our initial selection cycle as second and third rounds were performed on agarose and magnetic beads, conjugated with avidin. This matrix alternation clearly reduced the level of the unspecific lower area in the selection output. We observed a narrow homogeneous band, corresponding to fragments of 200-250 bp.

Having this issue solved we proceeded with round-by-round characterization of the selection output by Illumina sequencing. After mapping the obtained reads to the reference genome of *S. aureus* FP_SA_ST25, described in chapter 2 we observed gradual enrichment of reads mapping to the target loci *sbi* and *spa* while all other genomic regions were depleted. Importantly, fragments belonging to the two genes became dominant in the selection outputs as early as after second round of selection. At domain level, all domains of protein SpA were found specifically enriched. Concerning protein Sbi, we were able to recover fragments encoding only one of its Fc-binding domains. Surprisingly, the second Fc binding domain was not identified even when we prepared dual-gene library, consisting of genes *sbi* and *spa* only, and screened it against Fc fragment. Interestingly, although the discoverers of Sbi protein had predicted the second Fc-bonding domain, they were also not able to identify it out of neither whole-genome nor single gene libraries using shotgun phage display (Jacobsson & Frykberg 1995; Zhang et al. 1998). While the Fc-binding activity of the domain II has been recently confirmed experimentally (Atkins et al. 2008), we have not been able to explain the failure of shotgun phage display and our approach in identifying it.

Major concern about the specificity of our method was the high risk for generation of artificial combinatorial sequences (nucleotide or protein) which might have an affinity for some component of the selection system, because this would lead to enrichment of irrelevant sequences. However, so far we have not observed such kind of events. As indicated by our results, GeXplore is specific enough to allow screening of genome-encoded natural ORFs.

Chapter 5

Application of GeXplore

5.1. Introduction.

Pathogens utilize a multitude of diverse molecules to attach to the host, evade its immune system and disseminate over its tissues, eventually leading to an infectious disease. By its side, the host immune system counteracts the intruding microbes at cellular as well as molecular level. Therefore, the mounted humoral immune response could be seen as "immunological blueprint" which can be used for genome-wide search of the corresponding antigenic/ immunogenic proteins (Weichhart et al. 2003). Pathogen-specific antibodies such as IgG and IgA, derived from patient sera have been used in display technologies like ANTIGENome for identification of immune-relevant proteomes of human pathogens (Meinke et al. 2005). These proteomes have been then screened for potential vaccine candidates. Additionally, identified pathogen-specific proteins could be used for the development of diagnostic tools. An important observation obtained from these studies is that previously described virulence factors of the studied pathogens were being identified in the selection output (Etz et al. 2002). This result underscores the potential of display technologies for being used as an immune-capture assay for discovery of unknown virulence factors out of pathogen genomic libraries. Currently the virulence arsenal of certain human pathogens remains poorly understood. Especially interesting groups are the opportunists, which exert their pathogenic properties in compromised host only (Brown et al. 2012).

Streptococcus gallolyticus ssp. *gallolyticus*, formerly known as *Streptococcus bovis* biotype I is a Gram-positive gastrointestinal inhabitant in animals and humans (Schlegel et al. 2003). However,

it has also been recognized as an opportunistic pathogen due to its ability to cause septicaemia in birds, mastitis in cattle as well as bacteraemia and/or infective endocarditis in humans. Strikingly, close association between the bacterium and the incidence of colorectal cancer has been multiply reported during the last 4 decades and two recent reports have critically reviewed the available information (Abdulamir et al. 2011; Boleij et al. 2011). Accordingly, up to 60% of the patients with infective endocarditis and 80% of patients with bacteraemia had also colorectal neoplasia. The information about the virulence arsenal of this bacterium is scarce. A range of virulence features which might mediate the association of *S. gallolyticus* with colorectal cancer has been recently proposed (Boleij & Tjalsma 2013), such as its ability to translocate through the intestinal epithelium, to bind collagen with high affinity and to potentially evade the host immune system. However, despite the extensive research on the topic, the reason for this association remains unknown. Indeed, it is still unknown whether *S. gallolyticus* is just able to thrive in the setting of compromised colonic epithelium or, attached to the enterocytes, it contributes to the cancer initiation or progression, which is topic of our team in Nantes.

Mycobacterium ulcerans is the causative agent of the neglected tropical disease Buruli ulcer. The condition is the third most-common mycobacteriosis in the world after tuberculosis and leprosy and has been diagnosed in more than 30 countries over the world (Merritt et al. 2010). The main virulence factor of this bacterium is macrolide exotoxin called mycolactone leading to extensive skin ulceration due to its cytotoxicity. If left untreated, the disease leads to severe skin and limb mutilation. (George et al. 1999). Treatment include surgical removal of the affected area and more recent combination of rifampicin with streptomycin (Bolz et al. 2016). Nonetheless, the incidence of Buruli ulcer could be high, especially in poor rural area. The disease is most common in young teenagers and adults above 50 years (Bratschi et al. 2013). Current diagnostic methods involve microscopy, culture, histopathology and IS*2404*-specific PCR. However, all these laboratory

methods require skilled staff and technical equipment, both of which are difficult to provide with in remote and poor rural regions (Sakyi et al. 2016). Therefore, a cost-effective and field-friendly diagnostic tool is needed. Additionally, no vaccine is available up to date. Thus, according to our collaboration with Angers University (ATOMycA, INSERM Avenir, CRCNA U892) in the frame of the regional project "Alliance de Recherche sur les Maladies Infectieuses Nantes-Angers" (ARMINA), this atypic mycobacterium was also a relevant choice.

Therefore, we decided to use our optimized and validated approach, GeXplore, for identification of immune-relevant proteins of *S. gallolyticus* and *M. ulcerans.* These proteomes might contain unknown virulence factors and immunodominant proteins. Representative sets of sera from patients with *S. gallolyticus* (n=15) and *M. ulcerans* (n=24) were collected and used for purification of IgG and IgA antibodies. Genomic libraries of *S. gallolyticus* NTS31106099 and *M. ulcerans* S4018 were screened against purified IgG/IgA pools. After three rounds of selection, the obtained outputs were sequenced with Illumina NGS platform and analyzed. A representative set of enriched regions were selected, based on their read coverage after the final round of selection.

5.2. Materials and methods.

All manipulations with commercial kits were performed according to the manufacturer recommendations unless when stated otherwise. All primers used in the present work are enlisted in the table on page 21.

5.2.1. Determination of natural domain length distribution.

Domain length distributions for *S. aureus*, *S. gallolyticus* and *M. ulcerans* were determined using all their available proteomes in Pfam database version 27 (23, 3 and 1 entries, respectively as of June 2013). These data sets consist of complete list of Pfam domains (experimentally-confirmed

or predicted by Hidden Markov Models and multiple sequence alignments) which are found in the complete genome of a certain organism. The lists for *S. gallolyticus* and *M. ulcerans* were available until the release of Pfam version 29 in December 2015. The domain length distribution was determined in Microsoft Excel by calculating the frequency of domains in the proteomes with size L within a range of given scores for L. About 23 L-scores were set, ranging from 20 to 1000 (corresponding to the domain length in amino acids) with a step of 10 for 0-50 and 50 for the rest till 1000. The distributions were visualized by plotting domain frequency as a function of domain length.

5.2.2. Collection of human serum samples.

Serum samples from 15 patients (age = 57-84 years) who presented with *S. gallolyticus* blood-stream infection at Nantes University Hospital were provided by Stéphane Corvec, as 12 of them suffered infective endocarditis (IE, n=6), colorectal cancer (CRC, n=2) or both (n=4). Serum control samples from healthy blood donors matching the patients by sex and age were obtained from the Nantes Blood Bank. Serum samples from 26 patients (age = 3-65 years) who presented with an active Buruli ulcer were provided by Laurent Marsollier, together with serum from 24 healthy individuals who have been in contact with Buruli ulcer patients (termed "exposed").

5.2.3. Serum titration.

Titration of sera from patients with *S. gallolyticus* infection was performed by whole-cell ELISA. Briefly, overnight culture of *S. gallolytcius* NTS31106099 was prepared in BHI broth as explained in chapter 2. Cells were then harvested by centrifugation, washed three times with 1 x PBS and diluted to 1 x 10^9 cells (OD_{600} = 1.0). MaxiSorp plates (Nunc, Denmark) were then coated with 100 µL/ well of the obtained bacterial suspension, followed by overnight incubation at 4°C. After

being washed three times with 1 x PBS, the plates were blocked with 1% casein solution for 1 h at RT and washed three additional times with 1 x PBS. The wells were then loaded with 100 µL of control and patient sera in triplicate at 1/1000 times dilution in 1 x PBS. After incubation for 1 h at RT the plates were washed three times with 1 x PBS and loaded with 100 µL of anti-human IgG-HRP conjugate at 1/5000 dilution or anti-human IgA-HRP conjugate at 1/4000 in 1 x PBS. Following 1 h incubation at RT, the plates were washed six times with 1 x PBS and revealed for 15 min at RT with 100 µL of o-phenylenediamine dihydrochloride (OPD) substrate at 1 mg/mL in OPD buffer (100 mM citrate buffer, pH 5, 100 mM citrate buffer, pH5, containing 0,05% hydrogen peroxide). The plates were then read at 450 nm in a microplate-reader instrument (Tecan Infinite M200 Pro). Data were analyzed in Microsof Excel.

5.2.4. Purification of IgG/IgA from human sera.

About 115-200 µL of each serum set were pooled, resulting in 4 sample pools (*S. gallolyticus* patients and controls; *M ulcerans* patients and exposed) with a volume of 3 mL. The pools were pre-processed according to Fritzer *et al* (Fritzer et al. 2010). Briefly, the samples were heated for 30 min at 56°C to precipitate thermo-sensitive proteins, followed by centrifugation for 10 min at top speed to remove precipitates. The supernatants were then passed through 0.22 µm filter (Sartorius) and used for purification of IgG using HiTrap Protein G HP column (GE Healthcare) at a flow rate of 1 mL/min using a Bio-Rad BioLogicDuoFlow 10 system with PBS as running buffer. The flow-through fractions were diluted 3 times in water and used for IgA purification using Peptide M-agarose (Invivogen). After being purified, 50 µL aliquots of the IgG and IgA eluates were dialyzed against 500 mL of PBS in Slide-A-Lyzer MINI Dialysis Devices (Thermo Fisher Scientific) and analyzed on 10% SDS/ 15% polyacrylamide gels. Finally, the obtained antibodies

were biotinylated *in vitro* using EZ-Link Sulfo-NHS-Biotin according to the protocol explained in chapter 4.

5.2.5. Screening of random genomic libraries against disease-relevant human antibodies.

Short- and long-fragment libraries of *S. gallolyticus* NTS31106099 and *M. ulcerans* S4018 were prepared essentially according to the optimized protocol described in chapter 3. Prior selection the two libraries of each pathogen were pooled together in order to cover the natural domain distribution. Identification of antigenic proteins was performed following two approaches. The first selection approach was performed exclusively according to the optimized polystyrene/avidin-agarose/magnetic beads protocol described in chapter 4, except that 6 washes only were performed throughout the whole selection. A series of two samples were used per pathogen – patients + healthy controls for *S. gallolytcius* and patients + exposed for *M. ulcerans*. A pool of 2 µg of biotinylated IgG and IgA were coated per sample. Three rounds of selection were performed plus repetition of the third selection round in the presence of erythromycin at 1 ng/mL to differentiate between PCR-biased and selection-enriched genomic regions. The second approach is a modification of our basic protocol according to Weichhart *et al* (Weichhart et al. 2003). First, polystyrene plates at first round were replaced by Bio-Adembeads StreptaDivin (Ademtech, 50 µL/tube). Second, the amount of antibodies was increased from 2 to 10 µg. And finally, antigen binding was performed in solution – biotinylated antibodies were added directly to the translation reaction, followed by incubation for 1 h at 4°C with mild agitation and then the formed complexes were captured for 1 h at 4°C with bead conjugates (agarose or magnetic, 50 µL/tube) prior washing. Output processing, NGS and data analysis were performed essentially as explained for the GeXplore validation in chapter 4.

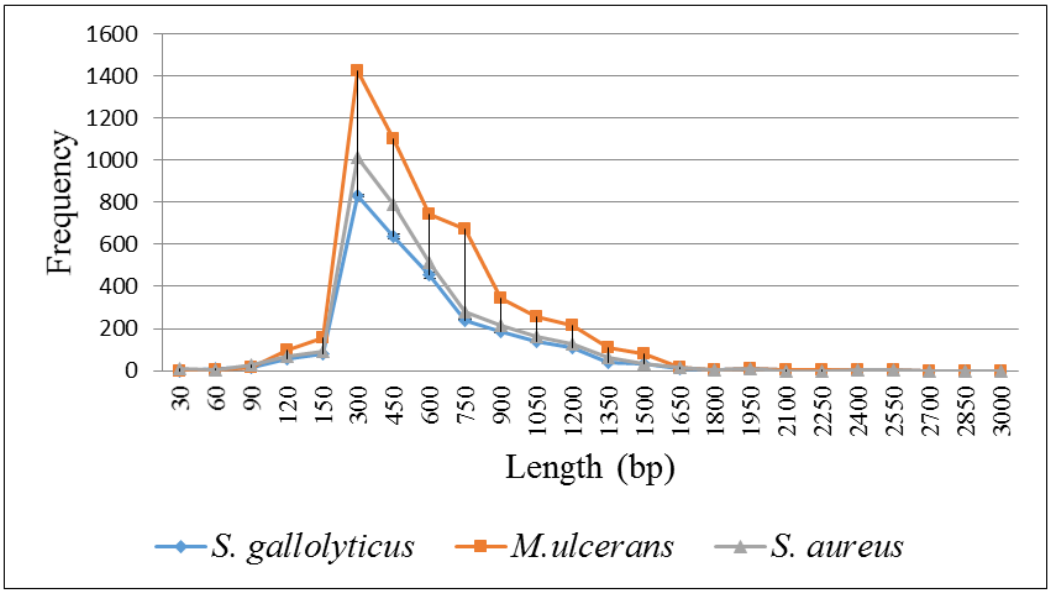
5.3. Results.

5.3.1. Natural domain length distribution.

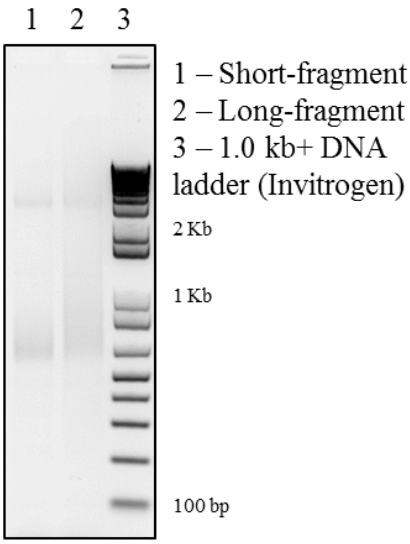
Pfam database is a huge collection of protein families which provides predicted proteomes for many complete genomes available in the public databases. The provided proteome datasets include, among other information, an estimated number and size of the protein domains encoded by the particular genome. Such proteome datasets were available for the three species included in our experiments until the release of Pfam version 29.0 (December, 2015) when the dataset for *S. gallolyticus* and *M. ulcerans* were removed due to rearrangement in its sequence database (Finn et al. 2014). We used the available datasets to get an idea about the domain length distribution over the three genomes of interest. Provided with the length of the predicted domains, we determined their frequency over the entire proteome as a function of their size.

The obtained results are presented in the chart on page 153 (Panel A). Accordingly, we observed a normal domain length distribution. Most of the domains were found to lie in the range between 20 and 550 residues, corresponding to 60-1650 bp. In order to follow the observed trend, we fragmented each of the three genomes of interest in a series of two batches – short-fragment (100-300 bp) and long-fragment (200-1500 bp). The obtained genomic fragments were used for the preparation of short- and long-fragment random genomic libraries. The *S. gallolyticus* NTS31106099 and *M. ulcerans* S4018 libraries, prepared according to our optimized protocol discussed in chapter 3 and used for the experiments discussed in this chapter are presented on panel B. Due to the fragment size heterogeneity, it is highly probable that shorter genomic fragments have been ligated and amplified preferentially over longer ones during ligation and subsequent conversion of circular constructs into linear expression library by PCR.

A) Domain length distribution for the three species of interest.



B) Short- and long-fragment library of *S. gallolytics* NTS31106099.



C) Short- and long-fragment library of *M. ulcerans* S4018.

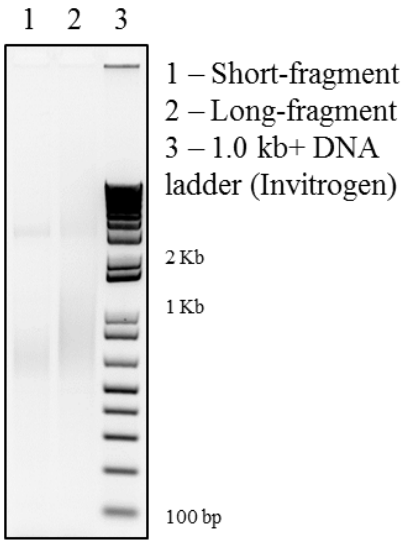


Figure 5.1. Determination of fragment length range.

However, after library preparation, we observed that the profile of our long-fragment libraries reaches about 1650 bp which correspond to an inserts of 1200 bp thus covering a significant fraction of the encoded domains.

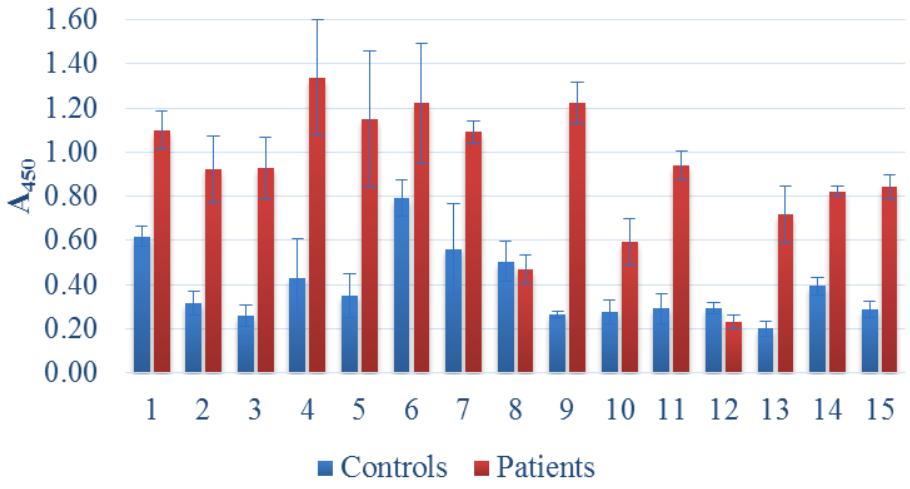
5.3.2. Serum titration.

We performed serum titration using whole-cell ELISA in order to have an idea about the anti-pathogen reactivity of our serum sets. We coated polystyrene plates with cells of *S. gallolyticus* NTS31106099 and then reacted the plates with appropriately diluted sera from patients and healthy blood donors. The obtained results are shown on page 155. As evident from the two charts presented, 13 and 10 out of the 15 sera tested showed IgG and IgA titers, respectively, which were higher than those of the corresponding controls (age and sex matches).

This finding correlates with previous reports, investigating the humoral response against *S. gallolyticus* using defined recombinant pilus proteins or cell-wall extracts (Abdulamir, 2009, Boleij, 2012, Butt et al, 2015).

Unfortunately, we were not able to develop a proper ELISA to titer the *M. ulcerans* serum sets. We attempted to perform a whole-cell ELISA in solution using suspension of *M. ulcerans*, but we did not obtain any significant results. We detected quite low signal in the two test groups, exposed and infected individuals as both showed similar levels. Additionally, many of the exposed healthy individual sera showed higher (even if negligible) signal than the patient ones. Therefore, we proceeded with antibody purification without further attempts to titer the *M. ulcerans* serum sets.

A) IgG titers in control and patient serum sets for *S. gallolyticus*.



B) IgA titers in control and patient serum sets for *S. gallolyticus*.

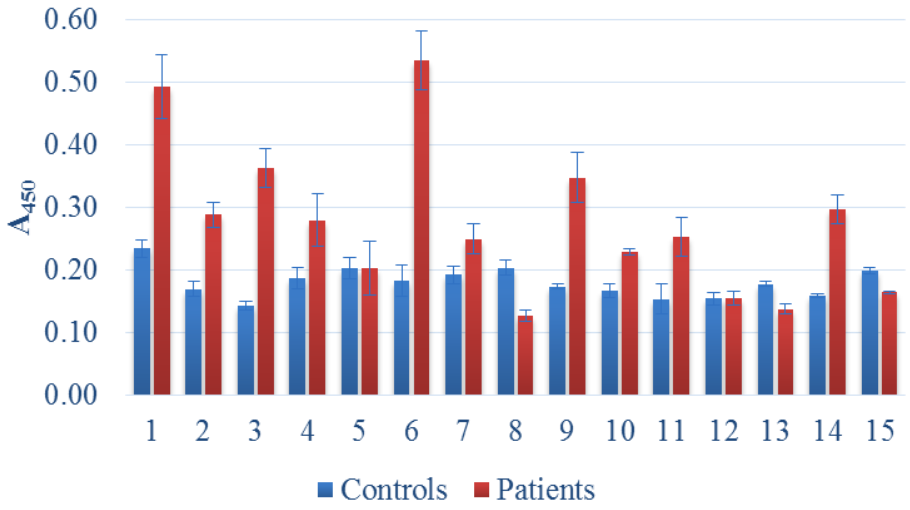


Figure 5.2. Titration of *S. gallolyticus* serum sets

5.3.3. IgG/IgA purification.

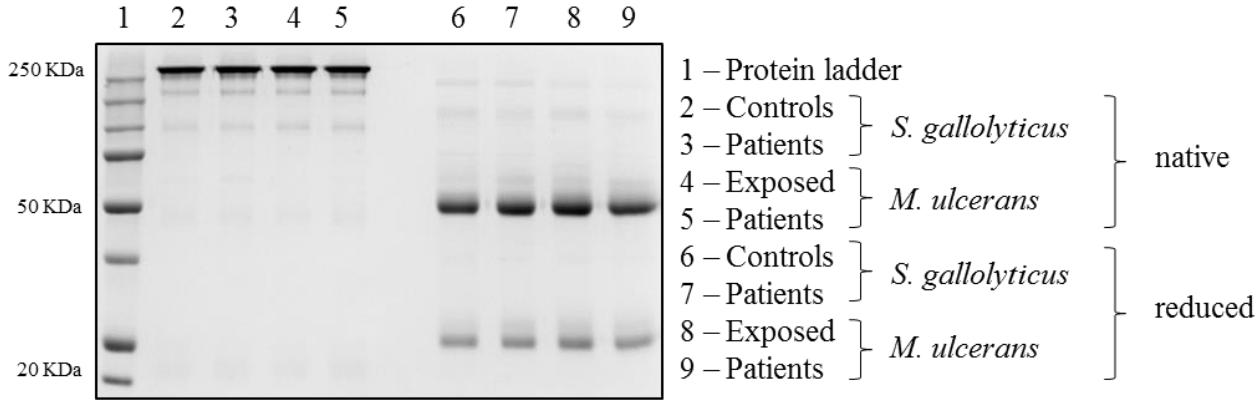
We purified total IgG and IgA fractions from the selected serum sets in order to use them as multi-ligand selection reagents. A single 3-mL pool was prepared from every serum set, which was then loaded on a Protein G column for purification of total IgG. The flow-through fractions was then directly used for purification of total IgA. Panel A on page 157 shows the calculated yields for each immunoglobulin fraction. In a general adult population IgG and IgA levels have been found to be at 11.18 ± 2.5 mg/mL and 2.6 ± 1.2 mg/mL, respectively (Gonzalez-Quintela et al, 2008). As evident from the table, while we were able to purify good amounts of IgG from every serum pool (7.4 – 8.6 mg/mL serum), we obtained significantly lower amounts of IgA than expected (0.65 – 1.15 mg/mL serum) considering their normal serum levels, reported in the literature. Panels B and C represent the purity of the obtained immunoglobulin fractions. As evident from the two electrophoregrams, IgG eluates had superior purity than the IgA ones.

A) Yields after IgG/IgA purification.

Type	*S. gallolyticus* serum sets		*M. ulcerans* serum sets	
	Controls	Patients	Exposed	Patients
IgG (mg/ mL)	7.42	7.51	8.65	8.55
IgA (mg/ mL)	0.65	1.15	0.86	0.68

Table 5.1.

B) IgGs, purified using Protein G column.



C) IgAs, purified using Peptide M column.

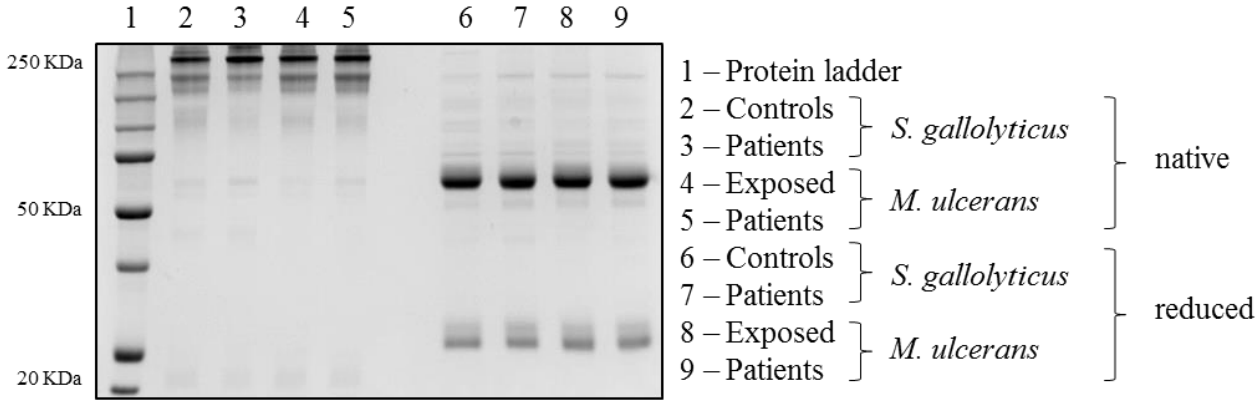


Figure 5.3. Purification of IgG/ IgA

5.3.4. Identification of immune-relevant proteins of *S. gallolyticus* NTS31106099.

Initially, we performed multi-ligand selection following our optimized conditions, described in chapter 4. Short-and long libraries of *S. gallolyticus* NTS31106099 were pooled and screened against IgG/IgA pool, obtained from healthy blood donors or patients with *S. gallolyticus* infection. After Illumina NGS sequencing, the obtained reads were mapped to the reference sequence of strain NTS31106099, described in chapter 2. Panel A on page 159 represents the obtained results. The ring diagram contains five rings. The innermost ring (pale-red) represents the short-fragment input library *Sg*si, characterized in chapter 3.

The second and third innermost rings (dark-red) correspond to outputs of third selection round (against control and patient antibodies, respectively). The last two rings (lighter red), correspond to the same samples, however from a third selection round performed in the presence of erythromycin. As evident from the figure, many genomic regions (> 100) have been retained and enriched during the selection. Enriched regions were then analyzed in detail using IGV software as we focused on regions with the highest enrichment only. The IGV program visualizes the reads mapped to the reference genome sequence together with a read coverage indication. We used this value to score the observed enriched regions and to select the regions with the highest enrichment. A list of 31 genomic regions was created as they were selected in a clockwise consecutive manner over the reference sequence. Panel A on page 160 represents the list of selected regions. It contains the following information – an arbitrary region number, size of the enriched region in base pairs as estimated visually in IGV, coordinates of the enriched region on the reference genome, coordinates of corresponding CDS on the reference genome, the functional annotation of the corresponding CDS and the read mapping depth for the two samples analyzed, controls and patients, as calculated by IGV. Accordingly, we observed the following results. The size of the enriched genomic regions was found to range between 130 and 364 bp (average = 230 bp).

A)	Input and outputs mapped against the genome of *S. gallolyticus* NTS31106099.

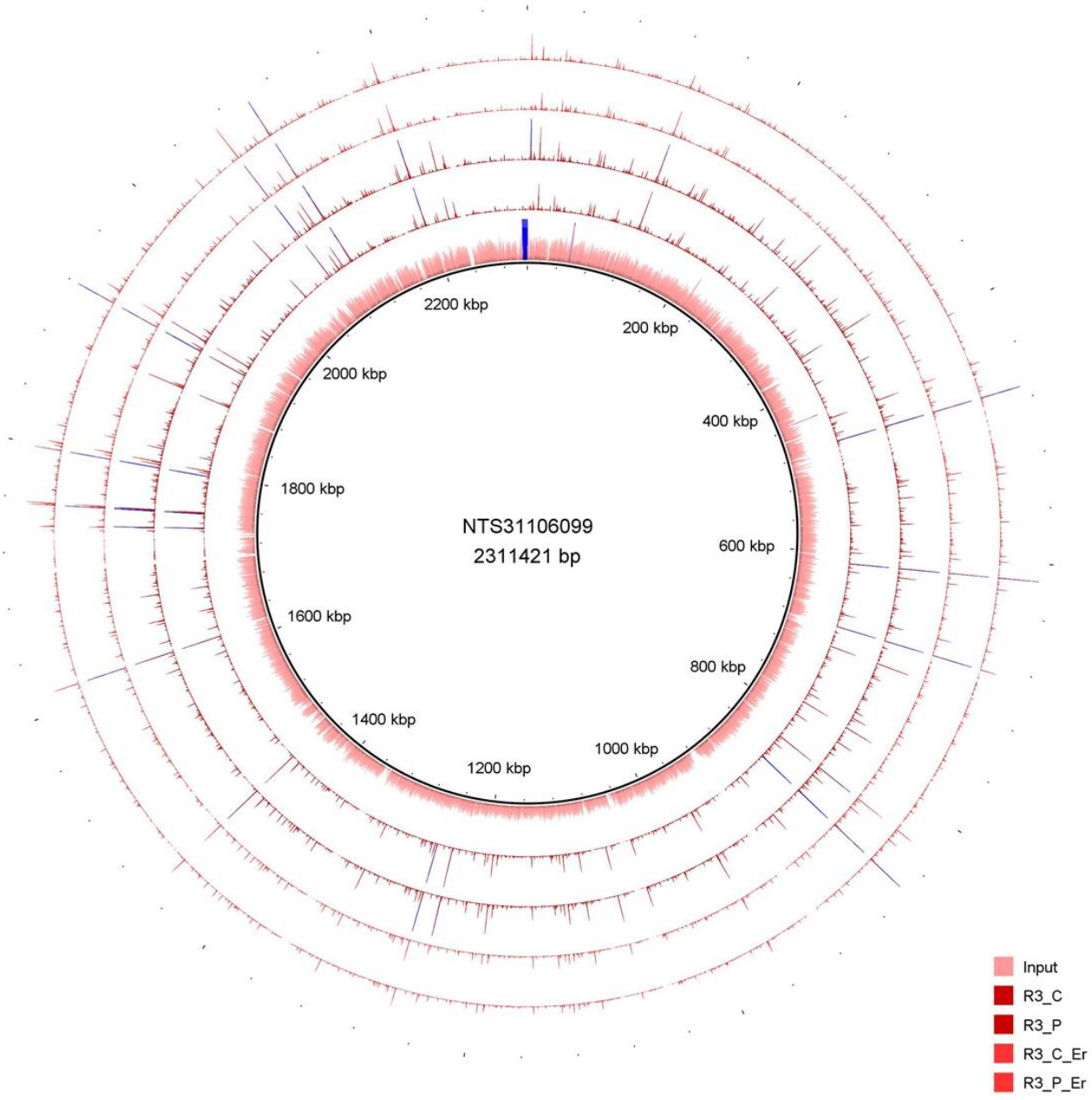


Figure 5.4. Enriched genomic regions of *S. gallolyticus* NTS31106099.

A) Representative subset of *S. gallolyticus* NTS31106099 regions with highest enrichment.

#	Size (bp)	Selected region	CDS coordinates	CDS annotation	Controls	Patients
1	343	3430..3772	3155..4537	UG96_00105, peptidoglycan hydrolase	0	26269
2	364	12150..12513	10084..13809	UG96_00140, phosphoribosylformylglicineamidine synthase	6442	8176
3	301	129280..129580	128295..130478	UG96_00845, maltose/glucose-specific PTS	10958	15266
4	161	292240..292400	292215..293099	UG96_01645, protease	6987	9265
5	220	368640..368859	368316..368897	UG96_02030, peptidase	6520	8200
6	300	471340..471640	471004..472749	UG96_02560, multidrug ABC-transporter ATP-binding protein	26377	32980
7	210	612700..612910	612761..613684	UG96_03265, memrane protein	81247	141406
8	210	685490..685700	685114..685734	UG96_03555, uridine kinase	12327	11448
9	140	708960..709100	708488..709747	UG96_03660, UDP-N-acetylglucosamine 1-carboxyvinyltransferase	6616	8168
10	178	779100..779278	779171..780520	UG96_03965, hypothetical protein	7111	7358
11	280	826480..826750	826414..828165	UG96_04205, ABC transporter	10964	15217
12	140	856980..857120	856722..857906	UG96_04340, MFS transporter	125470	211986
13	220	960420..960640	959684..961453	UG96_04855, mannitol-specific PTS transporter subunit IIB	5497	6786
14	190	1035950..1036140	1034901..1036226	UG96_05185, multidrug transporter, MatE	3823	4573
15	330	1065580..1065850	1064614..1068455	UG96_05325, ABC-trasporter permease protein	6086	6770
16	230	1194600..1194830	1194441..1195874	UG96_05965, aryl-phospho-beta-D-glucosidase	5591	7450
17	300	1341260..1341560	1241352..1242008	UG96_06195, hemolysin III	12326	16199
18	190	1259280..1259470	1259191..1259817	UG96_06295, glycine/ betaine ABC transporter permease	14401	16132
19	320	1453430..1453750	1450589..1453639	UG96_07295, collagen-binding protein	9756	13196
20	230	1614710..1614940	1613543..1615060	UG96_08095, fimbrial protein	17039	20788
21	130	1739190..1739320	1738678..1739514	UG96_08675, peptidylprolyl isomerase	12411	13152
22	180	1755160..1755340	1754762..1756144	UG96_08760, MFS transporter	33740	46311
23	240	1797840..1798080	1797173..1798504	UG96_09000, hemolysin	60459	61163
24	190	1879950..1880140	1879571..1880959	UG96_09415, MFS transporter	8096	9445
25	200	1920050..1920250	1919717..1920661	UG96_09595, magnesium transporter	23887	29568
26	220	1930800..1931020	1930401..1931174	UG96_09670, ABC transporter permease	10557	12846
27	180	2034070..2034250	2033409..2035556	UG96_10195 ABC transporter ATP-binding protein	5856	6875
28	270	2069700..2069970	2068490..2069899	UG96_10320, glucan-binding protein	10667	17303
29	250	2100050..2100300	2099042..2100208	UG96_10485, MFS transporter	136875	198864
30	251	2193750..2194000	2191759..2193390	UG96_10945, PTS system trehalose-specific transporter subunits IIBCA	25153	18831
31	190	2221050..2221240	2219250..2221193	UG96_11060, DNA mismatch repair protein MutL	5704	7630

Table 5.2.

B) The selected regions of *S. gallolyticus* NTS31106099 divided in functional categories.

Category	Regions	Number	% of the total
Membrane transport	3, 6, 7, 11, 12, 13, 14, 15, 18, 22, 24, 25, 26, 27, 29, 30	16	51.6
Virulence	17, 19, 20, 23	4	12.9
Cell-wall biosynthesis	1, 9, 28	3	9.7
Hypothetical	1	1	3.2
Other	2, 4, 5, 8, 16, 21, 31	7	22.6

Table 5.3.

All of the 31 selected regions were found to encode a protein. No intergenic region were selected. Based on their function we organized the selected proteins in 5 categories – membrane transport (51%, n=16), virulence (13%, n=4), cell-wall biosynthesis (10%, n=3), hypothetical (3%, n=1) and other (23%, n=7). Panel B shows the grouping of the selected regions into categories. Taken together these data show that 74% of the most enriched regions encode protein which could be associated with the cell surface (membrane/cell wall). Evidently, the selection output is dominated by fragments, encoding various membrane transporters.

Four of the regions (17, 19, 20 and 23) were found to encode proteins that might contribute to the pathogenesis of *S. gallolyticus*. Regions 17 and 23 encode for putative hemolysin. Region 20 encodes for putative fimbrial protein. Importantly, region 19 was found to encode the N-terminal part of the putative collagen-binding protein UG96_07295 discussed in chapter 2 and located on Tn*6263*. About 23% of the identified regions were found to encode neither surface-exposed nor virulence-associated proteins. These include proteins involved in sugar (n=1) and nucleotide (n=2) metabolism, proteases (n=2), a peptidyl-prolyl isomerase and MutL DNA mismatch repair protein. Read depth of the chosen region was in the range of 3823x..136785x and 4573x..211986x for the control and patient samples, respectively (panel A on page 162). Three regions (7, 12 and 29) presented with significantly higher read depth than the rest of the regions. All three regions encode membrane transport proteins. Regions 22 and 23 had intermediate read depth. They encode a MFS transporter and a hemolysin, respectively. The read depth difference for all other regions was not so significant. Although the patient and control read depth values series were quite similar, for almost all selected regions the patient series had higher depth than the control one. This difference was the most profound for the 3 dominant regions. Surprisingly, region 1 was enriched in only one of the samples – it encodes a putative peptidoglycan hydrolase and was enriched only in the patient sample. The same region was missing in the control sample.

A) Coverage comparison between the selected regions of *S. gallolyticus* NTS31106099.

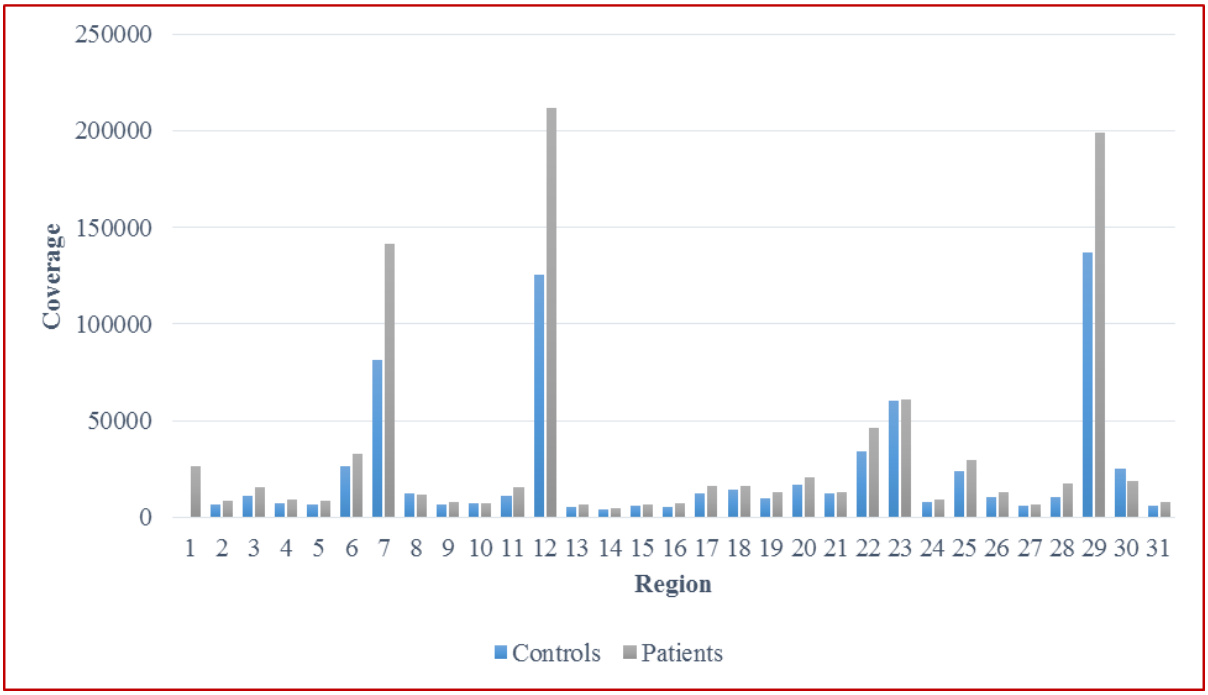


Figure 5.5.

5.3.4. Identification of immune-relevant proteins of *M. ulcerans* S4018.

Identical multi-ligand selection was performed using pooled genomic libraries of *M. ulcerans* S4018. After NGS, the reads were mapped to the draft genome of *M. ulcerans* S4018, discussed in chapter 2. Due to the poor quality of the draft sequence, we also mapped the reads to the publicly available complete genome of strain Agy99. The obtained results are shown on page 164. Note that only the mapping to strain S4018 is shown, since we did not observe any difference in the two alignments. As evident from the figure, multiple genomic regions were again enriched, while others were depleted or missing. Identical detailed analysis was performed using IGV software in order to select the highest enriched regions. Notably, the overall region enrichment in this experiment was significantly lower than the one observed for *S. gallolyticus*. Furthermore, the enrichment level was not so easy to differentiate among all enriched regions. Therefore, we set a lower coverage limit of 1000x and then, we selected all regions with higher than this coverage. The selected subset is presented on page 165 and comparison of the coverage of the selected regions is shown in page 166. Accordingly, we selected 44 genomic regions. The size of the enriched regions was found to vary between 110 and 500 bp (average 209 bp). About 42 of the regions were found to encode a protein. Interestingly, 2 regions (25 and 36) were found to be intergenic. Additional BLASTp search of these regions revealed that region 25 encodes a hypothetical protein. Similarly to the previous selection we were able to divide the selected regions into the following categories – transport (27.3 %, n=12), conserved membrane proteins (18.2%, n=8), virulence (13.6%, n=6), cell-wall biosynthesis (9.1%, n=4), hypothetical (2.3%, n=1) and other (27.3%, n=12). Taken together these result indicate that about 68% (n=30) of the proteins, encoded by the selected regions could be associated to the cell surface (membrane/cell wall). Read coverage of the selected regions was found to range from 1000x to 5838x and 3352 x for exposed and patients, respectively. About three regions (39, 42 and 44) had significantly higher coverage than the rest.

A)	Input and outputs mapped against the genome of *M. ulcerans* S4018.

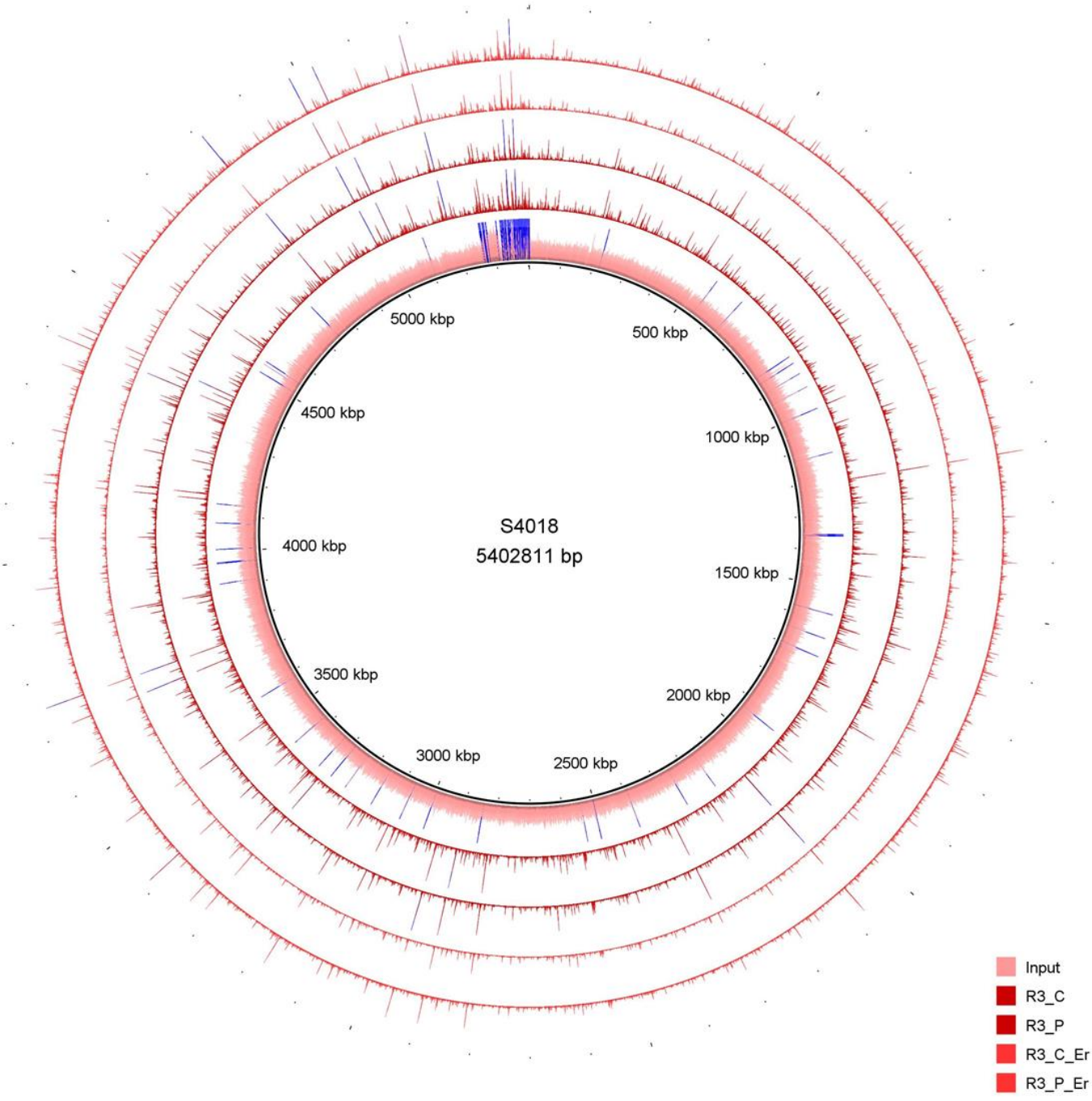


Figure 5.6. Enriched genomic regions of *M. ulcerans* S4018

A) Representative subset of *M. ulcerans* S4018 regions with highest enrichment.

#	size (bp)	Selected region	CDS coordinates	CDS annotation	Exposed	Patients
1	220	42530..42750	42428..42682	A3649_18800, 30S ribosomal protein S18	1088	983
2	200	75700..75900	75704..76018	A3649_00390, ethanolamine permease	1031	723
3	240	197700..197940	195565..197940	A3649_00980, magnesium-transporting ATPase	1566	1154
4	260	285280..285540	284830..285636	A3649_01420, 3-ketoacyl-ACP reductase	1206	907
5	240	430060..430300	428847..430271	A3649_02115, secreted protein	1082	954
6	230	487290..487520	487285..487776	A3649_00393, PE-PGRS protein	1499	1554
7	160	543130..540290	543089..544957	A3649_02665, molecular chaperone DnaK	1110	829
8	310	1207150..1207460	1206809..1208029	A3649_05600, membrane peptidase	1932	1639
9	230	1403770..1404000	1403411..1405936	A3649_06505, ABC transporter/ ATP-binding protein	1310	1218
10	500	1815500..1816000	1815089..1816480	A3649_08205, conserved membrane protein	1174	1427
11	180	1967800..1967980	1967182..1967973	A3649_08895, electron transfer flavorpotein subunit beta	1295	895
12	150	2076350..2076500	2075953..2076783	A3649_09320, peptide ABC transporter permease	3040	3133
13	220	2192000..2192220	2191099..2192304	A3649_09770, uroporphyrinogen-III C-methyltransferase	1198	936
14	110	2304640..2304750	2303960..2305012	A3649_10305, membrane protein of unknown function	1798	1751
15	260	2604960..2605220	2604307..2605353	A3649_11645, potassium transporter Kef	1012	566
16	270	2637080..2637350	2636256..2637776	A3649_11800, PPE protein	552	1299
17	130	2814370..2814500	2814360..2815622	A3649_12530, membrane-associated esterase	2047	1641
18	260	2950340..2950600	2950350..2952122	A3649_13170, fatty-acid-CoA ligase	1954	3143
19	210	3067290..3067500	3067333..3067632	A3649_13780, limonene-1,2-epoxide hydrolase	1082	906
20	150	3088270..3088430	3088367..3088747	A3649_13890, membrane transport protein	1523	1154
21	220	3177530..3177750	3177599..3178501	A3649_14250, ESX secretion-associated protein EspG	1205	1374
22	200	3313000..3313200	3312827..3313444	A3649_14780, glycogen phosphorylase	2510	1123
23	210	3342370..3342580	3342234..3342896	A3649_14930, type I restrcition-modification system subunit	1584	1488
24	115	3415465..3415580	3415073..3416383	A3649_15310, transmembrane protein	1072	927
25	128	3518376..3518504	3518376..3518504	intergenic region 1/ hypothetical protein	1770	1661
26	210	3713280..3713490	3712277..3713449	A3649_16760_magnesium and cobalt transporter	1880	2256
27	235	3891425..3891660	3890697..3892184	A3649_17440, diacylglycerol O-acyltransferase	1532	1650
28	140	4141480..4141620	4140775..4141752	A3649_18585, peptide ABC transporter permease	1516	1241
29	310	4155450..4155760	4155086..4156825	A3649_18640, ABC transporter ATP-binding protein	1795	1241
30	200	4235230..4235430	4234936..4235721	A3649_19025, short-chain dehydrogenase	1121	1032
31	130	4262610..4262740	4262163..4264184	A3649_19160, electron transfer flavoprotein	1066	714
32	180	4349890..4350070	4349100..4350044	A3649_19595, UDP-glucose 4-epimerase	1149	1123
33	190	4359530..4359720	4358860..4360440	A3649_19645, adenylyl cyclase	1023	981
34	220	4382860..4383080	4392564..4393697	A3649_19800, anion transporter	1636	2451
35	141	4422700..4422841	4422473..4422841	A3649_19935, glutamate-cycteine ligase	2169	1168
36	134	4515826..4515960	4515769..4515915	intergenic region 2	1171	1147
37	180	4569800..4569980	4569087..4570895	A3649_20635, isomerase	1719	1517
38	170	4809370..4809540	4807540..4809501	A3649_21710, conserved membrane protein	2491	3229
39	280	4984260..4984530	4982277..4984550	A3649_22490, copper-translocating P-type ATPase	4104	3235
40	170	5027520..5027690	5026500..5027807	A3649_22675, copper-translocating P-type ATPase	3230	2784
41	200	5093260..5093460	5093268..5093684	A3649_22990, conserved membrane protein	1329	1182
42	280	5182470..5182750	5182636..5183328	A3649_23375, immunogenic protein MBP64	4943	3352
43	270	5274080..5274350	5271442..5274612	A3649_23765, type I polyketide modular synthase	1718	1058
44	165	5347485..5347650	5347284..5348351	A3649_24230, PE-PGRS protein	5838	3045

Table 5.4.

B) The selected regions of *M. ulcerans* S4018 divided in functional categories.

Category	Regions	Number	% of the total
Membrane transport	2, 3, 9, 12, 15, 20, 26, 28, 29, 34, 39, 40	12	27.3
Conserved membrane proteins	8, 10, 14, 17, 24, 27, 38, 41	8	18.2
Virulence	6, 16, 19, 21, 43, 44	6	13.6
Cell-wall biosynthesis	4, 18, 30, 32	4	9.1
Hypothetical	25	1	2.3
Other	1, 5, 7, 11, 13, 22, 23, 31, 33, 35, 37, 42	12	27.3

Table 5.5.

These regions encode a copper-translocating P-type ATPase, an immunogenic protein MBP64 and PE-PGRS protein, respectively. Eight of the selected regions had intermediate coverage (12, 18, 22, 26, 34, 35, 38 and 40). All other regions had similar coverage. Surprisingly, 77% (n=34) of the selected regions showed higher coverage for the exposed sample than the patient one. This trend was especially visible for regions 22, 29, 35, 39, 42, 43 and 44. Out of the 10 regions with opposite trend (higher coverage in patients than in controls), the most prominent were regions 18, 26, 34 and 38. These regions encode a fatty-acid-CoA ligase, magnesium-cobalt transporter, an anion transporter and conserved membrane protein of unknown function.

A) Coverage comparison between the selected regions of *M. ulcerans* S4018.

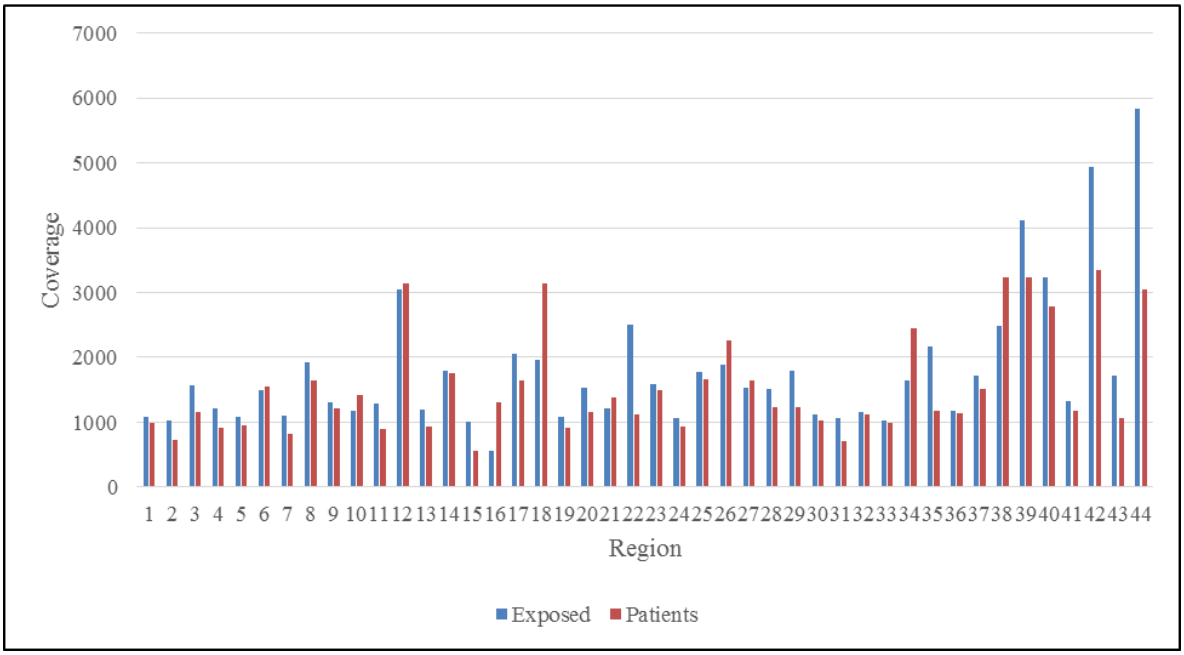


Figure 5.7.

5.4. Discussion.

We have used patient-derived antibodies to identify the immune-relevant proteomes of *S. gallolyticus* NTS31106099 and *M. ulcerans* S4018. Due to the immune-capture nature of the approach, it could be said that successful selection would depend on optimal antigen presentation. Antibodies are known to recognize two types of epitopes, conformational and sequential (Sela et al. 1967). Also, early and more recent studies have proposed that most of the epitopes are sequential (Barlow et al. 1986; Rubinstein et al. 2008). Furthermore, native protein conformation is important for the recognition of conformational epitopes by their cognate antibodies because key interacting amino acids are brought together by proper protein folding (Sela et al. 1967). Finally, protein domains could be considered as their independent folding units. Having all this in mind, we expected that many conformational epitopes might be disrupted if the genomic DNA used for library preparation is sheared into too short fragments. Indeed, truncated polypeptides might have different tertiary structure than the native protein conformation. Therefore, we decided to rationalize our choice for fragment length range of our genomic libraries. Shotgun phage display has been used mainly with fragment of 100-700 bp, but clones containing genomic fragments of up to 1500 bp have also been successfully selected (Jacobsson et al. 2003). ANTIGENome technology has been optimized for shorter fragments of 50 – 500 bp due to better expression of shorter peptides than full-length proteins (Etz et al. 2002). For our approach, we used the predicted proteomes of *S. gallolyticus* and *M. ulcerans* from Pfam database to get an idea about the natural size distribution of the encoded protein domains. We found that the domain size ranged from 20 to 550 amino acids as for both pathogens the domain frequency peaked at 100 amino acids. Those results correlated with the data from earlier reports studying the protein length distribution in pretty much every domain of life (Zhang 2000; Skovgaard et al. 2001; Brocchieri & Karlin 2005). Therefore, we created two libraries with different fragment length per pathogen – 100-300 and 200-

1200 bp and pooled them prior selection. However, analyses of the selection output indicated the same result as observed for *S. aureus* validation experiment, described in chapter 3 – only fragments of about 250-300 bp were selected. Again, we could propose the same explanation as for the *S. aureus*/Fc selection – it is possible that shorter fragments are better selected under our conditions than longer ones.

Intuitively, the use of patient-derived antibodies for identification of pathogen immune-reactive proteins requires, at first place, a measurable humoral response against the particular microbe of interest. Multiple studies have reported such response against *S. gallolyticus* (Darjee & Gibb 1993; Tjalsma et al. 2006; Abdulamir et al. 2009; Boleij et al. 2012; Butt et al. 2015). The unusual association of this bacterium with colorectal cancer has sparked this high interest in serological investigation because it has been hypothesized that increased response to *S. gallolyticus* might be used as a marker for an occult colonic malignancy (Darjee & Gibb 1993; Boleij & Tjalsma 2013). Accordingly, higher levels of serum IgGs have been indeed observed in patients with colorectal cancer (Tjalsma et al. 2006; Abdulamir et al. 2009). Furthermore, a recent multiplex serology assay has been developed consisting of 4 pilus proteins of *S. gallolyticus* (Boleij et al. 2012; Butt et al. 2015). In order to assess the reactivity of our serum sets, we first attempted to perform serum titration using cell-wall extracts of *S. gallolyticus* NTS31106099 according to Abdulamir *et al*. (Abdulamir et al. 2009). However, we were not able to reproduce the study's results – in fact, no difference between the patient and the control sera was observed with very high level of background. Therefore, we abandoned this approach and performed simple whole-cell ELISA which allowed us to detect higher levels of both IgGs and IgAs in almost all collected patient sera. However, it is important to mention that this result reflect neither an absolute *S. gallolyticus*-specific response nor protein-specific immune response. The main disadvantage of this experiment is that we did not treat the cells with sodium periodate to destroy the group D carbohydrate antigen

in order to detect protein-specific antibodies only. Additionally, *S. gallolyticus* shares common proteins with other streptococci of the gut which could lead to cross-reactivity.

Sero-epidemiological studies have also been reported for the other pathogen selected for our experiments – *M. ulcerans*. Elevated antibody response has been detected in patients with Buruli ulcer (Dobos et al. 2000; Gooding et al. 2001; Diaz et al. 2006; Yeboah-Manu et al. 2012; Röltgen et al. 2014), which has led to the suggestion that serology could be used for diagnosis of the disease (Dobos et al. 2000; Okenu et al. 2004). Another observation has suggested that protective immunity against *M. ulcerans* is possible – only some of the exposed people developed Buruli ulcer (Huygen et al. 2009; Bolz et al. 2016). This fact is particularly important for our approach since antigens recognized by healthy individuals only might be also protective from infections (Etz et al. 2002). Therefore, we selected sera from patients with active Buruli ulcer lesions and healthy individuals exposed to the microbe. Unfortunately, using whole-cell ELISA in suspension we were not able obtain meaningful results – very low signal was registered and no difference was observed against controls from European healthy blood donors. Therefore, we proceeded to antibody purification without tittering the collected sera. However, it should be mentioned that a better assay for this purpose is available based on an immunodominant 18-kDa shsp protein, specific for *M. ulcerans* (Diaz et al. 2006).

We chose *S. gallolyticus* and *M. ulcerans* as models to challenge the efficiency of GeXplore under multi-ligand conditions because although both pathogens represent significant social burden, their interaction with the host remains poorly understood. Since antibody response has been reported for both organisms, we decided to use our method in attempt to identify the underlying immune-relevant proteins. As evident from our results, after selection against disease-relevant human sera, for both pathogens we observed enrichment of multiple genomic regions. Importantly, most of these regions were found to encode surface-exposed and/or secreted proteins. These data are in

good agreement with the findings of earlier studies performed with the related ANTIGENome technology (Etz et al. 2002; Weichhart et al. 2003). The main advantage of GeXplore over these reports is the omitting of the demanding transformation step for library preparation and output enrichment. This modification streamlines the application of the method because: i) simplifies its overall performance by eliminating the need for multiple and tedious transformations which have been proposed as a bottleneck in display technologies (Jacobsson et al. 2003; Pluckthun 2012) and ii) allows the *in vitro* expression of larger bacterial genomes (5.4 Mb of *M. ulcerans*) due to the superior size of *in vitro*-expression libraries (up to 10^{12} molecules). For a comparison, transformation-based library of *S. aureus* COL (2.8 Mb) with a fragment length of 250 bp has shown 2-fold genome coverage only (Henics et al. 2003). Considering that only 1 in 18 clones is properly cloned in random genomic libraries (Mullen et al. 2006) it is obvious that this library size is insufficient for genome-wide screening. However, it is important to mention that GeXplore libraries are not perfect. As exemplified in chapter 3, a PCR-driven GC-bias is observed in libraries prepared from G+C poor genomes which is also associated with reduced genome coverage.

We restrain ourselves from discussing the identified proteins in details because their immuno-reactivity has not yet been confirmed. This validation is required because it is not excluded that some regions have been enriched due to unspecific interaction. Additionally, many proteins contain similar domains (*e.g.* the CnaB-domain of the surface-exposed proteins discussed in chapter 2 which is found on multiple mobile elements) which enrichment might be a result of cross-reactivity. Therefore, as demonstrated by Wechhard *et al*, the reactivity of the enriched peptides/proteins with the individual sera used as selective reagents needs to be validated by peptide ELISA.

As a perspective, the identified immune-relevant proteins of the two pathogens could be used in several directions. For example, several hypothetical mechanisms of the etiological association of

S. gallolyticus to colorectal cancer have been recently proposed including: i) carcinogenesis *via* cytokine-dependent inflammation, ii) characteristic adherence potential, iii) altering the profile of the bacterial flora, iv) promotion of early preneoplastic lesions, v)induction of uncontrolled cellular proliferation and vi) colonization of the colorectal mucosa (Abdulamir et al. 2011). Specifically, the bacterium has been found to stimulate the production of inflammatory cytokines in various cell lines including human colonic cancer cells (Biarc et al. 2004; Nguyen et al. 2006). Furthermore, it has been demonstrated that *S. gallolyticus* could trigger the production of angiogenic factors like interleukin (IL)-8, the overexpression of cyclooxygenase-2 and NF-κB (Biarc et al. 2004; Nguyen et al. 2006; Abdulamir et al. 2010; Abdulamir et al. 2009). Importantly, most of these studies have been performed using cell-wall extracted antigens of *S. gallolyticus*. Therefore, once we have validated the immune-reactivity of the identified *S. gallolyticus* proteins, we could test their impact on the mentioned phenotypes.

Regarding the immune-relevant proteins of *M. ulcerans*, once their serum reactivity has been confirmed, they could be used as potential vaccine candidates. Unfortunately, we did not identify proteins which are unique to this bacterium only which could be used for the development of serological diagnostic test.

General conclusions

We have succeeded in the development of a new, completely *in vitro* approach - GeXplore to studying host-pathogen interactions. Main contributions of our work are the elimination of the demanding transformation step required for library preparation and output enrichment in related display technologies, as well as the use of massively-parallel sequencing for output analysis. To our knowledge, this is the first report on characterization of random genomic libraries for *in vitro* selection (ribosome display) by Illumina platform. The development of GeXplore was challenged at several levels. First, we faced the major problem of significant vector recircularization at the library preparation step. While this problem has been solved for transformation-based approaches by vector dephosphorylation, this solution was not appropriate in our case because we aimed at completely *in vitro* library preparation based on PCR amplification. We overcame this complication by developing an alternative cloning strategy based on single-nucleotide G/C-assisted fragment ligation.

Importantly, elimination of the transformation step contributed to the fast performance of GeXplore. Starting from bacterial culture, the method allows performance of three selection rounds and submission of the obtained output for NGS analysis in just about 5 weeks.

Our approach also makes it possible to explore bigger genomes due to the larger size of *in vitro* expression libraries (up to 10^{12} independent members). Such library complexity would require hundreds to thousands transformations for an *in vivo* display system. We also observed that GeXplore seems not so sensitive to background as nearly only natural peptides are selected.

In a nutshell, our method has the potential to be highly versatile, since we have shown its use in different contexts such as “single ligand/gene library” and “single ligand/genomic library” but also

"multiple ligands/genomic library". We also anticipate that our approach would be applicable to non-protein ligands.

However, Gexplore has some limitations. For example, we observed reduced library coverage in GC-poor genomic regions due to the presence of PCR-driven bias associated with the overall G+C content of the studied genome. As discussed in chapter 3, regions with G+C content of 29% or lower are underrepresented in our GeXplore libraries. However, we still do not know if regions which are missing in the NGS output are also missing in the original input library. This point needs to be further investigated.

Unfortunately, the development of GeXplore took most of the time for the PhD thesis presented in this manuscript. Therefore, we could not further validate GeXplore and the selection outputs obtained from *S. gallolyticus* and *M. ulcerans*.

What could be done next? First of all, the proteins identified by GeXplore need to be validated by peptide ELISA in order to prove that serum antibodies specifically recognize them. In the case of *S. gallolyticus*, it will be interesting to further analyze the single protein detected with patient-derived sera only (peptidoglycan hydrolase). Furthermore, the proinflammatory potential of all selected proteins could be investigated since it is well accepted that chronic inflammation might predispose an individual to cancer development and pro-inflammatory properties of *S. gallolyticus* cell-wall extracts have been reported. In the case of *M. ulcerans*,the immune reactivity of the identified proteins also needs to be validated since they might serve as potential vaccine candidates.

To conclude, we think that our work has opened novel opportunities to use ribosome display as a genome-wide display technology. We also hope that the streamlined performance of GeXplore will contribute to the large field of host-pathogen interactions by allowing fast and efficient identification of unknown virulence factors and potential vaccine or diagnostic candidates.

Annexe

genomeA
Journals.ASM.org

Draft Genome Sequence of Erythromycin-Resistant *Streptococcus gallolyticus* subsp. *gallolyticus* NTS 31106099 Isolated from a Patient with Infective Endocarditis and Colorectal Cancer

Stanimir Kambarev,[a] Clément Caté,[a*] Stéphane Corvec,[b,c] Frédéric Pecorari[a]

Recherche en Oncologie Nucléaire, Centre de Recherche en Cancérologie de Nantes-Angers, INSERM UMR 892, CNRS 6299, Université de Nantes, Nantes, France[a]; Service de Bactériologie et Hygiène hospitalière, CHU de Nantes, Nantes, France[b]; Université de Nantes, EA3826 Thérapeutiques cliniques et expérimentales des infections, UFR de Médecine, Nantes, France[c]

* Present address: Clément Caté, Département des Sciences Biologiques and Centre de Recherche BioMed, Université du Québec à Montréal, Montréal, Québec, Canada.

***Streptococcus gallolyticus* subsp. *gallolyticus* is known for its close association with infective endocarditis and colorectal cancer in humans. Here, we report the draft genome sequence of highly erythromycin-resistant strain NTS 31106099 isolated from a patient with infective endocarditis and colorectal cancer.**

Received 13 March 2015 **Accepted** 17 March 2015 **Published** 23 April 2015

Citation Kambarev S, Caté C, Corvec S, Pecorari F. 2015. Draft genome sequence of erythromycin-resistant *Streptococcus gallolyticus* subsp. *gallolyticus* NTS 31106099 isolated from a patient with infective endocarditis and colorectal cancer. Genome Announc 3(2):00370-15. doi:10.1128/genomeA.00370-15.

Address correspondence to Frédéric Pecorari, frederic.pecorari@univ-nantes.fr.

Streptococcus gallolyticus subsp. *gallolyticus* (formerly *Streptococcus bovis* biotype I) is a common gut commensal in various animals and humans. However, the species is known for its ability to cause different diseases in birds and mammals as well as for its close association with infective endocarditis and colorectal cancer in humans (1–3). Despite the extensive research on this relationship, the underlying virulence features and pathomechanisms remain unclear (4, 5). Recommended antibiotic therapy for streptococcal endocarditis is a combination of penicillin and aminoglycoside. Although penicillin-resistant strains have not yet been isolated, resistances to kanamycin, streptomycin, and erythromycin have been reported and attributed to the presence of the genes *aph*(3')-*III*, *ant*(6)-*Ia*, and *ermB*, respectively (6–8). Nevertheless, such resistance determinants were not identified in the available genomes of *S. gallolyticus* subsp. *gallolyticus* (9–13). We report the draft genome of highly erythromycin-resistant *S. gallolyticus* subsp. *gallolyticus* NTS 31106099 isolated from a patient with infective endocarditis and colorectal cancer.

S. gallolyticus subsp. *gallolyticus* NTS 31106099 was grown overnight at 37°C on Columbia agar supplemented with 5% horse blood (Oxoid, United Kingdom) in an atmosphere of 5% CO_2. Genomic DNA extraction was accomplished using a DNeasy blood and tissue kit (Qiagen Gmbh, Germany) according to the manufacturer's recommendation. A sequencing library was prepared using Nextera XT (Illumina, USA) and sequenced using Illumina MiSeq (2 × 300 bp, pair-ends). A total of 10,190, 802 pair-end reads, corresponding to 2.1 Gb was used for *de novo* assembly in SPAdes 2.5.1 (14). Short and low-coverage contigs were filtered out, resulting in a set of 17 contigs between 857 and 583,716 bp with an average coverage of 235×. Annotation was performed by the NCBI Prokaryotic Genome Automatic Annotation Pipeline (PGAAP) (15). Reordering and comparisons were done using Mauve 2.3.1 (16), ACT 8 (17), and BLAST. Acquired antibiotic resistance genes were identified using ResFinder 2.1 (18).

The final assembly has a total length of 2,311,421 bp, an N_{50} of 226 kb, and a G+C content of 37.5%. Annotation revealed 2,198 coding sequences (CDS), 59 tRNAs, 38 pseudo genes, 6 rRNAs, and 1 noncoding RNA. Preliminary comparative analysis uncovered a 44.6-kb strain-specific island (JYKU01000013, UG96_07020-UG96_07300) inserted in a putative RNA methyltransferase gene (Gallo_1429 in UCN34 genome [10]). The element was predicted as a putative Tn*916*-like conjugative transposon and designated Tn*6263*, according to Roberts et al. (19). It contains about 50 CDS (involved in conjugal transfer, regulation, antibiotic resistance [*aph*(3')-*III* (UG96_07105), *ant*(6)-*Ia* (UG96_07115), and *ermB* (UG96_07135)], and virulence. About 33% of Tn*6263* shows 85% identity to CTn*7* of *Clostridium difficile* (20). Interestingly, about 76% of the element is 99% identical to contig 36 of recently released draft genome of vancomycin-resistant *Enterococcus faecium* VRE3 (JSET01000036.1). Future studies will shed light on the functionality and prevalence of Tn*6263*.

The draft genome of *S. gallolyticus* subsp. *gallolyticus* NTS 31106099 will be used for identification of virulence features associated with colorectal cancer and infective endocarditis.

Nucleotide sequence accession numbers. The draft sequence of *S. gallolyticus* subsp. *gallolyticus* NTS 31106099 studied in this project has been deposited at DDBJ/EMBL/GenBank under the accession no. JYKU00000000. The version described in this paper is JYKU01000000.

ACKNOWLEDGMENTS

This work was supported by the ARMINA (Alliance de Recherche sur les Maladies Infectieuses Nantes-Angers) consortium (grant 201209680) of La Région des Pays de la Loire, France.

We are grateful to Bo Segerman for his technical support and expertise.

REFERENCES

1. **Schlegel L, Grimont F, Ageron E, Grimont PA, Bouvet A, Grimont PAD, Bouvet A**. 2003. Reappraisal of the taxonomy of the *Streptococcus bovis*/*Streptococcus equinus* complex and related species: description of *Streptococcus gallolyticus* subsp. *gallolyticus* subsp. nov., *S. gallolyticus* subsp. *macedonicus* subsp. nov. and *S. gallolyticus* subsp. *pasteurianus* subsp. nov. Int J Syst Evol Microbiol **53**:631–645. http://dx.doi.org/10.1099/ijs.0.02361-0.
2. **Shibata Y, Tien LHT, Nomoto R, Osawa R**. 2014. Development of a multilocus sequence typing scheme for *Streptococcus gallolyticus*. Microbiology **160**:113–122. http://dx.doi.org/10.1099/mic.0.071605-0.
3. **Dumke J, Hinse D, Vollmer T, Knabbe C, Dreier J**. 2014. Development and application of a multilocus sequence typing scheme for *Streptococcus gallolyticus* subsp. *gallolyticus*. J Clin Microbiol **52**:2472–2478. http://dx.doi.org/10.1128/JCM.03329-13.
4. **Abdulamir AS, Hafidh RR, Abu Bakar F**. 2011. The association of *Streptococcus bovis/gallolyticus* with colorectal tumors: the nature and the underlying mechanisms of its etiological role. J Exp Clin Cancer Res **30**:11. http://dx.doi.org/10.1186/1756-9966-30-11.
5. **Boleij A, van Gelder MMHJ, Swinkels DW, Tjalsma H**. 2011. Clinical importance of *Streptococcus gallolyticus* infection among colorectal cancer patients: systematic review and meta-analysis. Clin Infect Dis **53**:870–878. http://dx.doi.org/10.1093/cid/cir609.
6. **Teng LJ, Hsueh PR, Ho SW, Luh KT**. 2001. High prevalence of inducible erythromycin resistance among *Streptococcus bovis* isolates in Taiwan. Antimicrob Agents Chemother **45**:3362–3365. http://dx.doi.org/10.1128/AAC.45.12.3362-3365.2001.
7. **Kimpe A, Decostere A, Martel A, Devriese LA, Haesebrouck F**. 2003. Phenotypic and genetic characterization of resistance against macrolides and lincosamides in *Streptococcus gallolyticus* strains isolated from pigeons and humans. Microb Drug Resist **9**(Suppl 1):S35–S38. http://dx.doi.org/10.1089/107662903322541874.
8. **Leclercq R, Huet C, Picherot M, Trieu-Cuot P, Poyart C**. 2005. Genetic basis of antibiotic resistance in clinical isolates of *Streptococcus gallolyticus* (*Streptococcus bovis*). Antimicrob Agents Chemother **49**:1646–1648. http://dx.doi.org/10.1128/AAC.49.4.1646-1648.2005.
9. **Sillanpää J, Nallapareddy SR, Qin X, Singh KV, Muzny DM, Kovar CL, Nazareth LV, Gibbs RA, Ferraro MJ, Steckelberg JM, Weinstock GM, Murray BE**. 2009. A collagen-binding adhesion, Acb, and ten other putative MSCRAMM and pilus family proteins of *Streptococcus gallolyticus* subsp. *gallolyticus* (*Streptococcus bovis* group, biotype I). J Bacteriol **191**:6643–6653. http://dx.doi.org/10.1128/JB.00909-09.
10. **Rusniok C, Couvé E, Da Cunha V, El Gana R, Zidane N, Bouchier C, Poyart C, Leclercq R, Trieu-Cuot P, Glaser P**. 2010. Genome sequence of *Streptococcus gallolyticus*: insights into its adaptation to the bovine rumen and its ability to cause endocarditis. J Bacteriol **192**:2266–2276. http://dx.doi.org/10.1128/JB.01659-09.
11. **Hinse D, Vollmer T, Rückert C, Blom J, Kalinowski J, Knabbe C, Dreier J**. 2011. Complete genome and comparative analysis of *Streptococcus gallolyticus* subsp. *gallolyticus*, an emerging pathogen of infective endocarditis. BMC Genomics **12**:400. http://dx.doi.org/10.1186/1471-2164-12-400.
12. **Lin IH, Liu TT, Teng YT, Wu HL, Liu YM, Wu KM, Chang CH, Hsu MT**. 2011. Sequencing and comparative genome analysis of two pathogenic *Streptococcus gallolyticus* subspecies: genome plasticity, adaptation and virulence. PLoS One **6**:e20519. http://dx.doi.org/10.1371/ journal-.pone.0020519.
13. **Romero-Hernández B, Tedim AP, Sánchez-Herrero JF, Librado P, Rozas J, Muñoz G, Baquero F, Cantón R, Del CR**. 2015. *Streptococcus gallolyticus* subsp. *gallolyticus* from human and animal origins: genetic diversity, antimicrobial susceptibility, and characterization of a vancomycin-resistant calf isolate carrying a *vanA*-Tn*1546*-like element. Antimicrob Agents Chemother **59**:2006–2015. http://dx.doi.org/10.1128/AAC.04083-14.
14. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA**. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol **19**:455–477. http://dx.doi.org/10.1089/cmb.2012.0021.
15. **Angiuoli SV, Gussman A, Klimke W, Cochrane G, Field D, Garrity G, Kodira CD, Kyrpides N, Madupu R, Markowitz V, Tatusova T, Thomson N, White O**. 2008. Toward an online repository of standard operating procedures (SOPs) for (meta)genomic annotation. Omics **12**:137–141. http://dx.doi.org/10.1089/omi.2008.0017.
16. **Darling AE, Mau B, Perna NT**. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One **5**:e11147. http://dx.doi.org/10.1371/journal.pone.0011147.
17. **Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J**. 2005. ACT: the Artemis comparison tool. Bioinformatics **21**:3422–3423. http://dx.doi.org/10.1093/bioinformatics/bti553.
18. **Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV**. 2012. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother **67**:2640–2644. http://dx.doi.org/10.1093/jac/dks261.
19. **Roberts AP, Chandler M, Courvalin P, Guédon G, Mullany P, Pembroke T, Rood JI, Smith CJ, Summers AO, Tsuda M, Berg DE**. 2008. Revised nomenclature for transposable genetic elements. Plasmid **60**:167–173. http://dx.doi.org/10.1016/j.plasmid.2008.08.001.
20. **Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, Thomson NR, Roberts AP, Cerdeño-Tárraga AM, Wang H, Holden MTG, Wright A, Churcher C, Quail MA, Baker S, Bason N, Brooks K, Chillingworth T, Cronin A, Davis P, Dowd L, Fraser A, Feltwell T, Hance Z, Holroyd S, Jagels K, Moule S, Mungall K, Price C, Rabbinowitsch E, Sharp S, Simmonds M, Stevens K, Unwin L, Whithead S, Dupuy B, Dougan G, Barrell B, Parkhill J**. 2006. The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. Nat Genet **38**:779–786. http://dx.doi.org/10.1038/ng1830.

References

Aarestrup, F.M. et al., 2012. Integrating Genome-based Informatics to Modernize Global Disease Monitoring, Information Sharing, and Response. *Emerging Infectious Disease journal*, 18(11). Available at: http://wwwnc.cdc.gov/eid/article/18/11/12-0453.

Abbott, J.C. et al., 2005. WebACT - An online companion for the Artemis Comparison Tool. *Bioinformatics*, 21(18), pp.3665–3666.

Abdulamir, A.S. et al., 2009. Investigation into the controversial association of Streptococcus gallolyticus with colorectal cancer and adenoma. *BMC cancer*, 9, p.403.

Abdulamir, A.S., Hafidh, R.R. & Abu Bakar, F., 2011. The association of Streptococcus bovis/gallolyticus with colorectal tumors: the nature and the underlying mechanisms of its etiological role. *Journal of experimental & clinical cancer research : CR*, 30(1), p.11. Available at: http://www.jeccr.com/content/30/1/11.

Abdulamir, A.S., Hafidh, R.R. & Bakar, F.A., 2010. Molecular detection, quantification, and isolation of Streptococcus gallolyticus bacteria colonizing colorectal tumors: inflammation-driven potential of carcinogenesis via IL-1, COX-2, and IL-8. *Molecular cancer*, 9, p.249. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2946291&tool=pmcentrez&rendertype=abstract.

El Adhami, W. & Stewart, P.R., 1997. Genome organisation of Staphylococcus aureus isolates from different populations. *J.Med.Microbiol.*, 46(4), pp.297–306.

Aird, D. et al., 2010. Analyzing and minimizing bias in Illumina sequencing libraries. *Genome Biology*, 11(Suppl 1), p.P3. Available at: http://genomebiology.com/2011/12/2/R18.

Alikhan, N.-F. et al., 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC genomics*, 12(1), p.402. Available at: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-12-402.

Almeida, R.A. et al., 2000. M-Like Proteins of Streptococcus Uberis. *Infection and Immunity*, 68(1), pp.294–302.

Altschul, S. et al., Basic Local Alignment Search Tool.

Angiuoli, S. V et al., 2008. Toward an online repository of Standard Operating Procedures (SOPs) for (meta)genomic annotation. *Omics : a journal of integrative biology*, 12(2), pp.137–141.

Anthonisen, I. et al., 2002. Organization of the Antiseptic Resistance Gene qacA and Tn 552 - Related β -Lactamase Genes in Multidrug- Resistant Staphylococcus haemolyticus Strains of Animal and Human Origins. *Antimicrob. Agents Chemother*, 46(11), pp.3606–3612.

Atkins, K.L. et al., 2008. S. aureus IgG-binding proteins SpA and Sbi: Host specificity and mechanisms of immune complex formation. *Molecular Immunology*, 45, pp.1600–1611.

Barbosa, E.G. V. et al., 2014. Value of a newly sequenced bacterial genome. *World J Biol Chem*, 5(2), pp.161–168. Available at: http://www.wjgnet.com/1949-8454/full/v5/i2/161.htm.

Barlow, D.J., Edwards, M.S. & Thornton, J.M., 1986. Continuous and discontinuous protein antigenic determinants. *Nature*, 322(21), pp.747–748.

Bartels, M.D. et al., 2014. Comparing whole-genome sequencing with sanger sequencing for spa typing of methicillin-resistant staphylococcus aureus. *Journal of Clinical Microbiology*, 52(12), pp.4305–4308.

Behar, G. et al., 2013. Tolerance of the archaeal Sac7d scaffold protein to alternative library designs: Characterization of anti-immunoglobulin G Affitins. *Protein Engineering, Design and Selection*, 26(4), pp.267–275.

Berggård, T., Linse, S. & James, P., 2007. Methods for the detection and analysis of protein-protein interactions. *Proteomics*, 7(16), pp.2833–2842.

Biarc, J. et al., 2004. Carcinogenic properties of proteins with pro-inflammatory activity from Streptococcus infantarius (formerly S.bovis). *Carcinogenesis*, 25(8), pp.1477–1484.

Bjerketorp, J. et al., 2002. A novel von Willebrand factor binding protein expressed by Staphylococcus aureus. *Microbiology*, 178, pp.2037–2044.

Boleij, A. et al., 2011. Novel clues on the specific association of Streptococcus gallolyticus subsp gallolyticus with colorectal cancer. *Journal of Infectious Diseases*, 203, pp.1101–1109.

Boleij, A. et al., 2012. Selective antibody response to Streptococcus gallolyticus pilus proteins in colorectal cancer patients. *Cancer Prevention …*, 5(2), pp.260–265.

Boleij, A. & Tjalsma, H., 2013. The itinerary of Streptococcus gallolyticus infection in patients with colonic malignant disease. *The Lancet Infectious Diseases*, 13(8), pp.719–724. Available at: http://dx.doi.org/10.1016/S1473-3099(13)70107-5.

Bolz, M. et al., 2016. Vaccination with the Surface Proteins MUL_2232 and MUL_3720 of Mycobacterium ulcerans Induces Antibodies but Fails to Provide Protection against Buruli Ulcer. *PLoS Neglected Tropical Diseases*, 10(2), pp.1–18.

Bratschi, M.W. et al., 2013. Geographic Distribution, Age Pattern and Sites of Lesions in a Cohort of Buruli Ulcer Patients from the Mapé Basin of Cameroon. *PLoS Neglected Tropical Diseases*, 7(6).

Brocchieri, L. & Karlin, S., 2005. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research*, 33(10), pp.3390–3400.

Brouwer, M.S.M. et al., 2011. Genetic organisation, mobility and predicted functions of genes on integrated, mobile genetic elements in sequenced strains of clostridium difficile. *PLoS ONE*, 6(8).

Brown, S.P., Cornforth, D.M. & Mideo, N., 2012. Evolution of virulence in opportunistic pathogens: Generalism, plasticity, and control. *Trends in Microbiology*, 20(7), pp.336–342. Available at: http://dx.doi.org/10.1016/j.tim.2012.04.005.

Buchanan, A., 2012. Evolution of protein stability using ribosome display. *Methods in molecular biology (Clifton, N.J.)*, 805, pp.191–212.

Burman, J.D. et al., 2008. Interaction of human complement with Sbi, a staphylococcal immunoglobulin-binding protein: Indications of a novel mechanism of complement evasion by Staphylococcus aureus. *Journal of Biological Chemistry*, 283(25), pp.17579–17593.

Butt, J. et al., 2015. Association of *Streptococcus gallolyticus* subspecies *gallolyticus* with colorectal cancer: Serological evidence. *International Journal of Cancer*, 0(June), p.n/a-n/a. Available at: http://doi.wiley.com/10.1002/ijc.29914.

Charbit, A. et al., 1988. Versatility of a vector for expressing foreign polypeptides at the surface of Gram-negative bacteria. *Gene*, 70(1), pp.181–189.

Chen, Y.-C. et al., 2013. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PloS one*, 8(4), p.e62856. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3639258&tool=pmcentrez&rendertype=abstract.

Ching, A.T.C. et al., 2012. Lepstospira interrogans shotgun phage display identified LigB as a heparin-binding protein. *Biochemical and Biophysical Research Communications*, 427(4), pp.774–779. Available at: http://dx.doi.org/10.1016/j.bbrc.2012.09.137.

Ciric, L. et al., 2000. The Tn 916 / Tn 1545 Family of Conjugative Transposons The Tn 916 Family of Conjugative Transposons ; An Ever Expanding Family of MGEs The Functions of the Transposon Encoded Proteins by Module. , pp.1–21.

Clark, J.M., 1988. Volume16Number 20n1988. , 16(9), pp.9677–9686.

Clark, J.M., Joyce, C.M. & Beardsley, G.P., 1987. Novel blunt-end addition reactions catalyzed by DNA polymerase I of Escherichia coli. *Journal of molecular biology*, 198, pp.123–127.

Cochetti, I. et al., 2007. New Tn916-related elements causing erm(B)-mediated erythromycin resistance in tetracycline-susceptible pneumococci. *Journal of Antimicrobial Chemotherapy*, 60(May), pp.127–131.

Correa, A. et al., 2014. Potent and specific inhibition of glycosidases by small artificial binding proteins (affitins). *PloS one*, 9(5), p.e97438.

Coulton, J., Mason, P. & DuBow, M., 1986. Molecular cloning of the fdhF gene of Escherichia coli K-12. *Journal of Bacteriology*, 156(3), pp.1315–1321. Available at: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1574-6968.1986.tb01430.x.

Crameri, R. & Suter, M., 1993. Display of biologically active proteins on the surface of filamentous phages: A cDNA cloning system for selection of functional gene products linked to the genetic information responsible for their production. *Gene*, 137(1), pp.77–83.

Dabney, J. & Meyer, M., 2012. Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *BioTechniques*, 52(2).

Darjee, R. & Gibb, a P., 1993. Serological investigation into the association between Streptococcus bovis and colonic cancer. *Journal of Clinical Pathology* , 46(12), pp.1116–1119. Available at: http://jcp.bmj.com/content/46/12/1116.abstract.

Darling, A.C.E. et al., 2004. Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements. , pp.1394–1403.

Deb, D.K. et al., 2002. Selective identification of new therapeutic targets of Mycobacterium tuberculosis by IVIAT approach. *Tuberculosis (Edinburgh, Scotland)*, 82(4–5), pp.175–182.

Denapaite, D. et al., 2015. Highly Variable Streptococcus oralis Strains Are Common Among Viridans Streptococci Isolated from Primates. *Ecological and Evolutionary Science*, 1(2), pp.1–23.

Diaz, D. et al., 2006. Use of the immunodominant 18-kilodalton small heat shock protein as a serological marker for exposure to Mycobacterium ulcerans. *Clinical and Vaccine Immunology*, 13(12), pp.1314–1321.

Diehl, F. et al., 2006. BEAMing: single-molecule PCR on microparticles in water-in-oil emulsions. *Nature methods*, 3(7), pp.551–559.

Van Dijk, E.L., Auger, H., et al., 2014. Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9).

Van Dijk, E.L., Jaszczyszyn, Y. & Thermes, C., 2014. Library preparation methods for next-generation sequencing: Tone down the bias. *Experimental Cell Research*, 322(1), pp.12–20. Available at: http://dx.doi.org/10.1016/j.yexcr.2014.01.008.

Dobos, K.M. et al., 2000. Serologic response to culture filtrate antigens of Mycobacterium ulcerans during Buruli ulcer disease. *Emerging Infectious Diseases*, 6(2), pp.158–164.

Doig, K.D. et al., 2012. On the origin of Mycobacterium ulcerans, the causative agent of Buruli ulcer. *BMC Genomics*, 13, p.258.

Douthwaite, J. & Jackson, R., 2012. *Ribosome display and related technologies: Methods and Protocols*, Available at: http://books.google.com/books?id=Ku2wPAAACAAJ.

Dumke, J. et al., 2014. Development and application of a multilocus sequence typing scheme for Streptococcus gallolyticus subsp. gallolyticus. *Journal of Clinical Microbiology*, 52, pp.2472–2478.

Edwards, D.J. & Holt, K.E., 2013. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microbial informatics and experimentation*, 3(i), p.2.

Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3630013&tool=pmcentrez&rendertype=abstract.

Ehrt, S. et al., 1997. A novel antioxidant gene from Mycobacterium tuberculosis [published erratum appears in J Exp Med 1998 Jan 5;187(1):141]. *J Exp Med*, 186(11), pp.1885–1896. Available at: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.jem.org/cgi/content/full/186/11/1885\nhttp://jem.rupress.org/cgi/reprint/186/11/1885.pdf.

El-Adhami, W., 1999. Expression of a clone specific DNA sequence from Staphylococcus aureus in Escherichia coli. *Journal of Biotechnology*, 73(2–3), pp.181–184.

Enright, M. & Day, N., 2000. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones ofStaphylococcus aureus. *Journal of clinical …*, 38(3), pp.1008–1015.

Etz, H. et al., 2002. Identification of in vivo expressed vaccine candidate antigens from Staphylococcus aureus. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10), pp.6573–6578.

Ewing, B. et al., 1998. Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment. *Genome research*, pp.175–185. Available at: http://genome.cshlp.org/content/8/3/175.short.

Finn, R.D. et al., 2014. Pfam: The protein families database. *Nucleic Acids Research*, 42(November 2013), pp.222–230.

Fleischmann, R.D. et al., 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (New York, N.Y.)*, 269(July), pp.496–512.

Fraser, C.M. et al., 1995. The Minimal Gene Complement of Mycoplasma genitalium. *Science*,

270(5235), pp.397–403. Available at: http://www.sciencemag.org/content/270/5235/397.abstract?ijkey=909d9aa79b0ea0694460ad8fe822df607df6fdff&keytype2=tf_ipsecsha.

Freudl, R., 1989. Insertion of peptides into cell-surface-exposed areas of the Escherichia coli OmpA protein does not interfere with export and membrane assembly. *Gene*, 82(2), pp.229–236.

Fritzer, A. et al., 2010. Novel conserved group A streptococcal proteins identified by the antigenome technology as vaccine candidates for a non-M protein-based vaccine. *Infection and Immunity*, 78, pp.4051–4067.

Galan, A. et al., 2016. Library- based display technologies: where do we stand? *Mol. BioSyst.* Available at: http://pubs.rsc.org/en/Content/ArticleLanding/2016/MB/C6MB00219F.

George, K.M. et al., 1999. Mycolactone: a polyketide toxin from Mycobacterium ulcerans required for virulence. *Science (New York, N.Y.)*, 283(February), pp.854–857.

Giefing, C. et al., 2008. Discovery of a novel class of highly conserved vaccine antigens using genomic scale antigenic fingerprinting of pneumococcus with human antibodies. *The Journal of experimental medicine*, 205(1), pp.117–131.

Gooding, T.M. et al., 2001. Immune Response to Infection with Mycobacterium ulcerans Immune Response to Infection with Mycobacterium ulcerans. *Society*, 69(3), pp.3–7.

Graham, J.E. & Clark-Curtiss, J.E., 1999. Identification of Mycobacterium tuberculosis RNAs synthesized in response to phagocytosis by human macrophages by selective capture of transcribed sequences (SCOTS). *Proceedings of the National Academy of Sciences of the United States of America*, 96(20), pp.11554–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=18072&tool=pmcentrez&rendertype=abstract.

Griffiths, A.D. & Tawfik, D.S., 2006. Miniaturising the laboratory in emulsion droplets. *Trends in Biotechnology*, 24(9), pp.395–402.

Handfield, M. et al., 2000. IVIAT: A novel method to identify microbial genes expressed specifically during human infections. *Trends in Microbiology*, 8(7), pp.336–339.

Hanes, J. & Plückthun, a, 1997. In vitro selection and evolution of functional proteins by using ribosome display. *Proceedings of the National Academy of Sciences of the United States of America*, 94(May), pp.4937–4942.

Haupt, K. et al., 2008. The Staphylococcus aureus protein Sbi acts as a complement inhibitor and forms a tripartite complex with host complement factor H and C3b. *PLoS Pathogens*, 4(12).

Heller, K. & Kadner, R.J., 1985. Nucleotide sequence of the gene for the vitamin B12 receptor protein in the outer membrane of Escherichia coli. *Journal of Bacteriology*, 161(3), pp.904–908.

Hempel, K. et al., 2011. Quantitative proteomic view on secreted, cell surface-associated, and cytoplasmic proteins of the methicillin-resistant human pathogen staphylococcus aureus under iron-limited conditions. *Journal of Proteome Research*, 10(3), pp.1657–1666.

Henics, T. et al., 2003. Small-fragment genomic libraries for the display of putative epitopes from clinically significant pathogens. *BioTechniques*, 35(1), pp.196–209.

Hildegard, E. et al., 2001. Bacterial phage receptors, versatile tools for display of polypeptides on the cell surface. *Journal of Bacteriology*, 183(23), pp.6924–6935.

Hinse, D. et al., 2011. Complete genome and comparative analysis of Streptococcus gallolyticus subsp. gallolyticus, an emerging pathogen of infective endocarditis. *BMC genomics*, 12(1), p.400. Available at: http://www.biomedcentral.com/1471-2164/12/400.

Holmberg, M. et al., 2012. Increased adsorption of histidine-tagged proteins onto tissue culture polystyrene. *Colloids and Surfaces B: Biointerfaces*, 92, pp.286–292. Available at:

http://dx.doi.org/10.1016/j.colsurfb.2011.12.001.

Holtfreter, S., Kolata, J. & Br??ker, B.M., 2010. Towards the immune proteome of Staphylococcus aureus - The anti-S. aureus antibody response. *International Journal of Medical Microbiology*, 300(2–3), pp.176–192.

Horvatić, A. et al., 2016. High-throughput proteomics and fight against pathogens. *Mol. BioSyst.* Available at: http://pubs.rsc.org/en/Content/ArticleLanding/2016/MB/C6MB00223D.

Hutchison, C.A., 2007. DNA sequencing: Bench to bedside and beyond. *Nucleic Acids Research*, 35(18), pp.6227–6237.

Huygen, K. et al., 2009. Buruli ulcer disease: Prospects for a vaccine. *Medical Microbiology and Immunology*, 198, pp.69–77.

Jacobsson, K. et al., 2003. Shotgun Phage Display - Selection for Bacterial Receptins or other Exported Proteins. *Biological procedures online*, 5(1), pp.123–135.

Jacobsson, K. & Frykberg, L., 1995. Cloning of Ligand-Binding Domains of Bacterial Receptors by Phage Display. *BioTechniques*, 18(5), pp.878–885.

Jacobsson, K. & Frykberg, L., 1996. Phage display shot-gun cloning of ligand-binding domains of prokaryotic receptors approaches 100% correct clones. *BioTechniques*, 20(6), pp.1070–1081.

Joensen, K.G. et al., 2014. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. *Journal of Clinical Microbiology*, 52(5), pp.1501–1510.

Kaatz, G.W., Seo, S.M. & Ruble, C.A., 1993. Efflux-mediated fluoroquinolone resistance in Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, 37(5), pp.1086–1094.

Kanagawa, T., 2003. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of bioscience and bioengineering*, 96(4), pp.317–323.

Kawasaki, G., 1997. Cell-free synthesis and isolation of novel genes and polypeptides. , US Patent, p.5643768.

Khan, S. et al., 2015. Draft Genome Sequence of Multidrug-Resistant Enterococcus faecium Clinical Isolate VRE3 , with a Sequence Type 16 Pattern and Novel Structural Arrangement of Tn 1546. , 3(4), p.2015.

Kingsford, C., Schatz, M.C. & Pop, M., 2010. Assembly complexity of prokaryotic genomes using short reads. *BMC Bioinformatics*, 11, p.21. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=20064276.

Knierim, E. et al., 2011. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. *PLoS ONE*, 6(11).

Korman, A.J. et al., 1982. cDNA clones for the heavy chain of HLA-DR antigens obtained after immunopurification of polysomes by monoclonal antibody. *Proceedings of the National Academy of Sciences of the United States of America*, 79(6), pp.1844–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=346077&tool=pmcentrez&rendertype=abstract.

Kraus, J.P. & Rosenberg, L.E., 1982. Purification of low-abundance messenger RNAs from rat liver by polysome immunoadsorption. *Proceedings of the National Academy of Sciences of the United States of America*, 79(13), pp.4015–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=346567&tool=pmcentrez&rendertype=abstract.

Kügler, J. et al., 2013. Oligopeptide M13 phage display in pathogen research. *Viruses*, 5(10), pp.2531–2545.

Kuroda, M. et al., 2001. Whole genome sequencing of meticillin-resistant Staphylococcus aureus.

Lancet (London, England), 357(9264), pp.1225–40. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11418146.

Land, M. et al., 2015. Insights from 20 years of bacterial genome sequencing. *Functional & integrative genomics*, 15(2), pp.141–61. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25722247\nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4361730.

Larsen, M. V. et al., 2012. Multilocus sequence typing of total-genome-sequenced bacteria. *Journal of Clinical Microbiology*, 50(4), pp.1355–1361.

Leclercq, R. et al., 2005. Genetic Basis of Antibiotic Resistance in Clinical Isolates of *Streptococcus gallolyticus* (*Streptococcus bovis*). *Antimicrob Agents Chemother*, 49(4), pp.1646–1648.

Lee, I. et al., 2016. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *International Journal of Systematic and Evolutionary Microbiology*, 66(2), pp.1100–1103.

Lei, L., 2012. Identification of candidate vaccine genes using ribosome display. *Methods in molecular biology (Clifton, N.J.)*, 805, pp.299–314.

Lei, L. et al., 2009. Screening of strain-specific Actinobacillus pleuropneumoniae genes using a combination method. *Journal of Microbiological Methods*, 77(2), pp.145–151. Available at: http://dx.doi.org/10.1016/j.mimet.2009.01.015.

Lei, L. et al., 2008. Selection of serotype-specific vaccine candidate genes in Actinobacillus pleuropneumoniae and heterologous immunization with Propionibacterium acnes. *Vaccine*, 26(49), pp.6274–6280.

Lewis, L. & Lloyd, C., 2012. Optimisation of antibody affinity by ribosome display using error-prone or site-directed mutagenesis. *Methods in molecular biology (Clifton, N.J.)*, 805,

pp.139–161.

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), pp.1754–1760.

Lima, S.S. et al., 2013. Adhesin activity of Leptospira interrogans lipoprotein identified by in vivo and in vitro shotgun phage display. *Biochemical and Biophysical Research Communications*, 431(2), pp.342–347. Available at: http://dx.doi.org/10.1016/j.bbrc.2012.12.095.

Lin, I.H. et al., 2011. Sequencing and comparative genome analysis of two pathogenic streptococcus gallolyticus subspecies: Genome plasticity, adaptation and virulence. *PLoS ONE*, 6(5).

Lindmark, H. & Guss, B., 1999. SFS, a novel fibronectin-binding protein from Streptococcus equi, inhibits the binding between fibronectin and collagen. *Infection and Immunity*, 67(5), pp.2383–2388.

Lowe, A.M., Beattie, D.T. & Deresiewicz, R.L., 1998. Identification of novel staphylococcal virulence genes by in vivo expression technology. *Molecular Microbiology*, 27(5), pp.967–976.

Lundberg, U. et al., 2013. Identification and characterization of antigens as vaccine candidates against Klebsiella pneumoniae. *Human Vaccines and Immunotherapeutics*, 9(March), pp.497–505.

Mahan, M., Slauch, J. & Mekalanos, J., 1993. Selection of Bacterial Virulence genes That Are Specifically Induced in Host Tissues. *Science*, 259(5095), pp.686–688.

Mattheakis, L.C., 1996. 195 [1 1] C e l l - F r e e S y n t h e s i s o f P e p t i d e L i b r a r i e s D i s p l a y e d on Polysomes. *Methods in Enzymology*, 267, pp.195–207.

Mattheakis, L.C., Bhatt, R.R. & Dower, W.J., 1994. An in vitro polysome display system for

identifying ligands from very large peptide libraries. *Proceedings of the National Academy of Sciences of the United States of America*, 91(19), pp.9022–9026.

Mavromatis, K. et al., 2012. The Fast Changing Landscape of Sequencing Technologies and Their Impact on Microbial Genome Assemblies and Annotation. *PLoS ONE*, 7(12), pp.1–6.

Mei, J.M. et al., 1997. Identification of Staphylococcus aureus virulence genes in a murine model of bacteraemia using signature-tagged mutagenesis. *Molecular microbiology*, 26, pp.399–407.

Meinke, A. et al., 2005. Antigenome technology: A novel approach for the selection of bacterial vaccine candidate antigens. *Vaccine*, 23, pp.2035–2041.

Meinke, A. et al., 2009. Composition of the ANTIGENome of Helicobacter pylori defined by human serum antibodies. *Vaccine*, 27, pp.3251–3259.

Merritt, R.W. et al., 2010. Ecology and transmission of buruli ulcer disease: A systematic review. *PLoS Neglected Tropical Diseases*, 4(12), pp.1–15.

Metzker, M.L., 2010. Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1), pp.31–46. Available at: http://dx.doi.org/10.1038/nrg2626.

Meyerhans, a, Vartanian, J.P. & Wain-Hobson, S., 1990. DNA recombination during PCR. *Nucleic acids research*, 18(7), pp.1687–1691.

Miller, J.R., Koren, S. & Sutton, G., 2010. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6), pp.315–327. Available at: http://dx.doi.org/10.1016/j.ygeno.2010.03.001.

Mouratou, B. et al., 2007. Remodeling a DNA-binding protein as a specific in vivo inhibitor of bacterial secretin PulD. *Proc Natl Acad Sci USA*, 104(46), pp.17983–17988.

Mouratou, B. et al., 2012. Ribosome display for the selection of Sac7d scaffolds. *Methods in molecular biology (Clifton, N.J.)*, 805, pp.315–331.

Mullen, L.M. et al., 2006. Phage display in the study of infectious diseases. *Trends in Microbiology*, 14(3), pp.141–147.

Muñoz-Provencio, Diego & Monedero, V., 2011. Shotgun phage display of Lactobacillus casei BL23 against collagen and fibronectin. *Journal of Microbiology and Biotechnology*, 21(2), pp.197–203.

Nguyen, I.S. et al., 2006. Streptococcus infantarius and colonic cancer: Identification and purification of cell wall proteins putatively involved in colorectal inflammation and carcinogenesis in rats. *International Congress Series*, 1289, pp.257–261.

Nilsson, M. et al., 1998. A fibrinogen-binding protein of Staphylococcus epidermidis. *Infection and immunity*, 66(6), pp.2666–2673.

Okenu, D.M.N. et al., 2004. Immunoglobulin M antibody responses to Mycobacterium ulcerans allow discrimination between cases of active Buruli ulcer disease and matched family controls in areas where the disease is endemic. *Clinical and diagnostic laboratory immunology*, 11(2), pp.387–391.

van Opijnen, T. & Camilli, A., 2013. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nature reviews. Microbiology*, 11(7), pp.435–42. Available at: http://www.nature.com.ezlibproxy1.ntu.edu.sg/nrmicro/journal/v11/n7/full/nrmicro3033.html#ref22.

Oyola, S.O. et al., 2012. Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes. *BMC Genomics*, 13(1), p.1. Available at: http://www.biomedcentral.com/1471-2164/13/1.

Pachuk, C.J. et al., 2000. Chain reaction cloning: A one-step method for directional ligation of multiple DNA fragments. *Gene*, 243, pp.19–25.

Palva, a et al., 1990. Nucleotide sequence of the tetracycline resistance gene of pBC16 from Bacillus cereus. *Nucleic acids research*, 18(6), p.1635. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=330541&tool=pmcentrez&rendertype=abstract.

Palzkill, T., Huang, W. & Weinstock, G.M., 1998. Mapping protein-ligand interactions using whole genome phage display libraries. *Gene*, 221(1), pp.79–83. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9852952.

Pan, W. et al., 2014. DNA polymerase preference determines PCR priming efficiency. *BMC biotechnology*, 14(1), p.10. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3937175&tool=pmcentrez&rendertype=abstract.

Patti, J.M. et al., 1992. Molecular characterization and expression of a gene encoding a Staphylococcus aureus collagen adhesin. *J Biol Chem*, 267(7), pp.4766–4772. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=1311320.

Pluckthun, A., 2012. Ribosome display: a perspective. *Methods in molecular biology (Clifton, N.J.)*, 805, pp.3–28.

Poljak, A. et al., 2012. Identification and characterization of Borrelia antigens as potential vaccine candidates against Lyme borreliosis. *Vaccine*, 30(29), pp.4398–4406. Available at: http://dx.doi.org/10.1016/j.vaccine.2011.10.073.

Resch, A. et al., 2005. Differential Gene Expression Profiling of Staphylococcus aureus Cultivated under Biofilm and Planktonic Conditions Differential Gene Expression Profiling of Staphylococcus aureus Cultivated under Biofilm and Planktonic Conditions. *Applied and environmental microbiology*, 71(5), pp.2663–76. Available at:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1087559&tool=pmcentrez&rendertype=abstract.

Rhoads, A. & Au, K.F., 2015. PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*, 13(5), pp.278–289. Available at: http://dx.doi.org/10.1016/j.gpb.2015.08.002.

Rice, L.B. et al., 2005. Tn5386, a novel Tn916-like mobile element in Enterococcus faecium D344R that interacts with Tn916 to yield a large genomic deletion. *Journal of Bacteriology*, 187(19), pp.6668–6677.

Rich, R.L. et al., 1998. Domain structure of the *Staphylococcus aureus* collagen adhesin. *Biochemistry*, 37(44), pp.15423–15433. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9799504.

Roberts, A.P. et al., 2008. Revised nomenclature for transposable genetic elements. *Plasmid*, 60(3), pp.167–173. Available at: http://dx.doi.org/10.1016/j.plasmid.2008.08.001.

Roberts, A.P. & Mullany, P., 2009. A modular master on the move: the Tn916 family of mobile genetic elements. *Trends in Microbiology*, 17(May), pp.251–258.

Roberts, A.P. & Mullany, P., 2011. Tn916-like genetic elements: A diverse group of modular mobile elements conferring antibiotic resistance. *FEMS Microbiology Reviews*, 35, pp.856–871.

Röltgen, K. et al., 2014. Late Onset of the Serological Response against the 18 kDa Small Heat Shock Protein of Mycobacterium ulcerans in Children. *PLoS Neglected Tropical Diseases*, 8(5).

Röltgen, K., Stinear, T.P. & Pluschke, G., 2012. The genome, evolution and diversity of Mycobacterium ulcerans. *Infection, Genetics and Evolution*, 12, pp.522–529.

Romero-Hernández, B. et al., 2015. Streptococcus gallolyticus subsp . gallolyticus from Human

and Animal Origins : Genetic Diversity , Antimicrobial Susceptibility , and Characterization of a Vancomycin-Resistant Calf Isolate Carrying a vanA -Tn 1546 -Like Element. *Antimicrob Agents Chemother*, 59(4).

Rosander, A., Guss, B. & Pringle, M., 2011. An IgG-binding protein A homolog in Staphylococcus hyicus. *Veterinary Microbiology*, 149(1–2), pp.273–276.

Rowland, S.J. & Dyke, K.G.H., 1990. Tn552, a novel transposable element from Staphylococcus aureus. *Molecular Microbiology*, 4(6), pp.961–975.

Rubinstein, N.D. et al., 2008. Computational characterization of B-cell epitopes. *Molecular Immunology*, 45, pp.3477–3489.

Rusniok, C. et al., 2010. Genome sequence of streptococcus gallolyticus: Insights into its adaptation to the bovine rumen and its ability to cause endocarditis. *Journal of Bacteriology*, 192, pp.2266–2276.

Russel, M. et al., 2004. Chapter 1 Introduction to phage biology and phage display. , pp.1–26.

Rutherford, K. et al., 2000. Artemis: sequence visualization and annotation. *Bioinformatics (Oxford, England)*, 16(10), pp.944–945.

Sakyi, S.A. et al., 2016. Clinical and Laboratory Diagnosis of Buruli Ulcer Disease: A Systematic Review. *Canadian Journal of Infectious Diseases and Medical Microbiology*, 2016, pp.1–10. Available at: http://www.hindawi.com/journals/cjidmm/2016/5310718/.

Sanger, F., Nicklen, S. & Coulson, a R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp.5463–5467.

Sarker, S.A. et al., 2015. Oral Phage Therapy of Acute Bacterial Diarrhea With Two Coliphage Preparations: A Randomized Trial in Children From Bangladesh. *EBioMedicine*, 4, pp.124–137.

Sato'o, Y. et al., 2013. A novel comprehensive analysis method for Staphylococcus aureus pathogenicity islands. *Microbiology and Immunology*, 57(2), pp.91–99.

Schlegel, L. et al., 2003. Reappraisal of the taxonomy of the Streptococcus bovis / Streptococcus equinus complex and related species : description of Streptococcus gallolyticus Printed in Great Britain. , (October 2002), pp.631–645.

Schlipköter, U.C. diseases: A. and challenges for public health. & Flahault, A., 2010. Communicable diseases: Achievements and challenges for public health. *Public Health Reviews*, 32(1), pp.90–119.

Schütze, T. et al., 2011. A streamlined protocol for emulsion polymerase chain reaction and subsequent purification. *Analytical Biochemistry*, 410, pp.155–157.

Schwarz, F. V, Perreten, V. & Teuber, M., 2001. Sequence of the 50-kb conjugative multiresistance plasmid pRE25 from Enterococcus faecalis RE25. *Plasmid*, 46(3), pp.170–87. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11735367.

Scietti, L. et al., 2016. Exploring host-pathogen interactions through genome wide protein microarray analysis. *Scientific reports*, 6(February), p.27996. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27302108.

Sebaihia, M. et al., 2006. The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. *Nature genetics*, 38(7), pp.779–786.

Sela, M. et al., 1967. Antibodies to Sequential and Conformational Determinants. *Cold Spring Harbor Symposia on Quantitative Biology*, 32(0), pp.537–545. Available at: http://symposium.cshlp.org/content/32/537.

Sen, R., Nayak, L. & De, R.K., 2016. A review on host-pathogen interactions: classification and prediction. *European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology*. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/27470504.

Shendure, J. & Ji, H., 2008. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10), pp.1135–1145. Available at: http://www.nature.com/doifinder/10.1038/nbt1486.

Sillanpää, J. et al., 2009. A collagen-binding adhesin, Acb, and ten other putative MSCRAMM and pilus family proteins of Streptococcus gallolyticus subsp. gallolyticus (Streptococcus bovis group, biotype I). *Journal of Bacteriology*, 191, pp.6643–6653.

Sinha, B. et al., 1999. Fibronectin-binding protein acts as Staphylococcus aureus invasin via fibronectin bridging to integrin alpha5beta1. *Cellular microbiology*, 1(2), pp.101–117.

Skovgaard, M. et al., 2001. On the total number of genes and their length distribution in complete microbial genomes. *Trends in Genetics*, 17(8), pp.425–428.

Smidt, M. et al., 2013. Comprehensive Antigen Screening Identifies Moraxella catarrhalis Proteins That Induce Protection in a Mouse Pulmonary Clearance Model. *PLoS ONE*, 8(5).

Smith, G.P., 1985. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science (New York, N.Y.)*, 228(4705), pp.1315–1317. Available at: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=4001944&retmode=ref&cmd=prlinks.

Steiner, D., Forrer, P. & Pl??ckthun, A., 2008. Efficient Selection of DARPins with Sub-nanomolar Affinities using SRP Phage Display. *Journal of Molecular Biology*, 382(5), pp.1211–1227.

Stinear, T. et al., 1999. Identification and characterization of IS2404 and IS2606: Two distinct repeated sequences for detection of Mycobacterium ulcerans by PCR. *Journal of Clinical Microbiology*, 37(4), pp.1018–1023.

Stinear, T.P. et al., 2007. Reductive evolution and niche adaptation inferred from the genome of

Mycobacterium ulcerans, the causative agent of Buruli ulcer. *Genome Research*, 17, pp.192–200.

Straus, D. & Ausubel, F.M., 1990. Genomic subtraction for cloning DNA corresponding to deletion mutations. *Proceedings of the National Academy of Sciences of the United States of America*, 87(5), pp.1889–93. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=53589&tool=pmcentrez&rendertype=abstract\nhttp://www.ncbi.nlm.nih.gov/pubmed/2408039\nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC53589.

Su, Y. a, He, P. & Clewell, D.B., 1992. Characterization of the tet (M) determinant of Tn916: evidence for regulation by transcription attenuation. *Antimicrobial agents and chemotherapy*, 36(4), pp.769–778.

Sullivan, M.J., Petty, N.K. & Beatson, S.A., 2011. Easyfig: A genome comparison visualizer. *Bioinformatics*, 27(7), pp.1009–1010.

Tatusova, T. et al., 2016. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*, 44(14), p.gkw569. Available at: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkw569.

Tenover, F.C. et al., 1995. Interpreting chromosomal DNA restriction patterns produced by pulsed- field gel electrophoresis: Criteria for bacterial strain typing. *Journal of Clinical Microbiology*, 33(9), pp.2233–2239.

Thompson, J.R., Marcelino, L. a & Polz, M.F., 2002. Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by "reconditioning PCR". *Nucleic acids research*, 30(9), pp.2083–2088.

Tjalsma, H. et al., 2006. Profiling the humoral immune response in colon cancer patients: Diagnostic antigens from Streptococcus bovis. *International Journal of Cancer*, 119(May),

pp.2127–2135.

Tolstorukov, M.Y. et al., 2001. Sequence-dependent B<-->A transition in DNA evaluated with dimeric and trimeric scales. *Biophysical journal*, 81(6), pp.3409–3421. Available at: http://dx.doi.org/10.1016/S0006-3495(01)75973-5.

Trouillet-Assant, S. et al., 2016. Adaptive processes of Staphylococcus aureus isolates during the progression from acute to chronic bone and joint infections in patients. *Cellular Microbiology*, 33.

Tuerk, C. & Gold, L., 1990. Systematic Evolution of Ligands By Exponential Enrichment - Rna Ligands To Bacteriophage-T4 Dna-Polymerase. *Science*, 249(4968), pp.505–510. Available at: <Go to ISI>://WOS:A1990DR81100032.

de Vries, L.E. et al., 2011. The gut as reservoir of antibiotic resistance: Microbial diversity of tetracycline resistance in mother and infant. *PLoS ONE*, 6(6).

Vytvytska, O. et al., 2002. Identification of vaccine candidate antigens of Staphylococcus aureus by serological proteome analysis. *Proteomics*, 2(5), pp.580–590.

Wang, Y., 2002. The Function of OmpA in Escherichia coli. *Biochemical and Biophysical Research Communications*, 292(2), pp.396–401. Available at: http://www.sciencedirect.com/science/article/pii/S0006291X0296657X.

Watson, J.D. & Crick, F.H.C., 1953. Molecular structure of nucleic acids. *Nature*, 171(4356), pp.737–738. Available at: http://www.nature.com/physics/looking-back/crick/\nhttp://www.ncbi.nlm.nih.gov/pubmed/13054692.

Weichhart, T. et al., 2003. Functional Selection of Vaccine Candidate Peptides from. *Infection and Immunity*, 71(8), pp.4633–4641.

Werner, G., Hildebrandt, B. & Witte, W., 2001. Aminoglycoside-streptothricin resistance gene cluster aadE-sat4-aphA-3 disseminated among multiresistant isolates of enterococcus

faecium. *Antimicrobial Agents and Chemotherapy*, 45(11), pp.3267–3269.

Williams, R. et al., 2006. Amplification of complex gene libraries by emulsion PCR. *Nature methods*, 3(7), pp.545–550.

Yandell, M. & Ence, D., 2012. A beginner's guide to eukaryotic genome annotation. *Nature reviews. Genetics*, 13(5), pp.329–42. Available at: http://www.nature.com/doifinder/10.1038/nrg3174\nhttp://www.ncbi.nlm.nih.gov/pubmed/22510764.

Yeboah-Manu, D. et al., 2012. Sero-epidemiology as a tool to screen populations for exposure to Mycobacterium ulcerans. *PLoS Neglected Tropical Diseases*, 6(1).

Zankari, E. et al., 2012. Identification of acquired antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy*, 67(July), pp.2640–2644.

Zerbino, D.R., 2011. Using the Velvet de novo assembler for short-read sequencing technologies. *Current Protocols in Bioinformatics*, pp.1–13.

Zerbino, D.R. & Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5), pp.821–829.

Zhang, J., 2000. Protein-length distributions for the three domains of life. *Trends in Genetics*, 16(3), pp.107–109.

Zhang, J. et al., 2011. The impact of next-generation sequencing on genomics. *Journal of Genetics*, 38(3), pp.95–109.

Zhang, L. et al., 1998. second IgG-binding protein in. *Microbiology (United Kingdom)*, 408(1998), pp.985–991.

Zhou, M.Y. & Gomez-Sanchez, C.E., 2000. Universal TA cloning. *Current issues in molecular biology*, 2, pp.1–7.

UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Stanimir KAMBAREV

GeXplore : développement d'une approche génomique pour l'étude des interactions hôte-pathogène.

GeXplore: development of a genome-wide approach to studying host-pathogen interactions.

Résumé

Les technologies d'études génomiques ont été utilisées avec succès pour l'identification de facteurs de virulence bactériens ou de vaccins potentiels. Malheureusement, la complexité et les limitations intrinsèques relatives à ces approches ont entravé leur large diffusion. Nous présentons le développement d'une technologie d'étude à l'échelle du génome entier, appelée GeXplore. Elle est basée sur le ribosome display et a été conçue pour fonctionner en conditions *in vitro*, afin d'éviter des inconvénients d'approches telles que le phage display. Nous avons utilisé la plate-forme Illumina pour obtenir la carte-génomique de 3 espèces : *S. aureus*, *S. gallolyticus* et *M. ulcerans*. Les génomes obtenus ont été caractérisés et utilisés comme références pour l'analyse des banques et produits de sélection. Nous avons développé et optimisé ensuite une stratégie de clonage alternative qui permet la préparation *in vitro* des banques et leurs amplifications. Nous avons analysé la représentativité des banques à l'aide du NGS et observé un biais pour les séquences pauvres en GC. Puis, nous avons amélioré la spécificité de notre méthode et l'avons validée en utilisant une interaction protéine-ligand bien caractérisée. GeXplore est ainsi capable d'identifier de multiples domaines de ligands au sein de banques génomiques provenant d'organismes pathogènes. Enfin, nous avons tenté d'identifier les immunoprotéomes de *S. gallolyticus* et *M. ulcerans* en utilisant des anticorps provenant de patients. Nous démontrons que notre méthode permet d'identifier de multiples domaines de protéines exposées à la surface présentant des propriétés potentiellement antigéniques / immunogènes.

Mots clés
GeXplore, interaction hôte-pathogène, génome entier, ribosome display

Abstract

Genome-wide display technologies have been successfully used for identification of multiple bacterial virulence factors and potential vaccine candidates. Unfortunately, the relative complexity and intrinsic limitations of such approaches have hampered their wide spread in the scientific community. Here, we present the development of a streamlined genome-wide display technology, which we term GeXplore. It is based on ribosome display and was designed to work under completely *in vitro* conditions, therefore aiming at avoiding several drawbacks of earlier approaches which involve *in vivo* step(s) such as cell surface or phage display. First, we used Illumina platform to obtain draft genomes of *S. aureus*, *S. gallolyticus* and *M.ulcerans* isolates. Obtained drafts were partially characterized and used as reference sequences for selection input and output analysis. Secondly, we developed an alternative GC-assisted cloning strategy which allows for completely *in vitro* preparation and amplification of random genomic libraries. We analysed the coverage of the obtained libraries using next-generation sequencing and report an important source of coverage bias. Thirdly, we improved the specificity our method which was then validated using a well-characterized protein-ligand interaction. We demonstrate that GeXplore is able to identify multiple ligand-biding domains out of pathogen-derived genomic libraries. Finally, we applied GeXplore to identify immune-relevant proteomes of *S. gallolyticus* and *M. ulcerans* using patient-derived antibodies. We demonstrate that our method allows the identification of multiple surface-exposed protein domains with potentially antigenic/immunogenic properties.

Key Words
GeXplore, host-pathogen interaction, whole genome, ribosome display