Creation of Diverse Microbial Libraries for Drug Discovery using MALDI-TOF MS / IDBac

Maria Sofia Ramos da Costa

Thesis for the degree of Philosophiae Doctor

June 2019

UNIVERSITY OF ICELAND
SCHOOL OF HEALTH SCIENCES
FACULTY OF PHARMACEUTICAL SCIENCES

Uppsetning bakteríustofnasafns til lyfjaþróunar með aðstoð MALDI-TOF MS / IDBac tækni

Maria Sofia Ramos da Costa

Ritgerð til doktorsgráðu

Háskóli Íslands
Heilbrigðisvísindasvið
Lyfjafræðideild

Júní 2019



UNIVERSITY OF ICELAND
SCHOOL OF HEALTH SCIENCES
FACULTY OF PHARMACEUTICAL SCIENCES

Thesis for a doctoral degree at the University of Iceland. All right reserved. No part of this publication may be reproduced in any form without the prior permission of the copyright holder.

© Maria Sofia Ramos da Costa 2019

ISBN 978-9935-9445-2-8

Printing by Háskolaprent.

Reykjavik, Iceland 2019

Author's address

Maria Sofia Ramos da Costa
Faculty of Pharmaceutical Sciences
School of Health Sciences
University of Iceland

Supervisors

Professor Sesselja Ómarsdóttir
Faculty of Pharmaceutical Sciences
School of Health Sciences
University of Iceland

Professor Brian T. Murphy
Dept. Medicinal Chemistry & Pharmacognosy
College of Pharmacy
University of Illinois at Chicago

Doctoral Committee

Professor Elín Soffía Ólafsdóttir
Faculty of Pharmaceutical Sciences
School of Health Sciences
University of Iceland

Professor Ólafur Sigmar Andrésson
Faculty of Life and Environmental Sciences
School of Engineering and Natural Sciences
University of Iceland

Professor Karl Kristinsson
Faculty of Medicine
School of Health Sciences
University of Iceland and National
University Hospital of Iceland

Opponents

Professor Thomas Ostenfeld Larsen
Department of Biotechnology and Biomedicine
Technical University of Denmark

Professor Ryan F. Seipke
Faculty of Bilogical Sciences
School of Molecular & Cellular Biology
University of Leeds

Ágrip

Þrátt fyrir að leit eftir nýjum sýklalyfjum úr náttúrunni hafi staðið yfir í nærri heila öld, hafa fáar nýjungar litið dagsins ljóss þegar kemur að söfnun baktería úr umhverfinu og uppsetningu sýnasafna í þeim tilgangi að leita lyfjavirkra efna úr þeim.

Flestar aðferðir og aðferðafræðin sem notuð er við leit að lyfjavirkum náttúruefnum hefur þróast umtalsvert og má þar t.d. nefna efnagreiningar náttúruefna, leitina að nýjum náttúruefnum með hjálp lífupplýsingafræði, greiningar á breytingum og tjáningu á erfðaklösum sem tjá náttúruefni, aðferðir við ákvörðun sameindabygginga auk aðferða sem miða að því að koma í veg fyrir enduruppgötvun þekktra efnasambanda (e. dereplication), ásamt skimunaraðferðum í lífvirkniprófunum.

Þrátt fyrir stórstígar framfarir á þessum sviðum undanfarna áratugi hafa litlar framfarir orðið í því hvernig bakteríusýnasöfn eru byggð upp sem má þó e.t.v. telja eitt mikilvægasta skrefið í þessu ferli.

Sökum einleitni tegunda og efnaflokka í þessum bakteríusöfnum hefur reynst erfitt að uppgötva ný og áður óþekkt sýklahemjandi efnasambönd úr þeim.

Markmið þessa verkefnis var að setja upp bakteríustofnasafn með lágmarksskörun hvað varðar bakteríutegundir og innihaldsefni. Bakteríurnar fengust úr fjölbreyttum sýnum sem safnað var úr náttúrunni og háhraða MALDI-TOF massagreiningaðferð og lífupplýsingatækni (IDBac) var nýtt til að flokka saman bakteríukólóníur samkvæmt tegundaröðun og aðgreina þær frekar eftir því hvaða náttúruefni þær framleiða in situ.

Þar að auki voru 1616 bakteríukóloníur úr 86 sýnum sem eiga rætur sínar að rekja til söfnunarferða um Ísland og úr sjónum í kringum Ísland, einangraðar af ræktunarskálum. Massagreining (MS) tengd lífupplýsingatækni (IDBac) var notuð til að flokka bakteríurnar niður eftir massagreiningarrófum af próteinum með mólþunga 3000-15000 Da. Þær voru síðan flokkaðar frekar niður í undirflokka með því að tengja saman massagreiningarróf af náttúrefnum sem þær framleiða á mólþungabilinu 200-2000 Da við fyrri róf. Að lokum voru þessar upplýsingar nýttar til að útbúa sýnasafn með 301 einangruðum bakteríum sem náðu yfir 54 ættkvíslir. Söfnun þessara niðurstaðna tók 25 tíma og tveir tímar fóru í greiningarvinnu.

Í annarri tilraun sem framkvæmd var á einangruðum bakteríustofnum úr fyrri söfnunarleiðöngrum á Íslandi, tókst að minnka sýnasafnið úr 833 í 233 sem samsvarar 72% fækkun stofna. Með þessu móti er hægt að lágmarka þann fjölda sem fer inn í bakteríustofnasýnasöfn og þar með þann kostnað sem fylgir, hámarka efnafræðilegan fjölbreytileika og lágmarka upphreinsun á efnum og skimun fyrir lífvirkni. Þessi verkferill sýnir að með því að útbúa betri sýnasöfn er mögulegt að bæta ferilinn við uppgötvun lyfjaefna úr bakteríum til muna og stuðla að markvissari útkomu við skimun sýnasafnanna.

Lykilorð:

Náttúruefni, massagreiningar, lífupplýsingafræði, metabólísk greining, IDBac

Abstract

In the course of a nearly century-long global effort to discover new bacterial-derived drugs from the environment, there have been few innovations to the way that researchers have collected samples and subsequently created microbial libraries sourced for therapeutic discovery. Virtually every stage of the natural products (NP) drug discovery process has undergone a renaissance in method innovation: analytical detection of NPs, bioinformatic-driven discovery of NPs, detection, modification and expression of NP biosynthetic gene clusters, structure elucidation and dereplication of NPs and NP biological activity screening. Despite these advances, the philosophy we use to create microbial strain libraries from the environment (arguably the most important step) has seen little change in nearly eight decades of biomedical research. As a result, it is difficult to discover novel antibiotic scaffolds due to the degree of taxonomic and chemical redundancy that exists in these strain libraries, which has led to a divestment in microbial-based natural product drug discovery. To address the need for creating libraries with minimal taxa and NP overlap from environmental samples, we developed a high-throughput matrix assisted laser desorption ionization mass spectrometry (MALDI-TOF MS) based bioinformatics pipeline (IDBac) that allows us to readily group bacterial colonies by putative taxonomic identity and further discriminate them based on *in situ* natural product production. In addition to this, using a collection expedition to Iceland as an example, we purified *all* distinguishable bacterial colonies (1,616 total) off of isolation plates derived from 86 environmental samples and we employed the mass spectrometry (MS) based-IDBac workflow on these isolates to form groups of taxa (based on protein MS fingerprints; 3,000-15,000 Da), and further distinguished taxa subgroups based on their degree of overlap within corresponding natural product spectra (200-2,000 Da). This informed the decision to create a library of 301 isolates spanning 54 genera. This process required only 25 hours of data acquisition and 2 hours of analysis.

In a separate experiment using isolates from previous expeditions to Iceland, we reduced the size of an existing library from 833 to 233 isolates, a 72.0% reduction. This pipeline minimizes library entries and costs associated with library generation, and maximizes chemical space entering into downstream biological screening efforts. This workflow represents a significant advance toward front-end discovery efforts, thereby removing serendipity from library generation.

Keywords:

Natural products, mass spectrometry, bioinformatics, metabolomics, IDBac

Acknowledgements

The work described in this thesis was mainly developed in the Deptartment of Medicinal Chemistry & Pharmacognosy, University of Illiniois at Chicago and partly in Faculty of Pharmaceutical Sciences, School of Health Sciences, University of Iceland. The project was funded by the Icelandic Center for Research (RANNÍS), which I am very grateful.

I would like to express my sincere gratitude to my supervisors, Professor Brian T. Murphy and Professor Sesselja Ómarsdóttir. To Professor Brian Murphy for giving me the opportunity to develop my work in his research group as a PhD student, for all the guidance, encouragement and inexhaustible patience during these years. To Professor Sesselja Ómarsdóttir, for giving me the opportunity to develop this project, encouragement, and support.

I would like to show my appreciation to my dissertion committee, Professor Elin Soffía Ólafsdóttir, Professor Ólafur S. Andrésson and Karl G. Kristinsson for their constructive criticism.

I would like to express gratitude to Dr. Laura Sanchez, in University of Illiniois at Chicago for giving me opportunity to use the autoFLEX Speed MALDI-TOF MS instrument in her lab and for all the mass spectrometry support. Also, an extensible thank you, to all Sanchez lab members, Valentina Petukhova, Jessica Cleary, Katherine Zink and Alanna Condren, for all the patience and assistance.

In 2015 and 2017, I was fortunate enough to venture out into remote areas of Iceland for field collections by SCUBA and aboard the Icelandic Coast Guard vessel Thor. For the opportunity, I would like to thank my supervisors, Brian T. Murphy and Sesselja Ómarsdottir, and the Marine Research Institute. To Erlendur Bogasson and Sævör Dagný Erlendsdóttir thanks for all the SCUBA assistance in their Dive Center (Strýtan) and companionship. To Eydis Einarsdóttir my sincere gratitude for all the organization and support on those expeditions.

I would like to recognize the past and present members in Murphy lab, who have contributed to my progress in the lab: Michael Mullowney, Vanessa Nepomuceno, Casey Murphy, Antonio Hernandez, Dr. Jeong Ho Lee, Erin Conley, Dr. Tuan Anh Tran, Lydia Davis, Taylor Kornfuehrer, Kevin Kim,

Cynthia Grim, Emma Ward, Radhika Dhakal, Linh Nguyen. To Chase Clark, thanks for being a great lab partner and for all the scientific discussions during our commom projects.

I would like to acknowledge the support from the Center of Biomolecular Sciences in UIC, especially members of Orjala, Federle, Mankin, Eustaquio and Burdette labs.

A special thank you to a group of enthusiastic ladies, Camila Crnkovic, Xue Jiang, Maryam Elfeki, Amira Kefi and Karina Lima, for the constant support during my time in Chicago.

To my colleagues in Hagi, for supporting me in the first months in Iceland, Maonian, Xiaxia, Chutimon, Phennapha and Zoltan. To Margarida Costa, for giving me the call in 2015 that changed my life and gave the opportunity to start all this adventure; and also, for all the support when I moved to Iceland. I would also like to thank Sigrún Sigurðardóttir for the amazing administrative support and to Auður Ágústsdóttir, who help me with all the necessary material for the expeditions.

Last but not the least, I would like to express my infinite gratitude to my mother Arminda and my father Adelino. Thank you for instilling me with a strong passion for learning and for doing everything possible to put me on the path to greatness. I am who I am today because of your effort, guidance, and support. Thank you for the best gift you could ever have given me: a wonderful education.

Finally, to Bruno, who has been a constant source of support and encouragement during this period, making me feel comfortable not only when I need it most, but every day. Thank you for being my best friend, and for being a caring and trustworthy partner. I am truly thankful for having you in my life.

Contents

Ágrip ..... **v**
Abstract ..... **vii**
Acknowledgements ..... **ix**
Contents ..... **xi**
List of abbreviations ..... **xiii**
List of figures ..... **xiv**
List of tables ..... **xvi**
List of original papers ..... **xvii**
Declaration of contribution ..... **xviii**
1 Introduction **1**
1.1 Natural products as a source of drug therapies 1
1.2 Microorganisms as a prolific providers of drug leads 1
1.3 Icelandic waters as a new source for drug lead discovery 2
1.4 IDBac: A MALDI-TOF MS method to innovate the front end of microbial natural products drug discovery 3
1.5 Microbial library generation and its associated flaws 6
2 Aims **9**
3 Materials and methods **11**
3.1 General experimental procedures 11
3.1.1 Sample expeditions and processing 11
3.2 Experimental procedures for chapter I 12
3.2.1 Sponge collection and processing 12
3.2.2 MALDI-TOF MS sample preparation 13
3.2.3 MALDI-TOF MS data acquisition 13
3.2.4 16S rRNA gene sequence analysis and data workup 15
3.2.5 Extraction of *Micromonospora* isolates 16
3.2.6 LC-MS/MS analysis of *Micromonospora* extracts 17
3.2.7 MALDI-TOF bioinformatic pipeline 17
3.2.8 Publication code and data availability 18
3.3 Experimental procedures for chapter II 18
3.3.1 Sample expedition and processing 18
3.3.2 MALDI-TOF MS procedure 22
3.3.3 16S rRNA sequencing 22
4 Results and discussion **41**

4.1 Coupling MALDI-TOF mass spectrometry protein and natural products analyses to rapidly discriminate bacterial function 41
4.1.1 MALDI-TOF MS metabolite data from two *Bacillus subtilis* genetic variants 42
4.1.2 MALDI-TOF MS protein and natural products analysis of isolates from a single *Micromonospora* species..... 44
4.1.3 MALDI-TOF MS protein and natural products of unknown bacteria isolated from a freshwater sponge 47
4.2 A new paradigm to maximize taxonomic and natural products space in microbial libraries using MALDI-TOF MS/IDBac 51
4.2.1 Collection and processing of environmental samples 51
4.2.2 Grouping and analysis of bacterial isolates into putative taxa based on MALDI-TOF MS spectra (3,000-15,000 Da)..... 54
4.2.3 Use of Metabolite Association Networks (MANs) to reduce inter- and intra-species natural product redundancy in a bacterial library..... 58
4.2.4 Use of IDBac to reduce redundancy in existing bacterial libraries..... 64
4.2.5 IDBac visualizes relationships based on user defined categories 64
5 Summary and conclusions..... 67
References 71
Paper I..... 81
Paper II..... 89
Paper III..... 105
Appendix 131

List of abbreviations

BGC	biosynthetic gene cluster
BLAST	basic local alignment search tool
BTS	bacterial test standard
CHCA	alpha-cyano-4-hydrocinnamic acid
$CHCl_3$	chloroform
DNA	deoxyribonucleic acid
EEZ	exclusive economic zone
EtBr	ethidium bromide
FTCIR	fourier-transform ion cyclotron resonance
GNPS	global natural products social molecular networking
HCD	higher-energy collisional dissociation
H_2O	water
HPLC	high performance liquid chromatography
LC-MS	liquid chromatography-mass spectrometry
MALDI	matrix-assisted laser desorption/ionization
m/z	mass-to-charge ratio
MeOH	methanol
mzML	standard file format for mass spectrometry data
NP	natural products
NRPS	nonribosomal peptide synthetase
OBUs	operational biosynthetic units
PCR	polymerase chain reaction
PONAR	sediment sampling device
ppm	parts per million
PKS	polyketide synthase
qTOF	quadrupole time of flight
TOF MS	time of flight mass spectrometry
RPLC	reversed-phase liquid chromatography
rRNA	ribosomal ribonucleic acid
SILVA	on-line resource of ribosomal sequences
sp.	species
TFA	trifluoroacetic acid
XCMS	cloud-based informatic platform

List of figures

Figure 1. Sampling locations of the Icelandic samples between 2013-2017. 11

Figure 2. Analysis of MALDI-TOF MS natural products data from two *Bacillus subtilis* genetic variants. 43

Figure 3. Annotation of surfactin analogs in *Bacillus subtilis* 3610 MS spectrum 44

Figure 4. Annotation of plipastatin analogs in *Bacillus subtilis* 3610 MS spectrum**.** 44

Figure 5. Annotation of *Micromonospora* B031 MALDI spectra showing intact acyl-desferrioxamine analogs (Spectrum annotation performed by co-author Chase Clark). 46

Figure 6. IDBac protein and natural products analysis of isolates from a single *Micromonospora* species. 47

Figure 7. IDBac provides rapid taxonomic and natural products fingerprinting of unknown environmental isolates. 48

Figure 8. Neighbor-joining phylogenetic tree of diversity plate 172. 49

Figure 9. Comparison between 16S rRNA gene sequencing and MALDI protein groupings using Euclidean distance and cosine distance. 50

Figure 10. Distribution of bacterial isolates as a function of media type. Samples are listed on the y-axis, and number of bacterial isolates are represented on the x-axis. 52

Figure 11. Collection trip workflow. 53

Figure 12. Dendrograms resulting from hierarchical clustering of MALDI-TOF MS protein spectra in IDBac of: 733 freshwater isolates (A), and 883 marine isolates (B). The dendrograms were “cut” at a threshold of 0.75 (dashed lines) to create groups of similar isolates. 55

Figure 13. Seeding of freshwater and marine datasets with *Micromonospora* sp. strains. 56

Figure 14. Seeding of freshwater and marine datasets with *Bacillus* sp. strains. 57

Figure 15. Four scenarios are possible when comparing environmental bacterial isolates on nutrient agar that are morphologically indistinguishable and share >99% 16S rRNA similarity. 60

Figure 16. (A) Protein MS clustering of 16 bacterial isolates **(B)** MAN generated for the group of selected strains, isolates colored by modularity score. Isolates chosen to be included in the library are colored in bold; transparent strains were discarded..... 63

Figure 17. Association of bacterial isolates and isolation media. 65

List of tables

Table 1. Media types (A1, ISP2, LWA, NZSG, Chitin) used to cultivate bacteria. 12

Table 2. Detailed MALDI-TOF MS acquisition parameters..... 14

Table 3. Collection coordinates and GenBank accession numbers of strains used in this study. 16

Table 4: Information about samples collected on the collection expedition (GPS coordinates, number of samples collected). 19

Table 5. Collection coordinates and GenBank accession numbers of strains used in this study. 23

Table 6. Understudied actinomycete genera represented in the finalized Iceland library..... 61

List of original papers

This thesis is based on the following original publications and manuscripts:

i. Coupling MALDI-TOF mass spectrometry protein and specialized metabolite analyses to rapidly discriminate bacterial function. *Clark, C. M.; ***Costa, M. S.**; Sanchez, L. M., Murphy, B. T. Proceedings of the National Academy of Sciences of the United States of America, 2018, 115, 4981-4986. *Authors share equal contribution.

ii. Using the open-source MALDI TOF MS/IDBac pipeline for analysis of microbial protein and specialized metabolite data. *Clark, C. M.; ***Costa, M. S.**; *Conley, E.; Ward E.; Sanchez, L. M., Murphy, B. T. Journal of Visualized Experiments, 2019 147, e59219. *Authors share equal contribution.

iii. A new paradigm to maximize taxonomic and natural product space in microbial libraries using MALDI-TOF MS/IDBac. ***Costa, M. S.**; *Clark, C. M.; Ómarsdottir, S.; Sanchez, L. M.; Murphy, B. T. Manuscript submitted to Journal of Natural Products. *Authors share equal contribution.

Declaration of contribution

With the guidance and supervision of Drs. Brian T. Murphy and Sesselja Ómarsdóttir, Maria Sofia Ramos da Costa planned the research, carried out the experiments, wrote the papers in close co-operation with all co-authors.

Paper I: Maria Sofia Ramos da Costa and Chase M. Clark contributed equally to the work. Maria Sofia Ramos da Costa, Chase M. Clark, Laura M. Sanchez and Brian T. Murphy designed the research, analyzed the data and wrote the paper.

Paper II: Maria Sofia Ramos da Costa, Chase M. Clark and Erin Conley contributed equally to the work. Maria Sofia Ramos da Costa, Chase M. Clark, Erin Conley, Emma Ward, Laura M. Sanchez and Brian T. Murphy designed the research, analyzed the data and wrote the paper.

Paper III: Maria Sofia Ramos da Costa, Chase M. Clark and Brian T. Murphy collected the samples during the expedition in 2017. Maria Sofia Ramos da Costa and Chase M. Clark contributed equally to the work. Maria Sofia Ramos da Costa, Chase M. Clark, Sesselja Omarsdottir, Laura M. Sanchez and Brian T. Murphy designed the research, analyzed the data and wrote the paper.

1 Introduction

1.1 Natural products as a source of drug therapies

Throughout the centuries, *Nature* has been a fruitful source of various medicinal products. The oldest records of medical treatments using natural products exist from ancient Egypt and Mesopotamia from around 2900 BC and 2600 BC[1,2]. The use of herbs, honey, and even animal feces for antibiotic treatments by ancient civilizations is an example of an important role of natural products in ancient societies[3]. Currently, several examples of these natural products are in use, *e.g*. the antibiotics: penicillin, tetracycline, erythromycin; the antimalarials: quinine and artemisinin; the lipid control agents: lovastatin and its analogs; the immune-suppressants: cyclosporine, rapamycin; and the anticancer drugs: taxol, doxorubicin, etoposide and teniposide[2,4]. These drugs have been approved by the United States Food and Drug Administration (US FDA) and the European Medicine Agency (EMA) for the treatment of human disorders. Overall, greater than one third of all pharmacologic agents are in some way derived from *Nature*[5], while between 60–80% of all antibiotics and anticancer drugs are naturally derived[2,5,6].

1.2 Microorganisms as a prolific providers of drug leads

Microorganisms are able to produce a variety of compounds that can play an important role for drug discovery. These microbes can be very different in terms of morphology, but most of them have the ability to produce natural products. These groups of microorganisms include phyla of bacteria such as Actinobacteria, Cyanobacteria, Firmicutes, and Proteobacteria. Actinomycetes are a class of Gram-positive bacteria that have been a paramount source of bioactive secondary metabolites for over 50 years, supplying humans with greater than half of the microbial antibiotics on the market[4,7]. Actinobacteria are ubiquitous in both terrestrial and marine environments. A wide range of bioactivities have been reported from these organisms such as anti-inflammatory, antimalarial, antiviral or anti-angiogenic[8]. The list of natural compounds produced by marine actinomycetes is immense and can be divided into several main chemical classes. Terpenoids, polyketides and peptides represent the most prominent classes of compounds that have been isolated from those organisms[8,9].

Cyanobacteria are well known for their ubiquity, morphological and physiological properties and for the production of toxic compounds. They are a promising source of new molecules, since they have been shown to produce unique metabolites with therapeutically relevant activities[10,11]. Secondary metabolites from cyanobacteria are a chemically diverse group of molecules, including, but not limited to, peptides, fatty acids, polyketides, alkaloids, porphinoids, and terpenoids. Many of these compounds have been found to be cytotoxic, antimicrobial, anti-inflammatory, enzyme inhibitors, antiviral, and anticancer agents[10,12]. One of the most commom groups of organisms occuring worldwide, also present in the gut microbiota, are Firmicutes. They usually produce the microbiota-derived lantibiotics such as lactocillin, epidermin and cytolysin[13]. The Firmicutes produce a variety of antimicrobial compounds (*e.g*., surfactins, plipastatins, etc)[14]. A diverse assemblage of Gram-negative bacteria belongs to the phylum Proteobacteria, which includes a high number of medically important species, such as the genera *Neisseria*, *Yersinia*, as well as technologically important genera such as *Agrobacterium*, *Rhizobium* and *Escherichia*[15]*.* Also, many species of Proteobacteria are often found in symbiotic associations with corals, sponges, and molluscs. Overall, these findings place microbes in a privileged position as a source for drug discovery[16].

1.3 Icelandic waters as a new source for drug lead discovery

Iceland has a unique geology and geographical position in the North Atlantic Ocean. The latitude and geothermal activity create unusual conditions in the waters surrounding Iceland, providing an environment that occurs in few other places on the planet. Importantly, the pharmacological potential of the chemical constituents in these extensive waters around Iceland is largely unexplored. Iceland is located in the northern part of the North Atlantic Ocean, just south of the Arctic circle and is the largest part of the Mid Atlantic Ridge that rises above sea level. It has an exclusive economic zone (EEZ) of 200 nautical miles (758,000 km^2). Many bays and deep fjords of various shapes and sizes indent the coastline, but the south shore is characterized by sandy beaches[17]. The volcanically active Reykjanes ridge reaches 300-400 nautical miles to the southwest into the North Atlantic Ocean and separates depths of 2000-3000 m on each side[18,19]. It is a natural boundary between different water masses originated from the north and south with temperature ranging from -0.9°C to approximately 12°C[19]. The continental shelf reaches from 20 to 100 km offshore and at the shelf break the depth

drops from a few hundred meters to 1000-1500 m[19,20]. The maximum depth in the Icelandic EEZ is around 3300 m.

Iceland is also an active geothermal area hosting warm-water adapted organisms of interest. The main hydrothermal vents are located on the Reykjanes Ridge (250-350 m)[21,22]. Numerous sites of hot springs that are connected to terrestrial based geothermal system have been found in intertidal areas around Iceland[23,24]. Marine organisms located at the geothermal fields in intertidal areas require warm temperatures for survival and they may have likely been genetically isolated for up to thousands of years[17,25].

One of the unique sites in Icelandic waters is the hydrothermal vent site in Eyjafjordur, a fjord in northern Iceland. These are two clusters of smectite cones in shallow waters (20-65 m depth) that were discovered in 1997 and 2004[24,26]. They are formed by precipitation of SiO_2-rich geothermal water (72-79°C, pH 10) flowing out of vent openings interacting with Mg-rich seawater[24,26]. Available information indicates a high diversity of algae and benthic invertebrates occurring on and sometimes covering the cones, except for the top vent. Microbes isolated from the geothermal fluid have been studied and include 50 thermophilic strains of terrestrial origin, as well as novel species[24]. These shallow water vent sites host quite different organisms from those in deep-waters and have the advantages of being easily accessible by SCUBA diving in contrast to the hydrothermal vent sites found in much deeper waters. Furthermore, Iceland has many fresh water lakes, ponds and rivers such as Kleifarvatn, Nesgja, Lón, Litla-á and Thingvallarvatn.

1.4 IDBac: A MALDI-TOF MS method to innovate the front end of microbial natural products drug discovery

Developing techniques to rapidly identify and discriminate between bacteria is critical toward front end discovery efforts. As described in a later section (1.5), bacterial libraries sourced for drug discovery are plagued with a high degree of taxonomic and natural product redundancy. In the past four decades, sequencing of the 16S ribosomal RNA (rRNA) gene has been instrumental to the identification and classification of bacteria due to its widespread presence in the Kingdom, degree of conservation, and size[27,28]. This and other genetic-based approaches such as pulsed field gel electrophoresis, multilocus sequence typing, and DNA-DNA hybridization have become commonplace[29]. However, several limitations to these

techniques including cost, turnaround time needed for sequencing and/or analysis, narrow windows of “universal” primers, and in some cases low species-level phylogenetic resolution, have limited their application. Most importantly, in the majority of cases these methods are unable to elucidate how microorganisms interact with one another and function *in situ*. To address this shortcoming, we developed a pipeline that allows rapid discrimination of bacteria based on mass spectral signatures of functional natural products production in complement to conserved ribosomal housekeeping proteins. This pipeline is particularly useful for using natural products production to distinguish strains that share >99% 16S rRNA gene sequence identity.

Natural products production represents functional traits in bacteria; this principle is exemplified in studies involving the marine obligate bacterial genus *Salinispora*[30–32]. To date, the suite of molecules described from *Salinispora* has served to complement existing phylogenetic methods to more precisely differentiate closely related species[33]. For example, extensive population-level biosynthetic gene cluster (BGC) diversity was found within three closely related species that shared 99% 16S rRNA gene sequence identity (*S. arenicola, S. pacifica, S. tropica*)[31]. Of 75 sequenced genomes within this group of three species, a surprising 229 distinct polyketide synthase (PKS) and nonribosomal peptide synthetase (NRPS) operational biosynthetic units (OBUs) were predicted, supporting that in some cases the acquisition of biosynthetic machinery to produce natural products is driven by a strain’s surrounding environmental pressures. These findings highlight the limitations of employing phylogenetic approaches based on 16S rRNA genes to infer bacterial function.

Shortly after the development of 16S rRNA gene sequencing, MALDI-TOF MS was implemented as a technique to identify large biomolecules[34,35]. Subsequent innovations in instrumentation led to the ability to obtain better resolved spectra of intact proteins in high throughput, facilitating the rapid and less-costly identification of bacteria based largely on ribosomal protein MS fingerprints[35–38]. Bruker[39] and bioMerieux[40] have successfully applied this technology in the clinical setting, while many others have employed it on relatively small strain groupings (from ten to a few hundred isolates) in a genus and species-specific manner for environmental bacteria, or to the classification of mammalian cells, as recently summarized by several reviews[41–44]. While MALDI MS protein profiling is useful to determine the putative genus and species-level groupings of bacterial isolates,[45,46] it does not provide information on their natural products, which are critical to survival

and adaption to the surrounding environment. Relatively little is known about the relationship between bacterial taxonomy and natural products production in the majority of bacteria isolated from the environment, yet these characteristics are central to researchers who study bacteria in both academic and industrial settings. Thus, a comprehensive analytical pipeline that allows simultaneous analyses of these factors has been a major obstacle to correlating microbial identity with natural products production (*e.g*. bacterial functional traits).

In the first study we present a significant innovation to previously described MS methods that analyze bacteria by utilizing the full capabilities offered by MALDI-TOF mass spectrometers. In addition to linear mode protein analysis, we employ a reflectron mode to analyze natural products, as the combination of both information-rich spectral regions has yet to be applied to existing MALDI-TOF MS analysis pipelines. Silva *et al.* recently provided a comprehensive history detailing the relatively limited use of MALDI MS to analyze natural products[47]. Utilizing reflectron mode provides increased resolving power and mass accuracy that allows us to inventory bacterial natural products production in seconds.

We employ consecutive protein and natural products MALDI-TOF MS analyses on cellular material scraped from single colonies of bacteria grown on agar plates in order to rapidly discriminate intra-species differences in functional chemistry, often on colonies that share indistinguishable colony morphology. To validate our pipeline, we demonstrate the ability to differentiate isolates within closely related groups of *Bacillus*, *Paenibacillus*, and *Micromonospora* species groupings and characterize them based on *in situ* antibiotic, siderophore, and motility factor production. Total acquisition, analysis, and visualization of MALDI-TOF MS data from *both* intact proteins and natural products of up to 384 bacterial colonies grown on agar can be performed in less than four hours and requires minimal MS expertise, compared to that required to operate qTOF or FTCIR mass spectrometers. This provides an alternative to laborious liquid cultivation, metabolite extraction, and chromatographic analyses that are the current standard of practice for studies that focus on natural products production. To our knowledge IDBac is the first attempt to couple MS analyses of protein content and *in situ* natural products production to afford detailed chemical profiles that allow the distinction of closely related bacterial colonies.

1.5 Microbial library generation and its associated flaws

Libraries of microorganisms have been a cornerstone of drug discovery efforts in the past century. Natural products (or their synthetic derivatives) isolated from these libraries have afforded us approximately 174 cancer drugs and greater than half of marketed anti-infective drugs[5,48–50]. In particular, the discovery of penicillin from a fungus in 1928 ushered in an unprecedented global effort to mine the environment for new microbial natural products[51]. This effort was highly successful and was driven by sampling expeditions whose aim was to amass libraries of cultivatable microorganisms from the environment. However, by the end of the "Golden Age" of antibiotic discovery midcentury, the re-isolation of known natural products became (and remains) a major problem. One cause was the high degree of strain duplication in these libraries, which resulted in unwanted natural product redundancy and lower coverage of chemical space[52,53]. The resulting need for larger microbial libraries and diminishing returns on this investment, among other factors, led the pharmaceutical industry to divest from microbial-based natural product drug discovery[7,49,53,54].

The redundancy of taxa and natural products in bacterial libraries is the result of a few circumstances: 1) strain libraries are often created on the basis of visual inspection of colony morphology with little consideration for natural product production[55]; 2) isolation efforts focused heavily on spore forming bacteria such as *Streptomyces* of the phylum Actinobacteria, and *Bacillus* of the phylum Firmicutes, which have a history of producing biologically active natural products. Further, large microbial natural product libraries can cost hundreds of thousands of dollars to create, and up to millions of dollars to maintain[49]. Few innovations have occurred in the process of microbial library creation during nearly eight decades, as serendipity and human biases have significantly limited the degree of available chemical space that enters biological screening efforts. These limitations have hampered therapeutic discovery from microorganisms.

In order to overcome the high rate of compound re-discovery, it is necessary to simultaneously reduce the high number of redundant taxa and the degree of overlapping inter- and intra-species chemical space in a microbial library. As discussed in section 1.4, in this thesis we developed a high-throughput matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS) data acquisition and bioinformatics pipeline – IDBac – that allows researchers to rapidly collect putative taxonomic and natural product information from single colonies of unknown environmental

bacteria[56]. From each colony, we generate two data sets: 1) protein fingerprints (3,000-15,000 Da, primarily ribosomal[57,58]) that are used to assign each isolate into putative genus/species groupings, and 2) natural product fingerprints (200-2,000 Da) of each colony to compare inter- and intra-species differences/similarities in natural product production. A few reviews and more recent studies highlight attempts to prioritize bacterial strains in a library with regard to either taxonomic identity or natural product potential[52,53,55,59–62].

In this second part of the thesis, we provide a template for the generation of a diverse microbial library using MALDI-TOF MS and IDBac. In order to demonstrate its effectiveness, we applied IDBac to samples collected during an expedition to Iceland in 2017. From 1,616 total bacterial isolates, we created a library of 301 strains that span 54 genera and exhibit minimal natural product overlap. This pipeline allowed the creation of a diverse library while requiring very limited resources (toothpick, bacterial colonies, access to a MALDI-TOF MS), and can be heavily tailored toward creating custom collections of individual taxa, targeting the isolation of understudied taxa, and studying patterns of pseudo-phylogeny and natural product production within large groups of microbial isolates. Use of this method forgoes the need for tedious and costly liquid fermentation, extraction, or chromatographic analyses that would be used to generate similar taxonomic and natural product information. Together, IDBac represents a significant innovation towards the creation of microbial libraries for drug discovery.

2 Aims

The main objective of this thesis was to innovate the way researchers generate microbial libraries for drug discovery by implementing an MS technique that provides knowledge of bacterial pseudo-phylogenetic origin and natural products capacity directly from a single colony on an agar plate. More specifically, the aims of this thesis can be described as follows:

- To develop a high-throughput MALDI-TOF MS/IDBac bioinformatics pipeline to rapidly group bacterial isolates based on their pseudo-phylogenetic identity, then define each group further based on natural products production.

- To implement the MALDI-TOF MS/IDBac platform to maximize taxonomic and natural products diversity in a library generated from an expedition to Iceland in May 2017.

3 Materials and methods

3.1 General experimental procedures

3.1.1 Sample expeditions and processing

Sponges, macroalgae, and sediment samples were collected from aquatic environments around Iceland in the time period from 2013-2017 (March 2013, May/June 2014, August/September 2015, May 2017) using sea excursions with the Marine Research Institute and by SCUBA diving (Figure 1). The samples were processed immediately after collection and unprocessed portions of each sample were frozen and preserved.

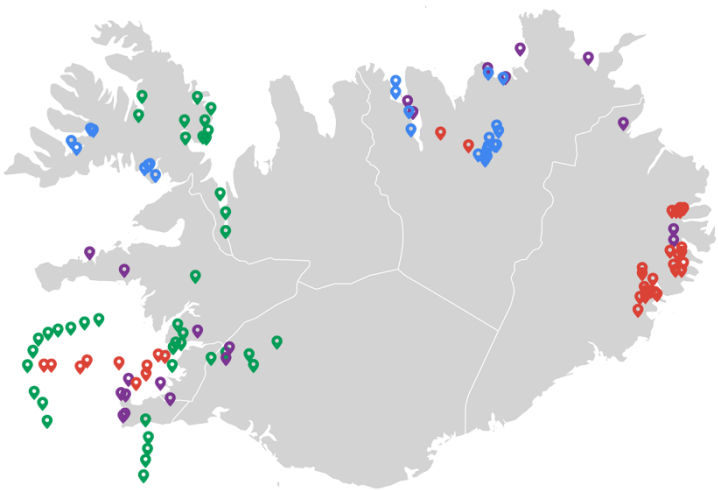


Figure 1. Sampling locations of the Icelandic samples between 2013-2017.

In violet represent the sampling locations from expedition in 2013; green – expedition in 2014; red – expedition in 2015; and in blue, expedition in 2017.

All marine and freshwater samples were submitted to heat treatment in sterile water or synthetic seawater for 8 minutes at 60°C. A few samples were not heat treated and were simply plated on nutrient media. After heat treatment, the samples were vortexed to remove cells directly attached to the sample. A solution of 10 µL of 1:10 dilution was inoculated in five different

types of solid media (each containing 28 µM of the antifungal agent cycloheximide, to prevent the growth of fungi in the sample) (Table 1). All the media plates were sealed with Parafilm and left at 27°C. After 30-60 days, the colonies on the plates were picked and plated in high nutrient medium (A1).

Table 1. Media types (A1, ISP2, LWA, NZSG, Chitin) used to cultivate bacteria.

Media Ingredient / 1 L	LWA	NZSG	A1	ISP2	Chitin
Malt Extract	0 g	0 g	0 g	10 g	0 g
Yeast Extract	0 g	5 g	4 g	4 g	3 g
Soluble Starch	0 g	20 g	10 g	0 g	0 g
Glucose	0 g	10 g	0 g	4 g	0 g
N-Z Amine A	0 g	5 g	0 g	0 g	0 g
Calcium Carbonate	0 g	3 g	0 g	0 g	0 g
Casitone	0 g	0 g	0 g	0 g	5 g
Peptone	0 g	0 g	2 g	0 g	0 g
Tryptone	0 g	0 g	0 g	0 g	0 g
NaCl	0 g	0 g	0 g	0 g	0 g
K_2HPO_4	0 g	0 g	0 g	0 g	700 mg
KH_2PO_4	0 g	0 g	0 g	0 g	300 mg
$MgSO_4.7H_20$	0 g	0 g	0 g	0 g	500 mg
8 mg/mL $FeSO_4.7H_2O$	0 g	0 g	0 g	0 g	625 µL
Chitin	0 g	0 g	0 g	0 g	4 g
Agar	15 g				

3.2 Experimental procedures for chapter I

3.2.1 Sponge collection and processing

A freshwater sponge sample was collected June 6th, 2016 from Marinette, Wisconsin (45º5'16.012"N, 87º35'10.468"W), from pilings near Red Arrow Beach at a depth of 3 m using SCUBA. The sponge was separated from

associated macro-organisms, rinsed with filter sterilized Lake Michigan water five times to remove bacteria from surrounding lake water, and most of the water expelled from the sponge by applying gentle pressure. A 1 cm^3 section of tissue and 10 mL of sterile 20% glycerol solution were ground for two minutes using an autoclaved mortar and pestle. A 60 ºC dry bath was utilized to pretreat a 500 µL aliquot for 9 minutes. The sample was then diluted 1:10 with 20% sterile glycerol solution. To an agar plate containing A1 nutrient media (5 g of soluble starch, 2 g of yeast extract, 1 g of peptone, 250 mL of filter-sterilized Lake Michigan water and 250 mL of distilled water), 50 µL of sample was added and spread across the surface. The plate was sealed with Parafilm and left at 27 ºC for 90 days.

3.2.2 MALDI-TOF MS sample preparation

For MALDI-TOF MS analysis, proteins were extracted using an extended direct transfer method that included a formic acid overlay[63]. Using a sterile toothpick, bacterial colonies that grew on nutrient agar were applied as a thin film onto a MALDI ground-steel target plate (Bruker Daltonics, Billerica, MA). Over each bacterial smear, 1 µL of 70% LC-MS grade formic acid (Optima, Fisher Chemical) was added and allowed to evaporate, followed by the addition and subsequent evaporation of 1 µL of 10 mg/mL α-cyano-4-hydroxycinnamic acid (CHCA; recrystallized from the 98% pure Sigma-Aldrich) solubilized in 50% acetonitrile, 2.5% trifluoroacetic acid and 47.5% water[63,64]. All solvents were HPLC or MS grade.

3.2.3 MALDI-TOF MS data acquisition

Measurements were performed using an Autoflex Speed LRF mass spectrometer (Bruker Daltonics) equipped with a smartbeam™-II laser (355 nm). Detailed instrument settings are available in table 2. Natural products spectra were recorded in positive reflectron mode (5000 shots; RepRate: 2000 Hz; delay: 9297 ns; ion source 1 voltage: 19 kV; ion source 2 voltage:16.55 kV; lens voltage: 8.3 kV; mass range: 50 Da to 2,700 Da, matrix suppression cutoff: 50 Da). Protein spectra were recorded in positive linear mode (1200 shots; RepRate: 1000; delay: 29793 ns; ion source 1 voltage: 19.5 kV; ion source 2 voltage: 18.2 kV; lens voltage: 7.5 kV; mass range: 1.9 kDa to 2.1 kDa, matrix suppression cutoff: 1.5 kDa). Protein spectra were corrected with an external Bruker Daltonics bacterial test standard (BTS). Natural products spectra were corrected with an external Bruker Daltonics peptide calibration standard and CHCA $[2M+H]^+$ (379.0930 Da). Automated data acquisitions were performed using flexControl software

v. 3.4.135.0 (Bruker Daltonics) and flexAnalysis software v. 3.4. Spectra were automatically evaluated during acquisition to determine whether a spectrum was of high enough quality to retain and add to the sum of the sample acquisition. The number of added spectra, quality requirements and other detailed acquisition settings are available in Table 2; for flexControl and flexAnalysis scripts, see the “Publication Code and Data Availability” section.

Table 2. Detailed MALDI-TOF MS acquisition parameters.

Parameter	Protein Spectra	Small Molecule Spectra
Instrument type	autoflex	autoflex
flexControl version	flexControl 3.4.135.0	flexControl 3.4.135.0
Type of digitizer	LeCroy LSA2000	LeCroy LSA2000
Number of shots	1200	5000
Retention time for Warp-LC	0 s	0 s
Digitizer bit depth	8	8
AIDA version number	AIDA4.7.373.7	AIDA4.7.373.7
Target type	0219793	0219793
Spectrum delay	29793 ns	9297 ns
SampleRate reciprocal	1.6 ns	0.2 ns
Spectrum size	42154 pts	253781 pts
Himass turbo mode	false	false
Laser repetition rate	1000 Hz	2000 Hz
Linear detector voltage	2.919 kV	2.919 kV
Reflector detector voltage	1.883 kV	1.883 kV
Voltage of high mass detector	0 kV	0 kV
Realtime smooth	high	off
AutoXecute method	IDBac_Protein_AutoX	IDBac_Small-Molecule_autoX
A flat line is created if the acceptance criteria of AutoXecute are not met	true	true
Flag indicating In Source Decay measurement	false	false

Flag indicating HPC usage	false	false
Calibration mass control list used	MBT_Standard	IDBac_Small_Molecule_Calibration
Sensitivity of digitizer	100 mV/fullscale	100 mV/fullscale
Analog Offset	1.4 mV	2 mV
Deflection pulser cal	0	0
PIE delay	220 ns	120 ns
Positive voltage polarity	POS	POS
Reflector voltage 2	0 kV	9.7 kV
Ion source voltage 1	19.5 kV	19 kV
Ion source voltage 2	18.2 kV	16.55 kV
Lens voltage	7.5 kV	8.3 kV
Reflector voltage 1	0 kV	21 kV
Matrix suppression mode	deflection	deflection
Matrix suppression cut off mass	1500	50

3.2.416S rRNA gene sequence analysis and data workup

Bacterial isolates were characterized by analysis of the 16S rRNA gene. Total genomic DNA was extracted using the Ultra Clean Microbial DNA Isolation kit (MOBIO Laboratories) according to the manufacturer's instructions. The primers 8F, FC27, 1100F, RC1492 and 519R were used for amplifying the 16S rRNA gene. The polymerase chain reaction (PCR) conditions were as follows: initial denaturation at 95°C for 5 minutes, followed by 35 cycles of denaturation at 95°C for 15 seconds, annealing at 60°C for 15 seconds, and extension at 72°C for 15 seconds, and a final extension step at 72°C for 2 minutes. PCR products were purified using QIAquick PCR Purification kit from Qiagen and the amplicons sequenced by Sanger sequencing. Geneious V10.0.9 software was used to assemble a consensus sequence of the 16S rRNA gene amplified for each bacterial isolate and the identity of the isolates determined using BLASTn. All sequences were submitted to GenBank and accession numbers along with information regarding each of the isolates' origin is available in Table 3. Phylogenetic trees were created by trimming aligned sequences (SILVA Incremental Aligner)[65] to equal length and using Geneious' "Tree Builder" with Jukes-Cantor and Neighbor Joining algorithms with 100 bootstrap replicates using a support threshold of 80%.

Table 3. Collection coordinates and GenBank accession numbers of strains used in this study.

Strain ID	Latitude	Longitude	Genus	Accession
B020	43°13'27.63"N	87°34'10.62"W	*Micromonospora*	KY858243
B022	43°13'27.63"N	87°34'10.62"W	*Micromonospora*	KY858245
B021	43°13'27.63"N	87°34'10.62"W	*Micromonospora*	KY858244
B012	43°13'27.63"N	87°34'10.62"W	*Micromonospora*	KY858247
B031	41°45'42.60"N	86°49'28.12"W	*Micromonospora*	KY858239
B029	43°13'27.63"N	87°34'10.62"W	*Micromonospora*	KY858242
B011	43°13'27.63"N	87°34'10.62"W	*Micromonospora*	KY858246
B001	43°13'27.63"N	87°34'10.62"W	*Micromonospora*	KY858238
172-1	45º5'16.012"N	87º35'10.468"W	*Rhodococcus*	KY858237
172-2	45º5'16.012"N	87º35'10.468"W	*Rhodococcus*	KY858236
172-3	45º5'16.012"N	87º35'10.468"W	*Rhodococcus*	KY858235
172-4	45º5'16.012"N	87º35'10.468"W	*Rhodococcus*	KY858234
172-5	45º5'16.012"N	87º35'10.468"W	*Rhodococcus*	KY858233
172-6	45º5'16.012"N	87º35'10.468"W	*Enterococcus*	KY858232
172-7	45º5'16.012"N	87º35'10.468"W	*Paenibacillus*	KY858231
172-8	45º5'16.012"N	87º35'10.468"W	*Rhodococcus*	KY858230
172-9	45º5'16.012"N	87º35'10.468"W	*Bacillus*	KY858229
172-10	45º5'16.012"N	87º35'10.468"W	*Paenibacillus*	KY858228
172-11	45º5'16.012"N	87º35'10.468"W	*Rhodococcus*	KY858227
A001	42°21'57.18"N	70º58'16.74"W	*Streptomyces*	MG188670
B003	43°13'27.63"N	87°34'10.62"W	*Streptomyces*	MG188671
3610	No information	No information	*Bacillus subtilis*	CP020102
B026	43°13'27.63"N	87°34'10.62"W	*Micromonospora*	KP009553
B030	43°13'27.63"N	87°34'10.62"W	*Micromonospora*	KY858241

3.2.5 Extraction of *Micromonospora* isolates

Extractions were performed from bacterial cultures growing on solid agar media (A1) following the protocol of Bligh, E. G. and Dyer, W.J[66]. Agar cultures were divided into 1 cm^3 pieces and 3 mm glass beads were added.

Extraction solvent was added in three steps with vigorous vortexing between steps 1) 1:2 (v/v) $CHCl_3$:MeOH, 2) $CHCl_3$ in 1/3 the added volume of step one, 3) H_2O in the same volume as chloroform was added in step two. From the resulting two-layer liquid partition, the organic (bottom) layer was retained for further analysis.

3.2.6 LC-MS/MS analysis of *Micromonospora* extracts

Micromonospora extracts were analyzed via LC-MS/MS with a method adapted from that described by Goering et al[67]. Experiments were performed on an Agilent 1200 workstation connected to a Thermo Fisher Scientific Q-Exactive mass spectrometer with an electrospray ionization source. Reversed-phase chromatography was performed by injection of 20 µL of 0.1 mg/mL of extract at a 0.3 mL/min flow rate across a Phenomenex Kinetex C18 RPLC column (150 mm x 2.1 mm i.d., 2 µm particle size). Mobile phase A was water with 0.1% formic acid and mobile phase B was acetonitrile with 0.1% formic acid. Mobile phase B was held at 15% for 1 minute, then adjusted to 95% over 12 minutes, where it was held for 2 minutes, and the system re-equilibrated for 5 minutes. The mass spectrometry parameters were as follows: scan range 200-2000 *m/z*, resolution 35,000, scan rate ~3.7 per second. Data were gathered in profile and the top 5 most intense peaks in each full spectrum were targeted for fragmentation that employed a collision energy setting of 25 eV for Higher-energy Collisional Dissociation (HCD) and isolation window of 2.0 m/z. Data were converted to mzXML and uploaded to the GNPS: Global Natural Products Social Molecular Networking platform for dereplication[68] and XCMS[69] for comparative metabolomics. For analysis in R, GNPS library matches were downloaded and converted to mzML, where necessary, with MSConvert. “Amphiphilic ferrioxamine” mzML files were manually edited in Notepad++ to the correct XML schema.

3.2.7 MALDI-TOF bioinformatic pipeline

In brief, MALDI-TOF MS raw data are first converted to the open-source mzXML format[70] using ProteoWizard’s command-line MSConvert[71]. The mzXML files are read into R through mzR[71] and custom code, and processed into peak lists using MALDIquant[72]. Lastly, the peak lists inform interactive data analyses and visualizations that are displayed through a default, offline internet browser with RStudio’s Shiny[73] package. The IDBac software is being distributed under a GNU General Public License, and links to the full code along with data acquisition and analysis tutorials can be found within the “Code and Data Availability” section. Easy to install IDBac software and

updates, method manuals and information on code contribution may be found at chasemc.github.io/IDBac and as a permanent snapshot at https://doi.org/10.17632/ysrtr9c5s7.1.

3.2.8 Publication code and data availability

IDBac software and updates, method manuals and information on code contribution may be found at chasemc.github.io/IDBac by following the respective links. The R code used to generate all figures and analyses for this publication along with the permanent, publication-version of IDBac is available for download at: https://doi.org/10.17632/ysrtr9c5s7.1. Mass spectra were deposited in MASSIVE for: MALDI (ftp://massive.ucsd.edu/MSV000081619), and LC-MS/MS (ftp://massive.ucsd.edu/MSV000081555) data.

3.3 Experimental procedures for chapter II

3.3.1 Sample expedition and processing

Sponges, macroalgae, marine and freshwater sediment samples were collected around Iceland in May 2017 (Table 4 and Figure 1). The samples were processed immediately after collection and unprocessed portions of each sample were frozen and preserved. The samples were processed according to section 3.1.1. After 30-60 days of growth, *all* the colonies on the plates were picked and plated in high nutrient medium (A1).

Table 4: Information about samples collected on the collection expedition (GPS coordinates, number of samples collected).

Location	Latitude	Longitude	Marine/Fresh	Sample Type
114A	66.0278306	-18.4062778	marine	sponge
114B	66.0278306	-18.4062778	marine	algae
114C	66.0278306	-18.4062778	marine	sediment
114D	66.0278306	-18.4062778	marine	algae
114E	66.0278306	-18.4062778	marine	water
115A	65.9843306	-18.3947889	marine	sea urchin eggs
115B	65.9843306	-18.3947889	marine	sediment
115C	65.9843306	-18.3947889	marine	sediment
115D	65.9843306	-18.3947889	marine	sediment
115E	65.9843306	-18.3947889	marine	sediment
116A	66.098072	-16.929879	freshwater	algae
116B	66.098072	-16.929879	freshwater	algae
116C	66.098072	-16.929879	freshwater	algae
116D	66.098072	-16.929879	freshwater	sediment
116E	66.098072	-16.929879	freshwater	sediment
116F	66.098072	-16.929879	freshwater	sediment
116G	66.098072	-16.929879	freshwater	sediment
116H	66.098072	-16.929879	freshwater	sediment
117A	66.076302	-16.688245	freshwater	sediment
117B	66.076302	-16.688245	freshwater	sediment
117C	66.076302	-16.688245	freshwater	algae
117D	66.076302	-16.688245	freshwater	sediment
117E	66.076302	-16.688245	freshwater	sediment
117F	66.076302	-16.688245	freshwater	sediment
117G	66.076302	-16.688245	freshwater	sediment

117H	66.076302	-16.688245	freshwater	sediment
117I	66.076302	-16.688245	freshwater	sediment
117J	66.076302	-16.688245	freshwater	sediment
118A	66.076305	-16.68827	marine	sediment
118B	66.076305	-16.68827	marine	algae
118C	66.076305	-16.68827	marine	sediment
118D	66.076305	-16.68827	marine	algae
119A	65.72176	-16.788817	freshwater	sediment
120A	65.721916	-16.792269	freshwater	sediment
121A	65.722159	-16.792697	freshwater	sediment
122A	65.721539	-16.786873	freshwater	sediment
122B	65.721539	-16.786873	freshwater	sediment
123A	65.701479	-16.771132	freshwater	sediment
124A	65.641517	-16.771132	freshwater	water
125A	65.641573	-16.809047	freshwater	water
126A	65.639585	-16.80899	freshwater	sediment
127A	65.640801	-16.80839	freshwater	sediment
128A	65.640207	-16.844106	freshwater	sediment
129A	65.638311	-16.910157	freshwater	sediment
130A	65.579718	-16.951147	freshwater	sediment
131A	65.561018	-16.947166	freshwater	sediment
132A	65.548551	-16.980315	freshwater	sediment
133A	65.57096	-17.073028	freshwater	sediment
134A	65.850981	-18.193052	marine	sediment
134B	65.850981	-18.193052	marine	sponge
134C	65.850981	-18.193052	marine	sponge
134D	65.850981	-18.193052	marine	sponge
135A	65.73357	-18.170398	marine	sediment

136A	65.444369	-22.206837	marine	sediment
137A	65.484244	-22.391858	marine	sponge
137B	65.484244	-22.391858	marine	sponge
137C	65.484244	-22.391858	marine	sponge
137D	65.484244	-22.391858	marine	jellyfish egg
137E	65.484244	-22.391858	marine	nudibranch
137F	65.484244	-22.391858	marine	tunicate
138A	65.512811	-22.304967	marine	algae
138B	65.512811	-22.304967	marine	algae
138C	65.512811	-22.304967	marine	sponge
138D	65.512811	-22.304967	marine	tunicate
138E	65.512811	-22.304967	marine	tunicate
138F	65.512811	-22.304967	marine	sponge
138G	65.512811	-22.304967	marine	sponge
138H	65.512811	-22.304967	marine	sponge
138I	65.512811	-22.304967	marine	sponge
138J	65.512811	-22.304967	marine	sponge
138K	65.512811	-22.304967	marine	algae
139A	65.508073	-22.330494	marine	sediment
139B	65.508073	-22.330494	marine	sediment
140A	65.738088	-23.24695	marine	sponge
140B	65.738088	-23.24695	marine	tunicate
140C	65.738088	-23.24695	marine	sediment
140D	65.738088	-23.24695	marine	sponge
140E	65.738088	-23.24695	marine	sediment
140F	65.738088	-23.24695	marine	sediment
141A	65.736375	-23.208858	marine	sediment
141B	65.736375	-23.208858	marine	algae

142A	65.622762	-23.467935	marine	sediment
142B	65.622762	-23.467935	marine	sediment
143A	65.647536	-23.542869	marine	nudibranch
143B	65.647536	-23.542869	marine	White eggs from beitukongur
143C	65.647536	-23.542869	marine	tunicate

3.3.2 MALDI-TOF MS procedure

From the plates described in the previous section, 1,616 bacterial isolates were isolated using toothpicks, onto 60-mm petri dishes. Each isolate was then transferred to a 48-well plate containing A1 medium supplemented with or without synthetic ocean salts. After 7 days of growth, each bacterial isolate was applied as a thin smear in three technical replicates onto a 384-spot MALDI target plate (Bruker Daltonics, Billerica, MA) using a sterile toothpick. The detailed experimental settings are described in section 3.2.2.

3.3.3 16S rRNA sequencing

DNA from 361 bacterial isolates was extracted using a DNeasy UltraClean Microbial Kit (Qiagen). The 16S rRNA gene was amplified using 27F (5'-CAGAGTTTGATCCTGGCT-3') and 1492R (5'-AGGAGGTGATCCAGCCGCA-3')[74] primers using polymerase chain reaction (PCR) under the following conditions: initial denaturation at 95 °C for 5 minutes; followed by 35 cycles of denaturation at 95 °C for 15 seconds, annealing at 60 °C for 15 seconds, and extension at 72 °C for 30 seconds; and a final extension step at 72 °C for 2 minutes. PCR products were purified using a QIAquick PCR Purification kit from Qiagen and the amplicons sequenced by Sanger sequencing. Data was analyzed by Geneious V11.1.4 software. Using the SILVA Alignment, Classification and Tree (ACT) Service, we aligned all 361 isolates and created a phylogenetic tree using Fastree and default settings[65]. All the sequences were deposited in the NCBI nucleotide database with the accession number MK143106 to MK143377; MK163385 to MK163438 and MK168020 to MK168054 (Table 5).

Table 5. Collection coordinates and GenBank accession numbers of strains used in this study.

Isolate	Latitude	Longitude	Closest match (Genus/Family)	Accession Number	Media Type	Sample Source
114A-2	66.0278306	-18.4062778	*Friedmanniella* sp.	MK143176	ISP2	sponge
114B-5	66.0278306	-18.4062778	*Aeromicrobium* sp.	MK143132	NZSG	algae
114B-8	66.0278306	-18.4062778	*Knoellia* sp.	MK143173	SWA	algae
114C-1	66.0278306	-18.4062778	*Nocardioides* sp.	MK143182	ISP2	sediment
114C-2	66.0278306	-18.4062778	*Kytococcus* sp.	MK143200	ISP2	sediment
114E-2	66.0278306	-18.4062778	*Kytococcus* sp.	MK143185	SWA	water
114E-7	66.0278306	-18.4062778	*Kineosporiaceae*	MK143184	A1	water
115B-4	65.9843306	-18.3947889	*Nocardioides* sp.	MK168020	ISP2	sediment
115B-7	65.9843306	-18.3947889	*Knoellia* sp.	MK143177	ISP2	sediment
115B-9	65.9843306	-18.3947889	*Micrococcus* sp.	MK143207	NZSG	sediment
115B-10	65.9843306	-18.3947889	*Bacillus* sp.	MK143296	NZSG	sediment
115D-1	65.9843306	-18.3947889	*Bacillus* sp.	MK143221	ISP2	sediment
115D-2	65.9843306	-18.3947889	*Bacillus* sp.	MK143298	ISP2	sediment
115E-1	65.9843306	-18.3947889	*Bacillus* sp.	MK143261	NZSG	sediment
116A-1	66.098072	-16.929879	*Bacillus* sp.	MK143283	A1	algae
116B-32	66.098072	-16.929879	*Paenibacillus* sp.	MK143245	A1	algae
116B-40	66.098072	-16.929879	*Paenibacillus* sp.	MK143235	A1	algae
116B-45	66.098072	-16.929879	*Paenibacillus* sp.	MK143236	A1	algae

116C-1	66.098072	-16.929879	*Arthrobacter* sp.	MK143111	NZSG	algae
116E-2	66.098072	-16.929879	*Paenibacillus* sp.	MK143260	Chitin	sediment
116E-3	66.098072	-16.929879	*Bacillus* sp.	MK143293	Chitin	sediment
116F-2	66.098072	-16.929879	*Micrococcus* sp.	MK143122	Chitin	sediment
116F-6	66.098072	-16.929879	*Micrococcus* sp.	MK143112	Chitin	sediment
116F-9	66.098072	-16.929879	*Micromonospora* sp.	MK143197	Chitin	sediment
116F-10	66.098072	-16.929879	*Bacillus* sp.	MK143252	NZSG	sediment
116F-12	66.098072	-16.929879	*Staphylococcus* sp.	MK143292	ISP2	sediment
116G-6	66.098072	-16.929879	*Bacillus* sp.	MK143291	A1	sediment
116H-8	66.098072	-16.929879	*Bacillus* sp.	MK143237	ISP2	sediment
116H-12	66.098072	-16.929879	*Streptomyces* sp.	MK143123	A1	sediment
116H-18	66.098072	-16.929879	*Oerskovia* sp.	MK163385	Chitin	sediment
117A-1	66.076302	-16.688245	*Arthrobacter* sp.	MK143208	LWA	sediment
117C-1	66.076302	-16.688245	*Micrococcus* sp.	MK143113	Chitin	algae
117D-1	66.076302	-16.688245	*Bacillus* sp.	MK143288	A1	sediment
117D-3	66.076302	-16.688245	*Paenibacillus* sp.	MK143281	A1	sediment
117D-4	66.076302	-16.688245	*Paenibacillus* sp.	MK143222	A1	sediment
117D-6	66.076302	-16.688245	*Bacillus* sp.	MK143277	NZSG	sediment
117D-7	66.076302	-16.688245	*Paenibacillus* sp.	MK143253	ISP2	sediment
117E-1	66.076302	-16.688245	*Brevundimonas* sp.	MK143369	LWA	sediment
117E-3	66.076302	-16.688245	*Rhizobium* sp.	MK143351	LWA	sediment

117F-1	66.076302	-16.688245	*Kocuria* sp.	MK143124	A1	sediment
117G-14	66.076302	-16.688245	*Micrococcus* sp.	MK143114	Chitin	sediment
117G-16	66.076302	-16.688245	*Arthrobacter* sp.	MK143115	LWA	sediment
117G-21	66.076302	-16.688245	*Arthrobacter* sp.	MK163386	LWA	sediment
117G-22	66.076302	-16.688245	*Arthrobacter* sp.	MK143163	LWA	sediment
117G-24	66.076302	-16.688245	*Arthrobacter* sp.	MK163387	LWA	sediment
117G-28	66.076302	-16.688245	*Pseudarthrobacter* sp.	MK168021	LWA	sediment
117G-33	66.076302	-16.688245	*Arthrobacter* sp.	MK143161	LWA	sediment
117G-35	66.076302	-16.688245	*Arthrobacter* sp.	MK163388	LWA	sediment
117G-38	66.076302	-16.688245	*Pseudarthrobacter* sp.	MK163389	LWA	sediment
117H-3	66.076302	-16.688245	*Micrococcus* sp.	MK143117	NZSG	sediment
117H-11	66.076302	-16.688245	*Bradyrhizobium* sp.	MK143330	A1	sediment
117H-14	66.076302	-16.688245	*Microbacterium* sp.	MK143116	A1	sediment
117H-15	66.076302	-16.688245	*Bradyrhizobium* sp.	MK143343	A1	sediment
117H-19	66.076302	-16.688245	*Microbacterium* sp.	MK143125	A1	sediment
117H-20	66.076302	-16.688245	*Bradyrhizobium* sp.	MK163390	A1	sediment
117H-21	66.076302	-16.688245	*Bradyrhizobium* sp.	MK143340	A1	sediment
117H-23	66.076302	-16.688245	*Microbacterium* sp.	MK143166	A1	sediment
117H-26	66.076302	-16.688245	*Bradyrhizobium* sp.	MK168022	A1	sediment
117H-31	66.076302	-16.688245	*Bradyrhizobium* sp.	MK143337	ISP2	sediment

117H-32	66.076302	-16.688245	*Bradyrhizobium* sp.	MK143338	ISP2	sediment
117H-37	66.076302	-16.688245	*Bradyrhizobium* sp.	MK163391	ISP2	sediment
117H-38	66.076302	-16.688245	*Bradyrhizobium* sp.	MK143341	ISP2	sediment
117H-39	66.076302	-16.688245	*Bradyrhizobium* sp.	MK143342	ISP2	sediment
117H-46	66.076302	-16.688245	*Xanthobacter* sp.	MK143365	ISP2	sediment
117H-59	66.076302	-16.688245	*Microbacterium* sp.	MK168023	LWA	sediment
117H-64	66.076302	-16.688245	*Bradyrhizobium* sp.	MK163392	LWA	sediment
117H-69	66.076302	-16.688245	*Bradyrhizobium* sp.	MK143345	LWA	sediment
117H-73	66.076302	-16.688245	*Bradyrhizobium* sp.	MK163393	Chitin	sediment
117H-81	66.076302	-16.688245	*Bradyrhizobium* sp.	MK143339	Chitin	sediment
117H-84	66.076302	-16.688245	*Bradyrhizobium* sp.	MK163394	Chitin	sediment
117H-85	66.076302	-16.688245	*Bradyrhizobium* sp.	MK168024	Chitin	sediment
117H-86	66.076302	-16.688245	*Bradyrhizobium* sp.	MK143346	Chitin	sediment
117H-90	66.076302	-16.688245	*Bradyrhizobium* sp.	MK143321	Chitin	sediment
117I-1	66.076302	-16.688245	*Staphylococcus* sp.	MK143254	NZSG	sediment
117J-1	66.076302	-16.688245	*Pseudarthrobacter* sp.	MK143193	LWA	sediment
118A-1	66.076305	-16.68827	*Bacillus* sp.	MK143223	A1	sediment
118B-32	66.076305	-16.68827	*Ruegeria* sp.	MK143344	ISP2	algae
118B-36	66.076305	-16.68827	*Maribacter* sp.	MK143311	ISP2	algae
118B-43	66.076305	-16.68827	*Leisingera* sp.	MK143358	ISP2	algae

118B-72	66.076305	-16.68827	*Loktanella* sp.	MK163395	SWA	algae
118C-1	66.076305	-16.68827	*Paenibacillus* sp.	MK143278	NZSG	sediment
118C-2	66.076305	-16.68827	*Psychrobacillus* sp.	MK143246	NZSG	sediment
118D-1	66.076305	-16.68827	*Bacillus* sp.	MK143272	ISP2	algae
118D-2	66.076305	-16.68827	*Kytococcus* sp.	MK163396	SWA	algae
119A-3	65.72176	-16.788817	*Cryobacterium* sp.	MK143186	A1	sediment
119A-9	65.72176	-16.788817	*Pseudarthrobacter* sp.	MK143134	A1	sediment
119A-10	65.72176	-16.788817	*Arthrobacter* sp.	MK168025	A1	sediment
119A-16	65.72176	-16.788817	*Variovorax* sp.	MK163397	ISP2	sediment
119A-17	65.72176	-16.788817	*Streptomyces* sp.	MK143126	ISP2	sediment
119A-18	65.72176	-16.788817	*Pseudomonas* sp.	MK143301	ISP2	sediment
119A-25	65.72176	-16.788817	*Pseudomonas* sp.	MK163398	ISP2	sediment
119A-31	65.72176	-16.788817	*Ochrobactrum* sp.	MK163399	NZSG	sediment
119A-37	65.72176	-16.788817	*Mesorhizobium* sp.	MK163400	LWA	sediment
119A-47	65.72176	-16.788817	*Streptomyces* sp.	MK143196	LWA	sediment
119A-48	65.72176	-16.788817	*Mesorhizobium* sp.	MK143302	LWA	sediment
119A-59	65.72176	-16.788817	*Phycicoccus* sp.	MK168026	Chitin	sediment
119A-69	65.72176	-16.788817	*Methylibium* sp.	MK143370	Chitin	sediment
119A-74	65.72176	-16.788817	*Deefgea* sp.	MK143331	Chitin	sediment
119A-75	65.72176	-16.788817	*Deefgea* sp.	MK143332	Chitin	sediment

119A-79	65.72176	-16.788817	*Deefgea* sp.	MK143333	Chitin	sediment
119A-82	65.72176	-16.788817	*Yersinia* sp.	MK143322	Chitin	sediment
120A-8	65.721916	-16.792269	*Microbacterium* sp.	MK143187	LWA	sediment
120A-11	65.721916	-16.792269	*Marmoricola* sp.	MK143135	LWA	sediment
120A-12	65.721916	-16.792269	*Microbacterium* sp.	MK143142	LWA	sediment
120A-16	65.721916	-16.792269	*Microbacterium* sp.	MK143158	LWA	sediment
120A-20	65.721916	-16.792269	*Microbacterium* sp.	MK163401	LWA	sediment
120A-22	65.721916	-16.792269	*Microbacterium* sp.	MK143128	LWA	sediment
120A-23	65.721916	-16.792269	*Microbacterium* sp.	MK143159	LWA	sediment
120A-24	65.721916	-16.792269	*Microbacterium* sp.	MK143157	LWA	sediment
120A-31	65.721916	-16.792269	*Nocardioides* sp.	MK163402	Chitin	sediment
120A-50	65.721916	-16.792269	*Nocardioides* sp.	MK143174	Chitin	sediment
120A-53	65.721916	-16.792269	*Nocardioides* sp.	MK163403	Chitin	sediment
121A-3	65.722159	-16.792697	*Microbacterium* sp.	MK143175	Chitin	sediment
121A-9	65.722159	-16.792697	*Microbacterium* sp.	MK168027	Chitin	sediment
121A-11	65.722159	-16.792697	*Microbacterium* sp.	MK143129	Chitin	sediment
121A-19	65.722159	-16.792697	*Microbacterium* sp.	MK143136	Chitin	sediment
122A-1	65.721539	-16.786873	*Streptomyces* sp.	MK143167	NZSG	sediment
122B-8	65.721539	-16.786873	*Thiomonas* sp.	MK143327	A1	sediment
122B-20	65.721539	-16.786873	*Thiomonas* sp.	MK143334	A1	sediment
122B-34	65.721539	-16.786873	*Thiomonas* sp.	MK143323	A1	sediment

122B-46	65.721539	-16.786873	*Thiomonas* sp.	MK143324	A1	sediment
122B-48	65.721539	-16.786873	*Thiomonas* sp.	MK143325	A1	sediment
122B-49	65.721539	-16.786873	*Thiomonas* sp.	MK143326	A1	sediment
123A-1	65.701479	-16.771132	*Bacillus* sp.	MK143282	ISP2	sediment
123A-3	65.701479	-16.771132	*Micrococcus* sp.	MK143118	A1	sediment
123A-4	65.701479	-16.771132	*Tibeticola* sp.	MK163404	A1	sediment
123A-8	65.701479	-16.771132	*Tibeticola* sp.	MK143328	A1	sediment
123A-17	65.701479	-16.771132	*Tibeticola* sp.	MK143376	A1	sediment
123A-25	65.701479	-16.771132	*Tibeticola* sp.	MK143377	A1	sediment
123A-40	65.701479	-16.771132	*Bacillus* sp.	MK143255	Chitin	sediment
123A-41	65.701479	-16.771132	*Bacillus* sp.	MK143276	Chitin	sediment
123A-45	65.701479	-16.771132	*Bacillus* sp.	MK168028	Chitin	sediment
123A-50	65.701479	-16.771132	*Bacillus* sp.	MK168029	Chitin	sediment
123A-53	65.701479	-16.771132	*Bacillus* sp.	MK143285	Chitin	sediment
123A-56	65.701479	-16.771132	*Agrococcus* sp.	MK143155	LWA	sediment
123A-58	65.701479	-16.771132	*Paenibacillus* sp.	MK163405	LWA	sediment
123A-60	65.701479	-16.771132	*Agrococcus* sp.	MK163406	LWA	sediment
123A-61	65.701479	-16.771132	*Agrococcus* sp.	MK143150	LWA	sediment
123A-62	65.701479	-16.771132	*Agrococcus* sp.	MK143149	LWA	sediment
123A-64	65.701479	-16.771132	*Agrococcus* sp.	MK143147	LWA	sediment
123A-65	65.701479	-16.771132	*Agrococcus* sp.	MK163407	LWA	sediment

123A-66	65.701479	-16.771132	*Agrococcus* sp.	MK143145	LWA	sediment
123A-67	65.701479	-16.771132	*Agrococcus* sp.	MK143106	LWA	sediment
123A-70	65.701479	-16.771132	*Agrococcus* sp.	MK163408	LWA	sediment
123A-71	65.701479	-16.771132	*Agrococcus* sp.	MK143144	LWA	sediment
123A-73	65.701479	-16.771132	*Agrococcus* sp.	MK143107	LWA	sediment
123A-74	65.701479	-16.771132	*Agrococcus* sp.	MK143154	LWA	sediment
123A-75	65.701479	-16.771132	*Agrococcus* sp.	MK143153	LWA	sediment
123A-77	65.701479	-16.771132	*Agrococcus* sp.	MK143156	LWA	sediment
123A-78	65.701479	-16.771132	*Dermacoccus* sp.	MK143205	LWA	sediment
123A-79	65.701479	-16.771132	*Dermacoccus* sp.	MK163409	LWA	sediment
123A-82	65.701479	-16.771132	*Dermacoccus* sp.	MK143108	LWA	sediment
123A-83	65.701479	-16.771132	*Agrococcus* sp.	MK143146	LWA	sediment
123A-84	65.701479	-16.771132	*Agrococcus* sp.	MK143152	LWA	sediment
123A-85	65.701479	-16.771132	*Agrococcus* sp.	MK143148	LWA	sediment
123A-86	65.701479	-16.771132	*Agrococcus* sp.	MK143141	LWA	sediment
123A-87	65.701479	-16.771132	*Tibeticola* sp.	MK143335	LWA	sediment
123A-90	65.701479	-16.771132	*Agrococcus* sp.	MK143140	LWA	sediment
123A-91	65.701479	-16.771132	*Agrococcus* sp.	MK143151	LWA	sediment
123A-93	65.701479	-16.771132	*Agrococcus* sp.	MK143143	LWA	sediment
123A-95	65.701479	-16.771132	*Agrococcus* sp.	MK143109	LWA	sediment
125A-2	65.641573	-16.809047	*Sporosarcina* sp.	MK143265	ISP2	water column

126A-1	65.639585	-16.80899	*Staphylococcus* sp.	MK143247	NZSG	sediment
126A-2	65.639585	-16.80899	*Kocuria* sp.	MK143209	ISP2	sediment
126A-9	65.639585	-16.80899	*Microbacterium* sp.	MK143194	A1	sediment
127A-1	65.640801	-16.80839	*Arthrobacter* sp.	MK168030	A1	sediment
128A-1	65.640207	-16.844106	*Sandaracinobacter* sp.	MK143313	A1	sediment
128A-6	65.640207	-16.844106	*Hydrogenophaga* sp.	MK143329	A1	sediment
128A-15	65.640207	-16.844106	*Bacillus* sp.	MK163410	NZSG	sediment
128A-35	65.640207	-16.844106	*Bosea* sp.	MK163411	ISP2	sediment
128A-46	65.640207	-16.844106	*Porphyrobacter* sp.	MK143366	ISP2	sediment
128A-49	65.640207	-16.844106	*Streptomyces* sp.	MK168032	ISP2	sediment
128A-51	65.640207	-16.844106	*Rhodopseudomonas* sp.	MK143357	ISP2	sediment
129A-1	65.638311	-16.910157	*Micromonospora* sp.	MK168031	Chitin	sediment
129A-2	65.638311	-16.910157	*Micromonospora* sp.	MK168033	Chitin	sediment
129A-3	65.638311	-16.910157	*Micromonospora* sp.	MK163412	Chitin	sediment
129A-7	65.638311	-16.910157	*Micromonospora* sp.	MK168034	Chitin	sediment
129A-13	65.638311	-16.910157	*Micromonospora* sp.	MK143137	Chitin	sediment
129A-18	65.638311	-16.910157	*Bacillus* sp.	MK143238	Chitin	sediment
129A-19	65.638311	-16.910157	*Micromonospora* sp.	MK168035	Chitin	sediment
129A-20	65.638311	-16.910157	*Micromonospora* sp.	MK163413	LWA	sediment
129A-21	65.638311	-16.910157	*Micromonospora* sp.	MK168035	LWA	sediment

129A-23	65.638311	-16.910157	*Micromonospora* sp.	MK163414	LWA	sediment
129A-24	65.638311	-16.910157	*Micromonospora* sp.	MK168036	LWA	sediment
129A-26	65.638311	-16.910157	*Micromonospora* sp.	MK163415	LWA	sediment
129A-30	65.638311	-16.910157	*Bacillus* sp.	MK143248	LWA	sediment
129A-35	65.638311	-16.910157	*Micromonospora* sp.	MK168037	LWA	sediment
129A-36	65.638311	-16.910157	*Bacillus* sp.	MK143297	LWA	sediment
129A-38	65.638311	-16.910157	*Tardiphaga* sp.	MK143352	A1	sediment
129A-39	65.638311	-16.910157	*Bacillus* sp.	MK143262	A1	sediment
129A-41	65.638311	-16.910157	*Mycobacterium* sp.	MK143110	A1	sediment
129A-42	65.638311	-16.910157	*Dyadobacter* sp.	MK143216	A1	sediment
129A-45	65.638311	-16.910157	*Psychrobacillus* sp.	MK143271	A1	sediment
129A-46	65.638311	-16.910157	*Paenibacillus* sp.	MK143266	ISP2	sediment
129A-49	65.638311	-16.910157	*Paenibacillus* sp.	MK143250	ISP2	sediment
129A-54	65.638311	-16.910157	*Bacillus* sp.	MK143224	NZSG	sediment
129A-55	65.638311	-16.910157	*Mycolicibacterium* sp.	MK163416	LWA	sediment
130A-5	65.579718	-16.951147	*Paenibacillus* sp.	MK143267	NZSG	sediment
130A-11	65.579718	-16.951147	*Leucobacter* sp.	MK143164	ISP2	sediment
130A-12	65.579718	-16.951147	*Cohnella* sp.	MK143287	ISP2	sediment
130A-13	65.579718	-16.951147	*Paenibacillus* sp.	MK168038	ISP2	sediment
130A-15	65.579718	-16.951147	*Paenibacillus* sp.	MK143284	ISP2	sediment
130A-16	65.579718	-16.951147	*Psychrobacter* sp.	MK143359	A1	sediment

130A-23	65.579718	-16.951147	*Bacillus* sp.	MK168039	A1	sediment
130A-25	65.579718	-16.951147	*Massilia* sp.	MK168040	A1	sediment
131A-2	65.561018	-16.947166	*Noviherbaspirillum* sp.	MK143356	NZSG	sediment
131A-6	65.561018	-16.947166	*Kytococcus* sp.	MK168041	A1	sediment
132A-26	65.548551	-16.980315	*Streptomyces* sp.	MK143168	LWA	sediment
132A-28	65.548551	-16.980315	*Streptomyces* sp.	MK163417	LWA	sediment
132A-30	65.548551	-16.980315	*Streptomyces* sp.	MK163418	A1	sediment
132A-31	65.548551	-16.980315	*Streptomyces* sp.	MK143169	A1	sediment
132A-32	65.548551	-16.980315	*Streptomyces* sp.	MK168042	A1	sediment
133A-1	65.57096	-17.073028	*Paenibacillus* sp.	MK163419	NZSG	sediment
133A-3	65.57096	-17.073028	*Kytococcus* sp.	MK143119	A1	sediment
133A-5	65.57096	-17.073028	*Bacillus* sp.	MK143299	A1	sediment
133A-7	65.57096	-17.073028	*Brachybacterium* sp.	MK143198	A1	sediment
134A-1	65.850981	-18.193052	*Nocardiopsis* sp.	MK143190	SWA	sediment
134A-3	65.850981	-18.193052	*Paenibacillus* sp.	MK143279	NZSG	sediment
134A-4	65.850981	-18.193052	*Rhodococcus* sp.	MK163420	NZSG	sediment
134A-5	65.850981	-18.193052	*Bacillus* sp.	MK168043	NZSG	sediment
134A-7	65.850981	-18.193052	*Bacillus* sp.	MK143275	NZSG	sediment
134A-16	65.850981	-18.193052	*Streptomyces* sp.	MK163421	Chitin	sediment
134A-17	65.850981	-18.193052	*Bacillus* sp.	MK143239	Chitin	sediment

134A-21	65.850981	-18.193052	*Bacillus* sp.	MK143273	ISP2	sediment
134A-22	65.850981	-18.193052	*Paenibacillus* sp.	MK143240	ISP2	sediment
134A-27	65.850981	-18.193052	*Bacillus* sp.	MK143241	ISP2	sediment
134A-39	65.850981	-18.193052	*Bacillus* sp.	MK143256	A1	sediment
134C-5	65.850981	-18.193052	*Salinicoccus* sp.	MK143226	ISP2	sponge
134C-9	65.850981	-18.193052	*Bacillus* sp.	MK143242	ISP2	sponge
134C-14	65.850981	-18.193052	*Bacillus* sp.	MK143225	SWA	sponge
134D-1	65.850981	-18.193052	*Shewanella* sp.	MK143303	A1	sponge
134D-18	65.850981	-18.193052	*Bacillus* sp.	MK168044	SWA	sponge
134D-22	65.850981	-18.193052	*Bacillus* sp.	MK143227	SWA	sponge
135A-2	65.73357	-18.170398	*Psychrobacillus* sp.	MK143270	A1	sediment
135A-6	65.73357	-18.170398	*Bacillus* sp.	MK143274	A1	sediment
135A-8	65.73357	-18.170398	*Saccharomonospora* sp.	MK143180	A1	sediment
135A-10	65.73357	-18.170398	*Streptomyces* sp.	MK143130	ISP2	sediment
135A-13	65.73357	-18.170398	*Psychrobacillus* sp.	MK143300	ISP2	sediment
135A-15	65.73357	-18.170398	*Streptomyces* sp.	MK163422	ISP2	sediment
135A-17	65.73357	-18.170398	*Bacillus* sp.	MK143228	ISP2	sediment
135A-18	65.73357	-18.170398	*Streptomyces* sp.	MK143170	NZSG	sediment
135A-21	65.73357	-18.170398	*Micromonospora* sp.	MK143203	NZSG	sediment
135A-23	65.73357	-18.170398	*Oerskovia* sp.	MK143160	NZSG	sediment

135A-27	65.73357	-18.170398	*Psychrobacillus* sp.	MK143249	NZSG	sediment
135A-29	65.73357	-18.170398	*Micromonospora* sp.	MK163423	NZSG	sediment
135A-30	65.73357	-18.170398	*Mycobacterium* sp.	MK143199	SWA	sediment
135A-32	65.73357	-18.170398	*Micromonospora* sp.	MK143206	SWA	sediment
135A-34	65.73357	-18.170398	*Rhodococcus* sp.	MK143191	SWA	sediment
135A-37	65.73357	-18.170398	*Streptomyces* sp.	MK168045	SWA	sediment
135A-41	65.73357	-18.170398	*Microbacterium* sp.	MK143165	Chitin	sediment
135A-42	65.73357	-18.170398	*Microbacterium* sp.	MK143162	Chitin	sediment
136A-17	65.444369	-22.206837	*Salinicoccus* sp.	MK143243	A1	sediment
136A-21	65.444369	-22.206837	*Salinicoccus* sp.	MK143229	A1	sediment
136A-24	65.444369	-22.206837	*Polaribacter* sp.	MK168046	A1	sediment
136A-28	65.444369	-22.206837	*Shewanella* sp.	MK143304	A1	sediment
136A-31	65.444369	-22.206837	*Shewanella* sp.	MK143305	A1	sediment
136A-33	65.444369	-22.206837	*Salinicoccus* sp.	MK143257	NZSG	sediment
136A-44	65.444369	-22.206837	*Cellulophaga* sp.	MK143213	SWA	sediment
136A-71	65.444369	-22.206837	*Cellulophaga* sp.	MK143211	SWA	sediment
136A-74	65.444369	-22.206837	*Cellulophaga* sp.	MK163424	SWA	sediment
136A-77	65.444369	-22.206837	*Micrococcus* sp.	MK163425	Chitin	sediment
136A-82	65.444369	-22.206837	*Shewanella* sp.	MK143336	Chitin	sediment
136A-84	65.444369	-22.206837	*Shewanella* sp.	MK143348	Chitin	sediment
136A-87	65.444369	-22.206837	*Shewanella* sp.	MK143306	Chitin	sediment

136A-88	65.444369	-22.206837	*Shewanella* sp.	MK143312	Chitin	sediment
136A-91	65.444369	-22.206837	*Micrococcus* sp.	MK143131	Chitin	sediment
136A-93	65.444369	-22.206837	*Micrococcus* sp.	MK168047	Chitin	sediment
136A-96	65.444369	-22.206837	*Salinicoccus* sp.	MK143251	Chitin	sediment
136A-99	65.444369	-22.206837	*Psychrobacter* sp.	MK143362	Chitin	sediment
137A-1	65.484244	-22.391858	*Shewanella* sp.	MK143355	NZSG	sponge
137A-17	65.484244	-22.391858	*Janibacter* sp.	MK143181	SWA	sponge
137A-20	65.484244	-22.391858	*Williamsia* sp.	MK143189	Chitin	sponge
137A-21	65.484244	-22.391858	*Rhodococcus* sp.	MK143178	Chitin	sponge
137A-25	65.484244	-22.391858	*Blastococcus* sp.	MK163426	Chitin	sponge
137B-1	65.484244	-22.391858	*Rhodococcus* sp.	MK163427	A1	sponge
137B-11	65.484244	-22.391858	*Rhodococcus* sp.	MK143183	A1	sponge
137B-18	65.484244	-22.391858	*Nocardioides* sp.	MK143132	A1	sponge
137B-27	65.484244	-22.391858	*Nocardioides* sp.	MK163428	A1	sponge
137B-30	65.484244	-22.391858	*Nocardioides* sp.	MK143138	A1	sponge
137B-47	65.484244	-22.391858	*Staphylococcus* sp.	MK163429	NZSG	sponge
137C-4	65.484244	-22.391858	*Mycolicibacterium* sp.	MK143192	ISP2	sponge
137C-23	65.484244	-22.391858	*Shewanella* sp.	MK143307	SWA	sponge
137D-3	65.484244	-22.391858	*Rhodococcus* sp.	MK163430	NZSG	jellyfish egg
137D-20	65.484244	-22.391858	*Aureimonas* sp.	MK143361	SWA	jellyfish egg
137F-8	65.484244	-22.391858	*Cellulophaga* sp.	MK143219	A1	tunicate

138A-7	65.512811	-22.304967	*Shewanella* sp.	MK143353	ISP2	algae
138A-12	65.512811	-22.304967	*Shewanella* sp.	MK163431	A1	algae
138A-16	65.512811	-22.304967	*Shewanella* sp.	MK143372	A1	algae
138A-25	65.512811	-22.304967	*Shewanella* sp.	MK143314	SWA	algae
138A-27	65.512811	-22.304967	*Shewanella* sp.	MK143350	SWA	algae
138A-39	65.512811	-22.304967	*Shewanella* sp.	MK168048	SWA	algae
138B-3	65.512811	-22.304967	*Vibrio* sp.	MK163432	ISP2	algae
138B-10	65.512811	-22.304967	*Cellulophaga* sp.	MK143217	A1	algae
138B-15	65.512811	-22.304967	*Algoriphagus* sp.	MK143214	SWA	algae
138B-23	65.512811	-22.304967	*Exiguobacterium* sp.	MK143264	SWA	algae
138B-43	65.512811	-22.304967	*Sulfitobacter* sp.	MK143349	SWA	algae
138D-1	65.512811	-22.304967	*Micromonospora* sp.	MK143204	ISP2	tunicate
138D-5	65.512811	-22.304967	*Shewanella* sp.	MK143315	SWA	tunicate
138E-3	65.512811	-22.304967	*Bacillus* sp.	MK143280	A1	tunicate
138H-4	65.512811	-22.304967	*Maribacter* sp.	MK143373	ISP2	sponge
138H-8	65.512811	-22.304967	*Loktanella* sp.	MK163433	NZSG	sponge
138H-11	65.512811	-22.304967	*Flavobacterium* sp.	MK143220	NZSG	sponge
138H-20	65.512811	-22.304967	*Sphingopyxis* sp.	MK143347	A1	sponge
138H-22	65.512811	-22.304967	*Sphingopyxis* sp.	MK163434	A1	sponge
138H-32	65.512811	-22.304967	*Dokdonia* sp.	MK143218	A1	sponge
138H-63	65.512811	-22.304967	*Aquimarina* sp.	MK143212	SWA	sponge

138H-65	65.512811	-22.304967	*Altererythrobacter* sp.	MK168049	SWA	sponge
138H-74	65.512811	-22.304967	*Aquimarina* sp.	MK168050	SWA	sponge
138H-75	65.512811	-22.304967	*Phaeobacter* sp.	MK143364	SWA	sponge
138H-81	65.512811	-22.304967	*Loktanella* sp.	MK163435	SWA	sponge
138H-88	65.512811	-22.304967	*Aquimarina* sp.	MK143215	SWA	sponge
138I-7	65.512811	-22.304967	*Shewanella* sp.	MK163436	A1	sponge
138I-19	65.512811	-22.304967	*Shewanella* sp.	MK143316	SWA	sponge
138J-13	65.512811	-22.304967	*Shewanella* sp.	MK143374	A1	sponge
138J-17	65.512811	-22.304967	*Shewanella* sp.	MK168051	A1	sponge
138K-12	65.512811	-22.304967	*Kytococcus* sp.	MK163437	A1	algae
138K-23	65.512811	-22.304967	*Rhizobium* sp.	MK143317	NZSG	algae
138K-25	65.512811	-22.304967	*Micrococcus* sp.	MK143195	NZSG	algae
138K-29	65.512811	-22.304967	*Shewanella* sp.	MK143375	SWA	algae
138K-41	65.512811	-22.304967	*Bacillus* sp.	MK143230	SWA	algae
138K-47	65.512811	-22.304967	*Rhizobium* sp.	MK143308	SWA	algae
138K-50	65.512811	-22.304967	*Shewanella* sp.	MK143309	SWA	algae
139A-4	65.508073	-22.330494	*Rhodococcus* sp.	MK143188	ISP2	sediment
139A-26	65.508073	-22.330494	*Streptomyces* sp.	MK168052	ISP2	sediment
139A-30	65.508073	-22.330494	*Bacillus* sp.	MK143290	ISP2	sediment
139A-32	65.508073	-22.330494	*Paenibacillus* sp.	MK143269	A1	sediment
139A-33	65.508073	-22.330494	*Paenisporosarcina*	MK143268	A1	sediment

			sp.			
139A-36	65.508073	-22.330494	*Bacillus* sp.	MK143231	A1	sediment
139A-37	65.508073	-22.330494	*Agrococcus* sp.	MK143201	A1	sediment
139A-43	65.508073	-22.330494	*Bacillus* sp.	MK143294	A1	sediment
139A-43B	65.508073	-22.330494	*Streptomyces* sp.	MK143171	A1	sediment
139A-55	65.508073	-22.330494	*Micromonospora* sp.	MK143202	SWA	sediment
139B-8	65.508073	-22.330494	*Shewanella* sp.	MK143318	A1	sediment
139B-14	65.508073	-22.330494	*Bacillus* sp.	MK143232	A1	sediment
139B-15	65.508073	-22.330494	*Rathayibacter* sp.	MK143139	A1	sediment
139B-19	65.508073	-22.330494	*Bacillus* sp.	MK143259	A1	sediment
139B-21	65.508073	-22.330494	*Yersinia* sp.	MK143360	ISP2	sediment
140B-1	65.738088	-23.24695	*Bacillus* sp.	MK143233	ISP2	tunicate
140E-2	65.738088	-23.24695	*Oceanobacillus* sp.	MK143286	NZSG	sediment
140E-3	65.738088	-23.24695	*Micrococcus* sp.	MK143120	NZSG	sediment
140E-5	65.738088	-23.24695	*Paenibacillus* sp.	MK143263	NZSG	sediment
140E-6	65.738088	-23.24695	*Streptomyces* sp.	MK143172	A1	sediment
140F-1	65.738088	-23.24695	*Bacillus* sp.	MK143234	ISP2	sediment
141AM-1	65.736375	-23.208858	*Pseudomonas* sp.	MK143367	NZSG	sediment
141BM-2	65.736375	-23.208858	*Micrococcus* sp.	MK168053	ISP2	algae
141BF-3	65.736375	-23.208858	*Pseudomonas* sp.	MK143320	A1	algae
141BF-13	65.736375	-23.208858	*Paenibacillus* sp.	MK143244	A1	algae

141BF-16	65.736375	-23.208858	*Pseudomonas* sp.	MK168054	ISP2	algae
141BF-17	65.736375	-23.208858	*Pseudomonas* sp.	MK143319	ISP2	algae
141BF-35	65.736375	-23.208858	*Pseudomonas* sp.	MK163438	ISP2	algae
142AF-4	65.622762	-23.467935	*Bacillus* sp.	MK143289	A1	sediment
142BM-4	65.622762	-23.467935	*Bacillus* sp.	MK143258	ISP2	sediment
142BM-5	65.622762	-23.467935	*Micrococcus* sp.	MK143121	ISP2	sediment
142BF-16	65.622762	-23.467935	*Rhodococcus* sp.	MK143133	LWA	sediment
142BF-17	65.622762	-23.467935	*Enterobacter* sp.	MK143363	LWA	sediment
142BF-21	65.622762	-23.467935	*Enterobacter* sp.	MK143371	LWA	sediment
142BF-27	65.622762	-23.467935	*Microbacterium* sp.	MK143179	LWA	sediment
142BF-42	65.622762	-23.467935	*Paenibacillus* sp.	MK143295	LWA	sediment
143B-EGS1	65.647536	-23.542869	*Pseudomonas* sp.	MK143354	A1	white eggs from beitukongur
143B-EGS6	65.647536	-23.542869	*Pseudomonas* sp.	MK143368	A1	white eggs from beitukongur
143B-EGS7	65.647536	-23.542869	*Pseudomonas* sp.	MK143310	A1	white eggs from beitukongur

4 Results and discussion

4.1 Coupling MALDI-TOF mass spectrometry protein and natural products analyses to rapidly discriminate bacterial function

As mentioned in section 1.4, the philosophy to create microbial libraries from the environment has seen little change over several decades. A major limitation of selecting bacterial colonies from microbial diversity plates was the visual inspection of colony morphology and subsequent use of subjectivity to decide which colonies to add to a library. As a result, microbial drug discovery platforms have been plagued with a high degree of taxonomic and chemical redundancy. But before we addressed this problem with relation to drug discovery, we had to first design a platform that allowed the simultaneous analysis of protein and natural product information from bacterial colonies. Thus, in order to visualize relational patterns between bacterial isolates, we created a bioinformatics pipeline designed to facilitate the multi-stage analysis of protein and natural products MS data. Generally speaking, we analyzed MS fingerprints of intact proteins (3,000-15,000 Da) and natural products (200-2,000 Da) in consecutive MALDI-TOF MS linear and reflectron mode acquisitions, respectively. Principal component analysis and hierarchical clustering of protein spectra placed bacterial isolates into putative genus- and species-level groups. Isolates within each grouping were then further discriminated based on differences in natural products production through analysis of Metabolite Association Networks (MANs). The IDBac software, written in R, may be downloaded and installed via a simple Windows installer (see “Code and Data Availability”). The software was designed for simplicity and ease-of-use, with documentation provided at each step in the workflow which, at the time of publication, provides hierarchical clustering, principle components analysis, and MAN analysis. However, whereas similar software has required users to convert raw data on their own and format these to software-dependent configurations, IDBac processes spectra directly from raw data. This provides a transparent and repeatable data-handling process that is less prone to user error or data tampering. As a byproduct, it bundles spectra, by sample, into the widely accepted and easily-shareable mzXML format. The IDBac software is available under a GNU General Public License, and links to the full code along with data acquisition

and analysis tutorials may be found within the “Code and Data Availability” section.

Although our platform is intended to visualize inter- and intra-species differences in natural products production between colonies rather than to identify the precise chemical structure of excreted natural products, it was important to validate that specific nodes (*m/z* features) in our MANs represented functional chemistry as opposed to matrix peaks, media components, or instrument noise. After binning peaks from our MALDI MS spectra, we accounted for the matrix and media ionizable compounds by subtracting a reference spectrum for a matrix and media control during each MALDI run. To account for instrument noise, we precluded signals from the peak picking algorithm that fell below a user-defined signal to noise ratio (i.e., 4:1).

Also, our pipeline is ideally suited for researchers that aim to a) compare functional chemistry between closely related isolates (*e.g*, comparing antibiotic production between two *B. subtilis* genetic variants), b) to probe relationships between taxonomic identity and environmental functionality (*e.g*., studying intra-species differences in *M. chokoriensis* siderophore production), or c) to assess the broad relatedness of unknown environmental bacterial isolates (*e.g*., visualizing phylogenetic and metabolic relatedness within a group of unknown bacteria). Next, each of these three aims are discussed individually.

4.1.1 MALDI-TOF MS metabolite data from two *Bacillus subtilis* genetic variants

Using the closely related lab-domesticated strains *Bacillus subtilis* 3610 and its frame-shift mutant *Bacillus subtilis* PY79, we demonstrated the ability of our MALDI-TOF MS bioinformatics pipeline – IDBac – to use natural products production to distinguish between nearly identical strains based on functional chemistry (Figure 2). The two strains were chosen based on their relation to one another; *B. subtilis* PY79[75,76] is deficient in the ability to produce the antibiotics surfactin and plipastatin due to a frameshift mutation in *sfp*, the 4'-phosphopantetheinyl transferase that mediates non-ribosomal peptide synthetase apoform activation[77].

B. subtilis 3610 and *B. subtilis* PY79 were grown on separate A1 nutrient agar plates, and their natural products were analyzed using MALDI-TOF MS. The differences in functional metabolite production between strains 3610 and PY79 were readily observed and visualized in our MALDI-generated MAN

(Figure 2b), where large nodes represent bacterial colonies and smaller nodes represent *m/z* values (singly charged molecules) present in their corresponding MALDI spectra. This MAN allowed for facile visualization of both the differences in natural products spectra between strains and the ions that were shared or unique to each strain. Four plipastatin (1-4) and three surfactin (5-7) analogs, including their resolved isotopes and adducts, were detected from strain 3610 (Figure 3 and 4) and not from PY79, while both strains shared the ability to produce partially characterized polyglutamate polymers, which have been implicated in species-specific functions such as virulence factor, biofilm formation, and sequestration of toxic metal ions[78,79]. The majority of *m/z* values in the MAN represent natural products, and IDBac successfully filtered out ions associated with matrix and media components. Importantly, MALDI-TOF MS analysis of the natural products region and data processing using IDBac correctly depicted subspecies antibiotic production differences between genetic variants of *B. subtilis* in under 30 minutes.

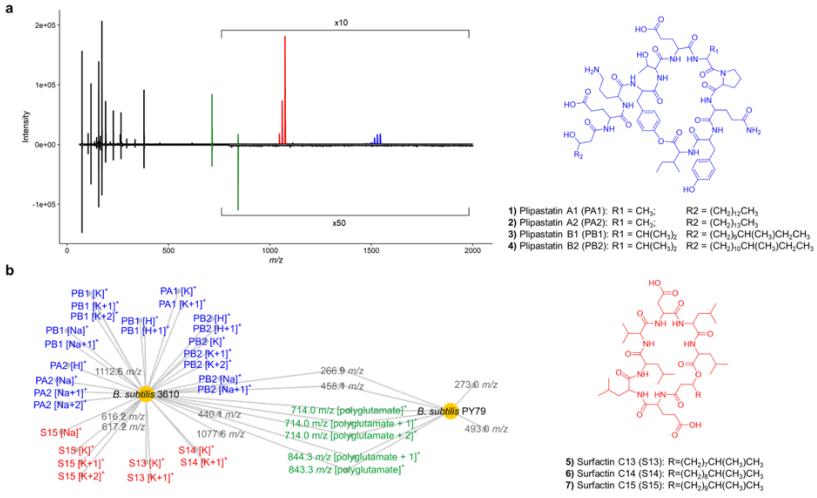


Figure 2. Analysis of MALDI-TOF MS natural products data from two *Bacillus subtilis* genetic variants.

a) Inverse spectrum comparison showing representative spectra for *B. subtilis* 3610 (positive spectrum) and *B. subtilis* PY79 (negative spectrum). **(b)** Metabolite Association Network (MAN) showing the differential production of surfactin and plipastatin antibiotic analogs between the two strains. Large nodes represent individual bacterial colonies, while smaller nodes represent individual *m/z* values in MALDI MS spectra that fall within our peak selection criteria.

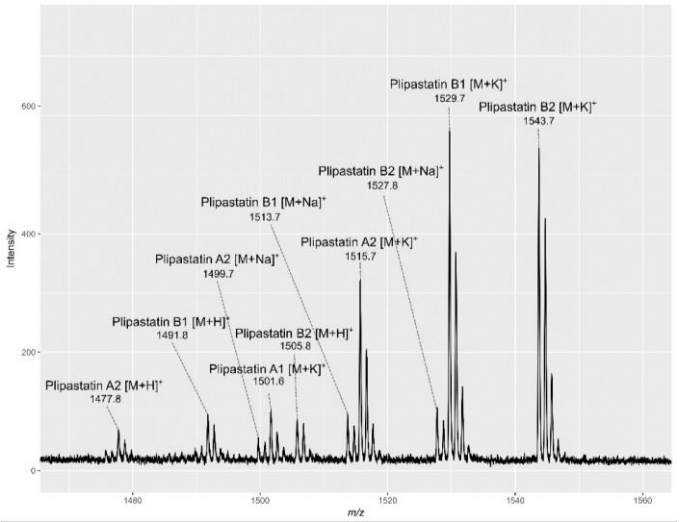


Figure 3. Annotation of surfactin analogs in *Bacillus subtilis* 3610 MS spectrum.

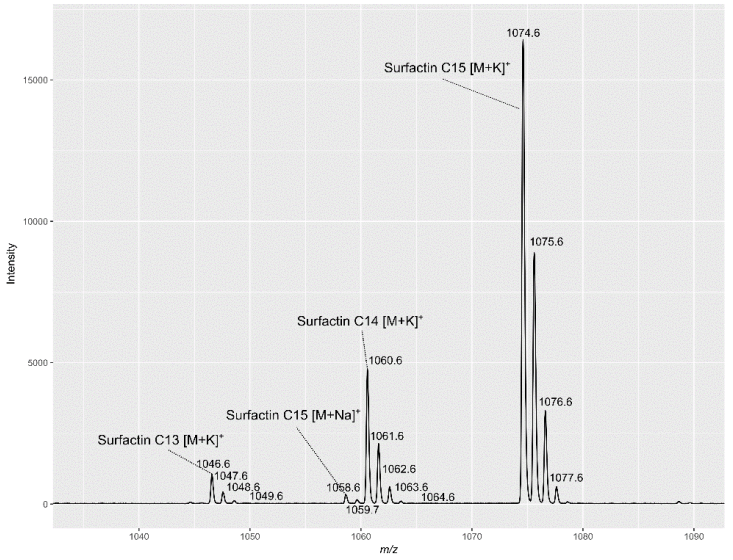


Figure 4. Annotation of plipastatin analogs in *Bacillus subtilis* 3610 MS spectrum**.**

4.1.2 MALDI-TOF MS protein and natural products analysis of isolates from a single *Micromonospora* species

Next, we demonstrated that intra-species strain groupings based on MALDI-TOF MS intact protein profiles can be further discriminated based on *in situ*

natural products production. From our in-house strain library, we selected eight *M. chokoriensis* isolates from two sediment samples collected nearly 175 km apart in Lake Michigan, USA. These isolates share greater than 99% 16S rRNA gene sequence identity across >1,460 nucleotides[80]. After cultivating each strain separately on high nutrient agar, we used MALDI-TOF MS to consecutively acquire MS spectra of the protein and natural products regions. We observed that hierarchical clustering of MALDI-TOF MS protein data correlated strongly with 16S rRNA similarity (Figure 6a), results consistent with many previous efforts that employed MALDI-TOF MS as an alternative to traditional sequence-based taxonomic classification methods, as recently summarized[41,42].

Importantly, while these MALDI-TOF MS protein fingerprints correlated well with corresponding phylogenetic groupings (Figure 6a), subtle differences in natural products production existed and could not be predicted through 16S rRNA phylogenetic analyses or geographic strain distribution patterns. However, analysis of IDBac's MAN rapidly revealed subtle, but significant variations in functional chemistry within these closely related *Micromonospora* isolates (Figure 6b), allowing us to assess the relationship of natural products production to both strain phylogeny and geographic origin. A major distinguishing pattern exhibited by seven of eight *M. chokoriensis* isolates was a group of features between 650-800 *m/z*. This pattern was characterized by successive 14 Da differences beginning at 673.5 *m/z* and extending to 771.6 *m/z,* along with sodium and potassium adducts of each, which we attributed to analogs differing in the addition/subtraction of methylene groups (Figure 5). An outlier strain, B001, did not share these features, indicating differential natural products production within this group of highly similar strains. We validated this finding by cultivating biological replicates in a random pattern across a 48-well microwell plate, resulting in eight independent cultures of B001 and four independent cultures of each of the seven remaining strains. We acquired MALDI data for ten replicates of each culture (Figure 6a, b) and confirmed this metabolite production pattern.

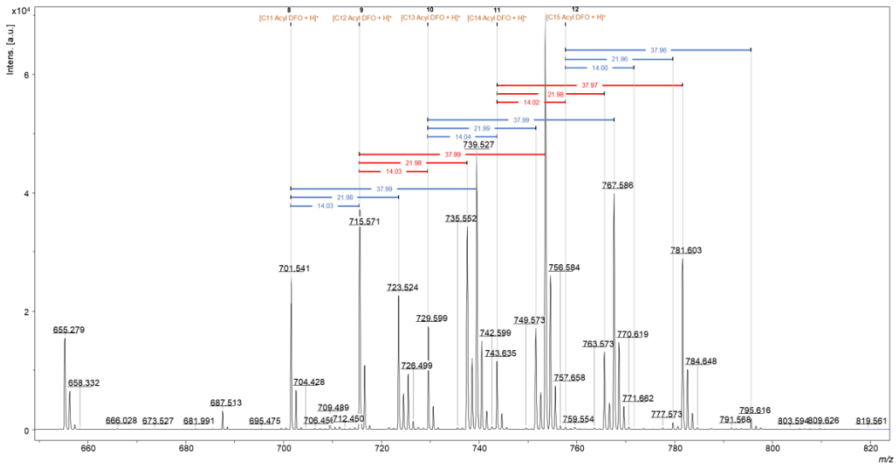


Figure 5. Annotation of *Micromonospora* B031 MALDI spectrum showing intact acyl-desferrioxamine analogs (Spectrum annotation performed by co-author Chase Clark).

To further validate the differential natural products production observed in the MAN, we employed LC-MS/MS analysis for evaluation with the Global Natural Products Social molecular networking platform[68] and comparative metabolomics XCMS analysis[81]. LC-MS/MS afforded an orthogonal means of analysis through chromatographic separation, a different mechanism of ionization, and tandem mass data that helped verify relationships between observed metabolites. Using GNPS, XCMS, and theoretical isotope abundance spectra, we identified a series of acylated desferrioxamine analogs. These belong to a class of siderophores that sequester the essential growth factor ferric iron from the environment. One of these, desferrioxamine B, is used clinically to treat metal poisoning. Interestingly, B001 lacks the capacity to produce these acylated analogs (Figure 6b), and this was readily highlighted upon analysis of the MAN. We also observed several other desferrioxamine analogs produced by these isolates. The LC-MS/MS based observation of strain-specific patterns of siderophore production corroborated our initial MALDI-TOF MS results that showed B001, while able to produce unmodified desferrioxamine B, was deficient in the ability to produce a series of acylated desferrioxamine B analogs (8-12). Interestingly, B001 was isolated from the same 1 cm^3 sediment sample as six of the seven strains that produced this compound series, highlighting that phylogenetic groupings and geographic location are not sufficient indicators of natural products production capacity. Our observations of desferrioxamine biosynthetic pathway promiscuity are consistent with previous studies[82–84]. However, each of these studies required extensive genome sequencing and/or liquid fermentation experiments followed by chromatographic analyses, whereas

we documented and visualized putative phylogenetic relationships *and* intra-species differences in functional chemistry using MALDI-TOF MS/IDBac analysis in a few hours, once each colony appeared on a Petri dish.

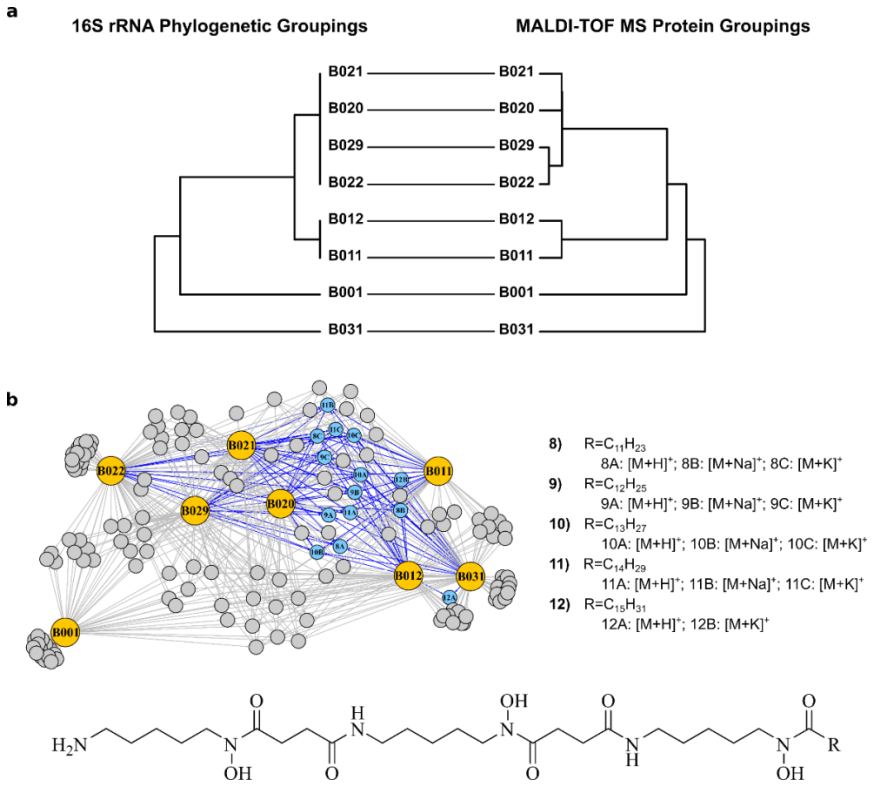


Figure 6. IDBac protein and natural products analysis of isolates from a single *Micromonospora* species.

(a) Tanglegram depicts a high degree of similarity between groupings of 16S rRNA gene sequence identity and MALDI-TOF MS protein data of *M. chokoriensis* isolates. **(b)** MAN of MALDI-TOF MS data from *M. chokoriensis* colonies highlights distinct intra-species differences in natural products production; this is due to differential production of a specific series of acylated desferrioxamine siderophores (shown as blue nodes), which B001 did not to produce.

4.1.3 MALDI-TOF MS protein and natural products of unknown bacteria isolated from a freshwater sponge

In many instances, the taxonomic identity and natural products production capacity of a group of bacteria are not well characterized or completely unknown, particularly in studies involving strains isolated from the environment. Thus, we tested the ability of the IDBac analysis pipeline to

rapidly extract protein and natural products information from unknown environmental bacteria cultivated from a freshwater sponge collected from Lake Huron. We cultivated sponge-associated bacteria from 1 cm^3 of tissue on high nutrient A1 medium (Figure 7a).

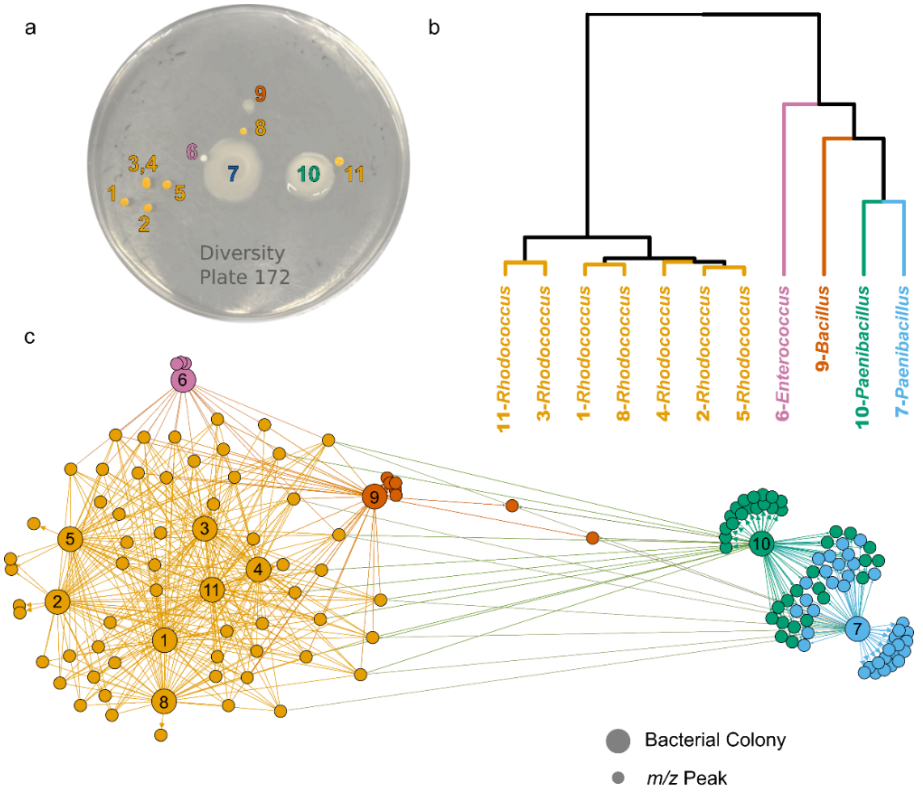


Figure 7. IDBac provides rapid taxonomic and natural products fingerprinting of unknown environmental isolates.

(a) Bacterial diversity plate obtained from placing freshwater sponge tissue on high nutrient A1 agar. **(b)** IDBac allowed for hierarchical clustering of MALDI-TOF MS protein spectra with the option to choose standard distance measures and clustering algorithms. For workflows requiring analysis of hundreds to thousands of strains, this allows for data-reduction *before* natural products analysis is performed. **(c)** MAN, colored via modularity analysis with default thresholds in Gephi[85], allowed for rapid decision-making based on gross *in situ* natural products production after matrix and media signals were subtracted automatically from the network in IDBac. *Rhodococcus* isolates shared a common set of core metabolites, while *Enterococcus* and *Bacillus* isolates exhibited less extensive but unique metabolomes. Significant outliers were the two *Paenibacillus* strains, which produced several shared and unique natural products.

Using a sterile toothpick, we selected all colonies that grew on the plate over a 90-day time period. These colonies were then subjected to MALDI-TOF MS analysis and the data were processed in IDBac. Principal component analysis and unsupervised hierarchical clustering of the bacterial protein range afforded four distinct groupings (Figure 9b). We confirmed that these MS protein groupings aligned with the genera *Enterococcus*, *Bacillus*, *Paenibacillus*, and *Rhodococcus* via 16S rRNA gene sequencing analysis of each of isolate (Figures 8 and 9).

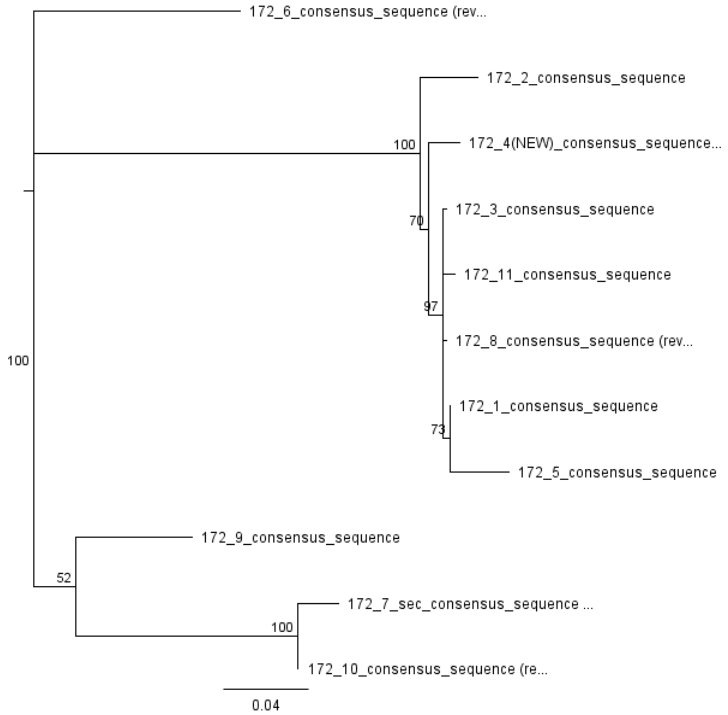


Figure 8. Neighbor-joining phylogenetic tree of diversity plate 172.

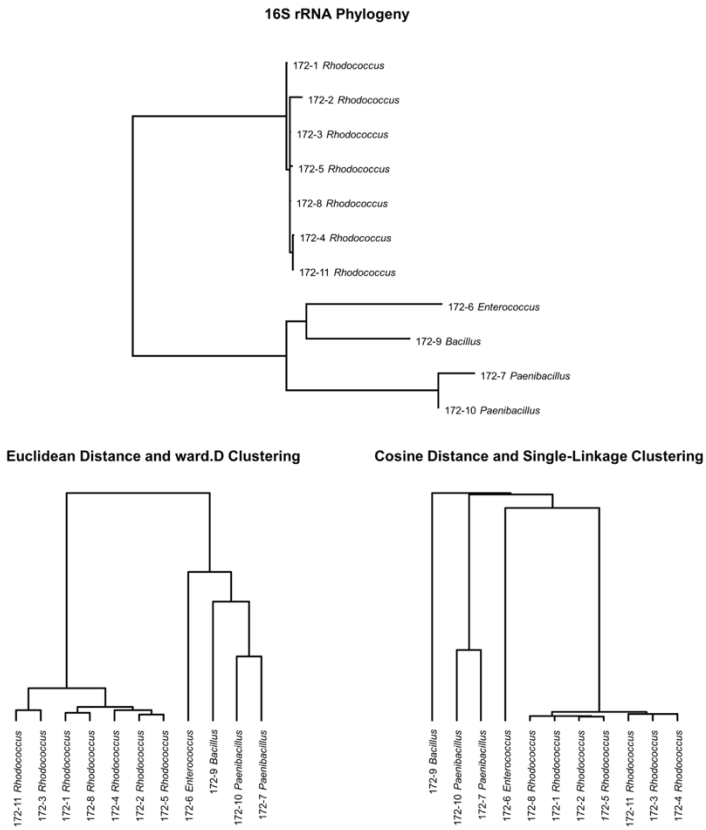


Figure 9. Comparison between 16S rRNA gene sequencing and MALDI protein groupings using Euclidean distance and cosine distance.

We next generated a MAN (Figure 7c) in order to further discriminate these groups based on natural products production. Interestingly, unsupervised subnetworking modularity analysis[86] separated the network into five groups of strains (colored in Figure 7c), which highly correlated to phylogenetic groupings. The seven *Rhodococcus* isolates exhibited a high degree of natural products associations, while the *Bacillus* and *Enterococcus* isolates were classified into unique modules. Importantly, our method quickly highlighted subtle differences in natural products production between two morphologically identical *Paenibacillus* strains and separated the two strains whose 16S rRNA gene sequences share a pairwise similarity of 99.87% (1496/1498 nucleotides)[80]. A close look at their natural products profiles shows subtle differences in production of a series of natural products ranging from *m/z* 900-1250. We wanted to confirm this difference was not due to variations in colony microenvironment, since nearby colonies can affect

natural products production through physical contact and chemical crosstalk[79,87,88]. To determine this, we grew each isolate by itself over three individual cultivation experiments and data acquisition events, and after observation of previously observed natural products patterns in the *m/z* 900-1250 range, we concluded that both isolates contained partially overlapping, but distinct metabolic capacities. These subtle differences in metabolic capacity were readily detected and visualized as a result of IDBac natural products analysis and could not be achieved through analysis of protein groupings alone. In just three hours our pipeline created statistically robust protein profile groupings of environmental isolates (11 strains, 10 MS replicates each, for a total of 110 MALDI spots) that mirrored phylogenetic groupings; it further distinguished colonies with nearly identical morphology and 16S rRNA gene sequence identity based on their capacity to produce natural products.

4.2 A new paradigm to maximize taxonomic and natural products space in microbial libraries using MALDI-TOF MS/IDBac

4.2.1 Collection and processing of environmental samples

In total, 86 samples from 30 locations, consisting of sediment (46), marine invertebrates (25), algae (12), and water column (3), were collected (Table 4). Each environmental sample was processed using common techniques that select for spore-forming and non-spore-forming bacteria, and plated onto five different agar-based nutrient media[89–91]. After incubation of up to 60 days, *every* distinguishable colony growing on nutrient media was isolated (Figure 10).

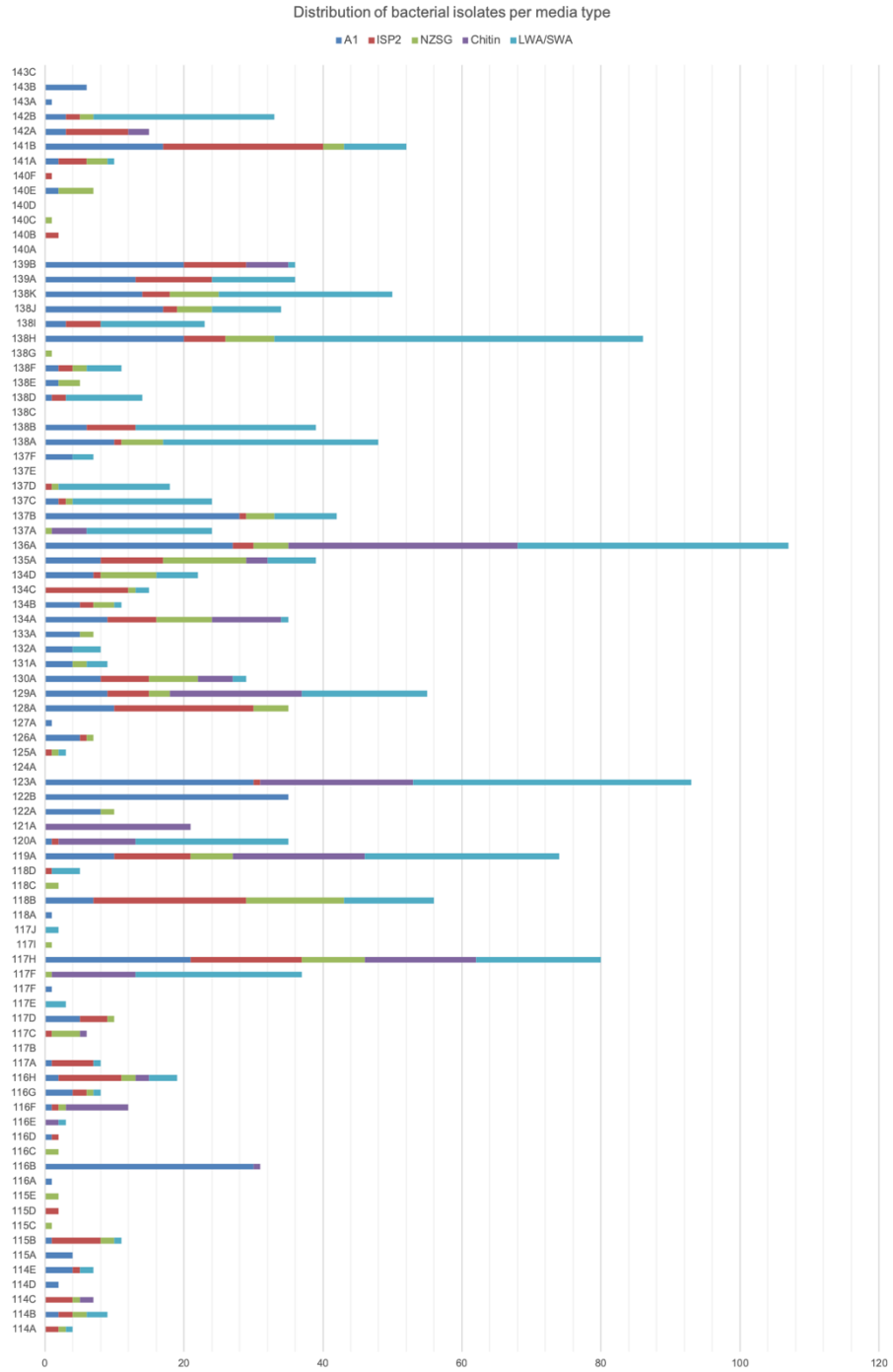


Figure 10. Distribution of bacterial isolates as a function of media type. Samples are listed on the y-axis, and number of bacterial isolates are represented on the x-axis.

All 1,616 isolates were re-plated in individual wells of a 48-well agar plate with high nutrient A1 medium to facilitate comparison using MALDI-TOF MS analysis (marine and freshwater-derived datasets were analyzed separately). Single colonies were then transferred to a MALDI target plate, data were acquired on an Autoflex Speed LRF MALDI-TOF mass spectrometer and analyzed using IDBac. The final step of re-plating, incubating, and acquiring data on bacterial isolates was performed in a three-week span in order to allow for simultaneous growth and triplicate data acquisition for each colony (4,848 total MALDI spots). With advanced instrumentation and automation (bacterial isolation and sample preparation), this *entire* process could conceivably be improved to a few days to one week (Figure 11).

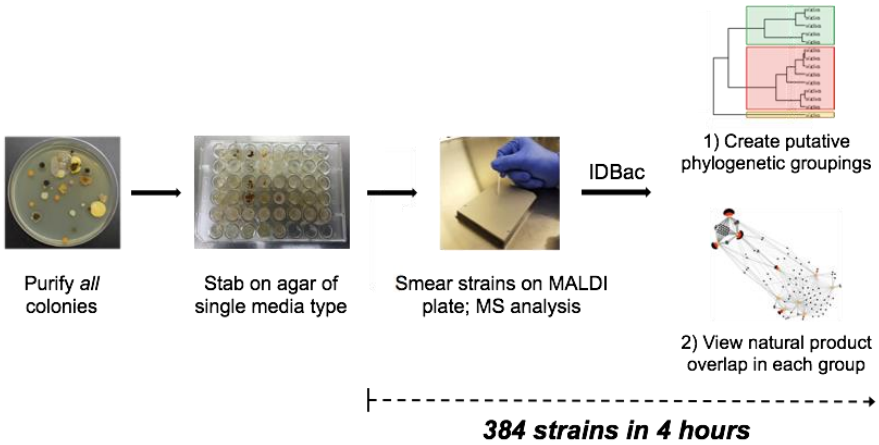


Figure 11. Collection trip workflow.

Above is the recommended order of procedures for minimizing natural product redundancy in a bacterial library. With a few basic laboratory supplies (nutrient agar, multiwell plates, toothpicks) and access to a MALDI-TOF MS, a researcher can build a library that doesn't rely solely on high numbers of entries to maximize natural product chemical space. This IDBac-generated library incorporates information from protein *and* natural product MS analyses, without requiring liquid fermentation, extraction, or chromatographic analyses.

Targeting of *all* distinguishable colonies represents a major deviation from past library generation practices that employed visual inspection of morphology and other selective colony picking techniques to build libraries. The ease of data collection and throughput of our MALDI-TOF MS/IDBac pipeline (sample preparation, data collection and visualization of 384 samples in under 4 hours) allowed us to incorporate high numbers of colonies that

otherwise would not be have been included using our previous protocols, which followed common practices.

4.2.2 Grouping and analysis of bacterial isolates into putative taxa based on MALDI-TOF MS spectra (3,000-15,000 Da)

MALDI-TOF MS is a well-established technique used to identify and group microbial isolates, based on analysis of MS spectra in the 3,000-15,000 Da range[92]. Since resulting MS fingerprints are largely due to the high copy number of ionized ribosomal proteins[57,58], pseudo-phylogenetic relationships can be inferred, often achieving species and sub-species level resolution. Application of this technique has been thoroughly reviewed elsewhere[36,57,93]. In particular, Dieckmann *et al.* used MALDI-TOF MS to dereplicate bacterial isolates from marine sponges, however, only protein spectra were taken into consideration, limiting isolate dereplication to the pseudo-phylogenetic level[94].

The 1,616 Iceland bacterial isolates were first separated into two sets: those isolated from marine and freshwater locations. Each set was grown on media supplemented with and without artificial ocean salts, respectively. Colonies were transferred to steel target plates and MALDI-TOF MS data acquired and subsequently processed in IDBac, as previously described[56].

IDBac offers several settings when creating hierarchical clustering plots based on protein MS spectra. We suggest pairing cosine-distance and average linkage clustering settings for library creation efforts. Average-linkage conserves the 0 to 1 scale of the cosine similarity score, allowing the resulting dendrogram to be more easily interpreted (Figure 12). To confirm that the resulting MALDI MS protein groupings were representative of phylogenetic similarity, we subsampled 361 isolates across both dendrograms (27.5% of freshwater and 17.8% of marine isolates) and sequenced their 16S rRNA genes (Table 5).

We observed that isolate groupings formed as a result of 16S rRNA phylogenetic analysis closely resembled those formed by MALDI-TOF/IDBac analysis (Appendix Note 1). Our results corroborate many previous efforts that document the use of MALDI-TOF MS to facilitate the grouping and/or taxonomic identification of environmental isolates[38,42,93,95].

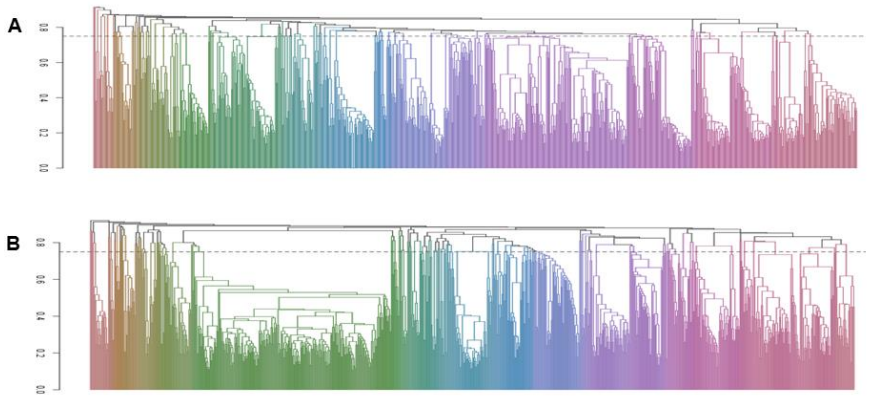


Figure 12. Dendrograms resulting from hierarchical clustering of MALDI-TOF MS protein spectra in IDBac of: 733 freshwater isolates (A), and 883 marine isolates (B). The dendrograms were "cut" at a threshold of 0.75 (dashed lines) to create groups of similar isolates.

It is important to note that while isolates often group at the species-level, the dendrograms created in IDBac are *not* phylogenetic trees. We noticed in our dendrograms that MALDI-TOF MS analysis of bacterial proteins generally results in groupings that represent genus taxonomic level and below but lose fidelity of clustering at higher-taxonomic levels. As recently shown by Seuylemezian *et al.*, while highly similar strains group together based on MALDI protein spectra similarity, the overall topology of clusters was slightly different than those found in a phylogenetic tree constructed using the 16S rRNA gene[96]. In our data, while genera such as *Shewanella*, *Bradyrhizobium*, *Thiomonas,* and *Salinococcus* (among others) form single groupings in the dendrogram, some larger, more diverse genera such as *Bacillus*, *Paenibacillus*, *Streptomyces*, and *Micromonospora* form multiple groupings distributed throughout each dendrogram. To confirm this wasn't due to outside variables (date spectra were acquired, batch of growth media used to purify colonies, etc.), we seeded the freshwater dataset with protein spectra from previously characterized strains. Each of three *Micromonospora* strains from our in-house collection (*Micromonospora aurantiaca* D053 and *M. humi* D077 from marine sediment off of the coast of Massachusetts, and *M. aurantiaca* B017 from freshwater sediment in Lake Michigan) matched with a

corresponding *Micromonospora* grouping in both freshwater and marine datasets (Figure 13). Similar coherence was observed when seeding our freshwater and marine datasets with characterized *Bacillus* strains (*Bacillus subtilis* 3610; *B. subtilis* NRRL B14596; Figure 14)

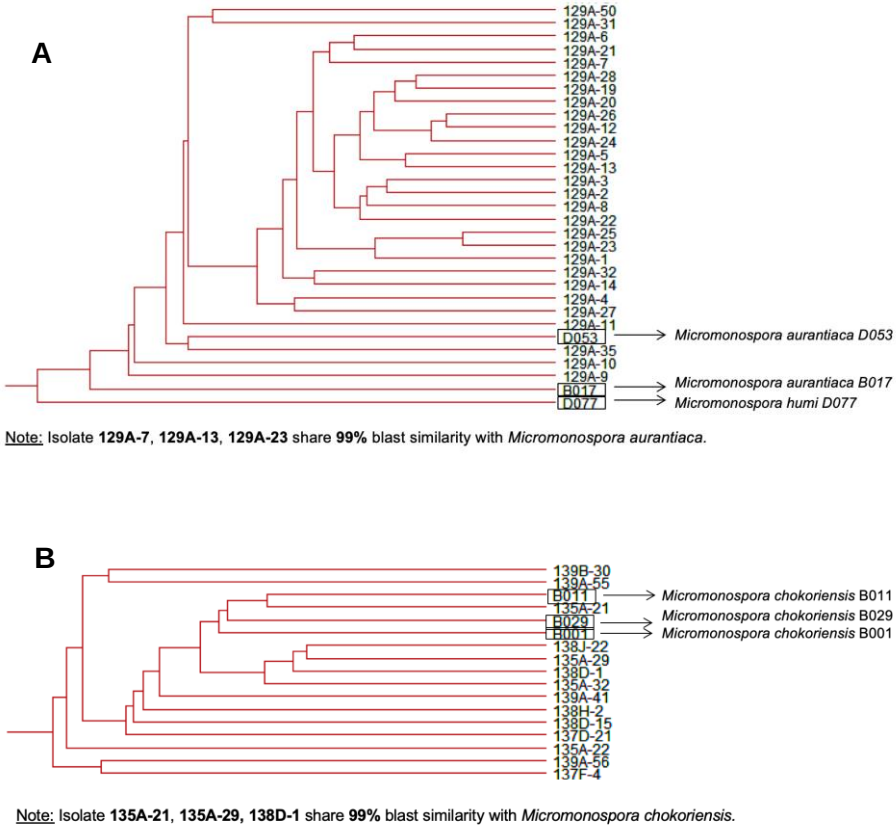


Figure 13. Seeding of freshwater and marine datasets with *Micromonospora* sp. strains.

(A) Seeding of the freshwater dendrogram with three *Micromonospora* strains from our in-house collection (*Micromonospora aurantiaca* D053 and *M. humi* D077 from marine sediment off of the coast of Massachusetts, and *M. aurantiaca* B017 from freshwater sediment in Lake Michigan). **(B)** Seeding of the marine dendrogram with *Micromonospora chokoriensis* B011, *M. chokoriensis* B029 and *M. chokoriensis* B001. Both figures represent an expansion of one section from the larger datase

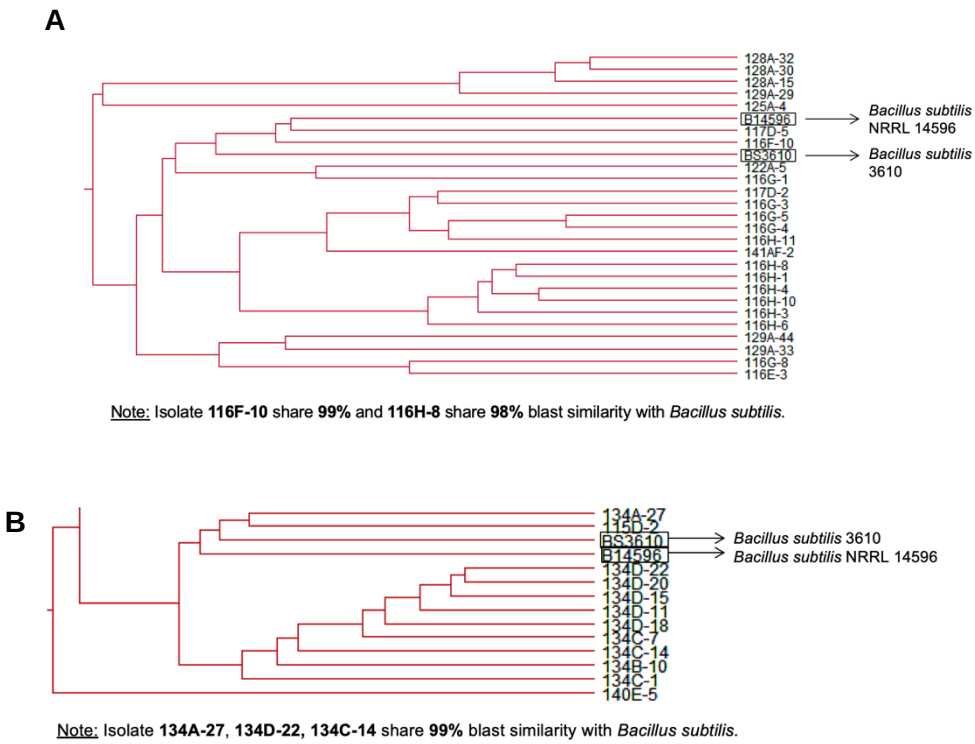


Figure 14. Seeding of freshwater and marine datasets with *Bacillus* sp. strains.

(A) Seeding of the freshwater dendrogram with *Bacillus subtilis* 3610 and *B. subtilis* NRRL B14596. **(B)** Seeding of the marine dendrogram with *Bacillus subtilis* 3610 and *B. subtilis* NRRL B14596. Both figures represent an expansion of one section from the larger dataset.

The spread of these genera across the dendrogram reflects, in some cases, the higher degree of resolution achieved by MALDI-TOF MS, since signals represent ionized proteins from more than just ribosomal origin. It also highlights the limitations of bacterial taxonomy-based methods, since two isolates may share high 16S rRNA sequence homology but low genomic DNA homology, as explained previously[96]. In the presence of a large reference database of protein spectra from characterized environmental microorganisms, the rapid taxonomic identification of unknown isolates is possible. A few studies describe the creation of custom MALDI-TOF protein MS databases to assist in the identification of specific bacterial taxa, however

the absence of an extensive database is a major roadblock to the use of MALDI-TOF MS as a means to identify bacteria isolated from the environment[41,96,97]. Currently, IDBac allows the user to build custom databases that can be readily searched, improved, and shared.

In the context of creating a diverse bacterial library for natural product discovery, although MALDI-TOF/IDBac analysis effectively groups isolates based on ribosomal MS fingerprints, it is not essential that the resulting hierarchical clustering perfectly mirrors phylogenetic groupings. Clustering isolates based on protein MS fingerprints primarily serves as a data reduction strategy for the following step, comparison of natural product production within closely related groups of isolates. Occasionally an isolate "mismatches" to a cluster that contains other genera of bacteria (e.g. a *Bacillus* isolate is paired with a group of *Streptomyces* isolates); this is generally due to the absence of close relatives in the dataset and a number of shared peaks emanating from non-ribosomal proteins. Regardless, in our experience it will stand as an obvious outlier when comparing natural product overlap between isolates within that cluster.

In the case of the Iceland protein MS dataset, a threshold of 0.75 was used to create 72 distinct freshwater and 101 distinct marine isolate groupings. Groups contained anywhere between 1 and 229 isolates. Each group was then manually selected within IDBac to generate a metabolite association network (MAN). MANs were used as the basis for selecting isolates with minimal overlapping natural product production for addition to the library. Strategies to analyze natural product overlap within these groups are discussed in the next section.

4.2.3 Use of Metabolite Association Networks (MANs) to reduce inter- and intra-species natural product redundancy in a bacterial library

It has been demonstrated that natural products production, broadly speaking, is conserved within individual genera[98,99]. However, coexisting with this phenomenon, some genera have exhibited extensive inter-species natural product biosynthetic diversity[100]. Supporting this, it is well known that on a species-level scale, highly similar strains can display a differential capacity to produce natural products, often due to biosynthetic gene clusters that were horizontally acquired[99,100]. We previously demonstrated a similar phenomenon in a group of *Micromonospora* strains isolated from the same 1 cm^3 of sediment[56]. In the context of IDBac, grouping bacteria by taxa is an effective data reduction strategy, and generally speaking, facilitates access to

different areas of chemical space. However, analysis of the MAN within each group is an effective strategy for revealing more subtle genus/species-level differences in natural product production, and accounts for both of the prior situations. As shown in Figures 15, larger nodes represent individual bacterial isolates, while smaller nodes represent *m/z* values detected in their corresponding MS spectra (200-2,000 Da). Matrix and media ions were subtracted from the resulting analysis. We previously demonstrated that the majority of *m/z* values in a MAN between two *Bacillus* strains (33 of 42; 78.6%) represented natural products, giving confidence in the ability of IDBac to filter out matrix signals[56]. It is worth nothing that the mass resolving power of MALDI-TOF MS does not allow for level -1 or -2 identifications of natural products[101]; further validation is necessary to elucidate a structure.

Since it is difficult to observe meaningful intra-species natural product production relationships in a MAN of all 1,616 isolates, it is recommended to view MANs in smaller subsets based on hierarchical clustering results (*e.g.*, groups that range from 10-50 isolates). Based on our data, groups between 1 and 50 isolates were optimal, though this range may vary depending on several factors within a dataset and user needs. It should be noted that "hairball" MANs generated when viewing large numbers of strains might be mitigated with high-resolution, high-mass accuracy MALDI-TOF MS instruments.

The purpose of using MANs is to make an informed judgement and quickly assess the natural product capacity of two seemingly identical bacterial colonies. How many colonies with similar morphology should be picked from the same collection site? From different collection sites? These fundamental questions have immense downstream effects on a microbial library sourced for drug discovery, and to date, researchers have been employing costly or dated technologies to address them. Figure 15 depicts four scenarios that would be expected to occur in the environment from any pair of isolates that share >99% 16S rRNA identity. Neither visual inspection or taxonomic classification strategies would inform the researcher about which scenario is occurring. IDBac provides this information and allows the educated choice of which bacterial isolates should be added to a library.

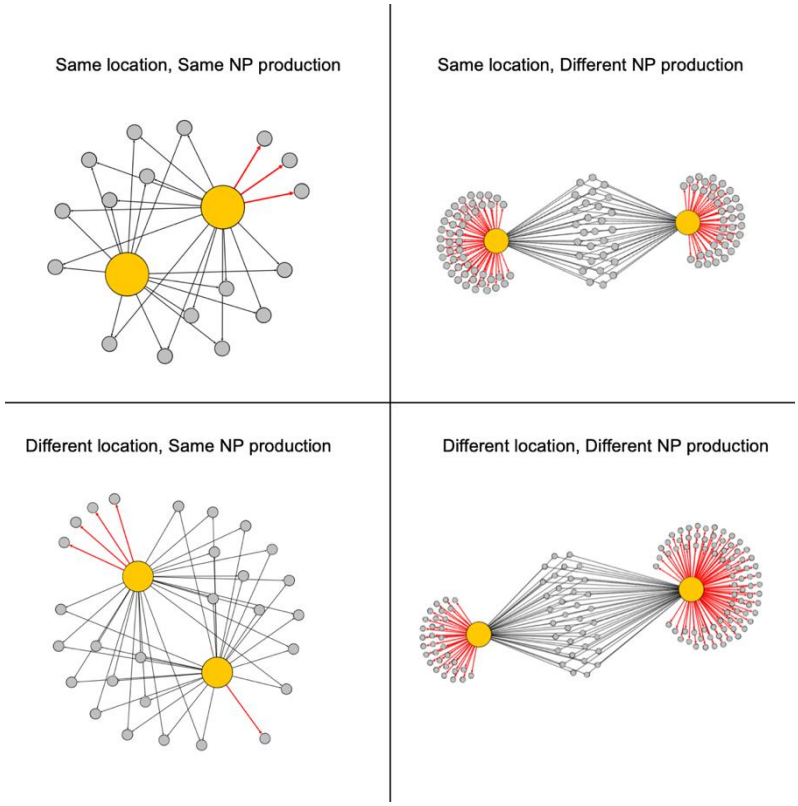


Figure 15. Four scenarios are possible when comparing environmental bacterial isolates on nutrient agar that are morphologically indistinguishable and share >99% 16S rRNA similarity.

IDBac allows the user to rapidly distinguish between these scenarios and make an informed decision as to whether colonies produce significantly overlapping populations of natural products. Large orange nodes represent individual bacterial isolates, while smaller gray nodes represent *m/z* values detected in their corresponding MS spectra (200-2,000 Da). "NP" denotes natural product.

We built our microbial library based on the analysis of MANs generated from 72 freshwater and 101 marine isolate groupings created from protein MS data. Each dendrogram grouping and corresponding MAN is provided in the Appendix Notes 2 and 3. It is up to the individual researcher and available resources to determine roughly how many isolates they wish to add to their library. Figure 16 depicts a grouping of 16 isolates (estimated to be *Bacillus* spp. from 16S rRNA gene sequencing of 116H-8 and 116F-10). The nine isolates clustered in pink exhibited a high degree of natural product overlap, therefore two isolates (116G-4 and 116H-1) were selected for addition to the library. The rest were discarded. Alternatively, isolates 116F-10 and 116H-8

showed a high degree of unique natural product signatures (non-overlapping with other isolates in the protein MS grouping), so they were also included in the library. Isolates 116G-1 and 122A-5, as well as 117D-5 and 117D-2 are a more ambiguous case. Although some natural product overlap exists, all strains in these groups contain a set of *m/z* values that are unique, which may represent a molecule(s) that exists exclusively in one bacterium over another, or simply differences in MS adducts. The researcher is now empowered to make an informed decision of whether to include one, both, or neither of the pair.

After approximately 25 hours of data acquisition and 2 hours of analysis, 301 isolates were selected for addition to the library out of a total 1,616. The 301 isolates span approximately 54 genera, which is an exceptionally high taxonomic diversity given the relatively small size of the library. In particular, included within the library are 25 isolates from 16 genera of the phylum Actinobacteria, that in total have an estimated 11 publications to date that report a natural product (table 6).

Table 6. Understudied actinomycete genera represented in the finalized Iceland library.

Genus/ Family	Estimated # publications reporting a natural product	References
Aeromicrobium sp.	2	Reeves et al., 2008[102]; Brikun et al., 2004[103]
Blastococcus sp.	0	-
Brachybacterium sp.	1	Kiran et al., 2014[104]
Cryobacterium sp.	0	-
Friedmanniella sp.	0	-
Janibacter sp.	1	Asolkar et al., 2004[105]
Knoellia sp.	0	-
Kytococcus sp.	1	Pospisil et al., 1998[106]
Leucobacter sp.	0	-

Marmoricola sp.	0	-
Oerskovia sp.	3	Bagiyan et al., 1997[107]; Yamada et al., 1977[108]; Yamada et al., 1976[109]
Phycicoccus sp.	0	-
Pseudoarthrobacter sp.	0	-
Rathayibacter sp.	1	Murray et al., 2017[110]
Kineosporiaceae	0	-
Williamsia sp.	2	Gao et al., 2017[111]; Xie et al., 2016[112]

Many of these understudied genera present colony morphologies that are atypical of traditional actinomycetes, which are often identified by their hard, leathery appearance, and presence of spores. Unbiased selection of *all* colonies in the early stages of bacterial isolation is credited with the inclusion of these in the library, and greatly enhances the available taxonomic and natural product space in future biological screening efforts.

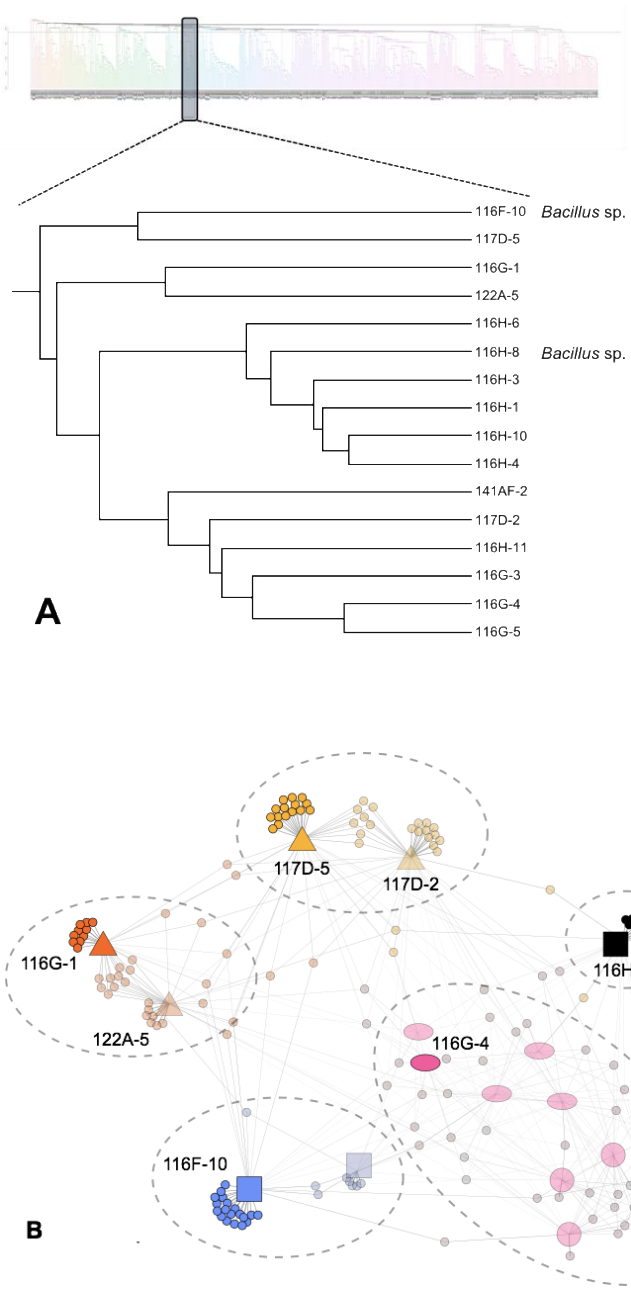


Figure 16. (A) Protein MS clustering of 16 bacterial isolates **(B)** MAN generated for the group of selected strains, isolates colored by modularity score. Isolates chosen to be included in the library are colored in bold; transparent strains were discarded.

4.2.4 Use of IDBac to reduce redundancy in existing bacterial libraries

We performed three previous expeditions to Iceland in 2013, 2014, and 2015. After each expedition, traditional morphology-guided colony picking strategies were employed (as opposed to picking *all* colonies in an unbiased manner). In total, the three previous expeditions yielded a library of 833 bacterial isolates. In order to assess the degree of redundancy in this library, the MALDI-TOF/IDBac pipeline was applied. Using the parameters and cutoffs described in the previous section, 600 isolates were discarded due to significant taxonomic and natural product redundancy, reducing the library to 233 isolates, a 72.0% reduction in library size. This reduction has significant downstream cost savings. There are several major obstacles to creating a microbial library for drug discovery, and these have been well documented[49,61,113]. Generally speaking, therapeutic discovery pipelines have relied on libraries of natural product extracts or fractions, a large percent of which were sourced from libraries of microorganisms. Each isolate in a library is processed through purification, liquid fermentation, extraction, and in some cases chromatographic separation to generate natural product material for the screening library. This process is costly and time consuming, as it may require tens to hundreds of thousands of dollars of investment to generate. Alternatively, the MALDI-TOF/IDBac pipeline provides a relatively rapid, easy, and cost-effective avenue for microbial library generation. It also precludes the need for DNA sequencing of isolates to determine phylogenetic similarity or chromatographic analyses of extracts or fractions to assess natural product overlap. Importantly, reducing the number of isolates in a library is a major cost saving step.

4.2.5 IDBac visualizes relationships based on user defined categories

IDBac was also designed to facilitate post-hoc analyses of field collections by providing visualization across the protein MS dendrograms. These visualizations occur in two forms, either in coloring the labels of the dendrogram or by graphing marks beside the dendrogram (Figure 17). In either case, users are able to associate metadata about isolates (e.g. geographic coordinates, sample source, water temperature, isolation media, etc.) to the dendrogram in order to quickly discern patterns across tens to thousands of isolates. For example, following the Iceland expedition we were able to quickly evaluate genus and species-level patterns of occurrence as a function of media formulation (Figure 17). This allows us to design more targeted and untargeted bacterial isolation procedures for future collection expeditions.

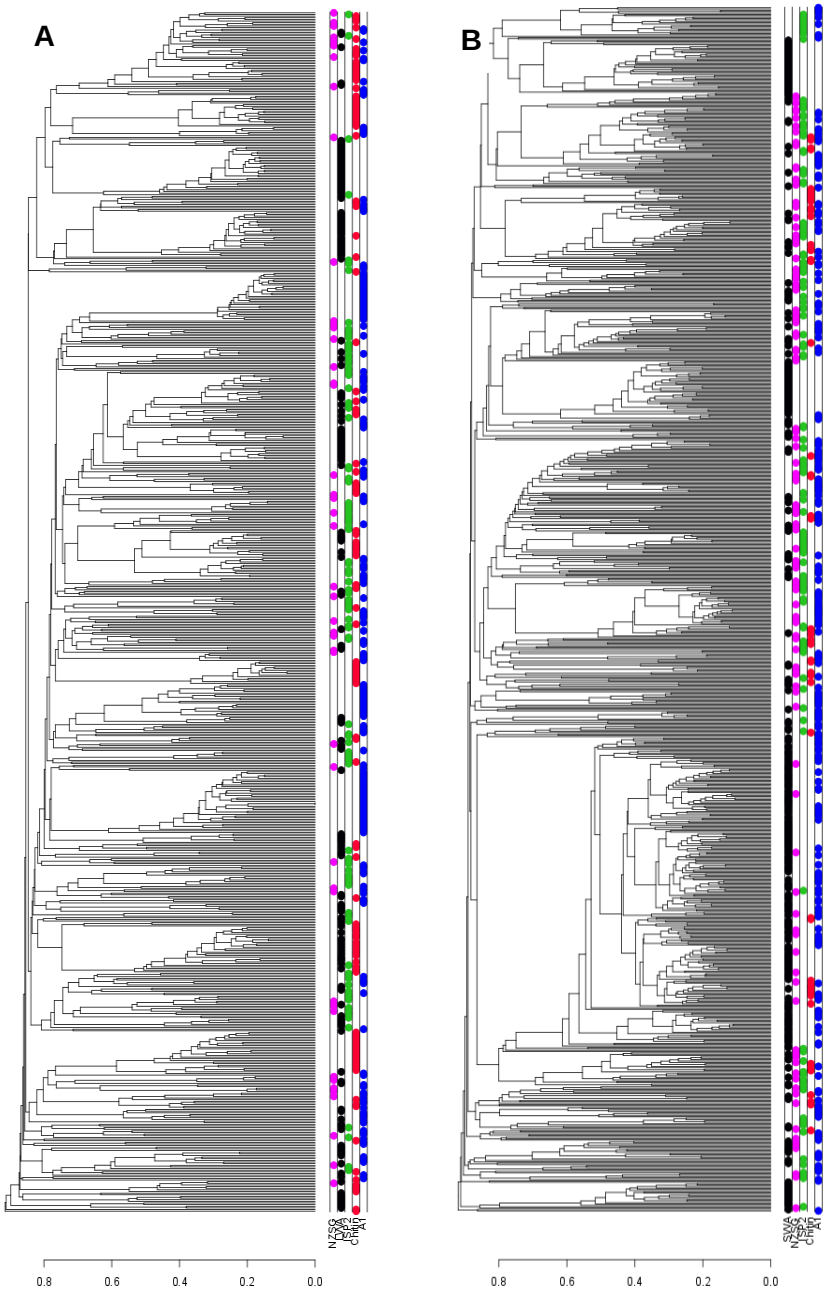


Figure 17. Association of bacterial isolates and isolation media.

(A) Freshwater dendogram. **(B)** Marine dendogram. This output of IDBac displays which of five media types that each of the isolates was isolated from. This is particularly useful when wanting to explore the selectivity of a media type for a specific taxon.

5 Summary and conclusions

In this thesis, we present a method and analysis pipeline to acquire and integrate MALDI-TOF MS data of *both* intact proteins and natural products from single bacterial colonies. Our technique is unique when compared to other MS-based methods that are commonly employed to analyze natural products. IDBac rapidly provides a data acquisition and analysis platform to group single bacterial colonies based on putative taxonomic identity, and further discriminates these groups using MS patterns of functional natural products. No existing MS platform provides the ability to analyze both a diverse array of proteins and non-volatile metabolites at the throughput of hundreds of strains in just a few hours.

One aspect of the IDBac pipeline that provides significant advantages over existing MS platforms is that it affords information on colony pseudo-phylogenetic identity and natural products production without the need for extraction and chromatographic analyses. The latter techniques rely on growing pure bacterial isolates (often in liquid culture), generating extracts, and using liquid chromatography analyses (generally coupled to MS), to separate and detect natural products. This is often a laborious, costly, and relatively time-consuming process.

IDBac is complementary to other innovative platforms such as GNPS, which aids in the dereplication of previously characterized and identification of potentially new natural products[68,114]. Successful integration of these orthogonal approaches was demonstrated in our analysis of *M. chokoriensis* isolates. We used IDBac to rapidly identify and discriminate between nearly identical isolates from a similar environment and detect subtle differences in siderophore production between isolates within the same species grouping. Subsequent growth of these strains, extraction, and LC-MS/MS analyses allowed us to use GNPS platform to dereplicate the structures of several iron scavenging desferrioxamine analogs. This highlights the individual strengths of orthogonal MS-based approaches to extract complementary information from a biological system.

Through the course of our studies we have documented a few potential limitations of our method. First, although IDBac can create accurate sub-species groupings of bacteria based on protein MS fingerprints, species-level identification is only possible in the presence of a searchable and extensive

protein fingerprint database. A greater community effort is required to document MALDI-TOF MS protein fingerprints into a publicly available database to maximize phylogenetic coverage. Such an effort would facilitate the rapid identification of unknown environmental bacteria, and elevate this process to be on par with existing commercial platforms used to identify clinical pathogens[41]. An effort to make a readily searchable/freely available bacterial protein MS database is ongoing in the Murphy laboratory and the first version of IDBac supports this mission by providing a simple interface for bundling and converting proprietary vendor-format data files to the open-access mzXML format for easy, reproducible sharing.

A second limitation is that bacterial natural products production is sensitive to external factors such as growth media, temperature, unintended microbial contamination, and the proximity of other microorganisms on a Petri dish[115]. For comparison of natural products profiles to be reliable, it is imperative that strains be cultivated under identical conditions and that several biological/technical replicates are acquired for each isolate. Fortunately, the throughput of our pipeline allows for multiple replicates of a sample to be analyzed rapidly. Conversely, MALDI-TOF MS protein fingerprints of bacterial colonies are robust and exhibit minimal fluctuation when the data are acquired on different growth media.

Finally, our method is limited by the resolution of the MALDI-TOF MS and would be improved through use of instruments with higher resolving power capabilities. High resolution instruments, such as MALDI-FT-ICR, and fragmentation data from a MALDI-TOF/TOF MS could aid in preliminary compound class dereplication via generation of accurate molecular formulas and structural information, respectively. However, this would increase the length of the experiments, data storage requirements, and would require additional technical expertise.

Nearly two decades ago, Goodfellow and colleagues wrote extensively on the importance of ensuring taxonomic diversity in microbial libraries. In summarizing the techniques needed to accomplish this, they noted the potential of MS techniques (pyrolysis MS in particular) to ensure rapid intraspecies pseudo-taxonomic dereplication of colonies growing directly on isolation plates. They opined that such a method would permit “the rational collection of colonies” and reduce “the requirement for time-consuming laboratory testing to distinguish duplicate colonies[113].” Almost a century has passed since researchers realized the potential of microorganisms to produce natural products, and despite advances such as colony picking

automation and DNA sequencing, researchers continue to create libraries that are plagued with redundancy because there has been no cost-effective way to rapidly assess both the bacterial identity and the natural product capacity of colonies in high throughput. In this thesis, we employed our freely available pipeline (IDBac – https://chasemc.github.io/IDBac) to group thousands of environmental bacterial isolates based on similarities in protein MS spectra, and coupled this with global comparison of their natural product profiles. When applied toward the process of generating a microbial library from environmental samples, this represents a significant advance towards removing the serendipity that in the past resulted in a large degree of taxonomic and natural product redundancy, which consequently had serious impacts toward the efficiency of drug discovery programs. Our MALDI-TOF MS/IDBac pipeline allows for the rapid generation of highly diverse microbial 'smart' libraries that do not necessitate high numbers of strains. Rather, it allows emphasis to be placed on minimizing entries into the library, significantly decreasing costs associated with library generation, and simultaneously maximizing available chemical space that enters into downstream biological screening efforts.

Finally, in fulfillment of Icelandic Research Fund grant #152336-051-3, we created a resource for the University of Iceland research community. We created a microbial library with a total of 534 strains spanning at least 54 different genera. The strains are from 183 different locations around Iceland and from a diverse variety of sources (algae, sediments, sponges, tunicates, jellyfish, nudibranchs, water column). All the samples were collected in the time period from 2013-2017 (March 2013, May/June 2014, August/September 2015, May 2017). These expeditions included sea excursions with the Marine Research Institute and SCUBA diving. The microbial library is kept in Faculty of Pharmaceutical Sciences – Hagi (Hofsvallagata 53, 107 Reykjavik). For more details on access to the library, Professor Sesselja Ómarsdóttir (sesselo@hi.is) can be contacted.

References

1. Cragg, G. M. & Newman, D. J. Plants as a source of anti-cancer agents. *Journal of Ethnopharmacology* **100**, (1–2) 72–79 (2005).

2. Cragg, G. M. & Newman, D. J. Natural products: A continuing source of novel drug leads. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1830**, (6) 3670–3695 (2013).

3. Gould, K. Antibiotics: from prehistory to the present day. *Journal of Antimicrobial Chemotherapy* **71**, (3) 572-575 (2016).

4. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the 30 years from 1981 to 2010. *Journal of Natural Products* **75**, (3) 311–335 (2012).

5. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs from 1981 to 2014. *Journal of Natural Products* **79**, (3) 629–661 (2016).

6. Lam, K. S. New aspects of natural products in drug discovery. *Trends in Microbiology* **15**, (6) 279–289 (2007).

7. Fenical, W. & Jensen, P. R. Developing a new resource for drug discovery: marine actinomycete bacteria. *Nature Chemical Biology* **2**, (12) 666–673 (2006).

8. Manivasagan, P., Kang, K.-H., Sivakumar, K., Li-Chan, E. C. Y., Oh, H.-M. & Kim, S.-K. Marine actinobacteria: An important source of bioactive natural products. *Environmental Toxicology and Pharmacology* **38**, (1) 172–188 (2014).

9. Jensen, P. R., Mincer, T. J., Williams, P. G. & Fenical, W. Marine actinomycete diversity and natural product discovery. *Antonie Van Leeuwenhoek* **87**, (1) 43–48 (2005).

10. Singh, R. K., Tiwari, S. P., Rai, A. K. & Mohapatra, T. M. Cyanobacteria: an emerging source for drug discovery. *The Journal of Antibiotics* **64**, (6) 401–412 (2011).

11. El-Elimat, T., Zhang, X., *et al.* Chemical diversity of metabolites from fungi, cyanobacteria, and plants relative to FDA-approved anticancer agents. *ACS Medicinal Chemistry Letters* **3**, (8) 645–649 (2012).

12. Chlipala, G. E., Mo, S. & Orjala, J. Chemodiversity in freshwater and terrestrial cyanobacteria - a source for drug discovery. *Current Drug Targets* **12**, (1) 1654–73 (2011).

13. Donia, M. S. & Fischbach, M. A. Human Microbiota. Small molecules

from the human microbiota. *Science (New York, N.Y.)* **349**, (6246) 1254766 (2015).

14. Stein, T. *Bacillus subtilis* antibiotics: structures, syntheses and specific functions. *Molecular Microbiology* **56**, (4) 845–857 (2005).

15. Gupta, R. S. The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. *FEMS Microbiology Reviews* **24**, (4) 367–402 (2000).

16. Baltz, R. H. Gifted microbes for genome mining and natural product discovery. *Journal of Industrial Microbiology & Biotechnology* **44**, (4–5) 573–588 (2017).

17. Ingólfsson, A. The distribution of intertidal macrofauna on the coasts of iceland in relation to temperature. *Sarsia* **81**, (1) 29–44 (1996).

18. Ulrich, J. Der Formenschats des Meerebodens-Neuere Forschungsergebnisse über das submarine Bodenrelief. *Geographische Rundhau* **15**, 136–148 (1963).

19. Malmberg, S. A. The Iceland Basin-Topography and Oceanographic Features. *Marine Research Institute, Reykjavik* **109**, 1–13 (2004).

20. Malmberg, S. A. M. G. Sea surface temperature and salinity in south Icelandic waters in period 1868-196. *Marine Research Institute, Reykjavik* 1–13 (1982).

21. German, C.R.B., J.; Chin, C.; Danielsen, M., Holland, S.; James, R.; Jonsdottir, A.; Ludford, E.; Moser, C.; Olafsson, J.; Palmer, M. R.; Rudnicki, M. D. Hydrothermal activity on the Reykjanes Ridge: the Steinahóll vent-field at 63°06´N. *Earh Planetary Science Letters* **121**, 647–654 (1994).

22. Ernst, G.G.J.C., R. R.; German, C. R.; Palmer, M. R.; Sparks, R. S. J. Vertical and lateral splitting of a hydrothermal plume at Steinaholl, Reykjanes Ridge, Iceland. *Earth Planetary Science Letters* **170**, 529–537 (2000).

23. Benjamínsson, J. Jarðhiti í sjó og flæðarmáli við Ísland. *Náttúrufræðingurin* **58**, 153–169 (1988).

24. Marteinsson, V. T., Kristjánsson, J. K., *et al.* Discovery and description of giant submarine smectite cones on the seafloor in Eyjafjordur, northern Iceland, and a novel thermal microbial habitat. *Applied and Environmental Microbiology* **67**, (2) 827–33 (2001).

25. Morritt & Ingólfsson Upper thermal tolerances of the beachflea *Orchestia gammarellus* (Pallas) (Crustacea: Amphipoda: *Talitridae*) associated with hot springs in Iceland. *Journal of Experimental Marine Biology and Ecology* **255**, (2) 215–227 (2000).

26. Geptner, A., Kristmannsdóttir, H., Kristjansson, J. & Marteinsson, V.

Biogenic saponite from an active submarine hot spring, Iceland. *Clays and Clay Minerals* **50**, (2) 174–185 (2002).

27. Woese, C. R. Bacterial evolution. *Microbiological Reviews* **51**, (2) 221–71 (1987).

28. Yarza, P., Yilmaz, P., *et al.* Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology* **12**, (9) 635–645 (2014).

29. Vandamme, P., Gillis, M., De Vos, P., Kersters, K. & Swings, J. Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiological Reviews* **60**, (2) 407–438 (1996).

30. Patin, N. V, Duncan, K. R., Dorrestein, P. C. & Jensen, P. R. Competitive strategies differentiate closely related species of marine actinobacteria. *The Multidisciplinary Journal of Microbial Ecology* (*ISME Journal)* **10**, (2) 478–490 (2016).

31. Ziemert, N., Lechner, A., Wietz, M., Millán-Aguiñaga, N., Chavarria, K. L. & Jensen, P. R. Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. *Proceedings of the National Academy of Sciences of the United States of America* **111**, (12) E1130-9 (2014).

32. Penn, K., Jenkins, C., *et al.* Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. *The Multidisciplinary Journal of Microbial Ecology (ISME Journal)* **3**, (10) 1193–1203 (2009).

33. Jensen, P. R. Natural Products and the Gene Cluster Revolution. *Trends in Microbiology* **24**, (12) 968–977 (2016).

34. Karas, M. & Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry* **60**, (20) 2299–2301 (1988).

35. Tanaka, K., Waki, H., *et al.* Protein and polymer analyses up to *m/z* 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry* **2**, (8) 151–153 (1988).

36. Sandrin, T. R., Goldstein, J. E. & Schumaker, S. MALDI TOF MS profiling of bacteria at the strain level: A review. *Mass Spectrometry Reviews* **32**, (3) 188–217 (2013).

37. Cain, T. C., Lubman, D. M., Weber, W. J. & Vertes, A. Differentiation of bacteria using protein profiles from matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry* **8**, (12) 1026–1030 (1994).

38. Holland, R. D., Wilkes, J. G., *et al.* Rapid identification of intact whole bacteria based on spectral patterns using matrix-assisted laser

desorption/ionization with time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry : RCM* **10**, (10) 1227–32 (1996).

39. Maier, T., Klepel, S., Renner, U. & Kostrzewa, M. Fast and reliable MALDI-TOF MS–based microorganism identification. *Nature Methods* **3**, (4) 68–71 (2006).

40. Dubois, D., Grare, M., Prere, M.-F., Segonds, C., Marty, N. & Oswald, E. Performances of the Vitek MS matrix-assisted laser desorption ionization-time of flight mass spectrometry system for rapid identification of bacteria in routine clinical microbiology. *Journal of Clinical Microbiology* **50**, (8) 2568–2576 (2012).

41. Rahi, P., Prakash, O. & Shouche, Y. S. Matrix-assisted laser desorption/ionization time-of-flight mass-spectrometry (MALDI-TOF MS) based microbial identifications: challenges and scopes for microbial ecologists. *Frontiers in Microbiology* **7**, 1359 (2016).

42. Popović, N. T., Kazazić, S. P., Strunjak-Perović, I. & Čož-Rakovac, R. Differentiation of environmental aquatic bacterial isolates by MALDI-TOF MS. *Environmental Research* **152**, 7–16 (2017).

43. Munteanu, B. & Hopf, C. Emergence of whole-cell MALDI-MS biotyping for high-throughput bioanalysis of mammalian cells? *Bioanalysis* **5**, (8) 885–893 (2013).

44. Cassagne, C., Normand, A.-C., L'Ollivier, C., Ranque, S. & Piarroux, R. Performance of MALDI-TOF MS platforms for fungal identification. *Mycoses* **59**, (11) 678–690 (2016).

45. Teramoto, K., Sato, H., *et al.* Phylogenetic classification of *Pseudomonas putida* strains by MALDI-MS using ribosomal subunit proteins as biomarkers. *Analytical Chemistry* **79**, (22) 8712–8719 (2007).

46. Arnold, R. J. & Reilly, J. P. Observation of *Escherichia coli* ribosomal proteins and their posttranslational modifications by mass spectrometry. *Analytical Biochemistry* **269**, (1) 105–112 (1999).

47. Silva, R., Lopes, N. P. & Silva, D. B. Application of MALDI Mass Spectrometry in Natural Products Analysis. *Planta Medica* **82**, 671–689 (2016).

48. Bérdy, J. Bioactive Microbial Metabolites. *The Journal of Antibiotics* **58**, (1) 1–26 (2005).

49. Katz, L. & Baltz, R. H. Natural product discovery: past, present, and future. *Journal of Industrial Microbiology & Biotechnology* **43**, (2–3) 155–176 (2016).

50. Wohlleben, W., Mast, Y., Stegmann, E. & Ziemert, N. Antibiotic drug

discovery. *Microbial Biotechnology* **9**, (5) 541–548 (2016).

51. Alexander Fleming Discovery and Development of Penicillin - Landmark - American Chemical Society. at <https://www.acs.org/content/acs/en/education/whatischemistry/landmarks/flemingpenicillin.html>

52. Larsen, T. O., Smedsgaard, J., Nielsen, K. F., Hansen, M. E. & Frisvad, J. C. Phenotypic taxonomy and metabolite profiling in microbial drug discovery. *Natural Product Reports* **22**, (6) 672–695 (2005).

53. Genilloud, O., González, I., Salazar, O., Martín, J., Tormo, J. R. & Vicente, F. Current approaches to exploit actinomycetes as a source of novel natural products. *Journal of Industrial Microbiology & Biotechnology* **38**, (3) 375–389 (2011).

54. Koehn, F. E. & Carter, G. T. The evolving role of natural products in drug discovery. *Nature Reviews Drug Discovery* **4**, (3) 206–220 (2005).

55. Knight, V., Sanglier, J.-J., *et al.* Diversifying microbial natural products for drug discovery. *Applied Microbiology and Biotechnology* **62**, (5–6) 446–458 (2003).

56. Clark, C. M., Costa, M. S., Sanchez, L. M. & Murphy, B. T. Coupling MALDI-TOF mass spectrometry protein and specialized metabolite analyses to rapidly discriminate bacterial function. *Proceedings of the National Academy of Sciences of the United States of America* **115**, (19) 4981–4986 (2018).

57. Ryzhov, V. & Fenselau, C. Characterization of the protein subset desorbed by MALDI from whole bacterial cells. *Analytical Chemistry* **73**, (4) 746–750 (2001).

58. Fernández-Álvarez, C., Torres-Corral, Y. & Santos, Y. Use of ribosomal proteins as biomarkers for identification of *Flavobacterium psychrophilum* by MALDI-TOF mass spectrometry. *Journal of Proteomics* **170**, 59–69 (2018).

59. Crüsemann, M., O'Neill, E. C., *et al.* Prioritizing natural product diversity in a collection of 146 bacterial strains based on growth and extraction protocols. *Journal of Natural Products* **80**, (3) 588–597 (2017).

60. Hubert, J., Nuzillard, J.-M. & Renault, J.-H. Dereplication strategies in natural product research: How many tools and methodologies behind the same concept? *Phytochemistry Reviews* **16**, (1) 55–95 (2017).

61. Liu, X., Ashforth, E., *et al.* Bioprospecting microbial natural product libraries from the marine environment for drug discovery. *The Journal of Antibiotics* **63**, (8) 415–422 (2010).

62. Floros, D. J., Jensen, P. R., Dorrestein, P. C. & Koyama, N. A metabolomics guided exploration of marine natural product chemical space. *Metabolomics* **12**, (9) 145 (2016).

63. Freiwald, A. & Sauer, S. Phylogenetic classification and identification of bacteria by mass spectrometry. *Nature Protocols* **4**, (5) 732–742 (2009).

64. Schumann, P. & Maier, T. MALDI-TOF Mass spectrometry applied to classification and identification of bacteria. *Methods in Microbiology* **41**, 275–306 (2014).

65. Pruesse, E., Peplies, J. & Glöckner, F. O. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, (14) 1823–1829 (2012).

66. Bligh, E. G. & Dyer, W. J. A rapid method of total lipid extraction and purification. *Canadian Journal of Biochemistry and Physiology* **37**, (8) 911–917 (1959).

67. Goering, A. W., McClure, R. A., *et al.* Metabologenomics: correlation of microbial gene clusters with metabolites drives discovery of a nonribosomal peptide with an unusual amino acid monomer. *ACS Central Science* **2**, (2) 99–108 (2016).

68. Wang, M., Carver, J. J., *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **34**, (8) 828–837 (2016).

69. Tautenhahn, R., Patti, G. J., Rinehart, D. & Siuzdak, G. XCMS Online: A web-based platform to process untargeted metabolomic data. *Analytical Chemistry* **84**, (11) 5035–5039 (2012).

70. Martens, L., Chambers, M., *et al.* mzML-a community standard for mass spectrometry data. *Molecular & Cellular Proteomics : MCP* **10**, (1) R110.000133 (2011).

71. Chambers, M. C., Maclean, B., *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology* **30**, (10) 918–920 (2012).

72. Gibb, S. & Strimmer, K. MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics* **28**, (17) 2270–2271 (2012).

73. Chang, W., Cheng, J., Allaire, J., Xie, Y. & McPherson, J. Shiny: Web Application Framework for R. (2016).

74. LANE & J., D. 16S/23S rRNA sequencing. *Nucleic acid techniques in bacterial systematics* 115–175 (1991).

75. Zeigler, D. R., Prágai, Z., *et al.* The origins of 168, W23, and other *Bacillus subtilis* legacy strains. *Journal of Bacteriology* **190**, (1) 6983–

6995 (2008).

76. Schroeder, J. W. & Simmons, L. A. Complete genome sequence of *Bacillus subtilis* strain PY79. *Genome Announcements* **1**, (6) (2013).

77. Stein, T. *Bacillus subtilis* antibiotics: structures, syntheses and specific functions. *Molecular Microbiology* **56**, (4) 845–857 (2005).

78. Thomas Candela & Agnès Fouet Poly-gamma-glutamate in bacteria. *Molecular Microbiology* **60**, (5) 1091–1098 (2006).

79. Yang, Y.-L., Xu, Y., *et al.* Connecting chemotypes and phenotypes of cultured marine microbial assemblages by imaging mass spectrometry. *Angewandte Chemie International Edition* **50**, (26) 5839–5842 (2011).

80. Yoon, S.-H., Ha, S.-M., *et al.* Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *International Journal of Systematic and Evolutionary Microbiology* **67**, (5) 1613–1617 (2017).

81. Gowda, H., Ivanisevic, J., *et al.* Interactive XCMS Online: simplifying advanced metabolomic data processing and subsequent statistical analyses. *Analytical Chemistry* **86**, (14) 6931–6939 (2014).

82. Arias, A. A., Lambert, S., *et al.* Growth of desferrioxamine-deficient *Streptomyces* mutants through xenosiderophore piracy of airborne fungal contaminations Anthony. *FEMS Microbiology Ecology* (7) (2015).

83. D'Onofrio, A., Crawford, J. M., *et al.* Siderophores from neighboring organisms promote the growth of uncultured bacteria. *Chemistry & Biology* **17**, (3) 254–264 (2010).

84. Bruns, H., Crüsemann, M., *et al.* Function-related replacement of bacterial siderophore pathways. *The Multidisciplinary Journal of Microbial Ecology (ISME Journal)* **12**, 320-329 (2018).

85. Bastian, M., Heymann, S. & Jacomy, M. Gephi : An open source software for exploring and manipulating networks visualization and exploration of large graphs. *International AAAI Conference on Weblogs and Social Media* 361–362 (2009).

86. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. (2008).

87. Gonzalez, D. J., Xu, Y., *et al.* Observing the invisible through imaging mass spectrometry, a window into the metabolic exchange patterns of microbes. *Journal of Proteomics* **75**, (16) 5069–5076 (2012).

88. Yang, Y.-L., Xu, Y., Straight, P. & Dorrestein, P. C. Translating metabolic exchange with imaging mass spectrometry. *Nature Chemical Biology* **5**, (12) 885–7 (2009).

89. Jensen, P. R., Gontang, E., Mafnas, C., Mincer, T. J. & Fenical, W. Culturable marine actinomycete diversity from tropical Pacific Ocean sediments. *Environmental Microbiology* **7**, (7) 1039–1048 (2005).

90. Gontang, E. A., Fenical, W. & Jensen, P. R. Phylogenetic diversity of gram-positive bacteria cultured from marine sediments. *Applied and Environmental Microbiology* **73**, (10) 3272–82 (2007).

91. Joseph, S. J., Hugenholtz, P., Sangwan, P., Osborne, C. A. & Janssen, P. H. Laboratory cultivation of widespread and previously uncultured soil bacteria. *Applied and Environmental Microbiology* **69**, (12) 7210–5 (2003).

92. Sauer, S., Freiwald, A., *et al.* Classification and identification of bacteria by mass spectrometry and computational analysis. *PloS one* **3**, (7) e2843 (2008).

93. Suarez, S., Ferroni, A., *et al.* Ribosomal proteins as biomarkers for bacterial identification by mass spectrometry in the clinical microbiology laboratory. *Journal of Microbiological Methods* **94**, (3) 390–6 (2013).

94. Dieckmann, R., Graeber, I., Kaesler, I., Szewzyk, U. & von Döhren, H. Rapid screening and dereplication of bacterial isolates from marine sponges of the Sula Ridge by Intact-Cell-MALDI-TOF mass spectrometry (ICM-MS). *Applied Microbiology and Biotechnology* **67**, (4) 539–548 (2005).

95. Croxatto, A., Prod'hom, G. & Greub, G. Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. *FEMS Microbiology Reviews* **36**, (2) 380–407 (2012).

96. Seuylemezian, A., Aronson, H. S., Tan, J., Lin, M., Schubert, W. & Vaishampayan, P. Development of a custom MALDI-TOF MS database for species-level identification of bacterial isolates collected from spacecraft and associated surfaces. *Frontiers in Microbiology* **9**, 780 (2018).

97. Strejcek, M., Smrhova, T., Junkova, P. & Uhlik, O. Whole-cell MALDI-TOF MS versus 16S rRNA gene analysis for identification and dereplication of recurrent bacterial isolates. *Frontiers in Microbiology* **9**, 1294 (2018).

98. Hoffmann, T., Krug, D., *et al.* Correlating chemical diversity with taxonomic distance for discovery of natural products in myxobacteria. *Nature Communications* **9**, (1) 803 (2018).

99. Grubbs, K. J., Bleich, R. M., *et al.* Large-scale bioinformatics analysis of *Bacillus* genomes uncovers conserved roles of natural products in bacterial physiology. *mSystems* **2**, (6) e00040-17 (2017).

100. Letzel, A.-C., Li, J., *et al.* Genomic insights into specialized

metabolism in the marine actinomycete *Salinispora*. *Environmental Microbiology* **19**, (9) 3660–3673 (2017).

101. Sumner, L. W., Amberg, A., *et al.* Proposed minimum reporting standards for chemical analysis chemical analysis working Group (CAWG) metabolomics standards initiative (MSI). *Metabolomics : Official journal of the Metabolomic Society* **3**, (3) 211–221 (2007).

102. Reeves, A. R., Seshadri, R., Brikun, I. A., Cernota, W. H., Gonzalez, M. C. & Weber, J. M. Knockout of the erythromycin biosynthetic cluster gene, eryBI, blocks isoflavone glucoside bioconversion during erythromycin fermentations in *Aeromicrobium erythreum* but not in S*accharopolyspora erythraea*. *Applied and Environmental Microbiology* **74**, (23) 7383–90 (2008).

103. Brikun, I. A., Reeves, A. R., Cernota, W. H., Luu, M. B. & Weber, J. M. The erythromycin biosynthetic gene cluster of *Aeromicrobium erythreum. Journal of Industrial Microbiology & Biotechnology* **31**, (7) 335–344 (2004).

104. Kiran, G. S., Sabarathnam, B., Thajuddin, N. & Selvin, J. Production of glycolipid biosurfactant from sponge-associated marine actinobacterium *Brachybacterium paraconglomeratum* MSA21. *Journal of Surfactants and Detergents* **17**, (3) 531–542 (2014).

105. Asolkar, R. N., Schröder, D., Heckmann, R., Lang, S., Wagner-Döbler, I. & Laatsch, H. Helquinoline, a new tetrahydroquinoline antibiotic from Janibacter limosus Hel 1+. *The Journal of Antibiotics* **57**, (1) 17–23 (2004).

106. Pospísil, S., Benada, O., Kofronová, O., Petrícek, M., Janda, L. & Havlícek, V. *Kytococcus sedentarius* (formerly *Micrococcus sedentarius*) and *Dermacoccus nishinomiyaensis* (formerly *Micrococcus nishinomiyaensis*) produce monensins, typical *Streptomyces cinnamonensis* metabolites. *Canadian Journal of Microbiology* **44**, (10) 1007–11 (1998).

107. Bagiyan, F. G., Eneyskaya, E. V, Kulminskaya, A. A., Savel'ev, A. N., Shabalin, K. A. & Neustroev, K. N. The action of alpha-mannosidase from *Oerskovia* sp. on the mannose-rich O-linked sugar chains of glycoproteins. *European Journal of Biochemistry* **249**, (1) 286–92 (1997).

108. Yamada, Y., Inouye, G., Tahara, Y. & Kondo, K. The structure of the menaquinones with a tetrahydrogenated isoprenoid side-chain. *Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism* **488**, (20) 280–284 (1977).

109. Yamada, Y., Inouye, G., Tahara, Y. & Kondo, K. On the chemical structure of menaquinones with the tetrahydrogenated isoprenoid side chain. *Biochimica et Biophysica Acta* **486**, (1) 195–203 (1976).

110. Murray, T. D., Schroeder, B. K., *et al. Rathayibacter toxicus*, other *Rathayibacter* species inducing bacterial head blight of grasses, and the potential for livestock poisonings. *Phytopathology* **107**, (7) 804–815 (2017).

111. Gao, Y., Liu, Q., *et al.* Inhibitory activities of compounds from the marine actinomycete *Williamsia* sp. MCCC 1A11233 variant on IgE-mediated mast cells and passive cutaneous anaphylaxis. *Journal of Agricultural and Food Chemistry* **65**, (49) 10749–10756 (2017).

112. Xie, C.-L., Niu, S.-W., Zhou, T.-T., Zhang, G.-Y., Yang, Q. & Yang, X.-W. Chemical constituents and chemotaxonomic study on the marine actinomycete *Williamsia* sp. MCCC 1A11233. *Biochemical Systematics and Ecology* **67**, 129–133 (2016).

113. Bull, A. T., Ward, A. C. & Goodfellow, M. Search and discovery strategies for biotechnology: the paradigm shift. *Microbiology and Molecular Biology Reviews: MMBR* **64**, (3) 573–606 (2000).

114. Yang, J. Y., Sanchez, L. M., *et al.* Molecular networking as a dereplication strategy. *Journal of Natural Products* **76**, (9) 1686–1699 (2013).

115. Zarins-Tutt, J. S., Barberi, T. T., *et al.* Prospecting for new bacterial metabolites: a glossary of approaches for inducing, activating and upregulating the biosynthesis of bacterial cryptic or silent natural products. *Natural Product Reports* **33**, (1) 54–72 (2016).

Paper I

Coupling MALDI-TOF mass spectrometry protein and specialized metabolite analyses to rapidly discriminate bacterial function

Chase M. Clark[a,1], Maria S. Costa[b,1], Laura M. Sanchez[a,2], and Brian T. Murphy[a,2,3]

[a]Department of Medicinal Chemistry and Pharmacognosy, College of Pharmacy, University of Illinois at Chicago, Chicago, IL; and [b]Faculty of Pharmaceutical Sciences, University of Iceland, Hagi, IS-107 Reykjavík, Iceland

Edited by Jerrold Meinwald, Cornell University, Ithaca, NY, and approved April 5, 2018 (received for review January 22, 2018)

For decades, researchers have lacked the ability to rapidly correlate microbial identity with bacterial metabolism. Since specialized metabolites are critical to bacterial function and survival in the environment, we designed a data acquisition and bioinformatics technique (IDBac) that utilizes in situ matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) to analyze protein and specialized metabolite spectra recorded from single bacterial colonies picked from agar plates. We demonstrated the power of our approach by discriminating between two *Bacillus subtilis* strains in <30 min solely on the basis of their differential ability to produce cyclic peptide antibiotics surfactin and plipastatin, caused by a single frameshift mutation. Next, we used IDBac to detect subtle intraspecies differences in the production of metal scavenging acyldesferrioxamines in a group of eight freshwater *Micromonospora* isolates that share >99% sequence similarity in the 16S rRNA gene. Finally, we used IDBac to simultaneously extract protein and specialized metabolite MS profiles from unidentified Lake Michigan sponge-associated bacteria isolated from an agar plate. In just 3 h, we created hierarchical protein MS groupings of 11 environmental isolates (10 MS replicates each, for a total of 110 spectra) that accurately mirrored phylogenetic groupings. We further distinguished isolates within these groupings, which share nearly identical 16S rRNA gene sequence identity, based on interspecies and intraspecies differences in specialized metabolite production. IDBac is an attempt to couple in situ MS analyses of protein content and specialized metabolite production to allow for facile discrimination of closely related bacterial colonies.

natural products | specialized metabolites | mass spectrometry | bioinformatics | metabolomics

For nearly two centuries researchers have studied bacteria to diagnose and treat diseases, elucidate intricate interspecies and intraspecies evolutionary processes, manage and develop agricultural biocontrol practices, and, broadly speaking, learn about the complex roles of microorganisms in the environment. Therefore, developing techniques to rapidly characterize and discriminate between bacteria has been paramount to these efforts. In the past four decades, sequencing of the 16S ribosomal RNA (rRNA) gene has been instrumental to the classification of bacteria due to its widespread presence in the kingdom, degree of conservation, and length (1, 2). This and other genetic-based approaches, such as pulsed field gel electrophoresis, multilocus sequence typing, and DNA–DNA hybridization, have become commonplace (3). However, several limitations to these techniques include cost, turnaround time needed for sequencing and/or analysis, narrow windows of "universal" primers, and, in some cases, low species-level phylogenetic resolution. Most importantly, in the majority of cases, these methods are unable to elucidate how microorganisms interact with one another and function in situ. To address this shortcoming, we developed a pipeline that allows rapid discrimination of bacteria based on mass spectral signatures of specialized metabolite production (molecules that are typically between 200 and 2,000 Da and not essential to the immediate growth or survival of an organism) in complement to conserved ribosomal housekeeping proteins.

Specialized metabolites represent functional traits in bacteria, and these molecules are useful for defining "what a bacterium does," rather than "who it is" in the context of its immediate environment (4–6). Ziemert et al. (5) analyzed 75 sequenced genomes within a group of three closely related species that shared 99% 16S rRNA gene sequence identity: *Salinispora arenicola*, *Salinispora pacifica*, and *Salinispora tropica*. The isolates were predicted to contain a surprising 229 distinct specialized metabolite biosynthetic gene clusters, the majority of which were acquired through recent horizontal gene transfer events and occurred in only one or two isolates in the group. Given this large potential chemical diversity harbored within *Salinispora* genomes, these findings highlight the limitations of using phylogenetic approaches based on 16S rRNA genes to infer bacterial function and stress the need for alternative approaches to rapidly assess chemical differences between closely related bacterial isolates.

Significance

Mass spectrometry is a powerful technique that has been used to identify bacteria by their protein content and to assess bacterial functional traits through analysis of their specialized metabolites. However, until now these analyses have operated independently, which has resulted in the inability to rapidly connect bacterial phylogenetic identity with potential environmental function. To bridge this gap, we designed a MALDI-TOF mass spectrometry data acquisition and bioinformatics pipeline (IDBac) to integrate data from both intact protein and specialized metabolite spectra directly from bacterial cells grown on agar. This technique organizes bacteria into highly similar phylogenetic groups and allows for comparison of metabolic differences of hundreds of isolates in just a few hours.

Author contributions: C.M.C., M.S.C., L.M.S., and B.T.M. designed research; C.M.C. and M.S.C. performed research; C.M.C., M.S.C., L.M.S., and B.T.M. analyzed data; and C.M.C., M.S.C., L.M.S., and B.T.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

Data deposition: Mass spectra data have been deposited in MASSIVE for MALDI (ftp://massive.ucsd.edu/MSV000081619) and LC-MS/MS (ftp://massive.ucsd.edu/MSV000081555). The 16S rRNA sequences of environmental bacteria used in this study were deposited in GenBank (see *SI Appendix*, Table S2) (accession nos. KY858227–KY858239, KY858241–KY858247, MG188670, MG188671, and KP009553).

[1]C.M.C and M.S.C. contributed equally to this work.

[2]L.M.S. and B.T.M. contributed equally to this work.

[3]To whom correspondence should be addressed. Email: btmurphy@uic.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1801247115/-/DCSupplemental.

Published online April 23, 2018.

Shortly after the development of 16S rRNA gene sequencing, matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) was implemented as a technique to identify large biomolecules (7, 8). Subsequent innovations in instrumentation led to the ability to obtain better-resolved spectra of intact proteins in high throughput, facilitating the rapid and less-costly identification of bacteria based largely on ribosomal MS fingerprints (8–11). Bruker (12) and bioMerieux (13) have successfully applied this technology in the clinical setting, while many others have used it on relatively small strain groupings (from 10 to a few hundred isolates) in a genus- and species-specific manner to environmental and/or clinical microorganisms or to the classification of mammalian cells, as recently summarized by the reviews in refs. 14–17. Despite the aforementioned applications, these methods do not provide information on bacterial specialized metabolite production. Relatively little is known about the relationship between bacterial taxonomy and specialized metabolite production in the majority of bacteria isolated from the environment, yet these characteristics are central to researchers who study bacteria in both academic and industrial settings. Thus, a comprehensive analytical pipeline that allows simultaneous analyses of these factors has been a major obstacle to correlating microbial identity with specialized metabolite production.

In this study, we present a significant innovation to described MS methods that analyze bacteria by utilizing the full capabilities offered by MALDI-TOF mass spectrometers. In addition to linear mode protein analysis, we use reflectron mode to analyze specialized metabolites, as the combination of both information-rich spectral regions has yet to be applied to existing MALDI-TOF MS analysis pipelines. Silva et al. (18) recently provided a comprehensive history detailing the underuse of MALDI MS to analyze specialized metabolites and the shortage of software to relieve current bioinformatics bottlenecks.

To validate our pipeline, we demonstrate the ability to differentiate isolates within closely related species groupings, often on colonies that share indistinguishable morphology, and characterize them based on in situ antibiotic (plipastatin), siderophore (desferrioxamine), and motility factor (surfactin) production. Total acquisition, analysis, and visualization of MALDI-TOF MS data from both intact proteins and specialized metabolites of up to 384 bacterial colonies can be performed in <4 h. Compared with other instrumentation that is commonly used to analyze specialized metabolites, such as quadrupole TOF, Orbitrap, and Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometers, MALDI-TOF MS requires minimal expertise to operate and is accompanied by facile sample preparation protocols. Our method provides an alternative to laborious liquid cultivation, metabolite extraction, and chromatographic experiments that are the current standard of practice for studies that focus on specialized metabolite analysis.

Results

To visualize relational patterns between bacterial isolates, we created a bioinformatics pipeline designed to facilitate the multistage analysis of protein and specialized metabolite MS data. Generally speaking, we analyzed MS fingerprints of intact proteins (3,000–15,000 Da) and specialized metabolites (200–2,000 Da) in consecutive MALDI-TOF MS linear and reflectron mode acquisitions, respectively. Principal components analysis and hierarchical clustering of protein spectra placed bacterial isolates into putative genus- and species-level groups. Isolates within each grouping were then further discriminated based on differences in specialized metabolite production through analysis of Metabolite Association Networks (MANs; see *SI Appendix*, *Text S1* for a description). The IDBac software, written in R, is available for download and installation via a simple Windows installer (*Publication Code and Data Availability*). The software was designed for simplicity and ease of use with documentation provided at each step in the workflow that at the time of publication provides inverted/mirror plots for comparison, hierarchical clustering, principle components analysis, and MAN analysis. However, whereas similar software has required users to convert raw data on their own and format these to software-dependent configurations, IDBac takes raw data as input. This provides a transparent and repeatable data-handling process that is less prone to user error or data-tampering and, as a by-product, bundles spectra by sample into the widely accepted and readily shareable mzML format. The IDBac software is available under a GNU General Public License, and links to the full code along with data acquisition and analysis tutorials are within *Publication Code and Data Availability*.

Although our platform is intended to visualize interspecies and intraspecies differences in specialized metabolite production between colonies rather than to identify the precise chemical structure of excreted specialized metabolites, it was important to validate that specific nodes (*m/z* features) in our MANs represented bacterial chemistry as opposed to matrix peaks, media components, or instrument noise. After preprocessing and peak-picking our MALDI-TOF MS spectra, we accounted for the matrix and media ionizable compounds by subtracting a reference peak list for a matrix and media control during each experimental run. To account for instrument noise, we precluded signals from the peak picking algorithm that fell below a user-defined signal-to-noise ratio (i.e., 4:1) and retained peaks occurring in a minimum of 70% of replicates. We demonstrated the ability of IDBac to use specialized metabolite production to distinguish between the nearly identical *Bacillus subtilis* 3610 and its frame-shift mutant *B. subtilis* PY79 (Fig. 1). The two strains were chosen based on their relation to one another: *B. subtilis* PY79 (19, 20) is deficient in the ability to produce the antibiotics surfactin and plipastatin due to a frameshift mutation in *sfp*, the 4′-phosphopantetheinyl transferase that mediates nonribosomal peptide synthetase apoform activation (21).

B. subtilis 3610 and *B. subtilis* PY79 were grown on separate A1 nutrient agar plates under identical conditions and their specialized metabolite regions (200–2,000 Da) were analyzed using MALDI-TOF MS (ten technical replicates each). The differences in specialized metabolite production between strains 3610 and PY79 were readily observed and visualized in our MAN (Fig. 1*B*), where large nodes represent bacterial colonies and smaller nodes represent *m/z* peaks (singly charged molecules) present in their corresponding MALDI-TOF spectra. This MAN allowed for facile visualization of both the differences in specialized metabolite spectra between strains and the peaks that were shared or unique to each strain. Four plipastatin (compounds **1**–**4**) and three surfactin (compounds **5**–**7**) analogs, including their resolved isotopologues and adducts, were detected from strain 3610, and not from PY79. Both strains shared the ability to produce partially characterized polyglutamate polymers (22), which have been implicated in species-specific functions such as virulence factor production, biofilm formation, and sequestration of toxic metal ions (22, 23). Importantly, the majority of *m/z* values (33 of 42; 78.6%) in the MAN represented specialized metabolites, and IDBac successfully filtered out ions associated with matrix and media components. MALDI-TOF MS analysis of the specialized metabolite region and data processing using IDBac correctly depicted subspecies antibiotic production differences between genetic variants of *B. subtilis* in <30 min.

Next, we demonstrated that intraspecies strain groupings based on MALDI-TOF MS intact protein profiles can be further discriminated based on in situ specialized metabolite production. From our in-house strain library, we selected eight *Micromonospora chokoriensis* isolates from two sediment samples collected nearly 175 km apart in Lake Michigan, US. These isolates share >99%

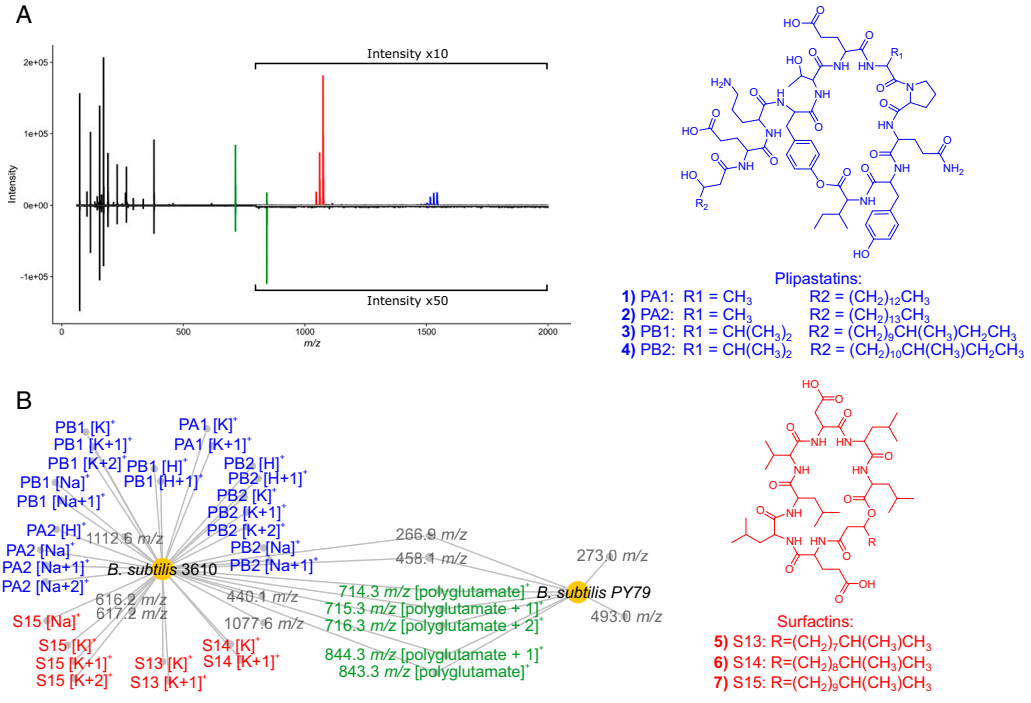


Fig. 1. Analysis of MALDI-TOF MS specialized metabolite data from two *B. subtilis* genetic variants (10 technical replicates each) shows distinct differences in plipastatin and surfactin analog production. (*A*) Inverse spectrum comparison showing representative spectra for *B. subtilis* 3610 (positive spectrum) and *B. subtilis* PY79 (negative spectrum). (*B*) MAN showing the differential production of surfactin and plipastatin antibiotic analogs between the two strains. Large nodes represent individual bacterial colonies, while smaller nodes represent individual *m/z* values in MALDI-TOF spectra that fall within our peak selection criteria. MALDI-TOF spectrum annotations can be found in *SI Appendix*, Fig. S1. Isotopologues are denoted as +1, +2. Based on their precedence in *B. subtilis* 3610 and PY79 in literature, we assigned several *m/z* peaks as polyglutamate-like polymers (22). The remaining peaks were not identified but may represent primary metabolites and/or yet-to-be-characterized specialized metabolites. Importantly, we assigned 33/42 *m/z* values (78.6%), providing evidence that the MAN is primarily composed of specialized metabolites.

16S rRNA gene sequence identity across >1,460 nucleotides (24). We cultivated biological replicates in a random pattern across a 48-well microwell plate, resulting in at least four independent biological replicates of each strain. With a minimum of eight technical replicates, we used MALDI-TOF MS to consecutively record MS spectra of the protein (3,000–15,000 *m/z*) and specialized metabolite regions (200–2,000 *m/z*). This resulted in at least 32 replicate spectra per isolate, with peaks required to be present in 70% or greater of replicates to be included in analyses (a user-defined parameter within IDBac). We observed that hierarchical clustering of MALDI-TOF MS protein data correlated strongly with 16S rRNA similarity (Fig. 2*A*), results consistent with many previous efforts that used MALDI-TOF MS as an alternative to traditional sequence-based taxonomic classification methods, as recently summarized (14, 15).

Importantly, while these MALDI-TOF MS protein fingerprints correlated nearly identically with corresponding phylogenetic groupings (Fig. 2*A*), differences in specialized metabolite production existed and could not be predicted through 16S rRNA phylogenetic analyses or geographic strain distribution patterns. However, analysis of IDBac's MAN rapidly discerned subtle, but significant, variations in bacterial chemistry within these closely related *Micromonospora* isolates (Fig. 2*B*), allowing us to assess the relationship of specialized metabolite production to both strain phylogeny and geographic origin. A major distinguishing pattern exhibited by seven of eight *M. chokoriensis* isolates was a group of features between 650 and 800 *m/z*. This pattern was characterized by successive 14-Da differences beginning at 673.5 *m/z* and extending to 771.6 *m/z*, along with sodium and potassium adducts of each, which we attributed to analogs differing in the addition/subtraction of methylene groups (*SI Appendix*, Fig. S2). An outlier strain, B001, did not share these features, indicating differential specialized metabolite production within this group of highly similar strains.

To further validate the differential specialized metabolite production observed in the MAN, we employed HPLC tandem MS (HPLC-MS/MS) analysis for evaluation with the Global Natural Products Social molecular networking platform (25) and comparative metabolomics XCMS analysis (26). HPLC-MS/MS afforded an orthogonal means of analysis through chromatographic separation, a different mechanism of ionization, and tandem mass data that helped verify relationships between observed metabolites. Using GNPS, XCMS, and simulated spectra for iron isotopologues (*SI Appendix*, Figs. S3–S7), we identified a series of acylated desferrioxamine analogs (compounds **8**–**12**). These belong to a class of siderophores that sequester the essential growth factor ferric iron from the environment. One of

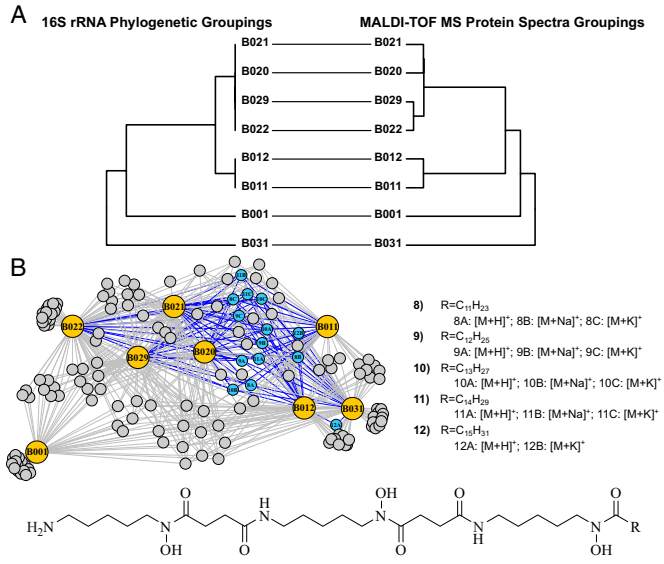


Fig. 2. IDBac protein and specialized metabolite analysis of isolates from a single *Micromonospora* species. (*A*) Tanglegram depicts a high degree of similarity between groupings of 16S rRNA gene sequence identity and MALDI-TOF MS protein data of *M. chokoriensis* isolates. (*B*) MAN of MALDI-TOF MS data from *M. chokoriensis* colonies highlights distinct intraspecies differences in specialized metabolite production; this is due to differential production of a specific series of acylated desferrioxamine siderophores (shown as blue nodes), which B001 did not produce.

these, desferrioxamine B, is used clinically to treat metal poisoning. Interestingly, according to our analyses, B001 lacks the capacity to produce these acylated analogs (Fig. 2*B*), and this was readily highlighted upon analysis of the MAN. We also observed several other desferrioxamine analogs produced by these isolates; links to these data are available in *SI Appendix*, Table S1. Of note, while the phylogenetic identity of *Micromonospora* isolate B031 is more closely related to B001, it overlaps the remaining six isolates in its ability to produce acylated desferrioxamines, highlighting the importance of our method to provide a means other than phylogeny to distinguish between in situ function of similar strains.

The HPLC-MS/MS–based observation of strain-specific patterns of siderophore production corroborated our initial MALDI-TOF MS results that showed B001, while able to produce unmodified desferrioxamine B, was deficient in the ability to produce a series of acylated desferrioxamine B analogs (compounds **8**–**12**). This is significant, as B001 was isolated from the same 1-cm^3 sediment sample as six of the seven strains that produced this compound series, highlighting that phylogenetic groupings and geographic location are not sufficient indicators of specialized metabolite production capacity. Our observations of desferrioxamine biosynthetic pathway promiscuity were consistent with previous studies (27–29). However, each of these studies required extensive genome sequencing and/or liquid fermentation experiments, followed by chromatographic analyses, whereas analysis through MALDI-TOF MS/IDBac was able to visualize putative phylogenetic relationships and intraspecies differences in specialized metabolism in a few hours, once each colony appeared on a Petri dish.

In many instances, the taxonomic identity and specialized metabolite production capacity of a group of bacteria are not well characterized or completely unknown, particularly in studies involving bacteria isolated from the environment. Thus, we tested the ability of the IDBac analysis pipeline to rapidly extract protein and specialized metabolite information from unknown environmental bacteria cultivated from a freshwater sponge collected in Lake Michigan. We cultivated sponge-associated bacteria from 1 cm^3 of tissue onto high-nutrient A1 medium (Fig. 3*A*; see *SI Appendix* for further details). Using a sterile toothpick, we selected all colonies that grew on the plate over a 90-d time period. These colonies were then subjected to MALDI-TOF MS analysis with 10 technical replicates, and the data were processed in IDBac. Principal components analysis and unsupervised hierarchical clustering of the bacterial protein range afforded four distinct groupings (Fig. 3*B*). We confirmed that these MS protein groupings aligned with the genera *Enterococcus*, *Bacillus*, *Paenibacillus*, and *Rhodococcus* via 16S rRNA gene sequencing analysis of each of isolate (*SI Appendix*, Figs. S8 and S9).

Next, we generated a MAN (Fig. 3*C*) to further discriminate these groups based on specialized metabolite production. Interestingly, unsupervised subnetworking modularity analysis (30, 31) separated the network into five groups of strains (colored in Fig. 3*C*), which highly correlated to phylogenetic groupings. The seven *Rhodococcus* isolates exhibited a high degree of specialized metabolite associations, while the *Bacillus* and *Enterococcus* isolates were classified as unique via modularity scoring. Importantly, our method quickly highlighted subtle differences in specialized metabolite production between two morphologically identical *Paenibacillus* strains and separated the two strains whose 16S rRNA gene sequences shared a pairwise similarity of 99.87% (1,496 of 1,498 nucleotides) (24). A careful look at their specialized metabolite profiles showed subtle differences in production of a series of specialized metabolite(s) ranging from 900 to 1,250 *m*/*z*. We wanted to confirm that this difference was not due to variations in colony microenvironment, since nearby colonies can affect specialized metabolite production through physical contact and

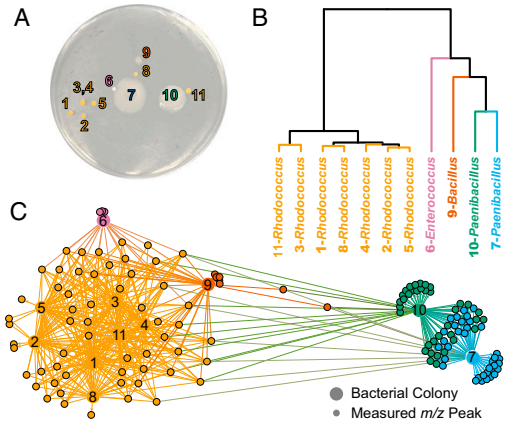


Fig. 3. Rapid protein and specialized metabolite fingerprinting of unknown environmental isolates using MALDI-TOF MS and IDBac. (*A*) Bacterial diversity plate obtained from placing freshwater sponge tissue on high-nutrient A1 agar. (*B*) IDBac allowed for hierarchical clustering of MALDI-TOF MS protein spectra with the option to choose standard distance measures and clustering algorithms. For workflows requiring analysis of hundreds to thousands of strains, protein grouping is essential for data reduction before specialized metabolite analysis is performed. (*C*) MAN, colored via modularity analysis with default thresholds in Gephi (31), allowed for rapid decision-making based on gross in situ specialized metabolite production after matrix and media signals were subtracted automatically from the network in IDBac. Significant outliers were the two *Paenibacillus* strains, which produced several shared and unique high molecular-weight specialized metabolites.

chemical cross-talk (22, 32, 33). To determine this, we grew each *Paenibacillus* isolate by itself over three individual cultivation experiments and MALDI-TOF MS data acquisition events, and after observation of previously observed specialized metabolite patterns in the 900–1,250 *m/z* range, we concluded that both isolates contained partially overlapping, but distinct, metabolic capacities (*SI Appendix*, Fig. S10). These subtle differences in metabolic capacity were readily detected and visualized as a result of the IDBac MAN and could not have been achieved through analysis of protein groupings alone. In just 3 h, our pipeline created statistically robust protein profile groupings of environmental isolates (11 strains, with 10 MS replicates each, for a total of 110 MALDI spots). It further distinguished colonies with nearly identical morphology and 16S rRNA gene sequence identity based on their capacity to produce specialized metabolites.

Discussion

One aspect of the IDBac pipeline that provides significant advantages over existing MS platforms is that it affords information on putative colony phylogeny and specialized metabolite production without the need for extraction and chromatographic analyses. The latter techniques rely on growing pure bacterial isolates (often in liquid culture), generating extracts, and using LC analyses (generally coupled to MS) to separate and detect specialized metabolites. This is often a laborious, costly, and relatively time-consuming process. Our pipeline is an extraction-free process and is ideally suited for researchers who aim to (*i*) compare functional chemistry between closely related isolates (e.g., comparing antibiotic production between two *B. subtilis* genetic variants), (*ii*) probe relationships between taxonomic identity and environmental functionality (e.g., studying intraspecies differences in *M. chokoriensis* siderophore production), or (*iii*) assess the broad relatedness of unknown environmental bacterial isolates (e.g., visualizing phylogenetic and metabolic relatedness within a group of unknown bacteria). The latter points highlight a strength of IDBac as an engine to generate research questions and hypotheses. For example, why have similar *Micromonospora* strains from the same 1 cm^3 of sediment (*SI Appendix*, Table S2) evolved differing capacities to scavenge iron? Is there any correlation between geographic location and desferrioxamine production in an expanded set of Lake Michigan sediment-derived *Micromonospora* isolates? Is there a synergistic role between the specialized metabolites that is unique to two phylogenetically similar *Paenibacillus* strains isolated from the same sponge? MALDI-TOF MS/IDBac analysis allows for easy visualization of global specialized metabolite patterns and, as a result, gives researchers the opportunity to ask these questions based on rapidly generated data visualizations.

IDBac is complementary to other innovative platforms such as GNPS, which aids in the dereplication of previously characterized and identification of potentially new specialized metabolites (25, 34). Toward this point, it is important to note that MALDI-TOF MS and various LC-MS systems use distinct modes of ionization and give rise to different observed features that hinder direct comparison of experimental data. Our work highlights the individual strengths of orthogonal MS-based approaches to extract complementary information from a biological system.

Through the course of our studies, we have documented a few potential limitations of our method. First, although IDBac can create accurate subspecies groupings of bacteria based on protein MS fingerprints, species-level identification is only possible in the presence of a searchable and extensive protein fingerprint database. A greater community effort is required to document MALDI-TOF MS protein fingerprints into a publicly available database to maximize phylogenetic coverage. Such an effort would facilitate the rapid identification of unknown environmental bacteria and elevate this process to be on par with existing commercial platforms used to identify clinical pathogens (14). An effort to make a readily searchable/freely available bacterial protein MS database is ongoing in our laboratory, and the first version of IDBac supports this mission by providing a simple interface for bundling and converting proprietary vendor-format data files to the open-access mzML format for easy, reproducible sharing.

A second limitation is that bacterial specialized metabolite production is sensitive to external factors such as growth medium, temperature, unintended microbial contamination, and the proximity of other microorganisms on a Petri dish (35). For comparison of specialized metabolite profiles to be reliable, it is imperative that strains are cultivated under identical conditions and that several biological/technical replicates are acquired for each isolate (both precautions were taken in this study). Fortunately, the throughput of our pipeline allows for multiple replicates of a sample to be analyzed rapidly. Conversely, MALDI-TOF MS protein fingerprints of bacterial colonies are robust and exhibit minimal fluctuation when the data are acquired on different growth media (*SI Appendix*, Fig. S11).

Finally, our method is limited by the resolution of the MALDI-TOF MS and would be improved through use of instruments with higher resolving power capabilities. High-resolution instruments, such as MALDI-FT-ICR and fragmentation data from a MALDI-TOF/TOF MS, could aid in preliminary compound class dereplication via generation of accurate molecular formulas and structural information, respectively. However, this would increase the length of the experiments and data storage requirements and would require additional technical expertise.

IDBac couples bacterial protein and specialized metabolite MS data to rapidly discriminate between isolates based on both

MICROBIOLOGY

their identity and potential environmental function. Our MS pipeline addresses a need in research communities that study microbial function (e.g., chemical ecology, pathogenesis, taxonomy, drug discovery, agriculture, and food safety). The pipeline is faster than existing MS methods used to analyze nonvolatile metabolite production within microorganisms and requires less technical experience to operate. The experiment requires 10^5 to 10^7 bacterial cells (36)—or approximately the tip of a toothpick—to generate MS profiles and is designed to be performed in high-throughput. Acquisition of MALDI-TOF MS data from both intact proteins and specialized metabolites of single bacterial colonies has been integrated in one single freely available bioinformatics pipeline. This work reports on coupling protein and specialized metabolite MS data in a semiautomated pipeline to make the rapid discrimination of bacteria accessible to the broad research community.

Methods

MALDI-TOF MS Sample Preparation and Data Acquisition. For MALDI-TOF MS analysis, proteins were extracted by using an extended direct transfer method that included a formic acid overlay (36, 37). Measurements were performed in linear and reflectron modes by using an Autoflex Speed LRF mass spectrometer (Bruker Daltonics) equipped with a smartbeam-II laser (355 nm). Detailed instrument settings are available in *SI Appendix*.

MALDI-TOF Data Bioinformatics Pipeline. In brief, MALDI-TOF MS raw data were first converted to the open-source mzML format (38) by using ProteoWizard's command-line MSConvert (39). The mzML files were read into R through mzR (39) and custom code and processed into peak lists by using MALDIquant (40). Lastly, the peak lists informed interactive data analyses and visualizations that were displayed through a default, offline internet browser with RStudio's Shiny (41) package. The IDBac software is being distributed under a GNU General Public License, and links to the full code along with data acquisition and analysis tutorials can be found within *Publication Code and Data Availability*.

Publication Code and Data Availability. IDBac software and updates, method manuals, and information on code contribution are available at chasemc.github.io/IDBac and permanent snapshots at https://zenodo.org/record/1115619. The flexControl, flexAnalysis, and R code used to generate all figures and analyses for this publication along with the publication-version of IDBac is available for download at https://data.mendeley.com/datasets/ysrtr9c5s7/1. Mass spectra were deposited in MASSIVE (accession nos. MSV000081619 and MSV000081555).

ACKNOWLEDGMENTS. We thank the following contributors: Russel Cuhel and crew of R. V. Neeskay (University of Milwaukee Wisconsin) for assistance with sediment collection; Keith Jung for collection of the freshwater sponge; Dr. Amanda Bulman (Bruker) for assistance with MALDI-TOF MS protein acquisition parameters; Dr. Terry Moore and Dr. Atul Jain for recrystallizing alpha-cyano-4-hydroxycinnamic acid matrix (CHCA); and Dr. Darrell McCaslin (University of Wisconsin Madison) for testing our method on legacy instruments. This work was supported by National Institute of Child Health and Human Development Grant K12HD055892; the National Institutes of Health Office of Research on Women's Health; National Geographic Grant CP-044R-17; Icelandic Research Fund Grant 152336-051; and University of Illinois at Chicago startup funds.

1. Woese CR (1987) Bacterial evolution. *Microbiol Rev* 51:221–271.
2. Yarza P, et al. (2014) Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* 12:635–645.
3. Vandamme P, et al. (1996) Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiol Rev* 60:407–438.
4. Patin NV, Duncan KR, Dorrestein PC, Jensen PR (2016) Competitive strategies differentiate closely related species of marine actinobacteria. *ISME J* 10:478–490.
5. Ziemert N, et al. (2014) Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. *Proc Natl Acad Sci USA* 111:E1130–E1139.
6. Penn K, et al. (2009) Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. *ISME J* 3:1193–1203.
7. Karas M, Hillenkamp F (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* 60:2299–2301.
8. Tanaka K, et al. (1988) Protein and polymer analyses up to *m/z* 100,000 by laser ionization time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom* 2: 151–153.
9. Sandrin TR, Goldstein JE, Schumaker S (2013) MALDI TOF MS profiling of bacteria at the strain level: A review. *Mass Spectrom Rev* 32:188–217.
10. Cain TC, Lubman DM, Weber WJ, Vertes A (1994) Differentiation of bacteria using protein profiles from matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom* 8:1026–1030.
11. Holland RD, et al. (1996) Rapid identification of intact whole bacteria based on spectral patterns using matrix-assisted laser desorption/ionization with time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom* 10:1227–1232.
12. Maier T, Klepel S, Renner U, Kostrzewa M (2006) Fast and reliable MALDI-TOF MS–Based microorganism identification. *Nat Methods* 3:68–71.
13. Dubois D, et al. (2012) Performances of the Vitek MS matrix-assisted laser desorption ionization-time of flight mass spectrometry system for rapid identification of bacteria in routine clinical microbiology. *J Clin Microbiol* 50:2568–2576.
14. Rahi P, Prakash O, Shouche YS (2016) Matrix-assisted laser desorption/ionization time-of-flight mass-spectrometry (MALDI-TOF MS) based microbial Identifications: Challenges and scopes for microbial ecologists. *Front Microbiol* 7:1359.
15. Popović NT, Kazazić SP, Strunjak-Perović I, Coz-Rakovac R (2017) Differentiation of environmental aquatic bacterial isolates by MALDI-TOF MS. *Environ Res* 152:7–16.
16. Munteanu B, Hopf C (2013) Emergence of whole-cell MALDI-MS biotyping for high-throughput bioanalysis of mammalian cells? *Bioanalysis* 5:885–893.
17. Cassagne C, Normand A-C, L'Ollivier C, Ranque S, Piarroux R (2016) Performance of MALDI-TOF MS platforms for fungal identification. *Mycoses* 59:678–690.
18. Silva R, Lopes NP, Silva DB (2016) Application of MALDI mass spectrometry in natural products analysis. *Planta Med* 82:671–689.
19. Zeigler DR, et al. (2008) The origins of 168, W23, and other *Bacillus subtilis* legacy strains. *J Bacteriol* 190:6983–6995.
20. Schroeder JW, Simmons LA (2013) Complete genome sequence of *Bacillus subtilis* strain PY79. *Genome Announc* 1:e01085-13.
21. Stein T (2005) *Bacillus subtilis* antibiotics: Structures, syntheses and specific functions. *Mol Microbiol* 56:845–857.
22. Yang Y-L, et al. (2011) Connecting chemotypes and phenotypes of cultured marine microbial assemblages by imaging mass spectrometry. *Angew Chem Int Ed Engl* 50: 5839–5842.
23. Candela T, Fouet A (2006) Poly-gamma-glutamate in bacteria. *Mol Microbiol* 60: 1091–1098.
24. Yoon S-H, et al. (2017) Introducing EzBioCloud: A taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int J Syst Evol Microbiol* 67: 1613–1617.
25. Wang M, et al. (2016) Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol* 34: 828–837.
26. Gowda H, et al. (2014) Interactive XCMS online: Simplifying advanced metabolomic data processing and subsequent statistical analyses. *Anal Chem* 86:6931–6939.
27. Arias AA, et al. (2015) Growth of desferrioxamine-deficient *Streptomyces* mutants through xenosiderophore piracy of airborne fungal contaminations. *FEMS Microbiol Ecol* 91:fiv080.
28. D'Onofrio A, et al. (2010) Siderophores from neighboring organisms promote the growth of uncultured bacteria. *Chem Biol* 17:254–264.
29. Bruns H, et al. (2018) Function-related replacement of bacterial siderophore pathways. *ISME J* 12:320–329.
30. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *J Stat Mech* 2008:P10008.
31. Bastian M, Heymann S, Jacomy M (2009) Gephi: An open source software for exploring and manipulating networks visualization and exploration of large graphs. *International AAAI Conference on Weblogs and Social Media* (Association for the Advancement of Artificial Intelligence, Menlo Park, CA), pp 361–362.
32. Gonzalez DJ, et al. (2012) Observing the invisible through imaging mass spectrometry, a window into the metabolic exchange patterns of microbes. *J Proteomics* 75: 5069–5076.
33. Yang Y-L, Xu Y, Straight P, Dorrestein PC (2009) Translating metabolic exchange with imaging mass spectrometry. *Nat Chem Biol* 5:885–887.
34. Yang JY, et al. (2013) Molecular networking as a dereplication strategy. *J Nat Prod* 76: 1686–1699.
35. Zarins-Tutt JS, et al. (2016) Prospecting for new bacterial metabolites: A glossary of approaches for inducing, activating and upregulating the biosynthesis of bacterial cryptic or silent natural products. *Nat Prod Rep* 33:54–72.
36. Freiwald A, Sauer S (2009) Phylogenetic classification and identification of bacteria by mass spectrometry. *Nat Protoc* 4:732–742.
37. Schumann P, Maier T (2014) MALDI-TOF mass spectrometry applied to classification and identification of bacteria. *Methods Microbiol* 41:275–306.
38. Martens L, et al. (2011) mzML–A community standard for mass spectrometry data. *Mol Cell Proteomics* 10:000133.
39. Chambers MC, et al. (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* 30:918–920.
40. Gibb S, Strimmer K (2012) MALDIquant: A versatile R package for the analysis of mass spectrometry data. *Bioinformatics* 28:2270–2271.
41. Chang W, Cheng J, Allaire J, Xie Y, McPherson J (2016) Shiny: Web Application Framework for R (RStudio, Boston).

Paper II

Video Article

Using the Open-Source MALDI TOF-MS IDBac Pipeline for Analysis of Microbial Protein and Specialized Metabolite Data

Chase M. Clark[1], Maria S. Costa[1,2], Erin Conley[1], Emma Li[1], Laura M. Sanchez[1], Brian T. Murphy[1]

[1]Department of Medicinal Chemistry and Pharmacognosy, College of Pharmacy, University of Illinois at Chicago

[2]Faculty of Pharmaceutical Sciences, University of Iceland

Correspondence to: Brian T. Murphy at btmurphy@uic.edu

URL: https://www.jove.com/video/59219
DOI: doi:10.3791/59219

Keywords: Biochemistry, Issue 147, Mass spectrometry, microbiology, natural products, specialized metabolites, MALDI-TOF MS, bioinformatics

Date Published: 5/15/2019

Citation: Clark, C.M., Costa, M.S., Conley, E., Li, E., Sanchez, L.M., Murphy, B.T. Using the Open-Source MALDI TOF-MS IDBac Pipeline for Analysis of Microbial Protein and Specialized Metabolite Data. *J. Vis. Exp.* (147), e59219, doi:10.3791/59219 (2019).

Abstract

In order to visualize the relationship between bacterial phylogeny and specialized metabolite production of bacterial colonies growing on nutrient agar, we developed IDBac—a low-cost and high-throughput matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) bioinformatics pipeline. IDBac software is designed for non-experts, is freely available, and capable of analyzing a few to thousands of bacterial colonies. Here, we present procedures for the preparation of bacterial colonies for MALDI-TOF MS analysis, MS instrument operation, and data processing and visualization in IDBac. In particular, we instruct users how to cluster bacteria into dendrograms based on protein MS fingerprints and interactively create Metabolite Association Networks (MANs) from specialized metabolite data.

Video Link

The video component of this article can be found at https://www.jove.com/video/59219/

Introduction

A major barrier to researchers who study bacterial function is the ability to quickly and simultaneously assess the taxonomic identity of a microorganism and its capacity to produce specialized metabolites. This has prevented significant advances in understanding the relationship between bacterial phylogeny and specialized metabolite production in the majority of bacteria isolated from the environment. Although MS-based methods that use protein fingerprints to group and identify bacteria are well described[1,2,3,4], these studies have generally been performed on small groups of isolates, in a species-specific manner. Importantly, information on specialized metabolite production, a major driver of microbial function in the environment, has remained unincorporated in these studies. Silva et al.[5] recently provided a comprehensive history detailing the underuse of MALDI-TOF MS to analyze specialized metabolites and the shortage of software to relieve current bioinformatics bottlenecks. In order to address these shortcomings, we created IDBac, a bioinformatics pipeline that integrates both linear and reflectron modes of MALDI-TOF MS[6]. This allows users to rapidly visualize and differentiate bacterial isolates based on both protein and specialized metabolite MS fingerprints, respectively.

IDBac is cost-effective, high-throughput, and designed for the lay user. It is freely available (chasemc.github.io/IDBac), and only requires access to a MALDI-TOF mass spectrometer (reflectron mode will be required for specialized metabolite analysis). Sample preparation relies on the simple “extended direct transfer” method[7,8] and data are collected with consecutive linear and reflectron acquisitions on a single MALDI-target spot. With IDBac, it is possible to analyze the putative phylogeny and specialized metabolite production of hundreds of colonies in under four hours, including sample preparation, data acquisition, and data visualization. This presents a significant time and cost advantage over traditional methods of identifying bacteria (such as gene sequencing), and analyzing metabolic output (liquid chromatography-mass spectrometry [LCMS] and similar chromatographic methods).

Using data obtained in linear mode analysis, IDBac employs hierarchical clustering to represent the relatedness of protein spectra. Since the spectra mostly represent ionized ribosomal proteins, they provide a representation of the phylogenetic diversity present in a sample. In addition, IDBac incorporates reflectron mode data to display specialized metabolite fingerprints as Metabolite Association Networks (MANs). MANs are bipartite networks that allow for easy visualization of shared and unique metabolite production between bacterial isolates. The IDBac platform allows researchers to analyze both protein and specialized metabolite data in tandem but also individually if only one data-type is acquired. Importantly, IDBac processes raw data from Bruker and Xiamen instruments, as well as txt, tab, csv, mzXML, and mzML. This eliminates the need for manual conversion and formatting of data sets, and significantly reduces the risk of user error or mishandling of MS data.

Copyright © 2019 Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License

Protocol

1. Preparation of MALDI matrix

1. Prepare 10 mg/mL MALDI-grade, and/or recrystallized α-cyano-4-hydroxycinnamic acid (CHCA) in MS-grade solvents: 50% acetonitrile (ACN), 47.5% water (H_2O), 2.5% trifluoroacetic acid (TFA). Example: 100 µL solution = 50 µL ACN + 47.5 µL H_2O + 2.5 µL TFA + 1 mg CHCA
    1. Prepare at least 1 µL of matrix solution per MALDI plate spot and vortex or sonicate until in solution (approximately 5 min sonication or no visible solids).
    CAUTION: TFA is a strong acid that should be handled in a chemical fume hood while wearing proper personal protective equipment, as it can damage skin, eyes, and airways with contact or inhalation.
    NOTE: CHCA is hygroscopic and light-sensitive and should be stored in amber vials in a desiccator. There are many MALDI matrix options available. CHCA is most common for protein profiling of bacteria, but also works for specialized metabolite analysis. Matrix selection depends on individual user/experiment needs.

2. Preparation of MALDI target plates

NOTE: See Sauer et al.[7], for more details.

1. Rinse MALDI plate with methanol (HPLC-grade or higher) and wipe dry with soft paper wipes. Do not use abrasive brushes when cleaning target plates, as this can permanently damage the surface of the target plate.
2. Assign protein and specialized metabolite calibrant spots. Organize calibration spots evenly across the sample population, to account for MALDI-plate-irregularities and instrument drift over time. Assign an appropriate number of media/matrix-blank spots for the study; these spots will contain only media and matrix, or only matrix.
3. Using a sterile toothpick, transfer a small portion of a bacterial colony to the appropriate spot on the MALDI plate. Spread the bacterial colony evenly over the spot. The spot should appear as flat as possible.
NOTE: It will be easier to flatten bacterial colonies that are more mucoid/amorphous. For more rigid/solid colonies, avoid leaving visible clusters of cell mass on the MALDI spot (**Figure 1**).

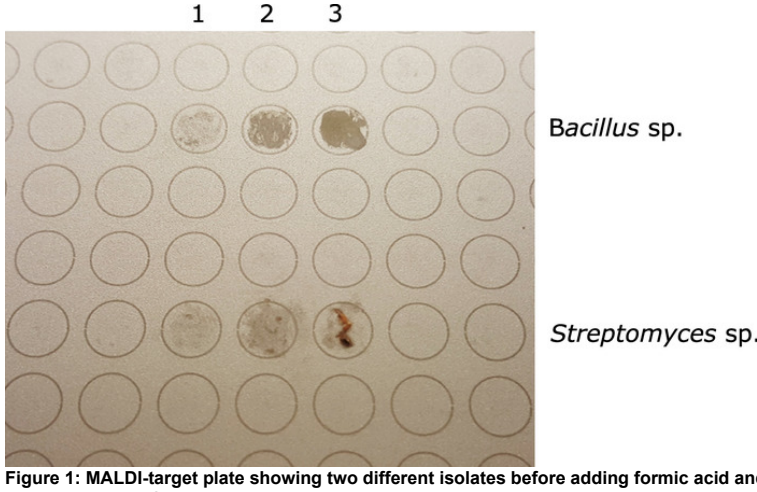


Figure 1: MALDI-target plate showing two different isolates before adding formic acid and MALDI matrix (top 3 spots - *Bacillus* sp.; bottom 3 spots - *Streptomyces* sp.). For both, column **3** represents excess sample; column **2** represents the appropriate amount of sample; column **1** represents insufficient sample for MALDI analysis. Please click here to view a larger version of this figure.

4. Prepare a matrix/media control by using a sterile toothpick to transfer a minimal amount of agar/media onto the appropriate spot(s) on the MALDI plate.
5. Overlay 1 µL of 70% mass spectrometry grade formic acid onto each sample spot, including the matrix control spots. Allow acid to air dry completely in a chemical fume hood (approximately 5 min).
CAUTION: Formic acid is a caustic chemical and should be handled in chemical fume hoods. It can damage airways if inhaled.
6. Add 1 µL of the prepared MALDI matrix solution to each sample spot, as well as to the matrix/media control spots. Allow matrix solution to air dry completely (approximately 5 min).
NOTE: It is possible to store the plate in a desiccator, in the dark, until it can be analyzed on a MALDI-TOF mass spectrometer. Allowable storage times may vary depending on sample stability.
7. Add 0.5–1.0 µL calibrant to the assigned calibration spots, followed by 1 µL MALDI matrix solution. Pipette the resulting solution up and down to mix. Allow all spots to air dry completely prior to introduction into the MALDI-TOF mass spectrometer.

Copyright © 2019 Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License

NOTE: The protein and specialized metabolite calibrants should be added within 30 min of MALDI analysis, as both are susceptible to degradation.

3. Data acquisition

NOTE: The general parameters for data acquisition are listed in **Table 1**.

Parameter	Protein	Specialized Metabolite
Mass Start (Da)	1920	60
Mass End (Da)	21000	2700
Mass Deflection (Da)	1900	50
Shots	500	1000
Frequency (Hz)	2000	2000
Laser Size	Large	Medium
MaxStdDev (ppm)	300	30

Table 1.

1. Following the protocols specific to the instrument being used, set up both protein and specialized metabolite calibrations.
2. Test a few separate target spots to determine the optimal laser power and detector gain to use when acquiring spectra (this will vary day-to-day and by instrument).
   NOTE: **Figure 2A** and **Figure 3A** show optimal spectra, while **Figure 2D** and **Figure 3D** are examples of poor-quality spectra.

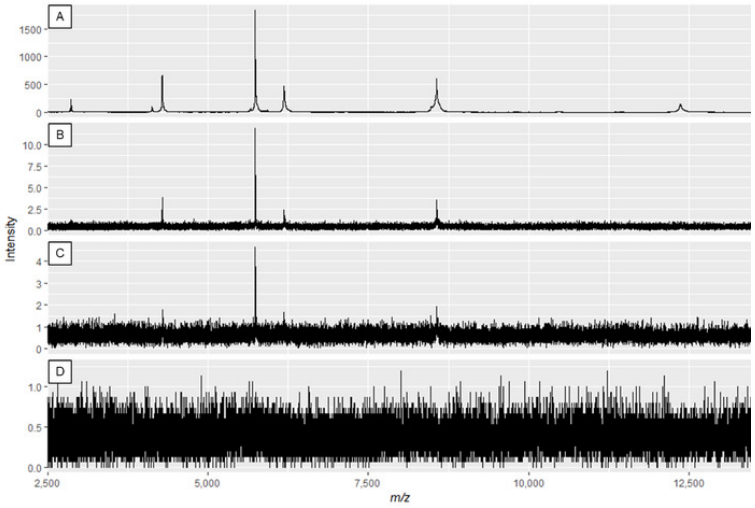


Figure 2: Example protein spectra displaying the effect of modifying laser power and detector gain. Spectra quality is best in panel **A**, and decreases until insufficient spectra quality in panels **C** and **D**. While the spectrum in panel **B** may result in useable peaks, panel **A** displays optimal data. Please click here to view a larger version of this figure.

Copyright © 2019 Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License

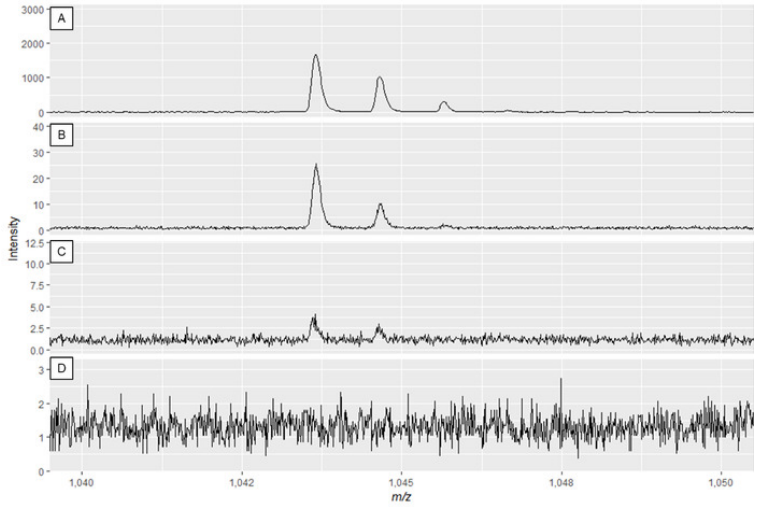


Figure 3: Example specialized metabolite spectra displaying the effect of modifying laser power and detector gain. Spectra quality is best in panel **A** and decreases until insufficient spectra quality in panels **C** and **D**. While the spectrum in panel **B** may result in useable peaks, panel **A** displays optimal data. Please click here to view a larger version of this figure.

3. Acquire spectra, saving protein spectra into one folder and specialized metabolite spectra into a second, separate folder.

4. Cleaning the MALDI target plate (adapted from Sauer et al.[7])

1. Remove the MALDI target plate from its holder and rinse with acetone.
2. Wash with a non-abrasive liquid soap to remove trace proteins and lipids, and soft paper wipes/soft-bristled toothbrush.
3. Rinse with de-ionized water for approximately 2 min to completely remove soap.
4. Sonicate the target plate in water (HPLC grade or higher) for 5 min.
5. Rinse the target plate with water (HPLC grade or higher).
6. Rinse the target plate with methanol (HPLC grade or higher).

5. Installing the IDBac Software

1. Download the IDBac software.
   NOTE: Permanent, versioned backups are also available for download (see the **Table of Materials**).
2. Double-click the downloaded “Install_IDBac.exe” to initiate the installer and follow the on-screen instructions.

6. Starting with Raw Data

NOTE: Detailed explanations and instructions of each data processing step are embedded within IDBac, however the main analyses and interactive inputs are described below.

1. Double-click the IDBac desktop shortcut to launch IDBac. IDBac will open on the **Introduction** tab by default.
2. Use the **Check for Updates** button to ensure that the most current version of IDBac is being used (requires internet access). If a newer version is available, IDBac will automatically download and install the update, after which IDBac will request to be restarted.
3. Click on the **Starting with Raw Data** tab and choose from the menu the type of data to be used with IDBac; continue by following the in-app instructions.
4. When setting-up the conversion and processing of data files, input a descriptive name for the experiment where prompted (see **Figure 4**). Experiments will later be displayed alphabetically, so a helpful strategy is to start experiment names with a group-attribute (e.g., “bacillus-trials_experiment-1”; “bacillus-trials_experiment-2”).

Copyright © 2019 Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License

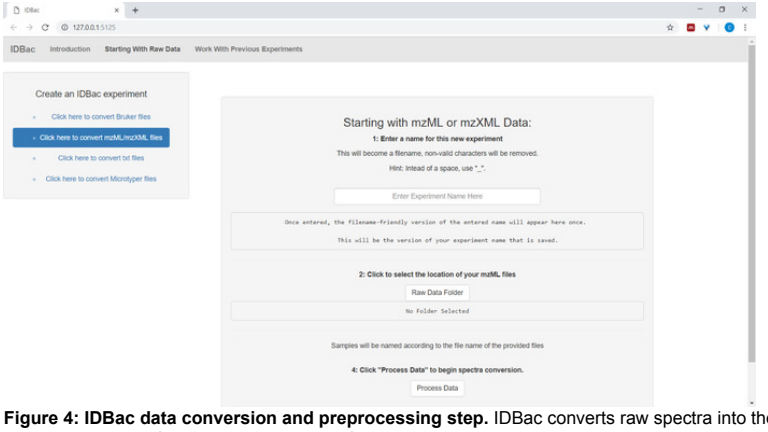


Figure 4: IDBac data conversion and preprocessing step. IDBac converts raw spectra into the open mzML format and stores mzML, peak lists, and sample information in a database for each experiment. Please click here to view a larger version of this figure.

7. Work with previous experiments

1. After converting files and processing them with IDBac, or anytime one wishes to reanalyze an experiment, navigate to the **Work with previous experiments** page and **Select an experiment to work with** (**Figure 5**).

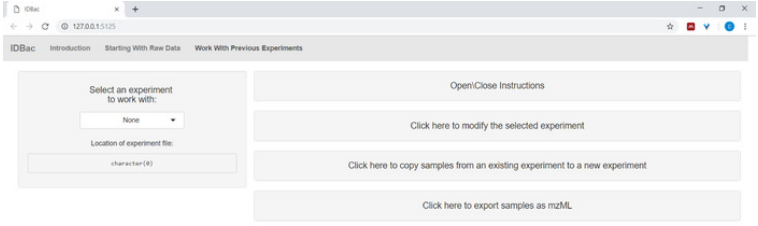


Figure 5: "Work with Previous Experiments" page. Use IDBac's "Work with Previous Experiments" page to select an experiment to analyze or modify. Please click here to view a larger version of this figure.

2. (Optional) Add information about samples using the menu **Click here to modify the selected experiment**. Input information into the auto-populated spreadsheet and press **Save** (**Figure 6**). This option allows the user to color-code data during analyses.

Copyright © 2019 Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License

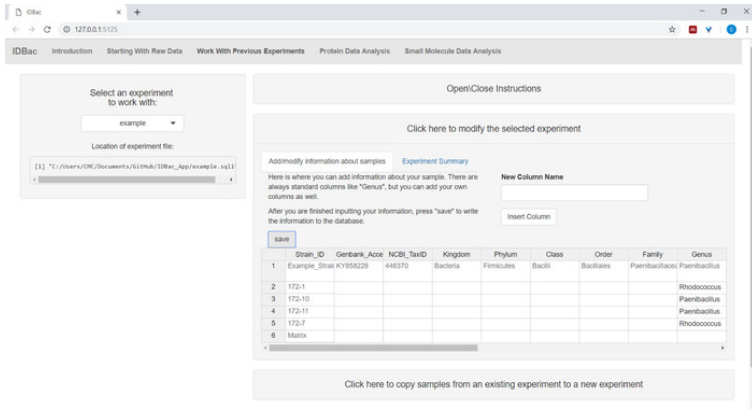


Figure 6: Input sample information. Within the "Work with Previous Experiments" page users can input information about samples such as taxonomic identity, collection location, isolation conditions, etc. Please click here to view a larger version of this figure.

3. (Optional) Transfer all, or a subset of samples to a new or another experiment by clicking **Transfer samples from previous experiments to new/other experiments** and following the provided instructions (**Figure 7**).

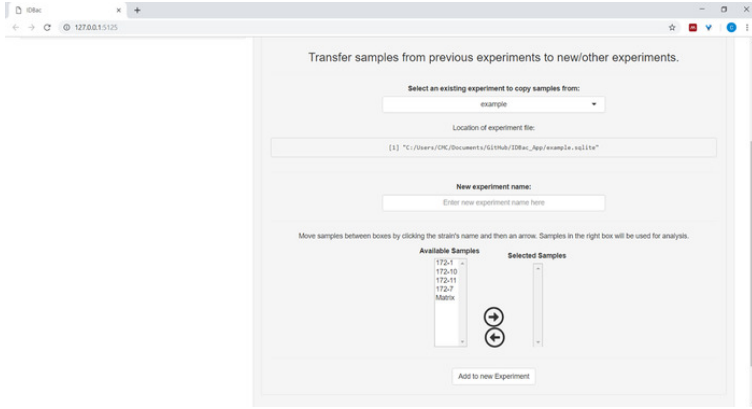


Figure 7: Transfer data. The "Work with Previous Experiments" page contains the option to transfer data between existing experiments and to new experiments. Please click here to view a larger version of this figure.

4. When ready to begin analysis, ensure the experiment to work with is selected. Select either **Protein Data Analysis** or **Small Molecule Data Analysis**.

8. Setting up protein data analysis and creating mirror plots

1. If analyzing protein data, first navigate to the **Protein Data Analysis** page. Choose peak-picking settings and evaluate protein spectra of samples via the displayed mirror plots (**Figure 8**).
   NOTE: In the mirror plots, a red peak signifies the presence of that peak only in the top spectrum, while blue peaks represent those occurring in both spectra.

Copyright © 2019 Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License

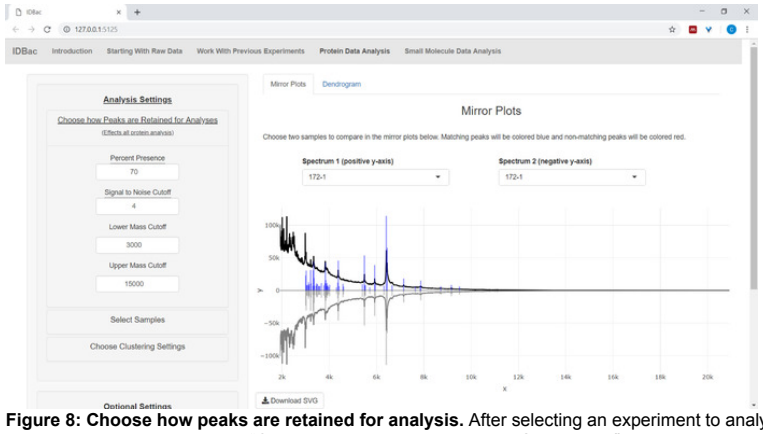


Figure 8: Choose how peaks are retained for analysis. After selecting an experiment to analyze, visiting the "Protein Data Analysis" page and subsequently opening the "Choose how Peaks are Retained for Analysis" menu allows users to choose settings like signal-to-noise ratio for retaining peaks. The displayed mirror plot (or dendrogram) will automatically update to reflect the chosen settings. Please click here to view a larger version of this figure.

2. Adjust the percentage of replicates in which a peak must be present in order for it to be included for analyses (e.g., if the threshold is set to 70% and a peak occurs in at least 7 out of 10 replicates, it will be included).
3. Using the mirror plots as visual guidance, adjust the signal to noise cutoff that retains the most "genuine" peaks and the least noise, noting that more replicates and a higher "percentage peak presence" value will allow selection of a lower signal to noise cutoff.
4. Specify the lower and upper *m/z* cutoffs, dictating the range of mass values within each spectrum to be used in further analyses by IDBac.

9. Clustering samples using protein data

1. Within the **Protein Data Analysis** page, select the **Dendrogram** tab. This allows for grouping samples into a dendrogram according to user-selected distance measures and clustering algorithms.
2. Click **Select Samples** on the menu and follow the instructions to select samples to include in the analyses. Only samples that contain protein spectra will be displayed within the **Available Samples** box (**Figure 9**).

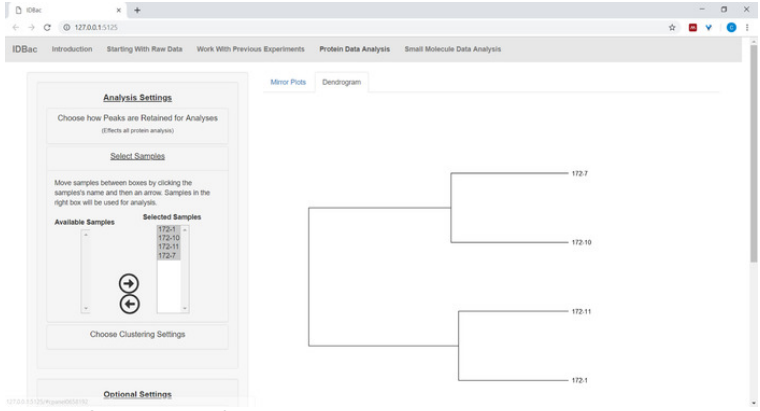


Figure 9: Select samples from the chosen experiment to include within the displayed dendrogram. Please click here to view a larger version of this figure.

3. Use the default values or, under **Choose Clustering Settings**, select the desired **Distance** and **Clustering** algorithms to be applied to the generation of the dendrogram.
4. Select **Presence/Absence** as input. Alternatively, if confident about the peak heights of the samples (e.g., after performing a study to assess variability of peak intensity), select **Intensities** as input.
   NOTE: At the time of publication, IDBac provides flexibility in settings for clustering, relying on users to choose the appropriate combinations. If unfamiliar with these options, it is suggested to pair either A: "cosine" distance, and "average (UPGMA)" clustering; or B: "Euclidean" distance, and "Ward.D2" clustering.
5. To display bootstrap values on the dendrogram, enter a number between 2 and 1000 under **Bootstraps**.
6. When reporting results, copy the text within the **Suggestions for Reporting Protein Analysis** paragraph. This provides the user-defined settings that generated the specific dendrogram.

Copyright © 2019 Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License

10. Customizing the protein dendrogram

1. To begin customizing the dendrogram, open the **Adjust the dendrogram** menu (**Figure 10**).

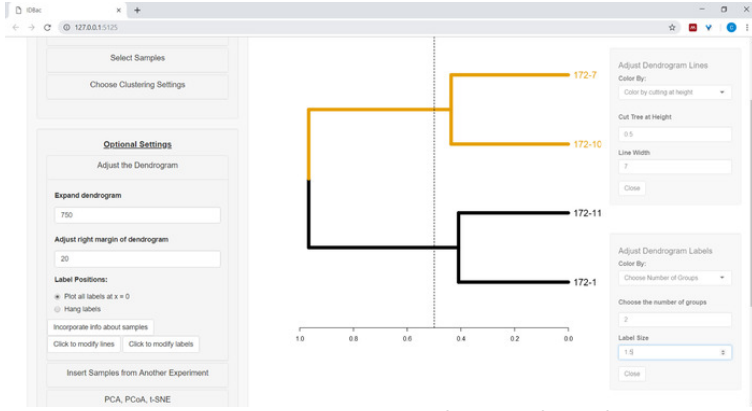


Figure 10: Adjust the dendrogram. IDBac provides a few options for modifying how the dendrogram looks, these may be found within the menu “Adjust the Dendrogram”. This includes coloring branches and labels by k-means, or by “cutting” the dendrogram at a user-provided height. Please click here to view a larger version of this figure.

2. To color the dendrogram’s lines and/or labels select the appropriate button: **Click to modify lines** or **Click to modify labels** and select the desired options.
3. To plot information from the spreadsheet next to the dendrogram (see step 7.2), select the button **Incorporate info about samples**. This will open a panel where a category (column in the spreadsheet) will self-populate based on the entered values (**Figure 11**).

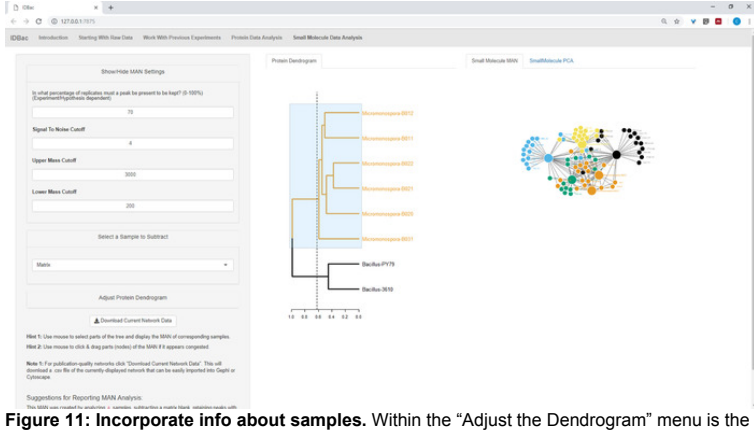


Figure 11: Incorporate info about samples. Within the “Adjust the Dendrogram” menu is the option “Incorporate info about samples”. Selecting this will allow plotting information about samples next to the dendrogram. Sample information is input within the “Work with Previous Experiments” page. Please click here to view a larger version of this figure.

11. Insert samples from a separate experiment into the dendrogram

1. To insert samples from another experiment, select the menu button **Insert Samples from Another Experiment**. Follow the directions in the newly-opened panel (**Figure 12**).

Copyright © 2019 Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License

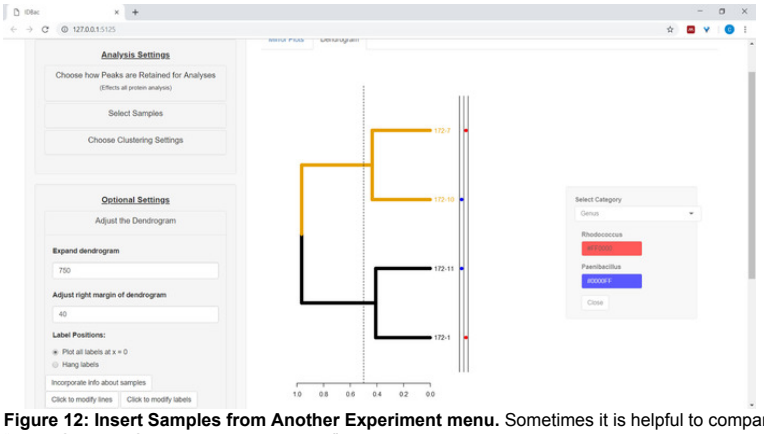


Figure 12: Insert Samples from Another Experiment menu. Sometimes it is helpful to compare samples from another experiment. Use the "Insert Samples from Another Experiment" menu to choose samples to include within the currently-displayed dendrogram. Please click here to view a larger version of this figure.

12. Analyzing specialized metabolite data and metabolite association networks (MANs)

1. Proceed to the **Metabolite Association Network (Small-Molecule)** page. This page allows for data visualization by principle components analysis (PCA) and MANs, which use bipartite networks to display the correlation of small molecule *m/z* values with samples.
2. If a protein dendrogram was created (section 9), it will be displayed on this page as well. Click-and-drag on the dendrogram to highlight select samples of interest to be analyzed. If no samples are highlighted or no protein dendrogram was created, a MAN of a either a random subset or all samples will appear, respectfully (**Figure 13**).

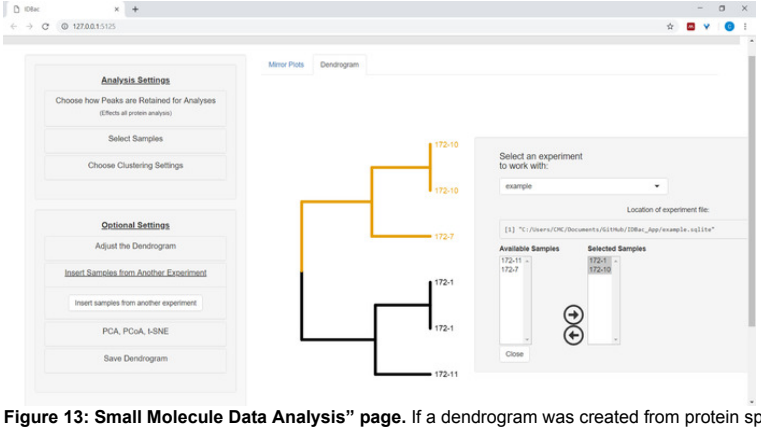


Figure 13: Small Molecule Data Analysis" page. If a dendrogram was created from protein spectra, it will be displayed within the "Small Molecule Data Analysis" page. This page will also display Metabolite Associate Networks (MANs) and Principle Components Analysis (PCA) for small molecule data. Please click here to view a larger version of this figure.

3. To subtract a matrix/media blank in the MAN, open the menu **Select a Sample to Subtract** and choose the appropriate sample to use as a blank.
4. Open the menu **Show/Hide MAN Settings** to select the desired values for percentage of peak presence in replicates, signal to noise, and upper and lower mass cutoffs, as was done for protein spectra in Section 9. Use the small molecule mirror plots to guide the selection of these settings.
5. Select "Download Current Network Data" to save the data of the MAN that is currently displayed. These data can be used in network analysis software other than IDBac.
6. For reporting results, copy the text within the **Suggestions for Reporting MAN Analysis** paragraph. This provides the user-defined settings used to generate the created MAN.

Copyright © 2019 Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License

13. Sharing data

1. Each IDBac "experiment" is saved as a single SQLite database. It contains the converted mzML raw spectra, detected peaks, and all user-input information about samples. Therefore, to share an IDBac experiment simply share the SQlite file that has the same name as the experiment (the file location is displayed on the **Working with Previous Experiments** page).

Representative Results

We analyzed six strains of *Micromonospora chokoriensis* and two strains of *Bacillus subtilis*, which were previously characterized[6], using data available at DOI: 10.5281/zenodo.2574096. Following directions in the **Starting with Raw Data** tab, we selected the option **Click here to convert Bruker files** and followed the IDBac-provided instructions for each dataset (**Figure 14**).

After the automated conversion and preprocessing/peak-peaking steps were completed, we proceeded to create a new combined IDBac experiment by transferring samples from the two experiments into a single experiment containing both *Bacillus* and *Micromonosopora* samples (**Figure 15**). The resulting analysis involved comparing protein spectra using mirror plots, as pictured in **Figure 16**, which was useful for evaluating spectra quality and adjusting peak-picking settings. **Figure 17** displays a screenshot of the protein clustering results with default settings selected. The dendrogram was colored by adjusting the threshold on the plot (appears as a dotted line). Of note is the clear separation between genera, with both *M. chokoriensis* and *B. subtilis* isolates clustering separately.

Figure 18, **Figure 19**, and **Figure 20** highlight the ability to generate MANs of user-selected regions by clicking and dragging across the protein dendrogram. With this we were able to rapidly create MANs to compare only the *B. subtilis* strains (**Figure 18**), only the *M. chokoriensis* strains (**Figure 19**), and all the strains simultaneously (**Figure 20**). The primary function of these networks is to provide researchers with a broad overview of the degree of specialized metabolite overlap between bacteria. With these data in hand, researchers now have the capacity to make informed decisions from only a small amount of material scraped from a bacterial colony.

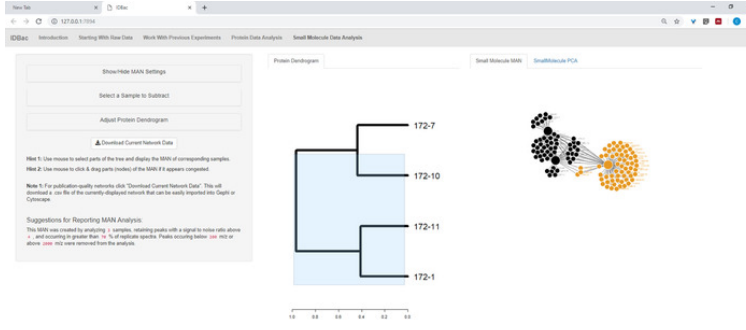


Figure 14: Spectra processing. Downloaded Bruker autoFlex spectra were converted and processed using IDBac. Please click here to view a larger version of this figure.

Copyright © 2019 Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License

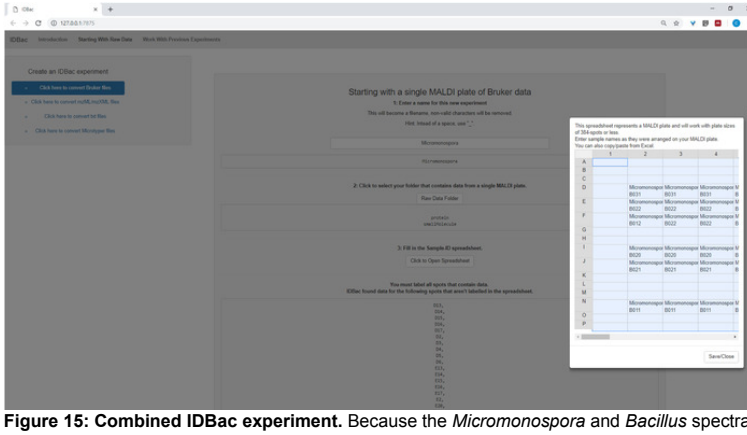


Figure 15: Combined IDBac experiment. Because the *Micromonospora* and *Bacillus* spectra were collected on different MALDI target plates, the two experiments were subsequently combined into a single experiment-"Bacillus_Micromonsopora". This was done within the "Work with Previous Experiments" tab, following directions within the menu "Transfer samples from previous experiments to new/other experiments". Please click here to view a larger version of this figure.

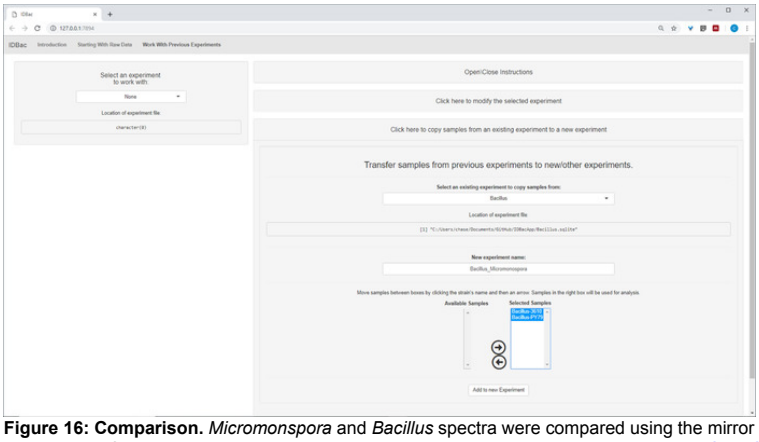


Figure 16: Comparison. *Micromonspora* and *Bacillus* spectra were compared using the mirror plots within the "Protein Data Analysis" page. Ultimately, default peak settings were chosen. Please click here to view a larger version of this figure.

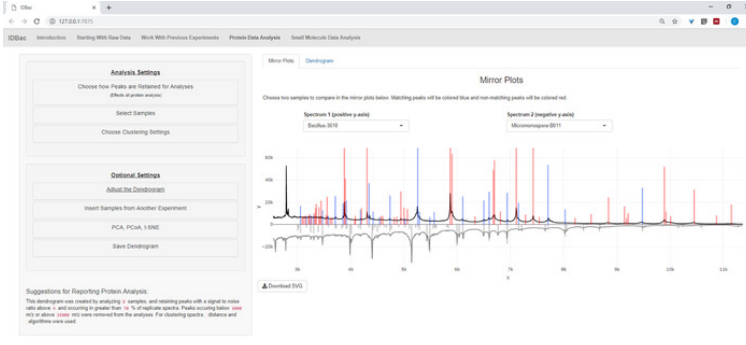


Figure 17: Hierarchical clustering. Hierarchical clustering, using default settings, correctly grouped *Bacillus* and *Micromonospora* isolates. The dendrogram was colored by "cutting" the dendrogram at an arbitrary height (displayed as a dashed-line) and 100 bootstraps used to show confidence in branching. Please click here to view a larger version of this figure.

Copyright © 2019 Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License



Figure 18: MAN created by selecting the *Bacillus* sp. strains from the protein dendrogram showed differential production of specialized metabolites. Please click here to view a larger version of this figure.

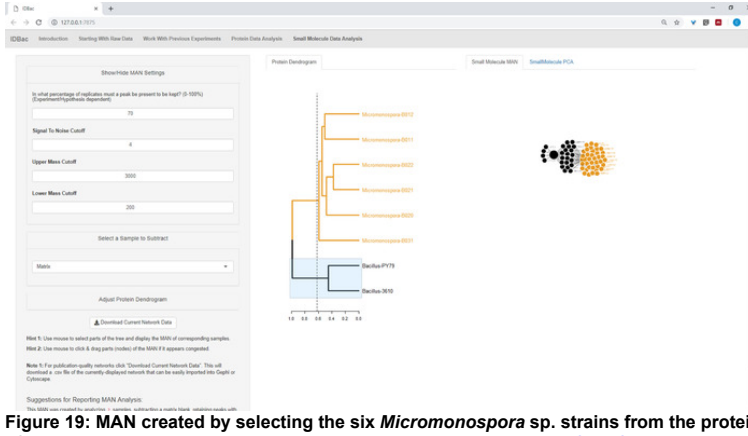


Figure 19: MAN created by selecting the six *Micromonospora* sp. strains from the protein dendrogram showed differential production of specialized metabolites. Please click here to view a larger version of this figure.

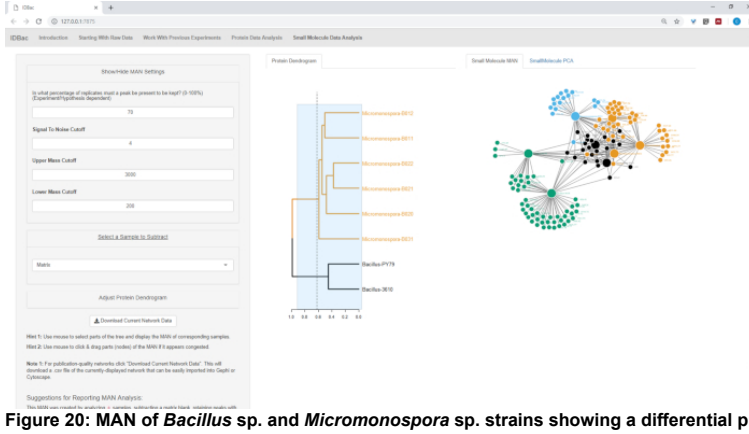


Figure 20: MAN of *Bacillus* sp. and *Micromonospora* sp. strains showing a differential production of specialized metabolites. Please click here to view a larger version of this figure.

Copyright © 2019 Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License

Discussion

The IDBac protocol details bacterial protein and specialized metabolite data acquisition and analysis of up to 384 bacterial isolates in 4 h by a single researcher. With IDBac there is no need to extract DNA from bacterial isolates or generate specialized metabolite extracts from liquid fermentation broths and analyze them using chromatographic methods. Instead, protein and specialized metabolite data are gathered by simply spreading material from bacterial colonies directly onto a MALDI target plate. This greatly reduces the time and cost associated with alternative techniques such as 16S rRNA gene sequencing and LCMS[9].

It is important to add a matrix blank and calibration spots to the MALDI plate, and we recommend using an appropriate number of replicates to ensure reproducibility and statistical confidence. The numbers of replicates will be experiment-dependent. For example, if a user intends to differentiate thousands of colonies from a collection of environmental diversity plates, fewer replicates may be necessary (our lab collects three technical replicates per colony). Alternatively, if a user wishes to create a custom database of strains from specific bacterial taxa to rapidly determine sub-species classifications of unknown isolates, then more replicates are appropriate (our lab collects eight biological replicates per strain).

IDBac is a tool for rapidly differentiating highly-related bacterial isolates based on putative taxonomic information and specialized metabolite production. It can complement or serve as a precursor to orthogonal methods such as in-depth genetic analyses, studies involving metabolite production and function, or characterization of specialized metabolite structure by Nuclear Magnetic Resonance spectroscopy and/or LC-MS/MS.

Specialized metabolite production (IDBac MANs) is highly susceptible to bacterial growth conditions, especially using different media, which is a potential limitation of the method. However these traits may be exploited by the user, as IDBac can readily generate MANs showing the differences in specialized metabolite production under a variety of growth conditions. It is important to note that while specialized metabolite fingerprints may vary by growth condition, we have previously shown that protein fingerprints remain relatively stable across these variables (see Clark et al.[6]). When dealing with environmental diversity plates, we recommend purifying bacterial isolates prior to analysis in order to reduce possible contributions from neighboring bacterial cross-talk.

Finally, the lack of a searchable public database of protein MS fingerprints is a major shortcoming in the use of this method to classify unknown environmental bacteria. We created IDBac with this in mind, and included automated conversion of data into a community-accepted open-source format (mzML)[10,11,12] and designed the software to allow searching, sharing, and creation of custom databases. We are in the process of creating a large public database (>10,000 fully characterized strains), which will allow for the classification of some isolates to the species-level, including links to GenBank accession numbers when available.

IDBac is open source and the code is available for anyone to customize their data analysis and visualization needs. We recommend that users consult an extensive body of literature (Sauer et al.[7], Silva et al.[5]) to help support and design their experimental goals. We host a forum for discussion at: https://groups.google.com/forum/#!forum/idbac and a means to report issues with the software at: https://github.com/chasemc/IDBacApp/issues.

Disclosures

The authors have nothing to disclose.

Acknowledgments

This work was supported by National Institute of General Medical Sciences Grant R01 GM125943, National Geographic Grant CP-044R-17; Icelandic Research Fund Grant 152336-051; and University of Illinois at Chicago startup funds. Also, we thank the following contributors: Dr. Amanda Bulman for assistance with MALDI-TOF MS protein acquisition parameters; Dr. Terry Moore and Dr. Atul Jain for recrystallizing alpha-cyano-4-hydroxycinnamic acid matrix (CHCA).

References

1. Sandrin, T. R., Goldstein, J. E., Schumaker, S. MALDI TOF MS profiling of bacteria at the strain level: A review. *Mass Spectrometry Reviews.* **32** (3), 188-217 (2013).
2. Cain, T. C., Lubman, D. M., Weber, W. J., Vertes, A. Differentiation of bacteria using protein profiles from matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry.* **8** (12), 1026-1030 (1994).
3. Holland, R. D., Wilkes, J. G., et al. Rapid identification of intact whole bacteria based on spectral patterns using matrix-assisted laser desorption/ionization with time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry.* **10** (10), 1227-1232 (1996).
4. Rahi, P., Prakash, O., Shouche, Y. S. Matrix-assisted laser desorption/ionization time-of-flight mass-spectrometry (MALDI-TOF MS) based microbial identifications: challenges and scopes for microbial ecologists. *Frontiers in Microbiology.* **7**, 1359 (2016).
5. Silva, R., Lopes, N. P., Silva, D. B. Application of MALDI mass spectrometry in natural products analysis. *Planta Medica.* **82**, 671-689 (2016).
6. Clark, C. M., Costa, M. S., Sanchez, L. M., Murphy, B. T. Coupling MALDI-TOF mass spectrometry protein and specialized metabolite analyses to rapidly discriminate bacterial function. *Proceedings of the National Academy of Sciences of the United States of America.* **115** (19), 4981-4986 (2018).
7. Freiwald, A., Sauer, S. Phylogenetic classification and identification of bacteria by mass spectrometry. *Nature Protocols.* **4** (5), 732-742 (2009).
8. Schulthess, B., Bloemberg, G. V, Zbinden, R., Böttger, E. C., Hombach, M. Evaluation of the Bruker MALDI Biotyper for identification of Gram-positive rods: development of a diagnostic algorithm for the clinical laboratory. *Journal of Clinical Microbiology.* **52** (4), 1089-97 (2014).

Copyright © 2019 Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License

9. Schumann, P., Maier, T. MALDI-TOF mass spectrometry applied to classification and identification of bacteria. *Methods in Microbiology.* **41**, 275-306 (2014).
10. Chambers, M. C., Maclean, B., et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology.* **30** (10), 918-920 (2012).
11. Kessner, D., Chambers, M., Burke, R., Agus, D., Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics.* **24** (21), 2534 (2008).
12. Martens, L., Chambers, M., et al. mzML-a community standard for mass spectrometry data. *Molecular & Cellular Proteomics.***10** (1), (2011).

Copyright © 2019 Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License

Paper III

A new paradigm to minimize taxonomic and natural product redundancy in microbial libraries using MALDI-TOF MS/IDBac.

Maria S. Costa [†, ‡ #], Chase M. Clark[‡#], Sesselja Ómarsdóttir[†], Laura M. Sanchez[‡] and Brian T. Murphy[‡*]

Affiliations:

[†]Faculty of Pharmaceutical Sciences, University of Iceland, Hagi, Hofsvallagata 53, IS-107 Reykjavík, Iceland;
[‡]Department of Medicinal Chemistry and Pharmacognosy, College of Pharmacy, University of Illinois at Chicago, 833 South Wood Street (MC 781), Room 539, Chicago, IL, United States
[#]Authors contributed equally to this work

*btmurphy@uic.edu

Tel: 312-413-9057

Fax: (312)-413-930

ABSTRACT

Libraries of microorganisms have been a cornerstone of drug discovery efforts since the mid-1950s, however the high degree of strain duplication in some libraries has resulted in unwanted natural product redundancy. In the current study, we implemented a workflow that minimizes both the natural product overlap and the total number of bacterial isolates in a library. Using a collection expedition to Iceland as an example, we purified *every distinct* bacterial colony (1,616 total) off of isolation plates derived from 86 environmental samples. We employed our mass spectrometry (MS) based IDBac workflow on these isolates to form groups of taxa based on protein MS fingerprints (3-15 kDa), and further distinguished taxa subgroups based on their degree of overlap within corresponding natural product spectra (0.2-2 kDa). This informed the decision to create a library of 301 isolates spanning 54 genera. This process required only 25 hours of data acquisition and 2 hours of analysis. In a separate experiment, we reduced the size of an existing library based on the degree of metabolic overlap observed in natural product MS spectra of bacterial colonies (from 833 to 233 isolates, a 72.0% reduction). Overall, our pipeline allows for the reduction of library size and costs associated with library generation, and attempts to minimize natural product redundancy entering into downstream biological screening efforts.

INTRODUCTION

Libraries of microorganisms have been a cornerstone of drug discovery efforts since the middle of the 1950s. Natural products (including their semi-synthetic derivatives) isolated from these libraries have afforded us more than 170 cancer drugs and greater than half of marketed anti-infective drugs.[1–4] In particular, discoveries of pyocyanase, penicillin, and tyrothricin (gramicidin and tyrocidine) ushered in an unprecedented global effort to mine the environment for new microbial natural products.[5–8] This effort was highly successful, and was driven by sampling expeditions whose aim was to amass libraries of cultivable microorganisms from the environment. Many method innovations in pharma involving automation, miniaturization, cultivation, and biological screening, were responsible for these successes.[9,10] Despite this, by the end of the "Golden Age" of antibiotic discovery in roughly the 1970s, the re-isolation of known natural products became (and remains) a major problem.[10,11] One cause of this was the high degree of strain duplication in these libraries, which resulted in unwanted natural product redundancy and lower coverage of chemical space.[12–15] This redundancy is the result of a few circumstances: 1) strain libraries were often created on the basis of visual inspection of colony morphology, which does not fully account for natural product production;[13,14,16] 2) bacterial isolation efforts focused heavily on spore forming bacteria such as Actinobacteria and Firmicutes, which have a history of producing biologically active natural products.[2] Diminishing returns on this investment and a shift toward combinatorial chemistry approaches, among several other factors, led most of the pharmaceutical industry to divest in microbial-based natural product drug discovery.[2,13,17–20]

In order to overcome the high rate of compound re-discovery, it is necessary to reduce the degree of overlapping inter- and intra-species chemical space in a microbial library. We previously developed a freely available high-throughput matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS) data acquisition and bioinformatics pipeline – IDBac – that allows researchers to rapidly collect putative phylogenetic and natural product information from single colonies of unknown environmental bacteria.[21] From each colony, we generate two data sets: 1) protein

fingerprints (3,000-15,000 Da, primarily ribosomal)[22,23] that are used to assign each isolate into putative genus/species groupings, and 2) natural product fingerprints (200-2,000 Da) of each colony to compare inter- and intra-species differences/similarities in natural product production. We recently published a video protocol detailing this procedure.[24] A few reviews and more recent studies highlight attempts to prioritize bacterial strains in a library with regard to either taxonomic identity or natural product potential; this list is not comprehensive.[12,13,15,16,25–30]

Herein, we provide a template for the generation of a diverse microbial library using MALDI-TOF MS and IDBac. We broadly define natural product diversity as natural products that exhibit different structures. To demonstrate the effectiveness of our method, we applied IDBac to samples collected during an expedition to Iceland in 2017. From 1,616 total bacterial isolates, we created a library of 301 isolates that span 54 genera and exhibited minimal natural product overlap, based on metabolic profiling. This pipeline allowed the creation of a diverse library while requiring limited resources (toothpick, bacterial colonies, access to a MALDI-TOF MS), and can be heavily tailored toward creating custom collections of individual taxa, targeting the isolation of understudied taxa, and studying patterns of pseudo-phylogeny and natural product production within large groups of microbial isolates. Use of this method provides an alternative to tedious and costly liquid fermentation, extraction, or chromatographic analyses that would be used to generate similar taxonomic and natural product information.

RESULTS AND DISCUSSION

Collection and processing of environmental samples. In total, 86 samples from 30 locations, consisting of sediment (46), marine invertebrates (25), algae (12), and water column (3), were collected (see Table S2 and Figure S1, Supporting Information). Each environmental sample was processed using common techniques that select for spore-forming and non-spore-forming bacteria, and plated onto five different agar-based nutrient media.[31–33] After incubation of up to 60 days, *every distinct* bacterial colony growing on nutrient media was isolated (those appearing to exist as a single phenotype;

Figure S2, Supporting Information). When all 1,616 isolates were obtained, each was re-plated in individual wells of a 48-well agar plate onto high nutrient A1 media to facilitate comparison on the same time scale (marine and freshwater-derived datasets were analyzed separately). Single colonies were then transferred to a MALDI target plate, data were acquired on an Autoflex Speed LRF MALDI-TOF mass spectrometer, and analyzed using IDBac. The final step of re-plating, incubating, and acquiring data on bacterial isolates was performed in a three-week span in order to allow for simultaneous growth and triplicate data acquisition for each colony (4,848 total MALDI spots). With advanced instrumentation and automation (bacterial isolation and sample preparation), this entire process could conceivably be improved to a few days to one week.

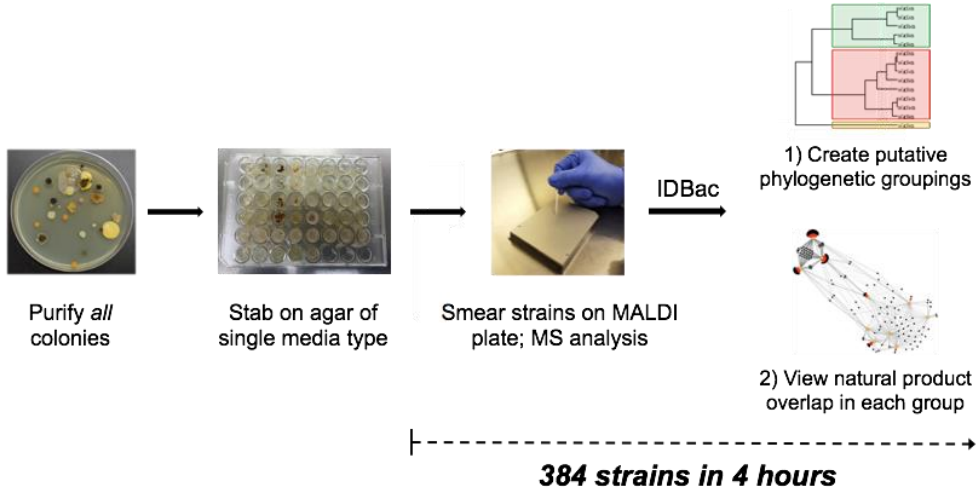


Figure 1. Collection trip workflow. Above is the recommended order of procedures for minimizing natural product redundancy in a microbial library. With a few basic laboratory supplies (nutrient agar, multi-well plates, toothpicks) and access to a MALDI-TOF MS, a researcher can build a diverse microbial library that doesn't rely on analysis based on phenotypic traits or high numbers of isolates. This IDBac-generated library incorporates information from protein *and* natural product MS analyses, without requiring liquid fermentation, extraction, or chromatographic analyses.

Targeting of *all* distinct colonies represents a major deviation from some past library generation practices that employed visual inspection of morphology and other selective colony picking techniques to build libraries. The ease of data collection and throughput of this pipeline (sample preparation, data collection and visualization of 384 samples in under 4 hours) allowed us to incorporate high numbers of colonies that otherwise would not have been included using our previous protocols, which followed common practices.

Grouping and analysis of bacterial isolates into putative taxa based on MALDI-TOF MS spectra (3,000-15,000 Da). MALDI-TOF MS is a well-established technique used to identify and group microbial isolates based on analysis of MS spectra in the 3,000-15,000 Da range.[34] Since resulting MS fingerprints are largely due to the high copy number of ionized ribosomal proteins,[22,23] pseudo-phylogenetic relationships can be inferred, often achieving species and sub-species level resolution. Application of this technique has been thoroughly reviewed elsewhere.[22,35,36] In relevance to the current study, Dieckmann *et al.* were among the first to use MALDI-TOF MS to dereplicate bacterial isolates from marine sponges, though only protein spectra were taken into consideration, limiting isolate dereplication to taxonomic groupings as opposed to natural product production.[37]

The 1,616 bacterial isolates from Iceland were first separated into two sets: those isolated from marine and freshwater locations. Each set was grown on media supplemented with and without artificial ocean salts, respectively. In our experience, a colony often exhibits different MS profiles in the presence of synthetic ocean salt in nutrient agar compared to its salt-free counterpart. Colonies were transferred to steel target plates and MALDI-TOF MS data were acquired and subsequently processed in IDBac, as previously described.[21,24] IDBac offers several settings when creating hierarchical clustering plots based on protein MS spectra. We suggest pairing cosine-distance and average-linkage-clustering settings for library creation efforts. Average-linkage conserves the 0 to 1 scale of the cosine similarity score, allowing the resulting dendrogram to be more easily interpreted (Figure 2). To confirm that the resulting MALDI MS protein groupings were representative of phylogenetic similarity, we

subsampled 361 isolates across both dendrograms (27.5% of freshwater and 17.8% of marine isolates) and sequenced their 16S rRNA genes (Table S2, Supporting Information). We observed that isolate groupings formed as a result of 16S rRNA phylogenetic analysis closely resembled those formed by MALDI-TOF/IDBac analysis (Figure S3, Supporting Information). Our results corroborate many previous efforts that document the use of MALDI-TOF MS to facilitate the grouping and/or taxonomic identification of environmental isolates.[35,38–40]

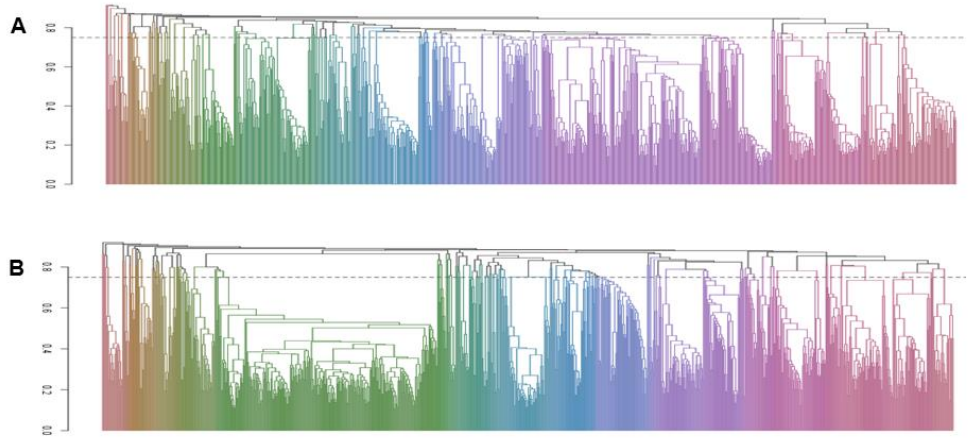


Figure 2. Dendrograms resulting from hierarchical clustering of MALDI-TOF MS protein spectra in IDBac of 733 freshwater isolates (A), and 883 marine isolates (B). The dendrograms were “cut” at a threshold of 0.75 (dashed lines) to create groups of similar isolates. Dendrograms were colored according to groups created by these cuts.

It is important to note that while isolates often group at the species-level, the dendrograms created in IDBac are *not* phylogenetic trees. We noticed in our dendrograms that MALDI-TOF MS analysis of bacterial proteins generally results in groupings that represent the genus- taxonomic level and below but lose fidelity of clustering at higher-taxonomic levels. As recently shown by Seuylemezian *et al.*, while highly similar isolates group together based on MALDI protein spectra similarity, the overall topology of clusters was slightly different than those found in a phylogenetic tree

constructed using the 16S rRNA gene.[41] In our data, while genera such as *Shewanella*, *Bradyrhizobium*, *Thiomonas,* and *Salinococcus* (among others) form single groupings in the dendrogram, some larger, more diverse genera such as *Bacillus*, *Paenibacillus*, *Streptomyces*, and *Micromonospora* form multiple groupings distributed throughout each dendrogram. To confirm this wasn't due to outside variables (date spectra were acquired, batch of growth media used to purify colonies, etc.), we seeded the freshwater dataset with protein spectra from previously characterized strains. Each of three *Micromonospora* strains from our in-house collection (*Micromonospora aurantiaca* D053 and *M. humi* D077 from marine sediment collected off the coast of Massachusetts, and *M. aurantiaca* B017 from freshwater sediment in Lake Michigan) matched with a corresponding *Micromonospora* grouping in both freshwater and marine datasets (Figure S4-S5, Supporting Information). Similar coherence was observed when seeding our freshwater and marine datasets with characterized *Bacillus* strains (*Bacillus subtilis* 3610; *B. subtilis* NRRL B14596; Figure S4-S5, Supporting Information). The spread of these genera across the dendrogram reflects, in some cases, the higher degree of resolution achieved by MALDI-TOF MS, since signals represent ionized proteins from more than just ribosomal origin. It also highlights the limitations of methods used to define bacterial taxonomy, since two isolates may share high 16S rRNA sequence homology but low genomic DNA homology, as explained previously.[41] In the presence of a large reference database of protein spectra from characterized environmental microorganisms, the rapid taxonomic identification of unknown isolates is possible. A few studies describe the creation of custom MALDI-TOF protein MS databases to assist in the identification of specific bacterial taxa, however the absence of an extensive database is a major roadblock to the use of MALDI-TOF MS as a means to identify bacteria isolated from the environment.[41–43] Currently, IDBac allows the user to build custom databases that can be readily searched, modified, and shared.

In the context of creating a diverse bacterial library for natural product discovery, it is not essential that hierarchical clustering perfectly mirrors phylogenetic groupings. Clustering isolates based on protein MS fingerprints primarily serves as a data reduction strategy for the following step, comparison of natural product production within closely-related

groups of isolates. Occasionally an isolate “mismatches” to a cluster that contains other genera of bacteria (e.g. a *Bacillus* isolate is paired with a group of *Streptomyces* isolates); this is generally due to the absence of close relatives in the dataset and a number of shared peaks emanating from non-ribosomal proteins. Regardless, in our experience it will stand as an obvious outlier when comparing natural product overlap between isolates within that cluster.

In the case of the Iceland protein MS dataset, a threshold of 0.75 was used to create 72 distinct freshwater and 101 distinct marine isolate groupings. Groups contained anywhere between 1 and 229 isolates. Each group was then manually selected within IDBac to generate a metabolite association network (MAN). MANs were used as the basis for selecting isolates with minimal overlapping natural product production for addition to the library. Strategies to analyze natural product overlap within these groups are discussed in the next section.

Use of Metabolite Association Networks (MANs) to reduce inter- and intra-species natural product redundancy in a bacterial library. It has been demonstrated that natural product production, broadly speaking, follows taxa-specific trends at the genus level.[44–46] Conversely, genera may also exhibit extensive natural product biosynthetic gene diversity at the genus level.[47,48] Further, it is well known that on a species-level, isolates with highly similar phenotypes can display a differential capacity to produce natural products.[45,47,49] We previously demonstrated a similar phenomenon in a group of *Micromonospora* strains isolated from the same 1 cm^3 of sediment.[21] In the context of IDBac, grouping bacteria by taxa is an effective means by which the user can detect these more subtle genus/species-level differences in natural product production, and account for molecules that deviate from taxa-specific patterns.

As shown in Figures 3 and 4B, larger nodes represent individual bacterial isolates, while smaller nodes represent *m/z* values detected in their corresponding MS spectra (200-2,000 Da). Matrix and media ions were automatically subtracted from the resulting analysis. We previously demonstrated that the majority of *m/z* values in a MAN between

two *Bacillus* strains (33 of 42; 78.6%) represented natural products, giving confidence in the ability of IDBac to filter out matrix signals.[21] It is worth nothing that the mass resolving power of MALDI-TOF MS does not allow for level-1 or -2 identifications of natural products;[50] further validation is necessary to elucidate a structure.

Since it is difficult to observe meaningful intra-species natural product production relationships in a MAN of all 1,616 isolates, it is recommended to view MANs in smaller subsets based on hierarchical clustering results (*e.g.*, groups that range from 10-50 isolates). Based on our data, groups between 1 and 50 isolates were optimal, though this range may vary depending on several factors within a dataset and user needs. It may be possible to mitigate “hairball” MANs that are generated when viewing large numbers of isolates, with high-resolution, high-mass accuracy MALDI-TOF MS instruments.

The purpose of using MANs is to quickly assess the natural product capacity of groups of bacterial colonies that share similar phylogenetic origin and make an informed judgement of which to add to a microbial library. How different is natural product production between colonies with similar morphology from the same collection site? From different collection sites? How similar is natural product production in ten morphologically identical colonies from the same isolation plate? These fundamental questions have immense downstream effects on microbial drug discovery efforts. Figure 3 depicts four scenarios that would be expected to occur in the environment from any pair of isolates that share >99% 16S rRNA gene sequence identity. Neither visual inspection nor taxonomic classification strategies would inform the researcher about which scenario is occurring. IDBac provides this information and allows a researcher to make an educated choice of which bacterial isolates they will add to a library.

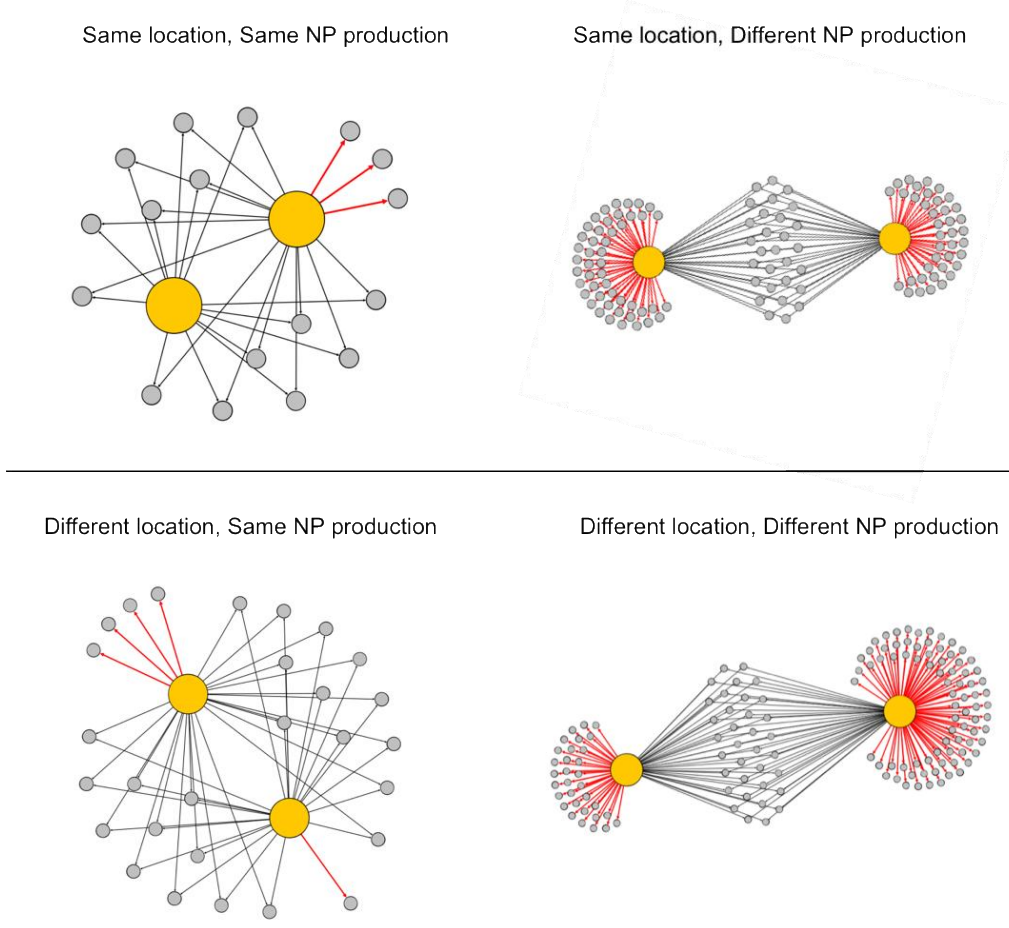


Figure 3. Four scenarios are possible when comparing environmental bacterial isolates on nutrient agar that are morphologically indistinguishable and share >99% 16S rRNA gene sequence similarity. The user can rapidly distinguish between these scenarios and make an informed decision as to whether colonies produce significantly overlapping populations of natural products. Large orange nodes represent individual bacterial isolates, while smaller gray nodes represent *m/z* values detected in their corresponding MS spectra (200-2,000 Da). “NP” denotes natural product.

We built our microbial library based on the analysis of MANs generated from 72 freshwater and 101 marine isolate protein MS groupings. Each dendrogram grouping and corresponding MAN is provided in the Supporting Information (Figure S8-S9). It is up to the individual researcher and available resources to determine roughly how many isolates they wish to add to a library, and what constitutes significant overlap. Figure 4 depicts a grouping of 16 isolates (estimated to be *Bacillus* spp. from 16S rRNA gene sequencing of 116H-8 and 116F-10). The nine isolates clustered in pink exhibited a high degree of natural product overlap, therefore two isolates (116G-4 and 116H-1) were selected for addition to the library. The rest were discarded. Alternatively, isolates 116F-10 and 116H-8 showed a high degree of unique natural product signatures (non-overlapping with other isolates in the group), so they were also included in the library. Isolates 116G-1 and 122A-5, as well as 117D-5 and 117D-2 are a more ambiguous case. Although some natural product overlap exists, all isolates in these groups contain a set of *m/z* values that are unique, which may represent a molecule(s) that exists exclusively in one bacterium over another. The researcher is now empowered to make an informed decision of whether to include one, both, or neither of the pair.

After approximately 25 hours of data acquisition (data analysis time is researcher dependent), 301 isolates were selected for addition to the library out of a total 1,616. The 301 isolates span approximately 54 genera, which is an exceptionally high taxonomic diversity given the relatively small size of the library. In particular, included within the library are 25 isolates from 16 understudied genera of the phylum Actinobacteria, that in total have an estimated 11 publications to date that report a natural product (Table S3, Supporting Information). Many of these understudied genera present colony morphologies that are atypical of traditional actinomycetes, which are often identified by their hard, leathery appearance, and presence of spores. Unbiased selection of *all* colonies in the early stages of bacterial isolation is credited with the inclusion of these in the library.

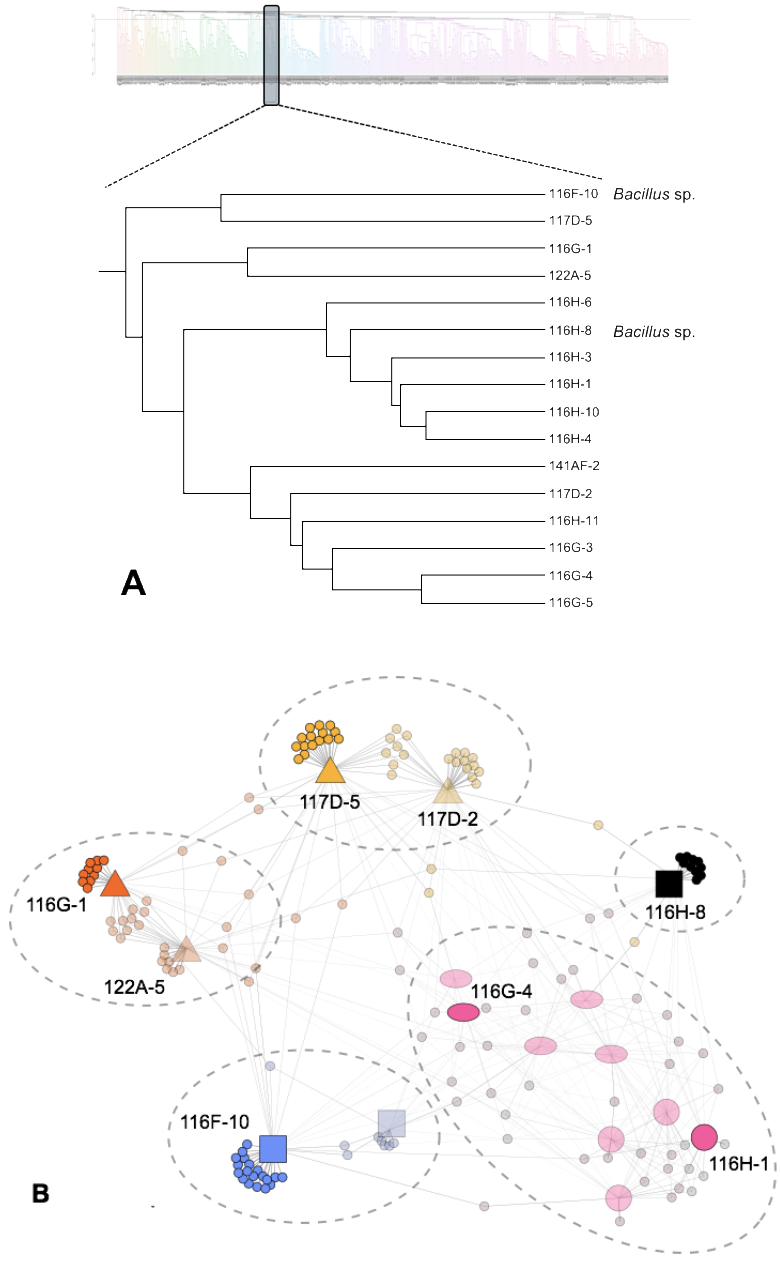


Figure 4. (A) Protein MS clustering of 16 bacterial isolates (B) MAN generated for the group of selected strains, isolates colored by modularity score. Isolates chosen to be

included in the library are colored opaque; transparent strains were discarded.

Use of IDBac to reduce redundancy in existing bacterial libraries. We performed three previous expeditions to Iceland in 2013, 2014, and 2015. After each expedition, traditional morphology-guided colony picking strategies were employed. In total, the three previous expeditions yielded 833 bacterial isolates. In order to assess the degree of redundancy in this library, the MALDI-TOF/IDBac pipeline was applied. Using the parameters and cutoffs described in the previous section and discretion of the researcher as to what constituted overlap, 600 isolates were discarded due to taxonomic and natural product redundancy, reducing the library to 233 isolates, a 72.0% reduction in library size. This reduction has significant downstream cost savings. There are several major obstacles to creating a microbial library for drug discovery, and these have been well documented.[2,27,51] Generally speaking, therapeutic discovery pipelines have relied on libraries of natural product extracts or fractions, some of which were sourced from libraries of microorganisms. Each isolate in a library is processed through purification, liquid fermentation, extraction, and in some cases chromatographic separation to generate natural product material for the screening library. This process is costly and time consuming, as it may require tens to hundreds of thousands of dollars of investment to generate.[52] The current work provides a relatively rapid, cost-effective avenue for microbial library generation. It also precludes the need for DNA sequencing of isolates to determine phylogenetic similarity or chromatographic analyses of extracts or fractions to assess natural product overlap.

Visualization of sample relationships based on user defined categories. IDBac was also designed to facilitate post-hoc analyses of field collections by providing visualizations across the protein MS dendrograms.[24] These visualizations occur in two forms, either in coloring the labels of the dendrogram or by graphing marks beside the dendrogram (Figures S6-S7, Supporting Information). In either case, users are able to associate metadata about isolates (*e.g*. geographic coordinates, sample source, water temperature, isolation media, etc.) to the dendrogram in order to quickly discern patterns across tens to thousands of isolates. For example, following the Iceland

expedition we were able to quickly evaluate genus and species-level patterns of occurrence as a function of media formulation (Figures S6-S7, Supporting Information). This allows us to design more targeted bacterial isolation procedures for future collection expeditions.

Limitations of IDBac for library creation. Limitations of the MALDI-TOF MS/IDBac method to external factors such as fluctuations in temperature and growth medium, and unintended microbial contamination, among other factors, were discussed previously.[21] In regard to the creation of microbial libraries, other limitations exist. The choice of MALDI matrix will affect which metabolites are ionized. The α-cyano-4-hydroxycinnamic acid matrix (CHCA) used in this study has been shown to ionize a variety of molecules such as PKS and PKS-NRPS-derived natural products,[53,54] alkaloids (200-400 Da), biosurfactants, and siderophores.[55] Regardless, MS profiles are not comprehensive of the full natural product producing capacity of each bacterial isolate. Cultivation of each isolate under multiple growth conditions followed by analysis in this pipeline will afford more comprehensive metabolic profiles. This is particularly feasible if the analysis is performed in multiwell plates, however one must be cognizant that gas phase natural products may diffuse into other wells and differentially influence natural product production. Regardless, it is recommended that each laboratory develop a standardized growth protocol to facilitate comparison of data across different strains, time points, and researchers.

IDBac is currently limited to qualitative manual assessment of MANs. A major limitation in past library creation methods was the implementation of visual bias while manually selecting colonies from petri dishes.[13] In the case of the current study, although a degree of bias remains, a researcher is basing the decision to keep/discard isolates based on both pseudo-phylogenetic groupings *and* natural product production capacity. This facilitates a more rigorous decision making process that can be experimentally verified across labs, based on spectral profiles. Further improvements include the development of algorithms to prioritize isolates directly from hierarchical clustering/MANs data. Lastly, it is possible that phylogenetically similar bacterial isolates

produce isobaric species, giving the false appearance of natural product overlap. This can be improved through use of instruments with higher resolving power capabilities such as MALDI-FT-ICR or MALDI Orbitrap, though increased data size and storage needs must be addressed. Particularly when processing high numbers of bacterial isolates, this pipeline *will* result in some degree of redundancy added to a library, or in the removal of isolates that may harbor unique natural product biosynthetic potential. Despite these shortcomings, a library constructed using the methodology described here is more likely to harbor a higher degree of natural product diversity compared to one formed on the basis of 16S rRNA gene sequencing or morphology-guided practices.

Nearly two decades ago, Goodfellow and colleagues wrote extensively on the importance of ensuring taxonomic diversity in microbial libraries. In summarizing the methods needed to accomplish this, they noted the potential of MS techniques (pyrolysis MS in particular) to ensure rapid intraspecies pseudo-taxonomic dereplication of colonies growing directly on isolation plates. They opined that such a method would permit “the rational collection of colonies” and reduce “the requirement for time-consuming laboratory testing to distinguish duplicate colonies.”[51] Nearly 70 years have passed since researchers began to rely on microorganisms to produce antibiotics for use in the clinic, and despite advances such as colony picking automation, cultivation practices, and DNA sequencing, among other innovations that led or could lead to the discovery of new antibiotics,[9,10,13,56–58] researchers continue to create libraries that are plagued with redundancy because there has been no cost effective way to rapidly and simultaneously assess both the bacterial identity and the natural product capacity of colonies in high-throughput. In the current work, we employed our freely available pipeline to group thousands of environmental bacterial isolates based on similarities in protein MS spectra, and coupled this with global comparison of their natural product profiles. This pipeline allows for the rapid generation of diverse microbial ‘smart’ libraries that do not necessitate high numbers of entries.[52,59] Rather, it minimizes entries into the library, reduces natural product redundancy entering into downstream biological screening efforts, and significantly decreases costs associated with library generation.

EXPERIMENTAL SECTION

Sample collection and processing. Sponges, macroalgae, marine and freshwater sediment samples were collected by SCUBA. Samples were processed immediately after collection. Approximately 1 cm^3 portions of all samples were submitted to heat in sterile water for 8 minutes at 60 °C. A similar portion of select samples was also plated on nutrient media without heat treatment. All samples were vortexed, diluted 1:10 in sterile, de-ionized water, and 10 µL used to inoculate five different agar-based nutrient media containing 28 µM cycloheximide (to inhibit the growth of fungi). Nutrient plates were sealed with Parafilm® and left at room temperature. After 30-60 days, all distinguishable colonies were purified on 60 mm petri dishes containing high nutrient media (A1).

MALDI-TOF MS analysis. From the petri dishes described in the previous section, 1,616 bacterial isolates were re-plated onto 48-well plates containing A1 media, and supplemented with or without synthetic ocean salts, depending on the sample source. After 7 days of growth three technical replicates of each bacterial isolate were applied, as a thin smear, to a 384-spot MALDI target plate (Bruker Daltonics, Billerica, MA) using a sterile toothpick. The number and type of replicates will vary depending on the researcher's specific needs, however both biological and technical replicates are recommended as MALDI-TOF MS takes seconds to acquire and can be averaged within the IDBac pipeline. After transferring the colonies to the MALDI target plate, 1 µL of 70% formic acid (Optima, Fisher Chemical) was overlaid by pipette and allowed to dry, followed by overlay and drying of 1 µL of 10 mg/mL α-cyano-4-hydroxycinnamic acid (CHCA; recrystallized from the 98% pure Sigma-Aldrich, part-C2020) matrix prepared in 50% acetonitrile, 47.5% water, and 2.5% trifluoroacetic acid. All solvents were MS grade.

MALDI-TOF MS analysis were performed using an Autoflex Speed LRF mass spectrometer (Bruker Daltonics) equipped with a smartbeam™-II laser (355 nm).

Automated data acquisitions were performed using flexControl software v. 3.4.135.0 (Bruker Daltonics) and flexAnalysis software v. 3.4. Protein spectra were recorded in positive linear mode (1200 shots; RepRate: 1000; delay: 29793 ns; ion source 1 voltage: 19.5 kV; ion source 2 voltage: 18.2 kV; lens voltage: 7.5 kV; mass range: 1.9 kDa to 2.1 kDa, matrix suppression cutoff: 1.5 kDa). Protein spectra were corrected with external Bruker Daltonics bacterial test standard (BTS). Natural product spectra were recorded in positive reflectron mode (5000 shots; RepRate: 2000 Hz; delay: 9297 ns; ion source 1 voltage: 19 kV; ion source 2 voltage:16.55 kV; lens voltage: 8.3 kV; mass range: 50 Da to 2,700 Da, matrix suppression cutoff: 50 Da). Natural Product spectra were corrected with external Bruker Daltonics peptide calibration standard and CHCA $[2M+H]^+$ (379.0930 Da). Detailed experimental settings are described in Clark and Costa *et al.*[21]

16S rRNA sequencing. DNA from 361 bacterial isolates was extracted using a DNeasy UltraClean Microbial Kit (Qiagen). The 16S rRNA gene was amplified using 27F (5'-CAGAGTTTGATCCTGGCT-3') and 1492R (5'-AGGAGGTGATCCAGCCGCA-3')[60] primers using polymerase chain reaction (PCR) under the following conditions: initial denaturation at 95 °C for 5 minutes; followed by 35 cycles of denaturation at 95 °C for 15 seconds, annealing at 60 °C for 15 seconds, and extension at 72 °C for 30 seconds; and a final extension step at 72 °C for 2 minutes. PCR products were purified using a QIAquick PCR Purification kit from Qiagen and the amplicons sequenced by Sanger sequencing. Data were analyzed by Geneious V11.1.4 software. Using the SILVA Alignment, Classification and Tree (ACT) Service, all 361 isolates were aligned and a phylogenetic tree using Fastree[61] was created, using default settings.[62] All the sequences were deposited in the NCBI nucleotide database (Table S2, Supporting Information).

Data availability.

MALDI-TOF MS data were deposited in MASSIVE (doi:10.25345/C5261K, MASSIVE accession: MSV000083461). The 16S rRNA sequences of environmental bacteria used in this study were deposited in GenBank with the accession numbers MK143106 to

MK143377; MK163385 to MK163438; and MK168020 to MK168054 (Table S2 Supporting Information).

Software availability

IDBac is available at https://chasemc.github.io/IDBac and the source code of each release backed-up in permanence at DOI:10.5281/zenodo.1115619

AUTHOR INFORMATION

Maria S. Costa and Chase M. Clark contributed equally to the work.

ACKNOWLEDGMENTS

The authors would like to acknowledge Erlendur Bogason and his team at the Strytan Dive Center and Eydis Einarsdottir for assistance in sample collection, and Mark Sadek, Ying Gao, and Casey Murphy for assistance with microbial cultivation. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM125943 (BTM and LMS). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was also supported by Icelandic Research Fund Grant 152336-051 (BTM and SO) and UIC startup funds (LMS).

REFERENCES

(1) Bérdy, J. *J. Antibiot.* **2005**, *58* (1), 1–26.

(2) Katz, L.; Baltz, R. H. *J. Ind. Microbiol. Biotechnol.* **2016**, *43* (2–3), 155–176.

(3) Newman, D. J.; Cragg, G. M. *J. Nat. Prod.* **2016**, *79* (3), 629–661.

(4) Wohlleben, W.; Mast, Y.; Stegmann, E.; Ziemert, N. *Microb. Biotechnol.* **2016**, *9* (5), 541–548.

(5) The Discovery and Development of Penicillin – American Chemical Society https://www.acs.org/content/acs/en/education/whatischemistry/landmarks/flemingpenicillin.html (accessed Jan 13, 2019).

(6) Fleming, A. *Br. J. Exp. Pathol.* **1929**. *10* (3), 226–236.

(7) Gause, G. F.; G., B. M. *Nature* **1944**, *154* (3918), 703–703.
(8) Emmerich, R.; Löw, O. *Zeitschrift für Hyg. und Infect.* **1899**, *31* (1), 1–65.
(9) Pereira, D. A.; Williams, J. A. *Br. J. Pharmacol.* **2007**, *152* (1), 53–61.
(10) Leeds, J. A.; Schmitt, E. K.; Krastel, P. *Expert Opin. Investig. Drugs* **2006**, *15* (3), 211–226.
(11) Lewis, K. *Nat. Rev. Drug Discov.* **2013**, *12* (5), 371–387.
(12) Larsen, T. O.; Smedsgaard, J.; Nielsen, K. F.; Hansen, M. E.; Frisvad, J. C. *Nat. Prod. Rep.* **2005**, *22* (6), 672–695.
(13) Genilloud, O.; González, I.; Salazar, O.; Martín, J.; Tormo, J. R.; Vicente, F. *J. Ind. Microbiol. Biotechnol.* **2011**, *38* (3), 375–389.
(14) Okuda, T.; Ando, K.; Bills, G. Chapter 6 from *Handbook of Industrial Mycology*; CRC Press/Taylor and Francis Group, Boca Raton, FL, 2004.
(15) Fujimori, F.; Okuda, T. *J. Antibiot.* **1994**, *47* (2), 173–182.
(16) Knight, V.; Sanglier, J.-J.; DiTullio, D.; Braccili, S.; Bonner, P.; Waters, J.; Hughes, D.; Zhang, L. *Appl. Microbiol. Biotechnol.* **2003**, *62* (5–6), 446–458.
(17) Fenical, W.; Jensen, P. R. *Nat. Chem. Biol.* **2006**, *2* (12), 666–673.
(18) Koehn, F. E.; Carter, G. T. *Nat. Rev. Drug Discov.* **2005**, *4* (3), 206–220.
(19) Demain, A. L. *Nat. Biotechnol.* **2002**, *20* (4), 331.
(20) Barrett, J. F. *Curr. Opin. Microbiol.* **2005**, *8* (5), 498–503.
(21) Clark, C. M.; Costa, M. S.; Sanchez, L. M.; Murphy, B. T. *Proc. Natl. Acad. Sci.* **2018**, *115* (19), 4981–4986.
(22) Ryzhov, V.; Fenselau, C. *Anal. Chem.* **2001**, *73* (4), 746–750.
(23) Fernández-Álvarez, C.; Torres-Corral, Y.; Santos, Y. *J. Proteomics* **2018**, *170*, 59–69.
(24) Clark, C. M.; Costa, M. S.; Conley, E.; Li, E.; M., S. L.; Murphy, B. T. *J. Vis. Exp.* **2019**, in-press, e59219.
(25) Crüsemann, M.; O'Neill, E. C.; Larson, C. B.; Melnik, A. V.; Floros, D. J.; da Silva, R. R.; Jensen, P. R.; Dorrestein, P. C.; Moore, B. S. *J. Nat. Prod.* **2017**, *80* (3), 588–597.
(26) Hubert, J.; Nuzillard, J.-M.; Renault, J.-H. *Phytochem. Rev.* **2017**, *16* (1), 55–95.
(27) Liu, X.; Ashforth, E.; Ren, B.; Song, F.; Dai, H.; Liu, M.; Wang, J.; Xie, Q.; Zhang,

L. *J. Antibiot.* **2010**, *63* (8), 415–422.

(28) Floros, D. J.; Jensen, P. R.; Dorrestein, P. C.; Koyama, N. *Metabolomics* **2016**, *12* (9), 145.

(29) Wang, M.; Carver, J. J.; Phelan, V. V; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; et al. *Nat. Biotechnol.* **2016**, *34* (8), 828–837.

(30) Hindra; Huang, T.; Yang, D.; Rudolf, J. D.; Xie, P.; Xie, G.; Teng, Q.; Lohman, J. R.; Zhu, X.; Huang, Y.; et al. *J. Nat. Prod.* **2014**, *77* (10), 2296–2303.

(31) Jensen, P. R.; Gontang, E.; Mafnas, C.; Mincer, T. J.; Fenical, W. *Environ. Microbiol.* **2005**, *7* (7), 1039–1048.

(32) Gontang, E. A.; Fenical, W.; Jensen, P. R. *Appl. Environ. Microbiol.* **2007**, *73* (10), 3272–3282.

(33) Joseph, S. J.; Hugenholtz, P.; Sangwan, P.; Osborne, C. A.; Janssen, P. H. *Appl. Environ. Microbiol.* **2003**, *69* (12), 7210–7215.

(34) Sauer, S.; Freiwald, A.; Maier, T.; Kube, M.; Reinhardt, R.; Kostrzewa, M.; Geider, K. *PLoS One* **2008**, *3* (7), e2843.

(35) Suarez, S.; Ferroni, A.; Lotz, A.; Jolley, K. A.; Guérin, P.; Leto, J.; Dauphin, B.; Jamet, A.; Maiden, M. C. J.; Nassif, X.; et al. *J. Microbiol. Methods* **2013**, *94* (3), 390–396.

(36) Sandrin, T. R.; Goldstein, J. E.; Schumaker, S. *Mass Spectrom. Rev.* **2013**, *32* (3), 188–217.

(37) Dieckmann, R.; Graeber, I.; Kaesler, I.; Szewzyk, U.; von Döhren, H. *Appl. Microbiol. Biotechnol.* **2005**, *67* (4), 539–548.

(38) Holland, R. D.; Wilkes, J. G.; Rafii, F.; Sutherland, J. B.; Persons, C. C.; Voorhees, K. J.; Lay, J. O. *Rapid Commun. Mass Spectrom.* **1996**, *10* (10), 1227–1232.

(39) Popović, N. T.; Kazazić, S. P.; Strunjak-Perović, I.; Čož-Rakovac, R. *Environ. Res.* **2017**, *152*, 7–16.

(40) Croxatto, A.; Prod'hom, G.; Greub, G. *FEMS Microbiol. Rev.* **2012**, *36* (2), 380–407.

(41) Seuylemezian, A.; Aronson, H. S.; Tan, J.; Lin, M.; Schubert, W.; Vaishampayan,

P. *Front. Microbiol.* **2018**, *9*, 780.

(42) Strejcek, M.; Smrhova, T.; Junkova, P.; Uhlik, O. *Front. Microbiol.* **2018**, *9*, 1294.

(43) Rahi, P.; Prakash, O.; Shouche, Y. S. *Front. Microbiol.* **2016**, *7*, 1359.

(44) Hoffmann, T.; Krug, D.; Bozkurt, N.; Duddela, S.; Jansen, R.; Garcia, R.; Gerth, K.; Steinmetz, H.; Müller, R. *Nat. Commun.* **2018**, *9* (1), 803.

(45) Grubbs, K. J.; Bleich, R. M.; Santa Maria, K. C.; Allen, S. E.; Farag, S.; AgBiome Team, A.; Shank, E. A.; Bowers, A. A. *mSystems* **2017**, *2* (6), e00040-17.

(46) Adamek, M.; Alanjary, M.; Sales-Ortells, H.; Goodfellow, M.; Bull, A. T.; Winkler, A.; Wibberg, D.; Kalinowski, J.; Ziemert, N. *BMC Genomics* **2018**, *19* (1), 426.

(47) Letzel, A.-C.; Li, J.; Amos, G. C. A.; Millán-Aguiñaga, N.; Ginigini, J.; Abdelmohsen, U. R.; Gaudêncio, S. P.; Ziemert, N.; Moore, B. S.; Jensen, P. R. *Environ. Microbiol.* **2017**, *19* (9), 3660–3673.

(48) Klementz, D.; Döring, K.; Lucas, X.; Telukunta, K. K.; Erxleben, A.; Deubel, D.; Erber, A.; Santillana, I.; Thomas, O. S.; Bechthold, A.; et al. *Nucleic Acids Res.* **2016**, *44* (D1), D509–D514.

(49) Seipke, R. F. *PLoS One* **2015**, *10* (1), e0116457.

(50) Sumner, L. W.; Amberg, A.; Barrett, D.; Beale, M. H.; Beger, R.; Daykin, C. A.; Fan, T. W.-M.; Fiehn, O.; Goodacre, R.; Griffin, J. L.; et al. *Metabolomics* **2007**, *3* (3), 211–221.

(51) Bull, A. T.; Ward, A. C.; Goodfellow, M. *Microbiol. Mol. Biol. Rev.* **2000**, *64* (3), 573–606.

(52) Silver, L. L. *Clin. Microbiol. Rev.* **2011**, *24* (1), 71–109.

(53) Spraker, J. E.; Sanchez, L. M.; Lowe, T. M.; Dorrestein, P. C.; Keller, N. P. *ISME J.* **2016**, *10* (9), 2317–2330.

(54) Moree, W. J.; McConnell, O. J.; Nguyen, D. D.; Sanchez, L. M.; Yang, Y.-L.; Zhao, X.; Liu, W.-T.; Boudreau, P. D.; Srinivasan, J.; Atencio, L.; et al. *ACS Chem. Biol.* **2014**, *9* (10), 2300–2308.

(55) Phelan, V. V; Moree, W. J.; Aguilar, J.; Cornett, D. S.; Koumoutsi, A.; Noble, S. M.; Pogliano, K.; Guerrero, C. A.; Dorrestein, P. C. *J. Bacteriol.* **2014**, *196* (9), 1683–1693.

(56) Ling, L. L.; Schneider, T.; Peoples, A. J.; Spoering, A. L.; Engels, I.; Conlon, B. P.;

Mueller, A.; Schäberle, T. F.; Hughes, D. E.; Epstein, S.; et al. *Nature* **2015**, *517* (7535), 455–459.

(57) Berdy, B.; Spoering, A. L.; Ling, L. L.; Epstein, S. S. *Nat. Protoc.* **2017**, *12* (10), 2232–2242.

(58) Mahler, L.; Wink, K.; Beulig, R. J.; Scherlach, K.; Tovar, M.; Zang, E.; Martin, K.; Hertweck, C.; Belder, D.; Roth, M. *Sci. Rep.* **2018**, *8* (1), 13087.

(59) Bull A.T.; Stach J.E.; Ward A.C.; Goodfellow M. *Antonie Van Leeuwenhoek* **2005**, *87* (3), 65–79.

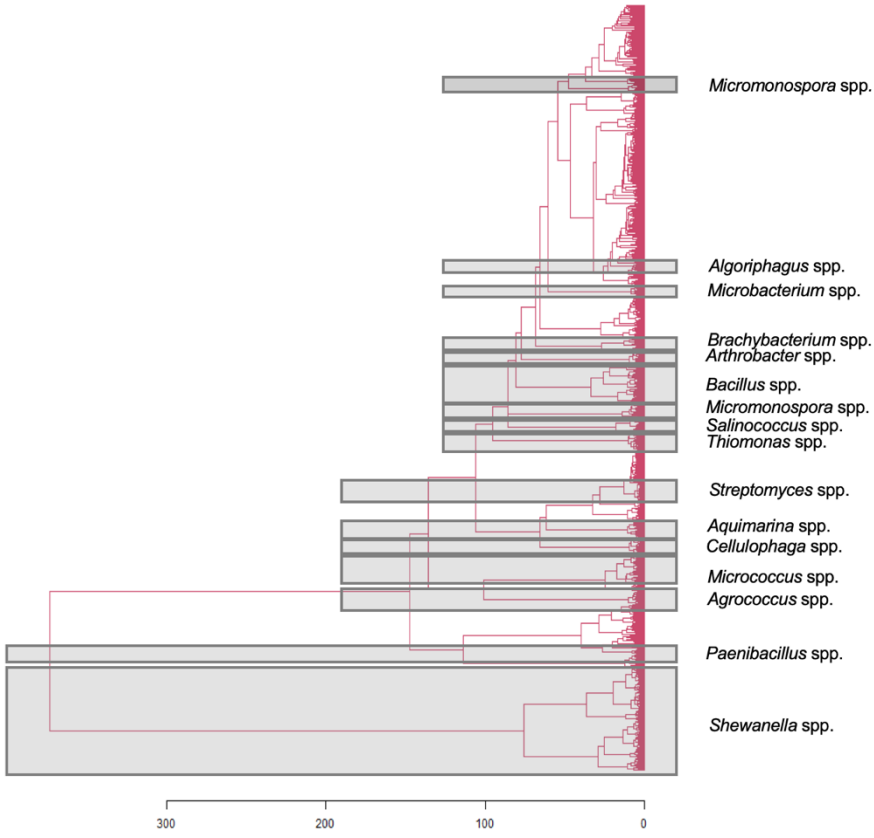
(60) Lane; J., D. *Nucleic Acid Tech. Bact. Syst.* **1991**, 115–175.

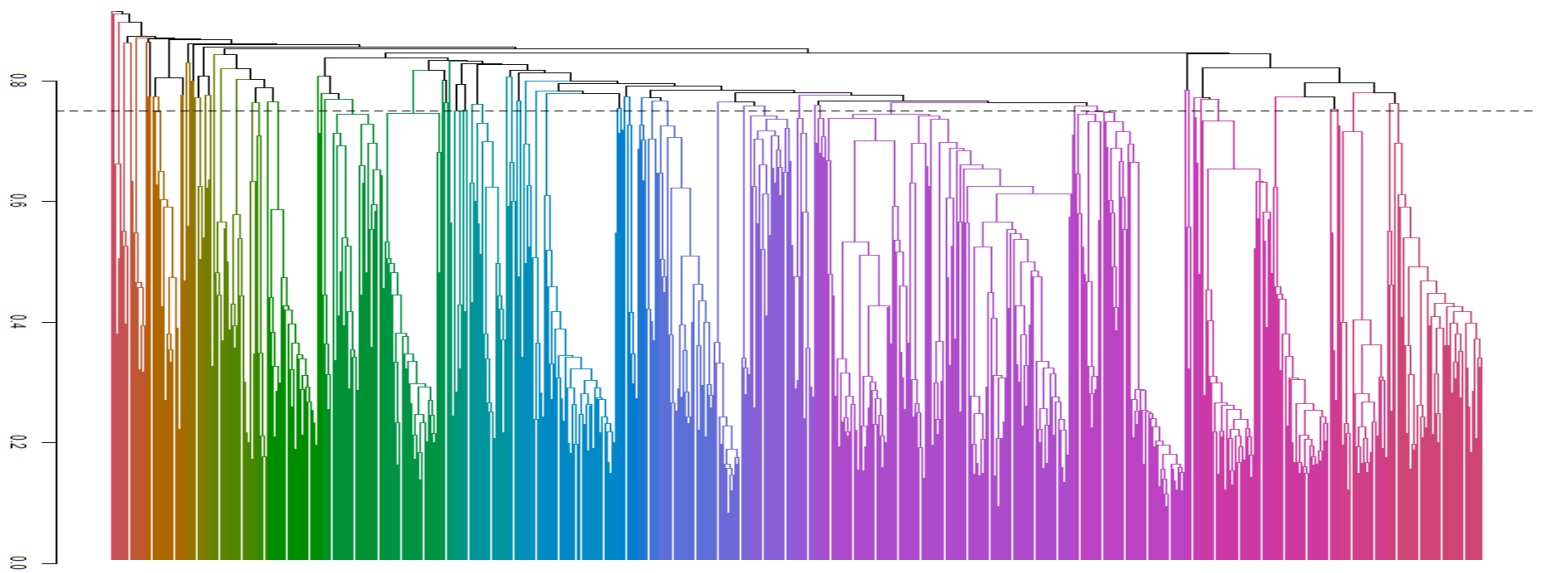
(61) Price, M. N.; Dehal, P. S.; Arkin, A. P. *Mol. Biol. Evol.* **2009**, *26* (7), 1641–1650.

(62) Pruesse, E.; Peplies, J.; Glöckner, F. O. *Bioinformatics* **2012**, *28* (14), 1823–1829.

Appendix

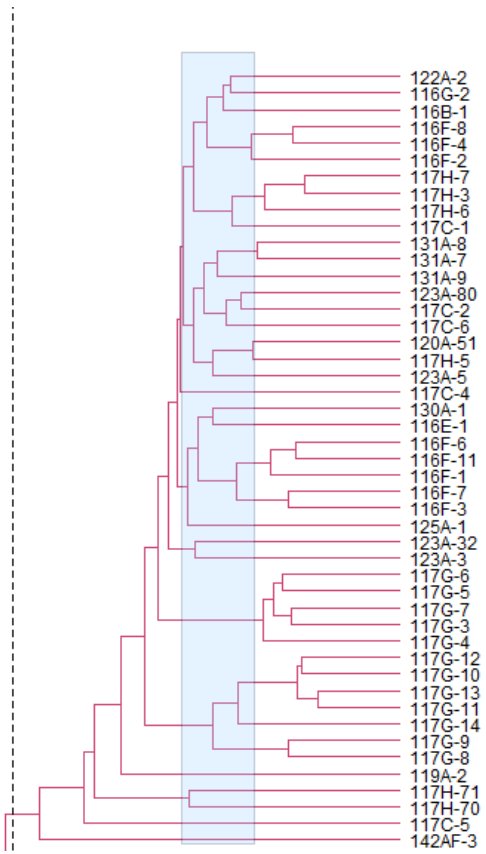
Appendix Note 1: Pseudophylogenetic groupings in the MALDI protein dendogram.

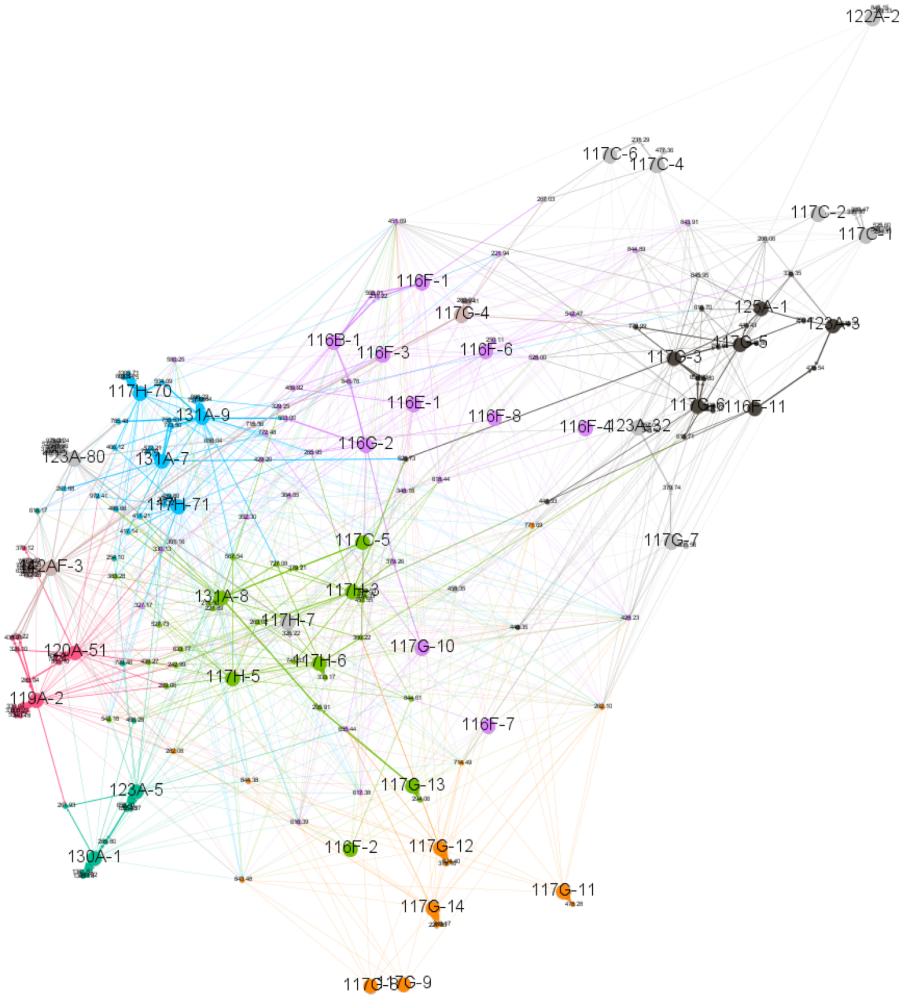




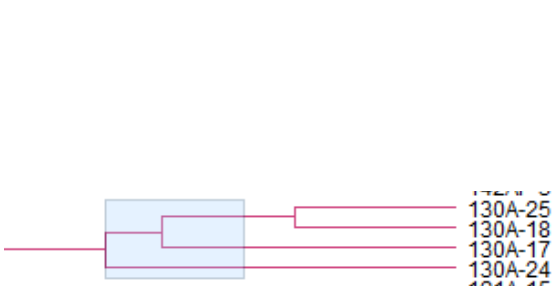
Appendix 2: Group1 to Group 72. Dendrogram resulting from hierarchical clustering of MALDI-TOF MS protein spectra of 733 freshwater isolates using IDBac. The dendrogram was "cut" at a threshold of 0.75 (dashed lines) to create groups of similar isolates. 72 distinct groups were created and the respective MANs are presented in the following figures.

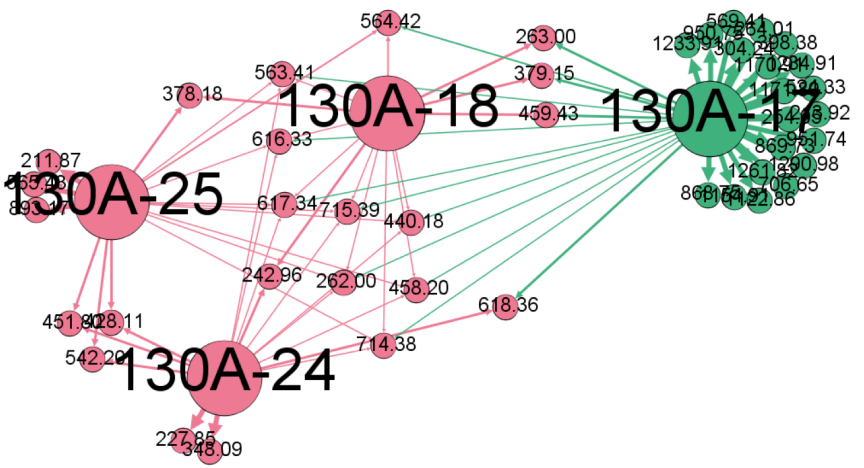
Group 1


122A-2
116G-2
116B-1
116F-8
116F-4
116F-2
117H-7
117H-3
117H-6
117C-1
131A-8
131A-7
131A-9
123A-80
117C-2
117C-6
120A-51
117H-5
123A-5
117C-4
130A-1
116E-1
116F-6
116F-11
116F-1
116F-7
116F-3
125A-1
123A-32
123A-3
117G-6
117G-5
117G-7
117G-3
117G-4
117G-12
117G-10
117G-13
117G-11
117G-14
117G-9
117G-8
119A-2
117H-71
117H-70
117C-5
142AF-3

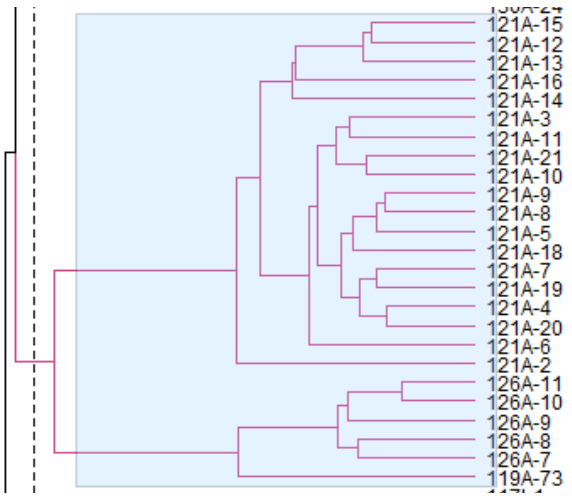

122A-2
117C-6
117C-4
117C-2
117C-1
116F-1
117G-4
120A-1
123A-8
116B-1
116F-3
116F-6
117G-5
117G-3
117H-70
131A-9
116E-1
116F-8
116F-11
116F-4
123A-32
123A-80
131A-7
116G-2
117H-71
117C-5
117G-7
142AF-3
131A-8
117H-3
117H-7
120A-51
117G-10
117H-6
117H-5
119A-2
116F-7
123A-5
117G-13
130A-1
116F-2
117G-12
117G-11
117G-14
117G-9

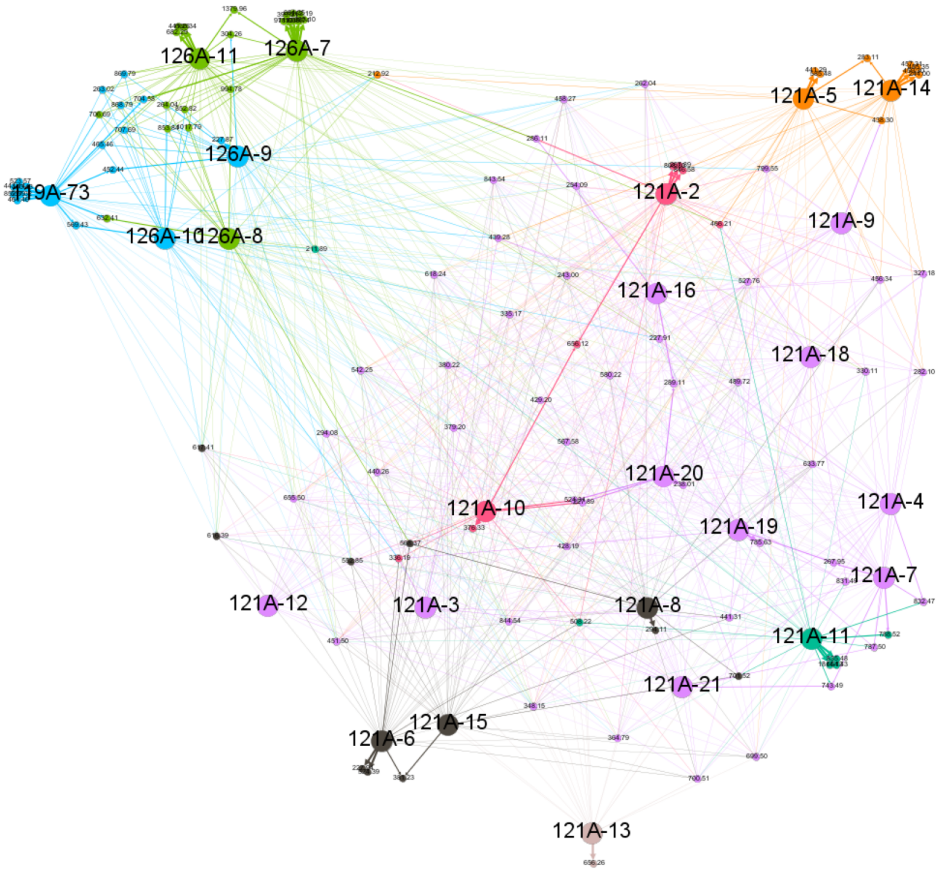
Group 2


130A-25
130A-18
130A-17
130A-24

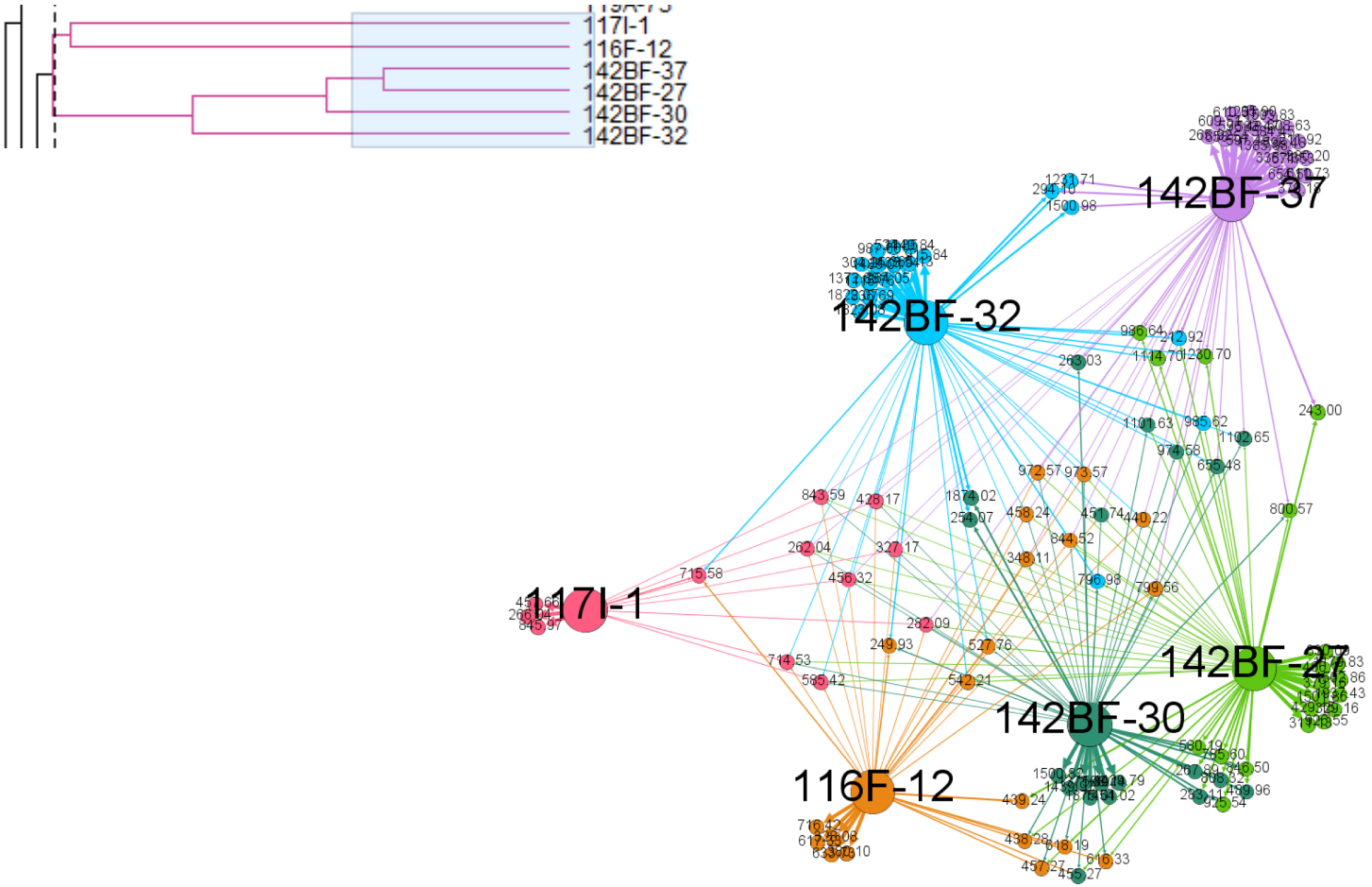

130A-18
130A-17
130A-25
130A-24
564.42
263.00
563.41
379.15
378.18
459.43
616.33
211.87
617.34
715.89
440.18
242.96
262.00
458.20
618.36
714.38
227.85
348.09

Group 3

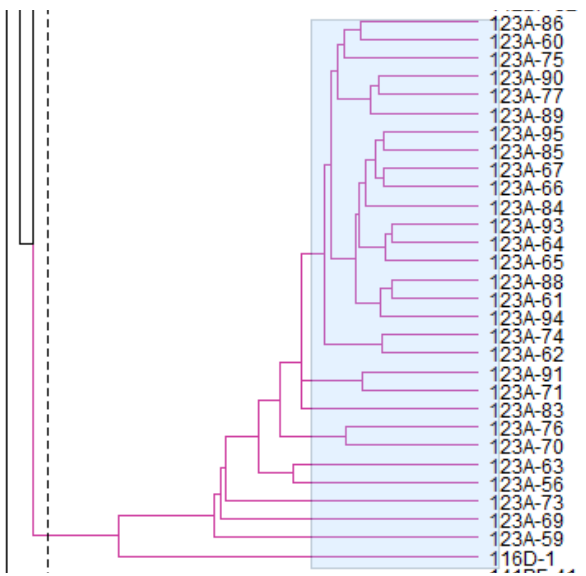

121A-15
121A-12
121A-13
121A-16
121A-14
121A-3
121A-11
121A-21
121A-10
121A-9
121A-8
121A-5
121A-18
121A-7
121A-19
121A-4
121A-20
121A-6
121A-2
126A-11
126A-10
126A-9
126A-8
126A-7
119A-73

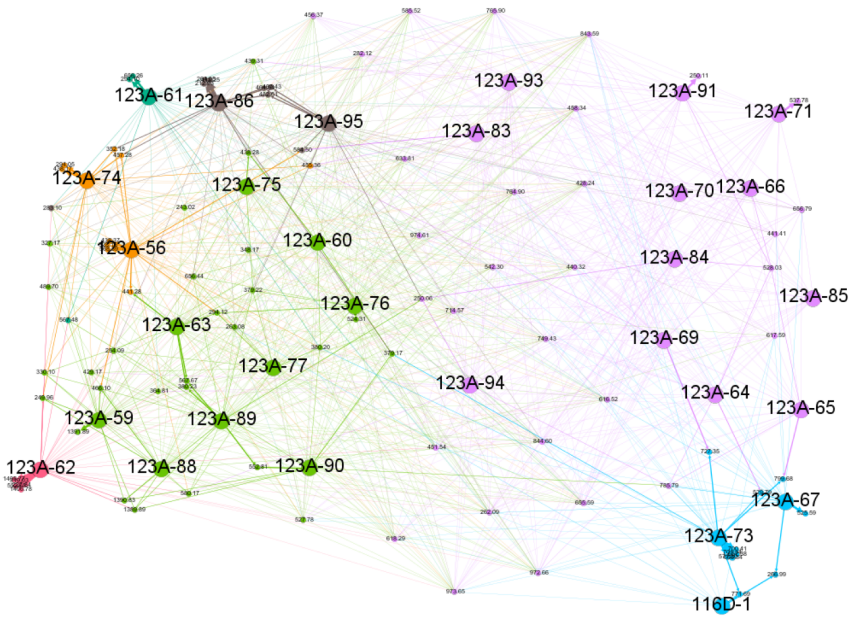

126A-11
126A-7
126A-9
119A-73
126A-10
126A-8
121A-5
121A-14
121A-2
121A-9
121A-16
121A-18
121A-20
121A-10
121A-4
121A-19
121A-7
121A-12
121A-3
121A-8
121A-11
121A-21
121A-15
121A-6
121A-13

Group 4

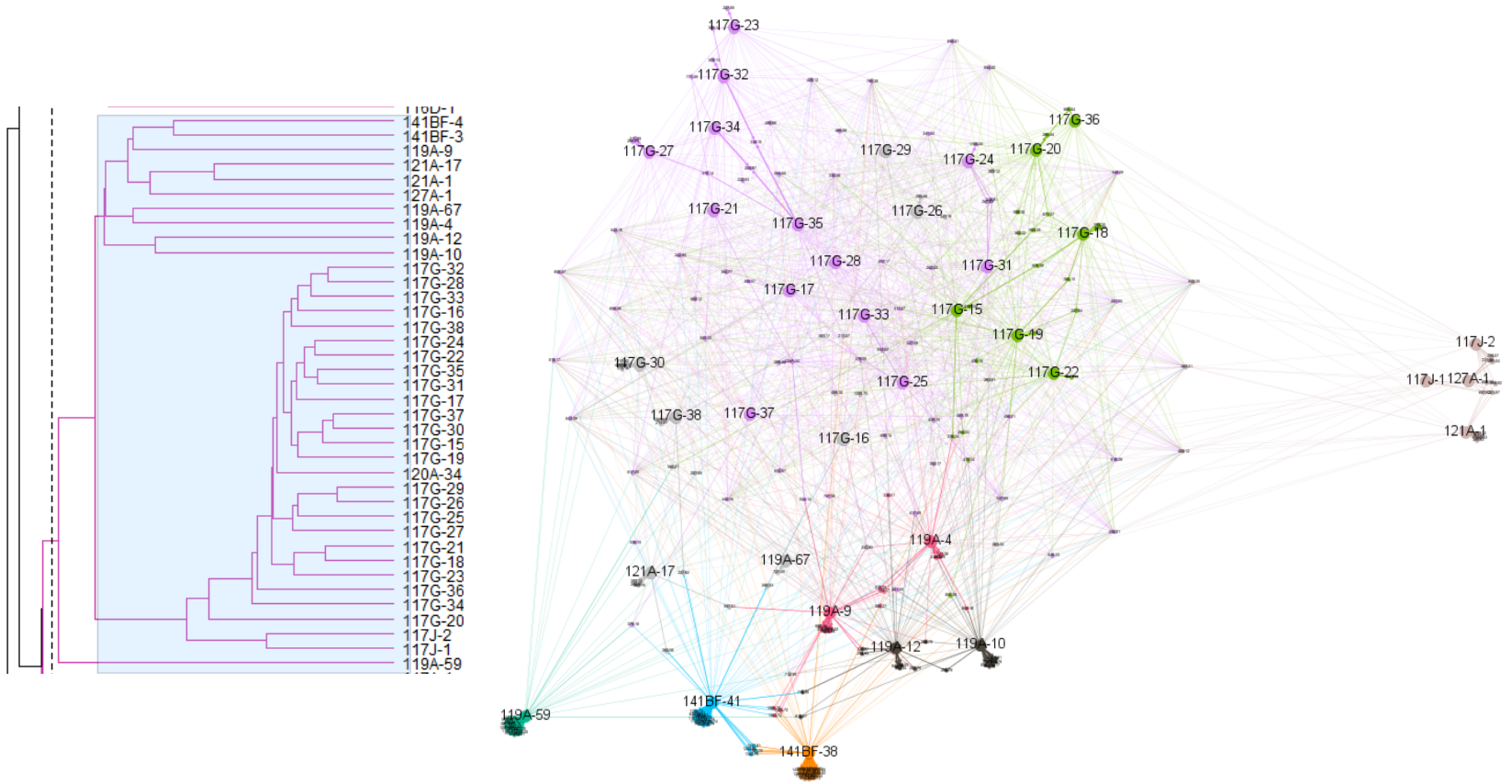


Group 5

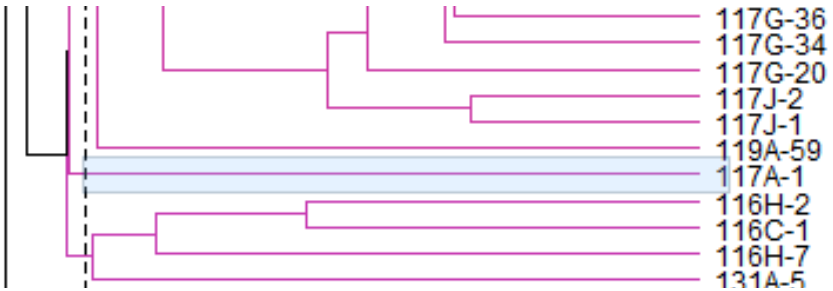

123A-86
123A-60
123A-75
123A-90
123A-77
123A-89
123A-95
123A-85
123A-67
123A-66
123A-84
123A-93
123A-64
123A-65
123A-88
123A-61
123A-94
123A-74
123A-62
123A-91
123A-71
123A-83
123A-76
123A-70
123A-63
123A-56
123A-73
123A-69
123A-59
116D-1

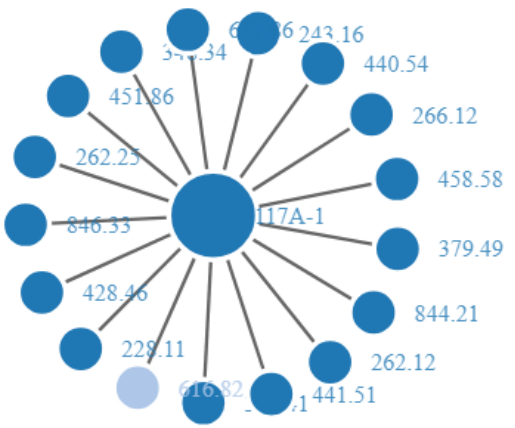

123A-61
123A-86
123A-95
123A-93
123A-91
123A-71
123A-83
123A-74
123A-75
123A-70
123A-66
123A-56
123A-60
123A-84
123A-85
123A-76
123A-63
123A-69
123A-77
123A-94
123A-64
123A-65
123A-59
123A-89
123A-62
123A-88
123A-90
123A-67
123A-73
116D-1

Group 6

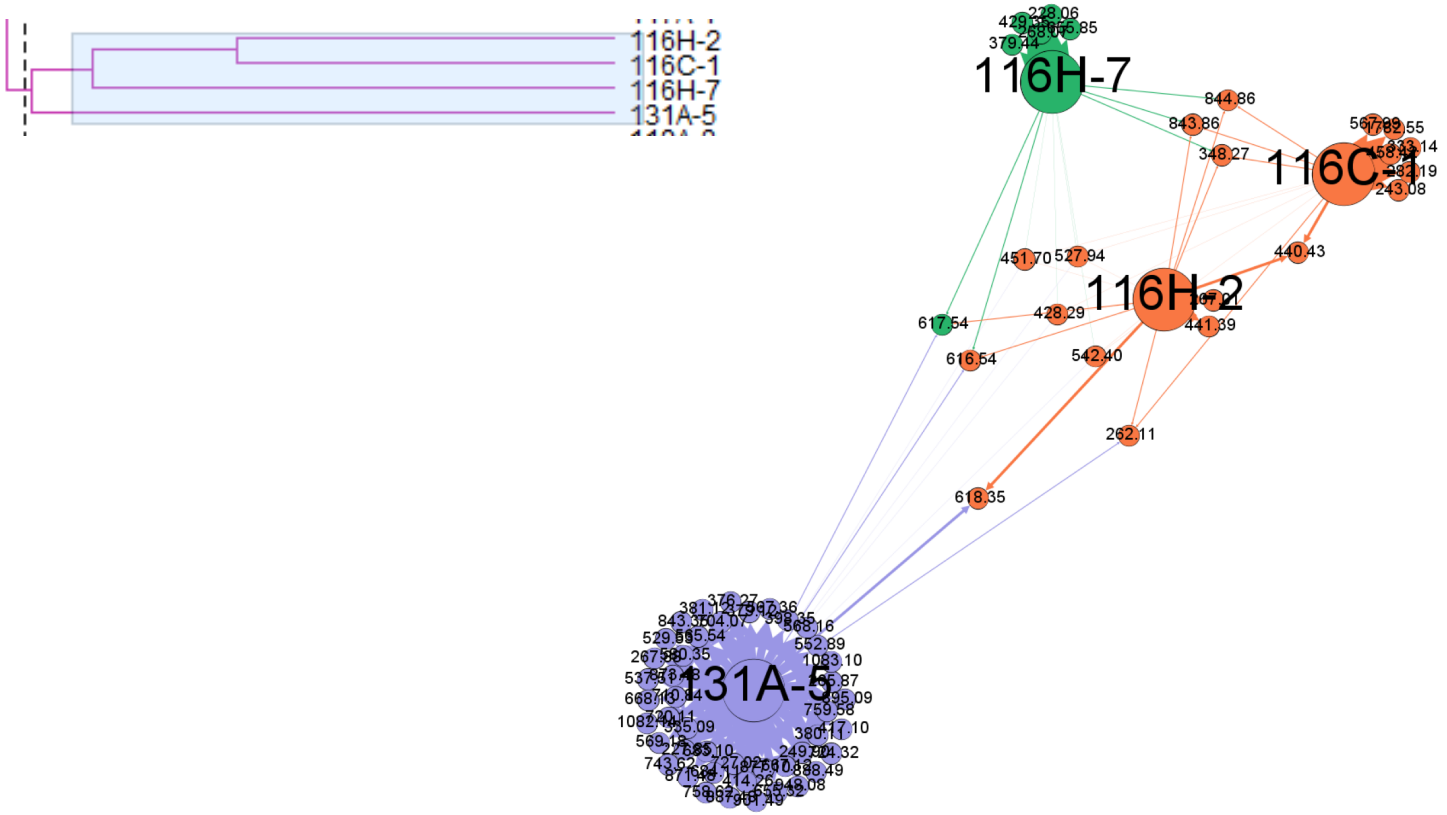

141BF-4
141BF-3
119A-9
121A-17
121A-1
127A-1
119A-67
119A-4
119A-12
119A-10
117G-32
117G-28
117G-33
117G-16
117G-38
117G-24
117G-22
117G-35
117G-31
117G-17
117G-37
117G-30
117G-15
117G-19
120A-34
117G-29
117G-26
117G-25
117G-27
117G-21
117G-18
117G-23
117G-36
117G-34
117G-20
117J-2
117J-1
119A-59
117G-23
117G-32
117G-34
117G-36
117G-27
117G-29
117G-24
117G-20
117G-21
117G-26
117G-35
117G-18
117G-28
117G-31
117G-17
117G-33
117G-15
117G-19
117G-30
117G-22
117G-25
117G-38
117G-37
117G-16
117J-2
117J-1
127A-1
121A-1
119A-4
119A-67
121A-17
119A-9
119A-12
119A-10
141BF-41
119A-59
141BF-38

Group 7

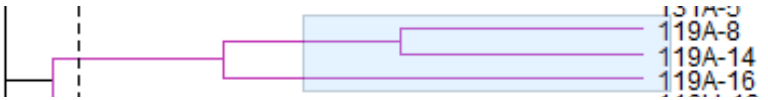

117G-36
117G-34
117G-20
117J-2
117J-1
119A-59
117A-1
116H-2
116C-1
116H-7

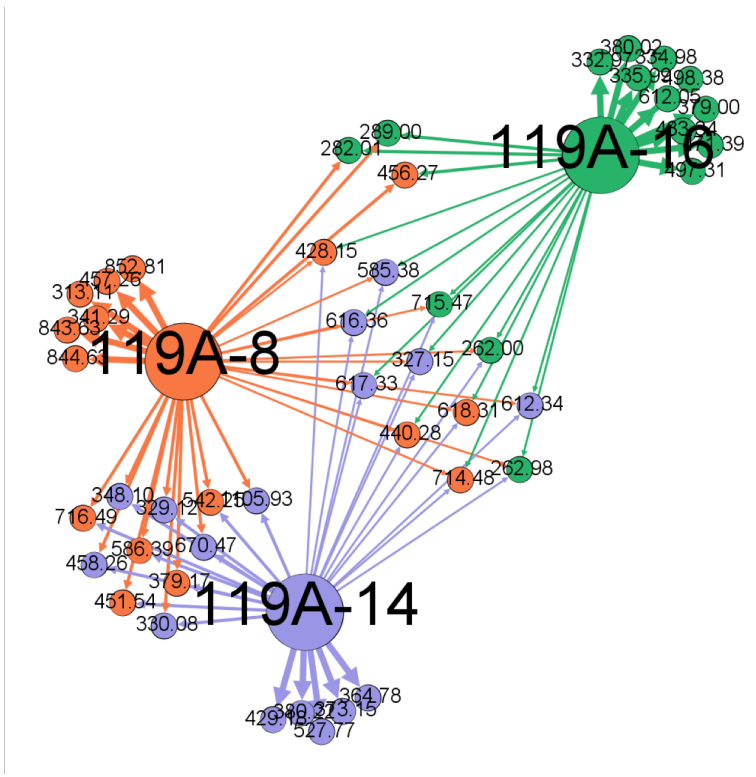

243.16
440.54
266.12
458.58
379.49
844.21
262.12
441.51
451.86
262.25
846.33
428.46
117A-1

Group 8

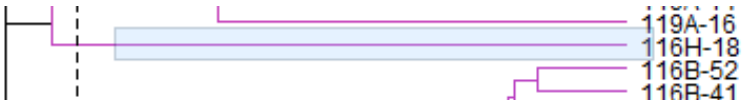

116H-2
116C-1
116H-7
131A-5
116H-7
116C-1
116H-2
131A-5
844.86
843.86
348.27
451.70
527.94
440.43
428.29
441.39
617.54
616.54
542.40
262.11
618.35
552.89
759.58
243.08

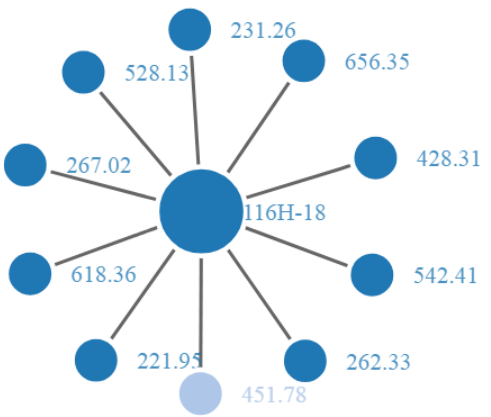
Group 9


119A-8
119A-14
119A-16

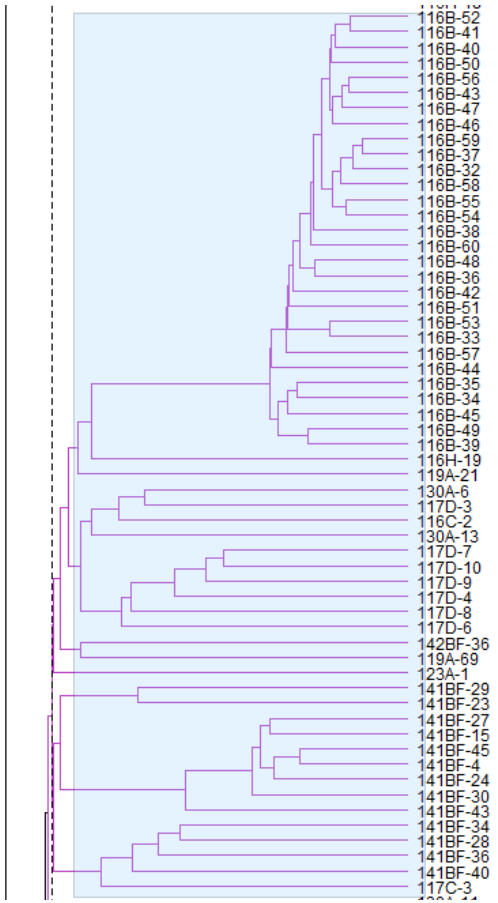

119A-16
119A-8
119A-14
289.00
282.01
456.27
428.15
585.38
715.47
616.36
262.00
327.15
617.33
618.31
612.34
440.28
714.48
262.98
852.81
313.11
341.29
844.61
348.10
716.49
329.12
542.21
505.93
586.39
670.47
458.26
379.17
451.54
330.08
364.78
527.77

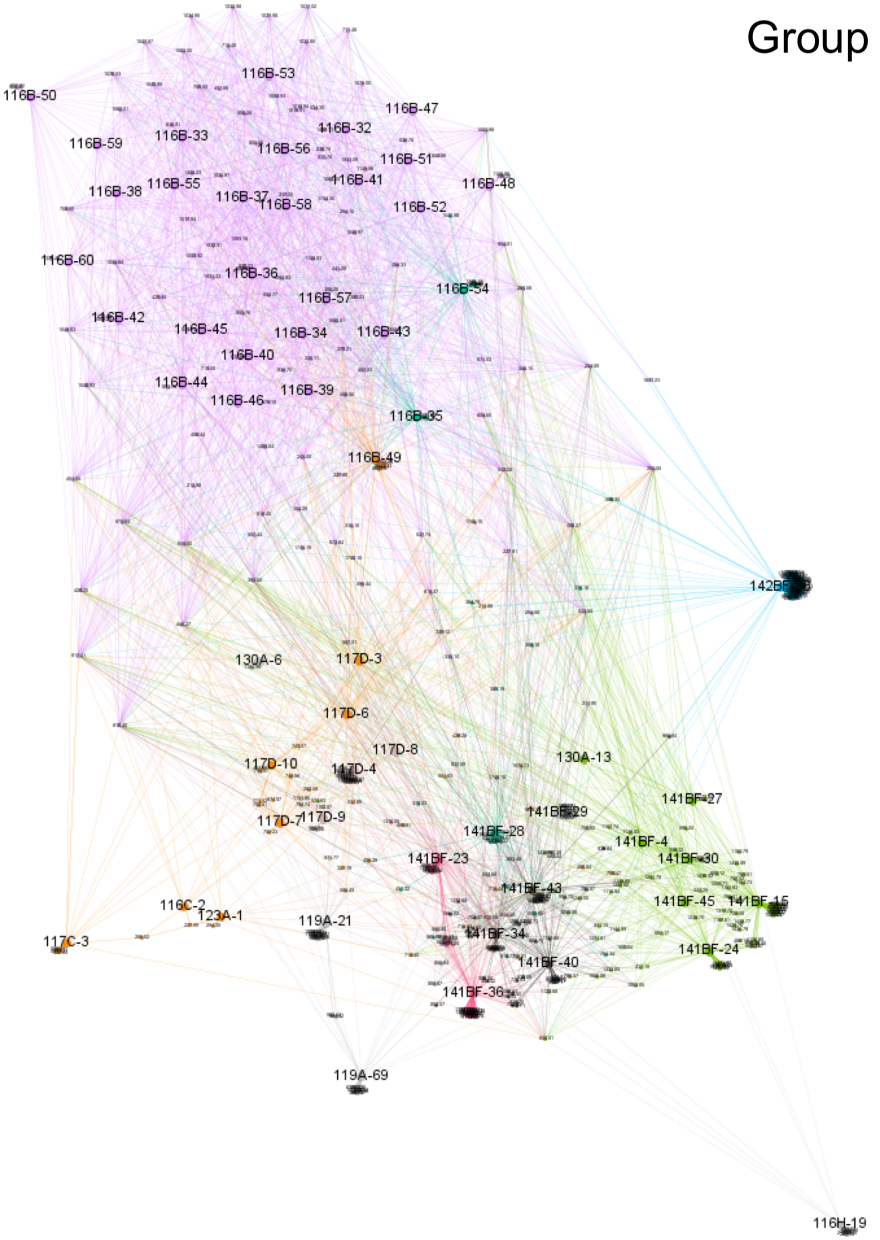
Group 10


119A-16
116H-18
116B-52

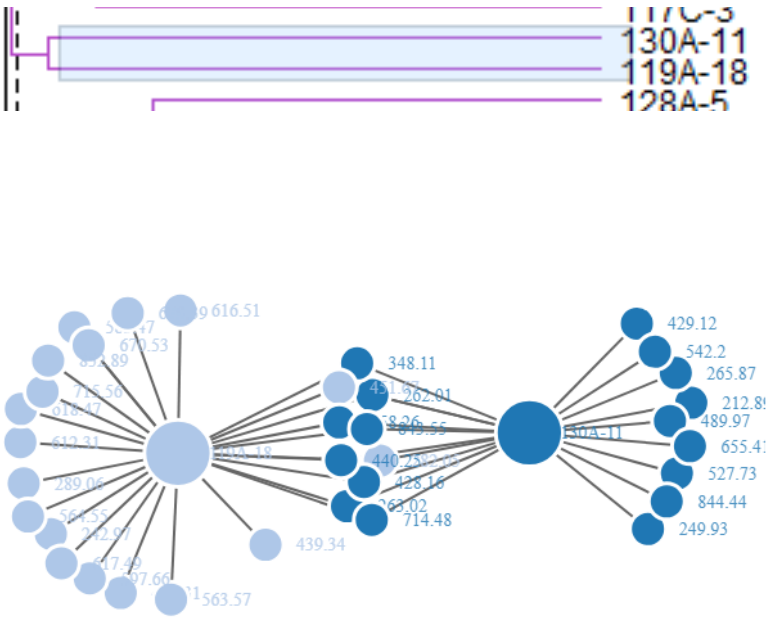

231.26
528.13
656.35
267.02
428.31
116H-18
618.36
542.41
221.95
262.33
451.78

Group 11

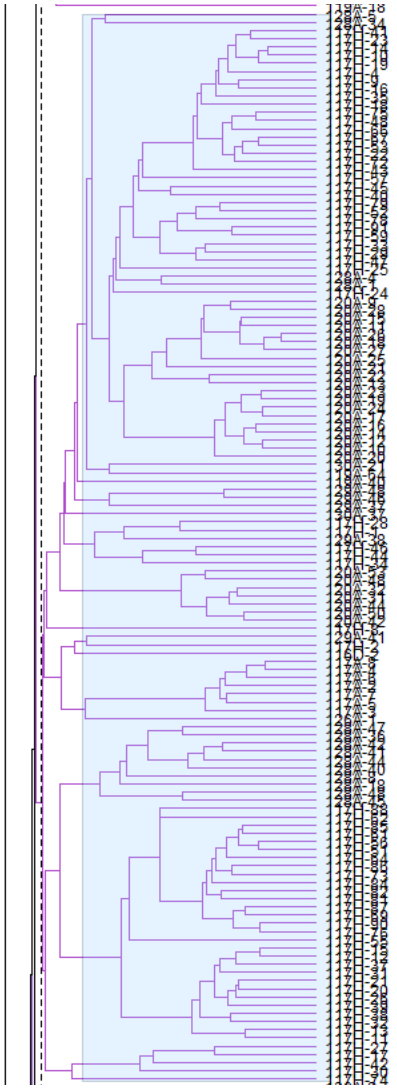

116B-52
116B-41
116B-40
116B-50
116B-56
116B-43
116B-47
116B-46
116B-59
116B-37
116B-32
116B-58
116B-55
116B-54
116B-38
116B-60
116B-48
116B-36
116B-42
116B-51
116B-53
116B-33
116B-57
116B-44
116B-35
116B-34
116B-45
116B-49
116B-39
116H-19
119A-21
130A-6
117D-3
116C-2
130A-13
117D-7
117D-10
117D-9
117D-4
117D-8
117D-6
142BF-36
119A-69
123A-1
141BF-29
141BF-23
141BF-27
141BF-15
141BF-45
141BF-4
141BF-24
141BF-30
141BF-43
141BF-34
141BF-28
141BF-36
141BF-40
117C-3

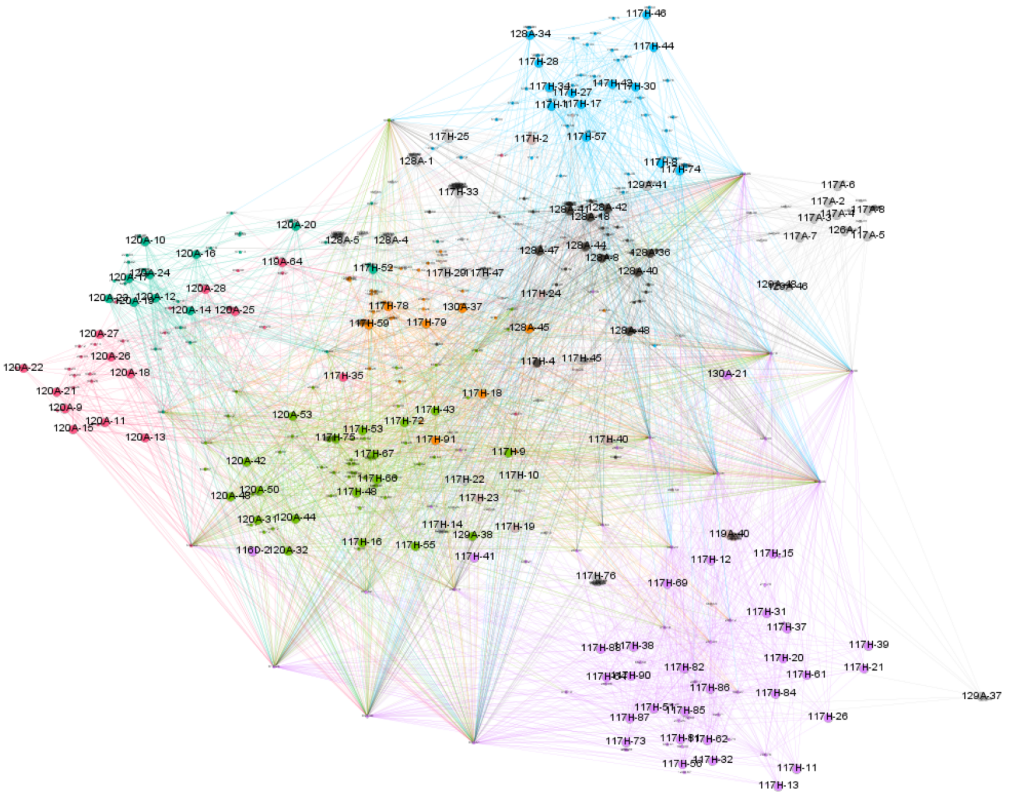

116B-53
116B-50
116B-47
116B-32
116B-59
116B-33
116B-56
116B-51
116B-41
116B-48
116B-38
116B-55
116B-37
116B-58
116B-52
116B-60
116B-36
116B-54
116B-57
116B-42
116B-45
116B-34
116B-43
116B-40
116B-44
116B-39
116B-46
116B-35
116B-49
142BF
130A-6
117D-3
117D-6
117D-8
130A-13
117D-10
117D-4
141BF-27
141BF-29
117D-7
117D-9
141BF-28
141BF-4
141BF-23
141BF-30
141BF-43
116C-2
141BF-45
141BF-15
123A-1
119A-21
117C-3
141BF-34
141BF-24
141BF-40
141BF-36
119A-69
116H-19

Group 12


130A-11
119A-18
616.51
348.11
429.12
542.2
265.87
489.97
655.4
527.73
844.44
249.93
714.48
439.34
563.57

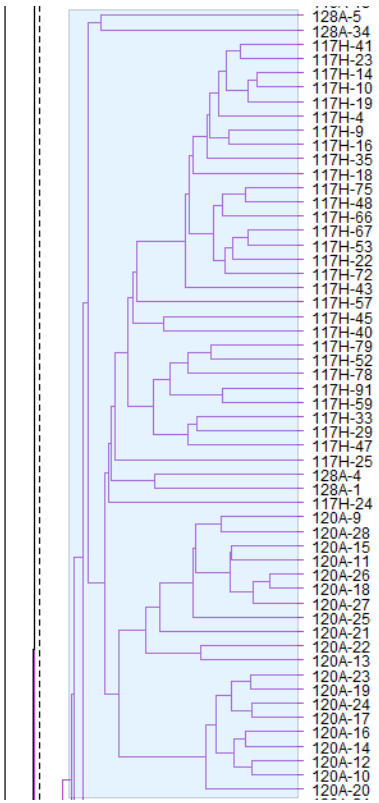
Group 13

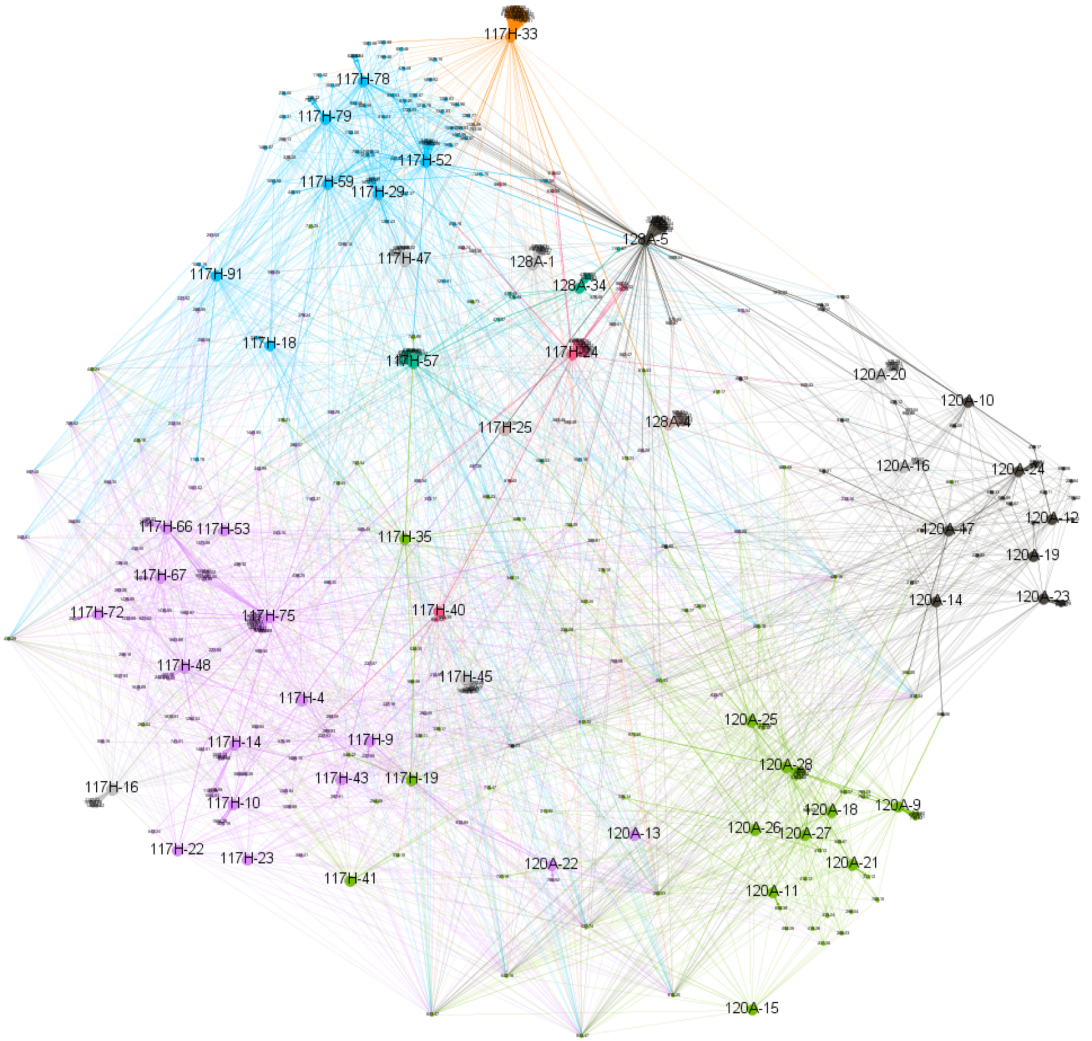




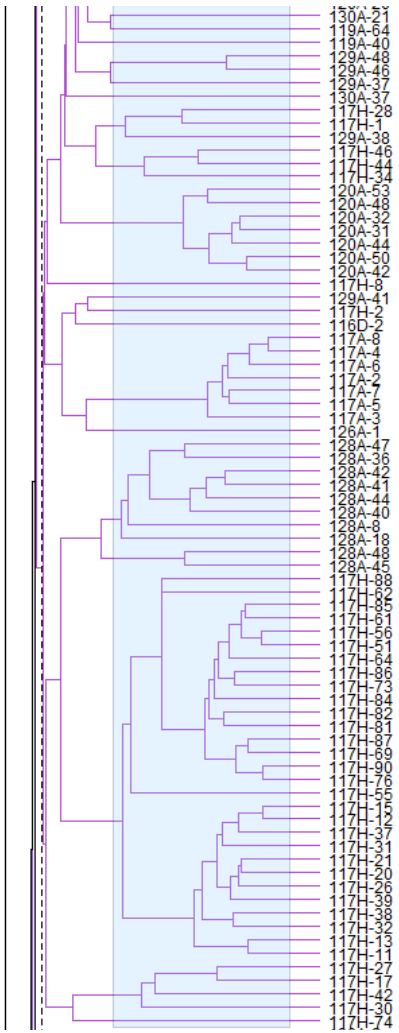
*Note: This group was subdivided into 2 groups, since it contained greater than 100 strains.

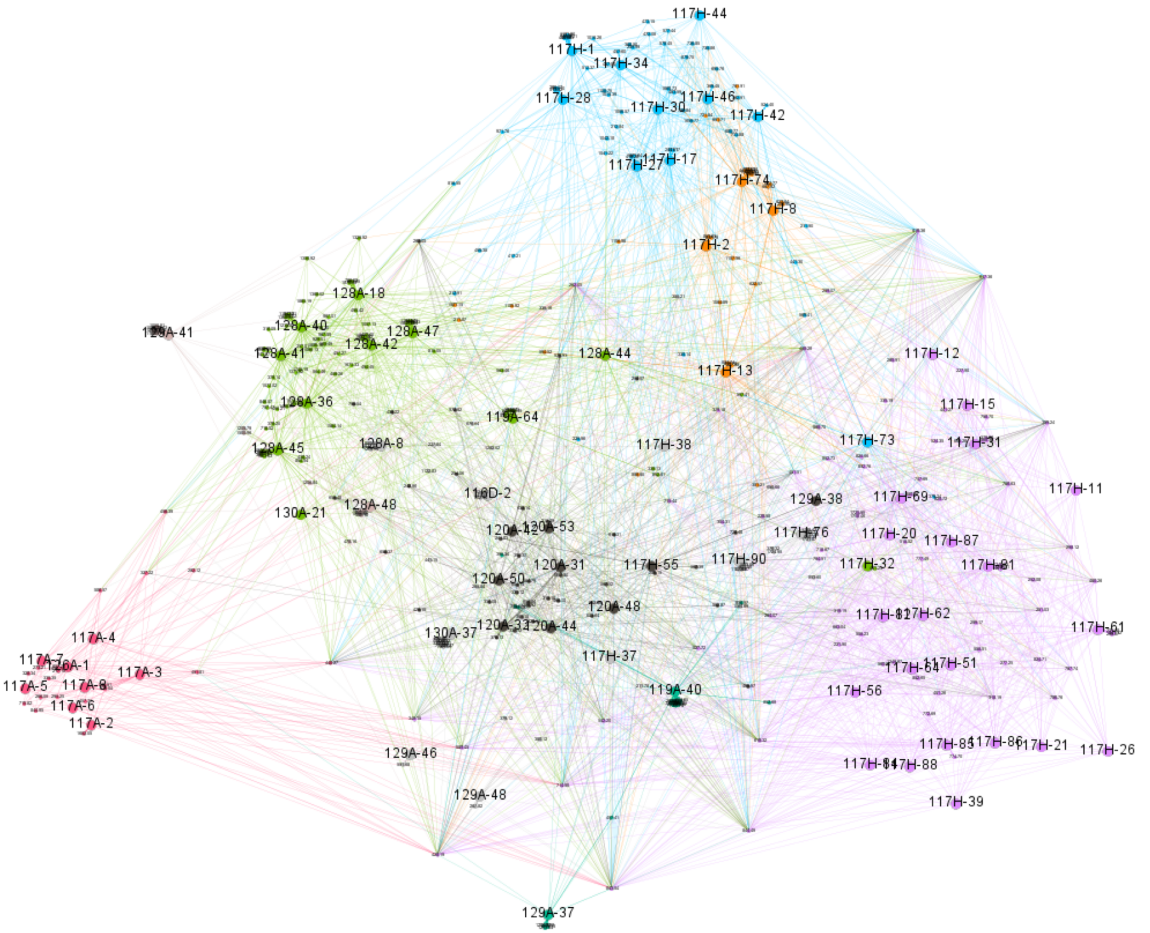
Group 13a


128A-5
128A-34
117H-41
117H-23
117H-14
117H-10
117H-19
117H-4
117H-9
117H-16
117H-35
117H-18
117H-75
117H-48
117H-66
117H-67
117H-53
117H-22
117H-72
117H-43
117H-57
117H-45
117H-40
117H-79
117H-52
117H-78
117H-91
117H-59
117H-33
117H-29
117H-47
117H-25
128A-4
128A-1
117H-24
120A-9
120A-28
120A-15
120A-11
120A-26
120A-18
120A-27
120A-25
120A-21
120A-22
120A-13
120A-23
120A-19
120A-24
120A-17
120A-16
120A-14
120A-12
120A-10
120A-20

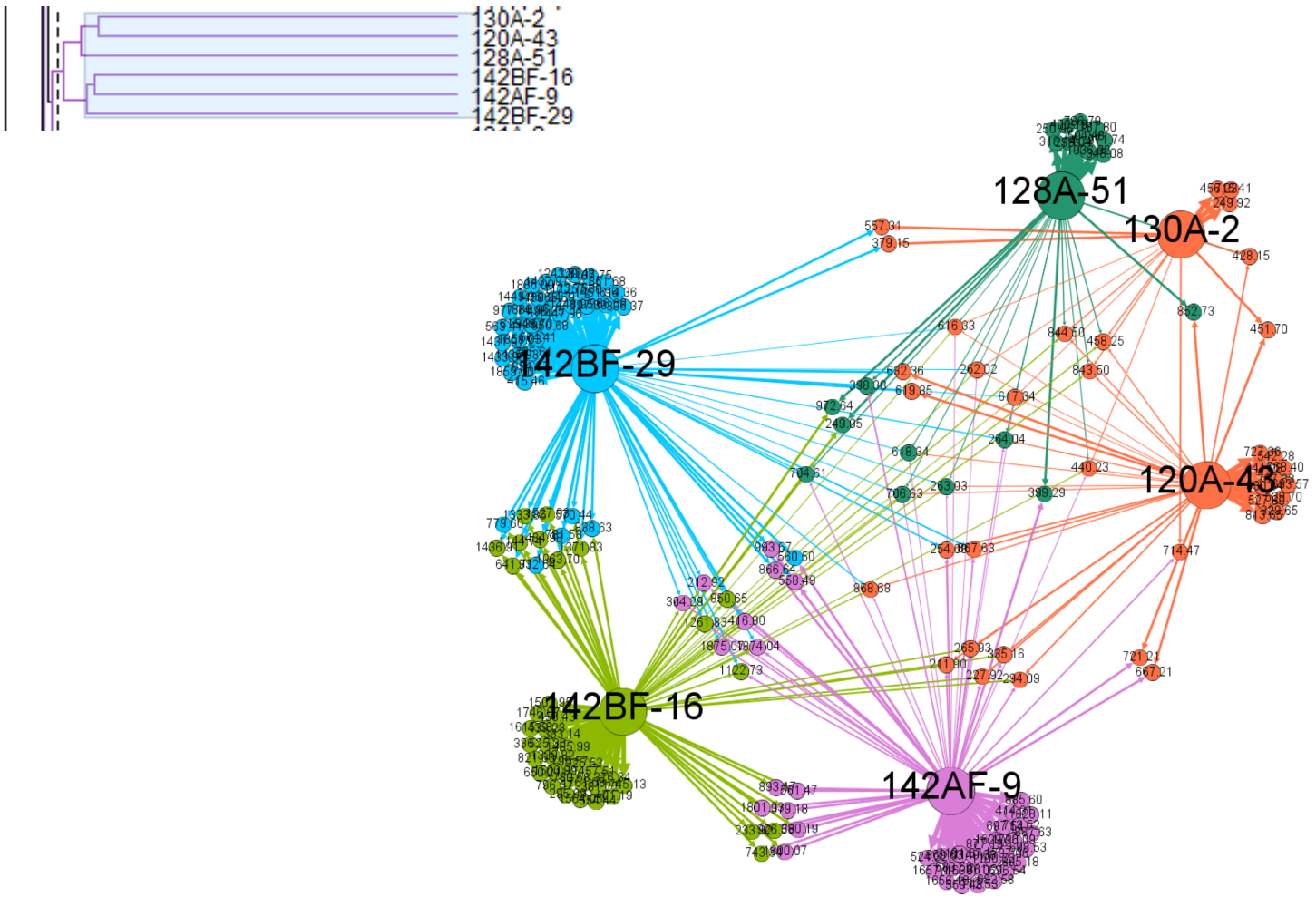

117H-33
117H-78
117H-79
117H-52
117H-59
117H-29
128A-5
117H-47
128A-1
117H-91
128A-34
117H-18
117H-57
117H-24
120A-20
120A-10
117H-25
128A-4
120A-16
120A-24
120A-12
117H-66
117H-53
117H-35
120A-47
120A-19
117H-67
120A-14
120A-23
117H-72
117H-75
117H-40
117H-48
117H-45
117H-4
120A-25
117H-14
117H-9
120A-28
117H-16
117H-43
117H-19
120A-18
120A-9
117H-10
120A-26
120A-27
120A-13
117H-22
117H-23
120A-22
120A-21
117H-41
120A-11
120A-15

Group 13b

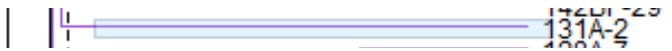

130A-21
119A-64
119A-40
129A-49
129A-46
129A-37
130A-37
117H-28
117H-1
129A-38
117H-46
117H-44
117H-34
120A-53
120A-48
120A-32
120A-31
120A-44
120A-50
120A-42
117H-8
129A-41
117H-2
116D-2
117A-8
117A-4
117A-6
117A-2
117A-7
117A-1
117A-5
117A-3
126A-1
128A-47
128A-36
128A-42
128A-41
128A-44
128A-40
128A-8
128A-18
128A-48
128A-45
117H-88
117H-62
117H-85
117H-61
117H-56
117H-51
117H-64
117H-86
117H-73
117H-84
117H-82
117H-81
117H-87
117H-69
117H-90
117H-76
117H-55
117H-15
117H-12
117H-37
117H-31
117H-21
117H-20
117H-26
117H-39
117H-38
117H-32
117H-13
117H-11
117H-27
117H-17
117H-42
117H-30
117H-74

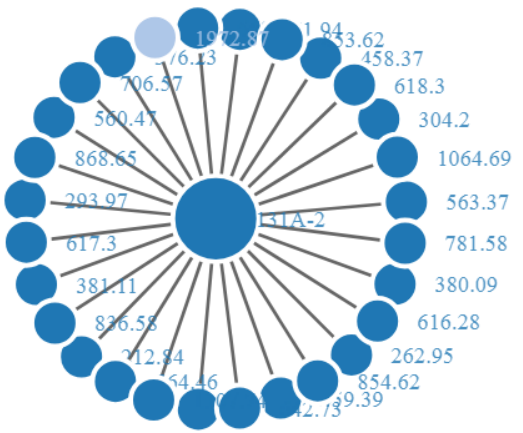

117H-44
117H-1
117H-34
117H-28
117H-46
117H-30
117H-42
117H-27
117H-17
117H-74
117H-8
117H-2
128A-18
128A-41
128A-40
128A-47
128A-42
128A-41
128A-44
117H-12
117H-13
128A-36
117H-15
119A-64
128A-45
128A-8
117H-38
117H-73
117H-31
116D-2
129A-38
117H-69
117H-11
130A-21
128A-48
120A-42
120A-53
117H-76
117H-20
117H-87
120A-31
117H-55
117H-90
117H-32
117H-81
120A-50
120A-48
117H-82
117H-62
130A-37
120A-33
120A-44
117H-61
117A-4
117H-37
117A-7
117A-1
117A-3
117H-64
117H-51
117A-5
117A-8
119A-40
117H-56
117A-6
117A-2
129A-46
117H-85
117H-86
117H-21
117H-26
117H-84
117H-88
129A-48
117H-39
129A-37

Group 14

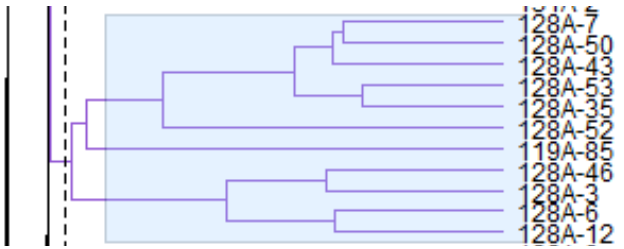

130A-2
120A-43
128A-51
142BF-16
142AF-9
142BF-29
128A-51
130A-2
120A-43
142BF-29
142BF-16
142AF-9

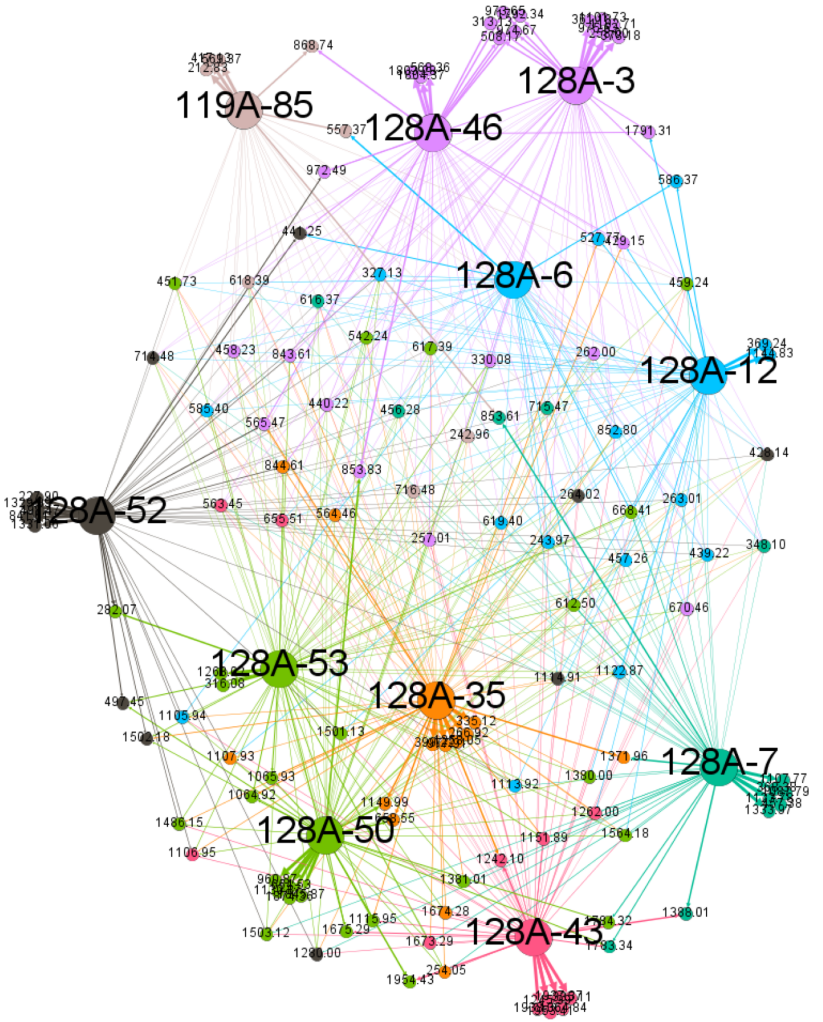
Group 15


131A-2

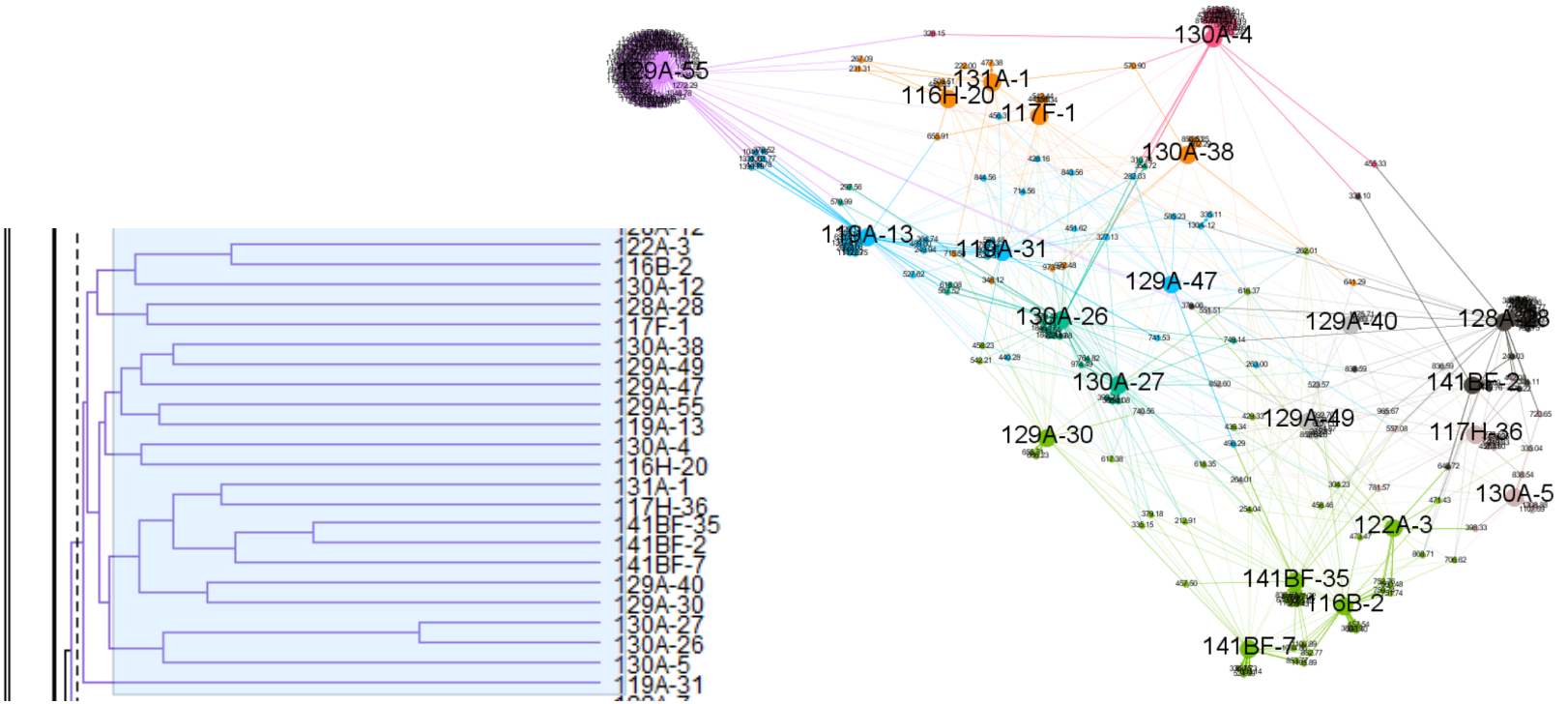

1.94
33.62
458.37
618.3
304.2
1064.69
563.37
781.58
380.09
616.28
262.95
854.62
706.57
560.47
868.65
293.97
617.3
381.11
836.58
131A-2

Group 16


128A-7
128A-50
128A-43
128A-53
128A-35
128A-52
119A-85
128A-46
128A-3
128A-6
128A-12

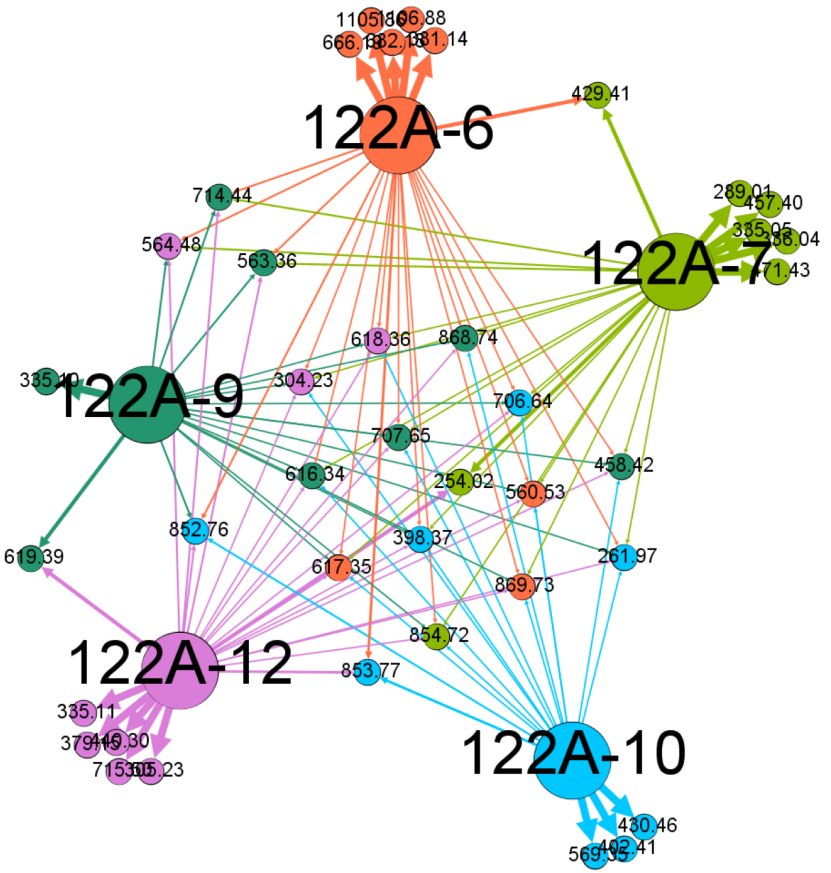

119A-85
128A-46
128A-3
128A-6
128A-12
128A-52
128A-53
128A-35
128A-7
128A-50
128A-43

Group 17

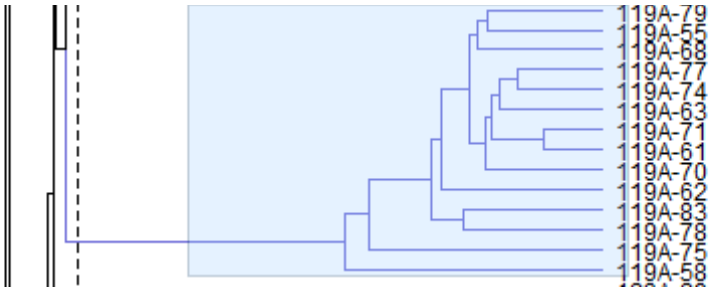

122A-3
116B-2
130A-12
128A-28
117F-1
130A-38
129A-49
129A-47
129A-55
119A-13
130A-4
116H-20
131A-1
117H-36
141BF-35
141BF-2
141BF-7
129A-40
129A-30
130A-27
130A-26
130A-5
119A-31
129A-55
131A-1
116H-20
117F-1
130A-4
130A-38
119A-13
119A-31
129A-47
130A-26
129A-40
128A-28
141BF-2
130A-27
129A-49
129A-30
117H-36
130A-5
122A-3
141BF-35
116B-2
141BF-7

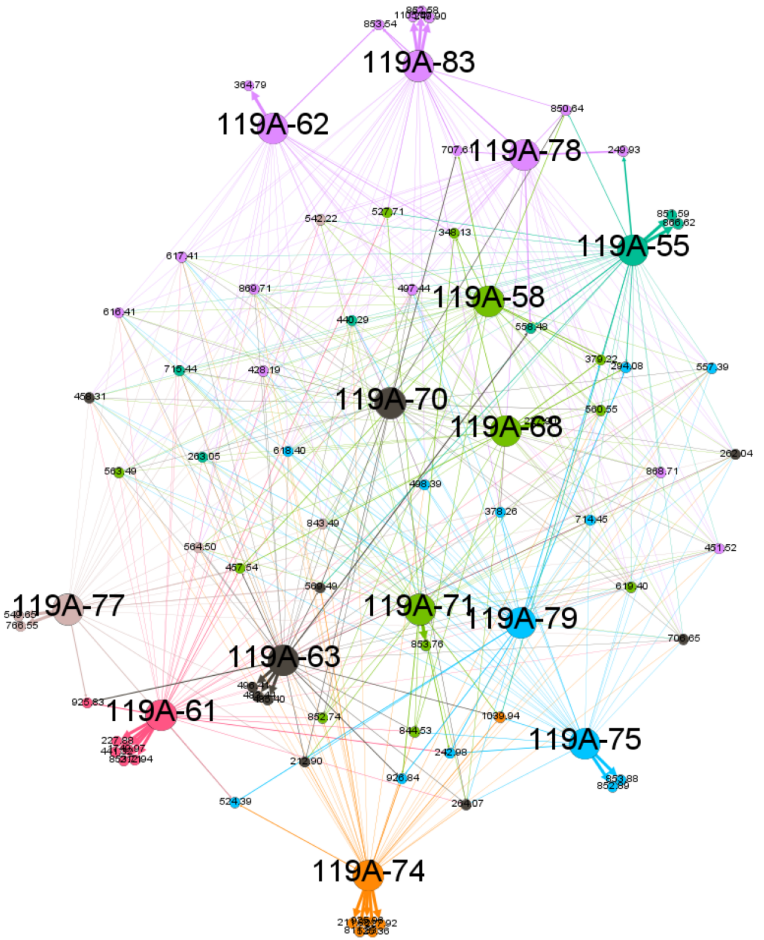
Group 18


122A-7
122A-6
122A-9
122A-10
122A-12

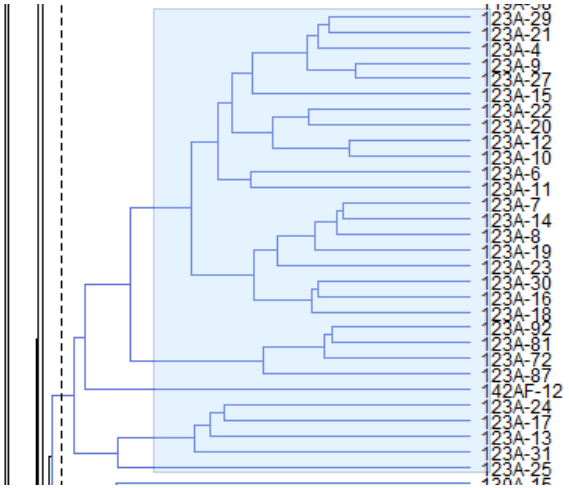

122A-6
122A-7
122A-9
122A-12
122A-10
429.41
714.44
564.48
563.36
618.36
868.74
304.23
706.64
707.65
616.34
254.02
560.53
458.42
852.76
398.37
619.39
617.35
261.97
869.73
854.72
853.77
335.11
430.46

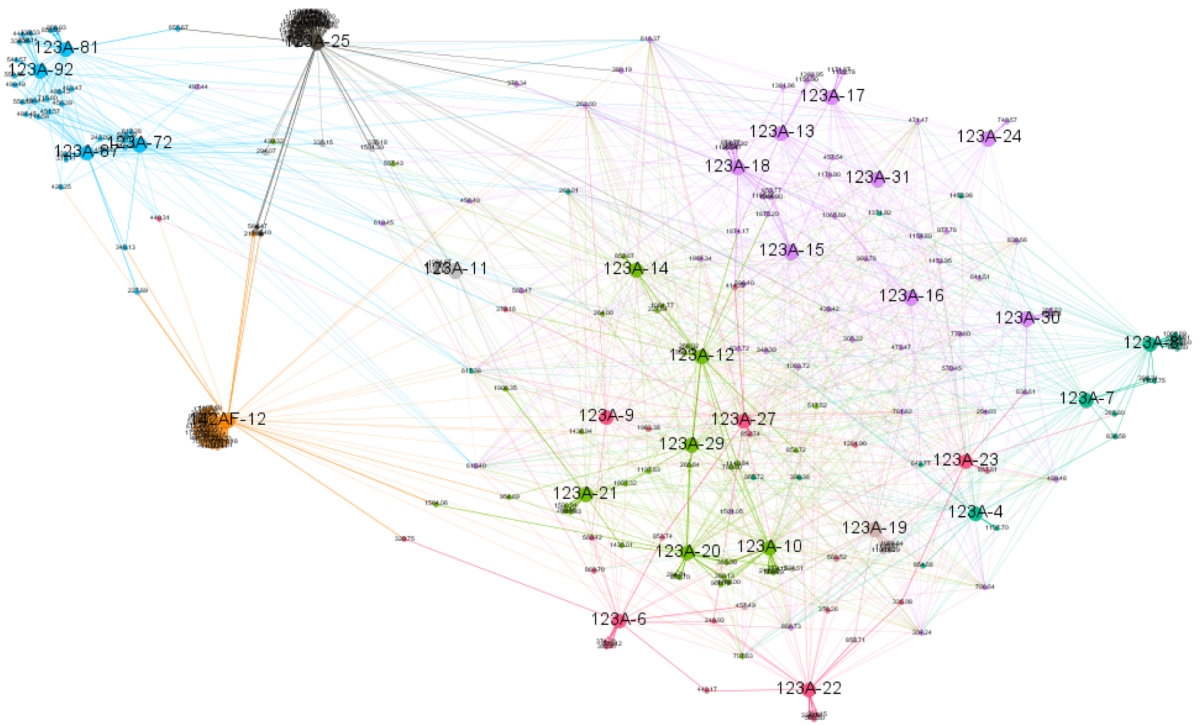
Group 19


119A-79
119A-55
119A-68
119A-77
119A-74
119A-63
119A-71
119A-61
119A-70
119A-62
119A-83
119A-78
119A-75
119A-58

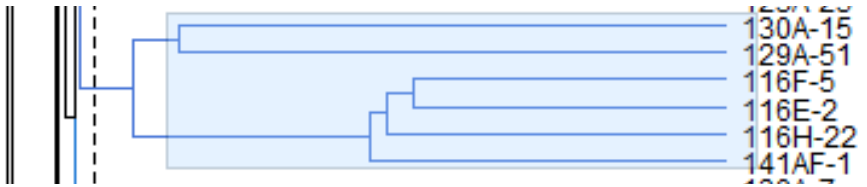

119A-83
119A-62
119A-78
119A-55
119A-58
119A-70
119A-68
119A-77
119A-71
119A-79
119A-63
119A-61
119A-75
119A-74

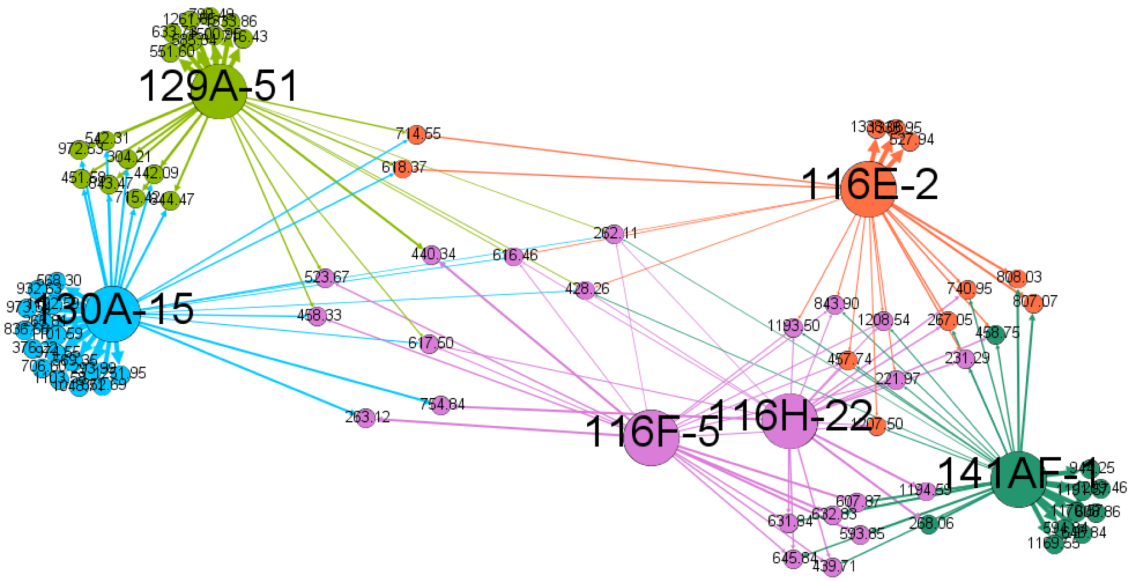
Group 20


123A-29
123A-21
123A-4
123A-9
123A-27
123A-15
123A-22
123A-20
123A-12
123A-10
123A-6
123A-11
123A-7
123A-14
123A-8
123A-19
123A-23
123A-30
123A-16
123A-18
123A-92
123A-81
123A-72
123A-87
142AF-12
123A-24
123A-17
123A-13
123A-31
123A-25


123A-81
123A-92
123A-72
123A-87
123A-25
123A-17
123A-13
123A-24
123A-18
123A-31
123A-15
123A-11
123A-14
123A-16
123A-30
123A-12
123A-7
142AF-12
123A-9
123A-27
123A-29
123A-23
123A-21
123A-4
123A-19
123A-20
123A-10
123A-6
123A-22

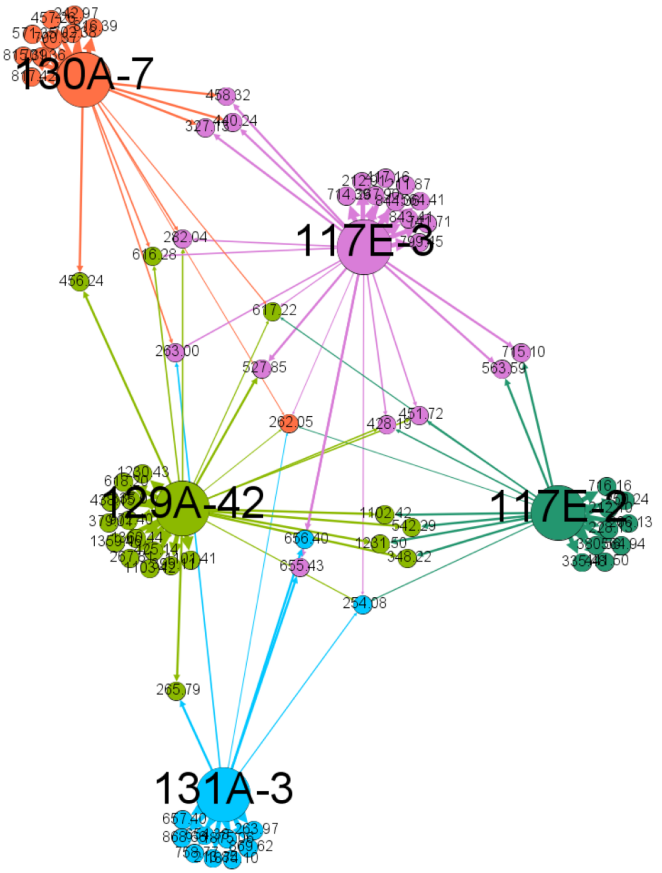
Group 21


130A-15
129A-51
116F-5
116E-2
116H-22
141AF-1


129A-51
116E-2
130A-15
116F-5
116H-22
141AF-1

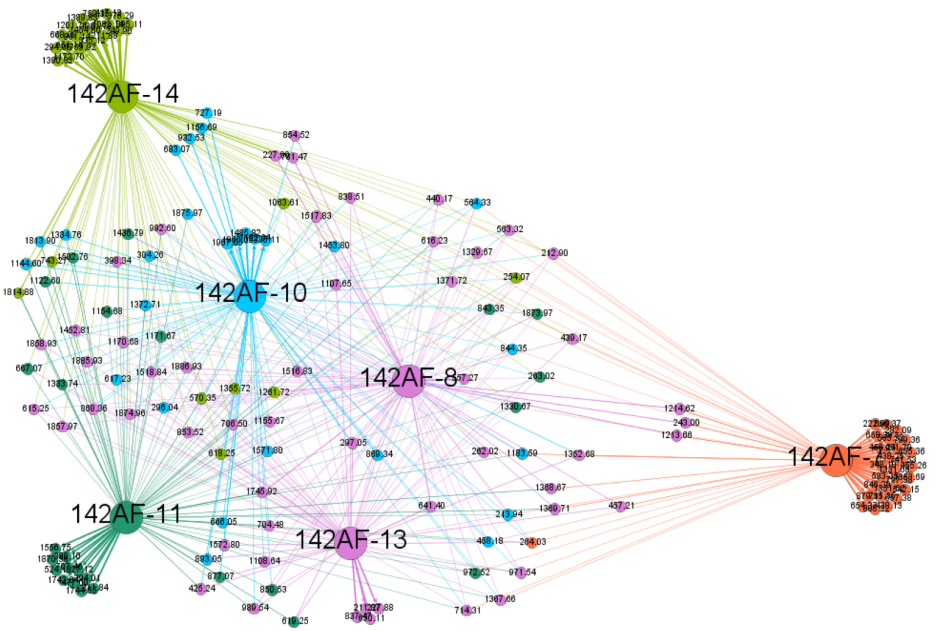
Group 22


130A-7
117E-3
117E-2
131A-3
129A-42

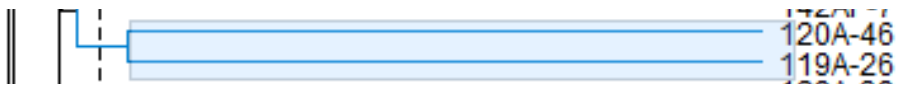

130A-7
117E-3
129A-42
117E-2
131A-3

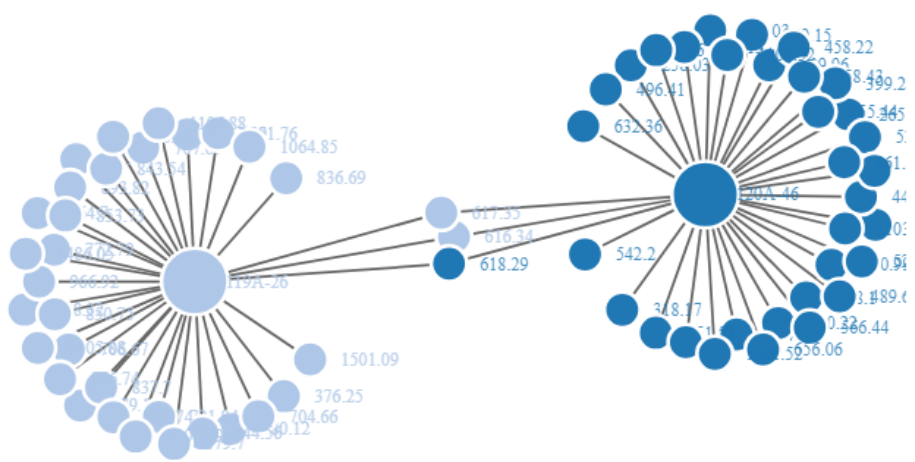
Group 23


142AF-8
142AF-11
142AF-10
142AF-13
142AF-14
142AF-7

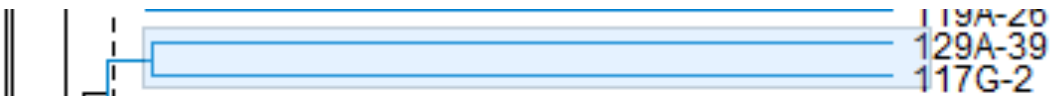

142AF-14
142AF-10
142AF-8
142AF-11
142AF-13

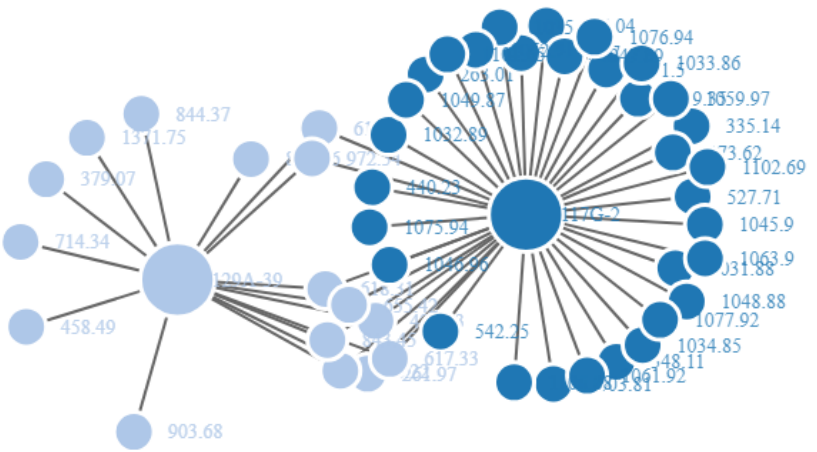
Group 24


120A-46
119A-26

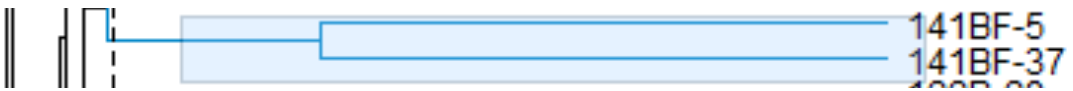

1064.85
836.69
617.35
616.34
618.29
1501.09
376.25
704.66
632.36
542.2
458.22
366.44

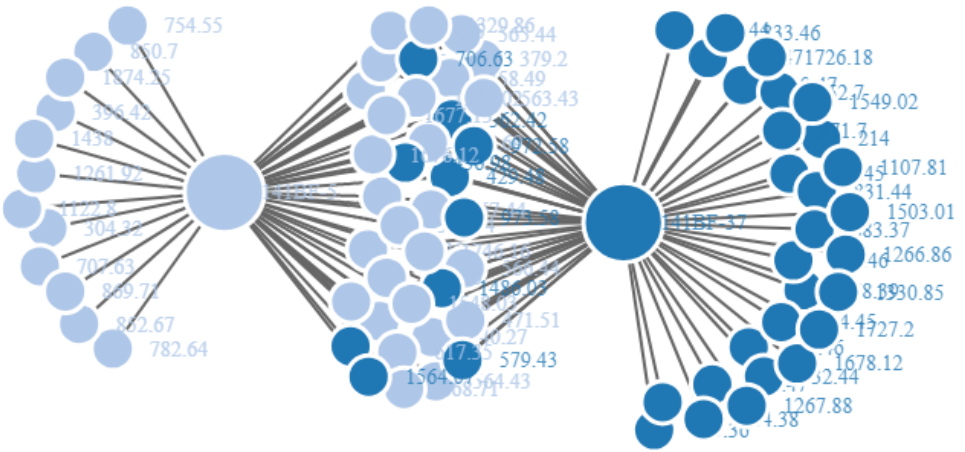
Group 25


129A-39
117G-2

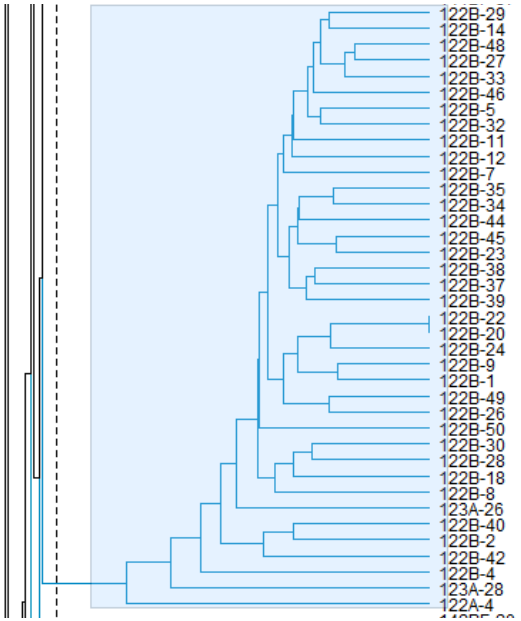

844.37
379.07
714.34
458.49
903.68
1076.94
1033.86
1049.87
1032.89
335.14
1102.69
527.71
1045.9
1063.9
1048.88
1077.92
1034.85
542.25
617.33

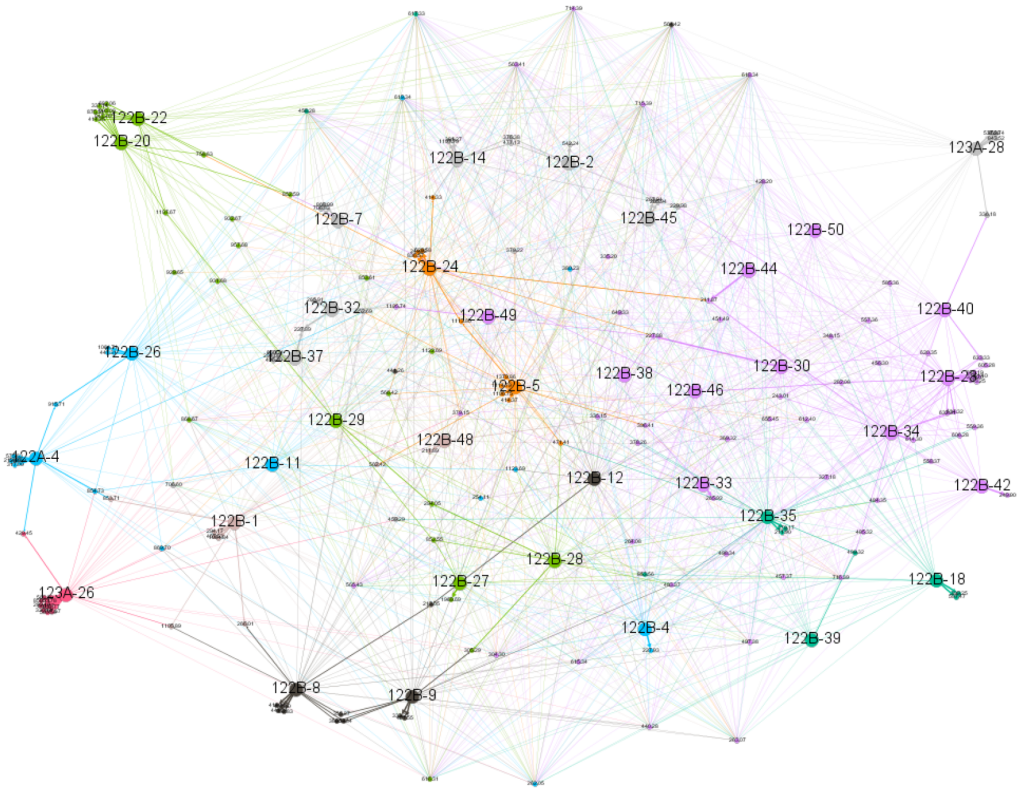
Group 26


141BF-5
141BF-37

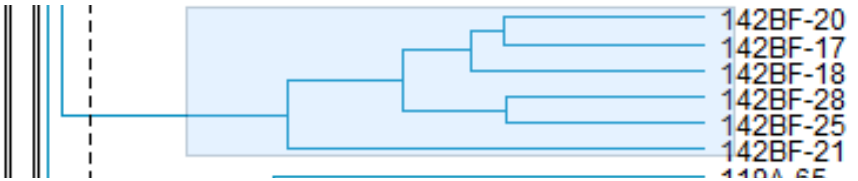

754.55
782.64
579.43
1549.02
1107.81
1503.01
1266.86
1727.2
1678.12
1267.88

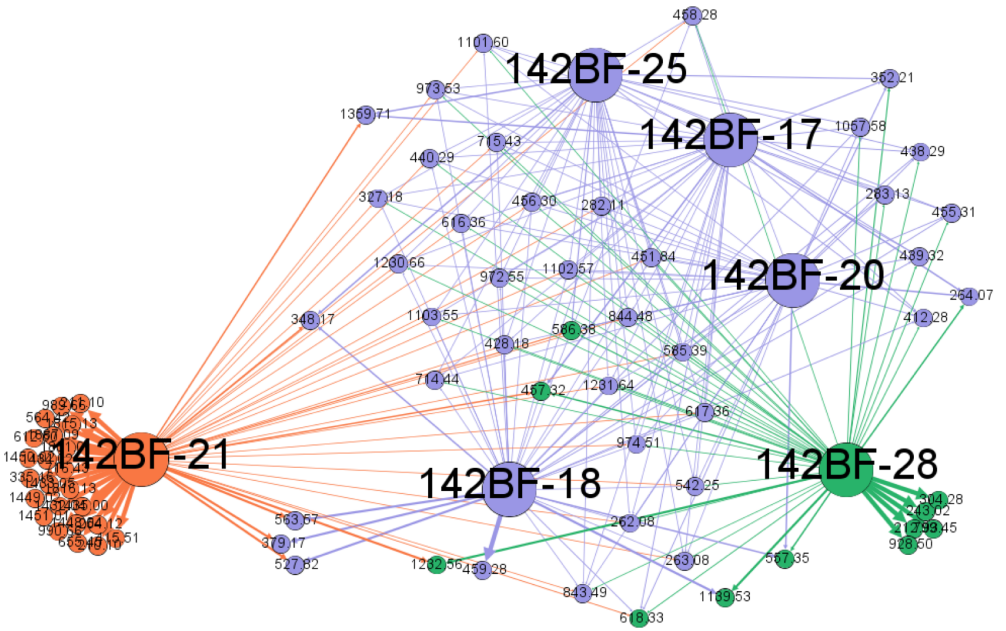
Group 27


122B-29
122B-14
122B-48
122B-27
122B-33
122B-46
122B-5
122B-32
122B-11
122B-12
122B-7
122B-35
122B-34
122B-44
122B-45
122B-23
122B-38
122B-37
122B-39
122B-22
122B-20
122B-24
122B-9
122B-1
122B-49
122B-26
122B-50
122B-30
122B-28
122B-18
122B-8
123A-26
122B-40
122B-2
122B-42
122B-4
123A-28
122A-4

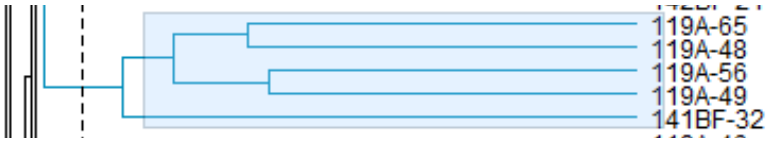

122B-22
122B-20
122B-14
122B-2
123A-28
122B-7
122B-45
122B-50
122B-24
122B-44
122B-32
122B-49
122B-40
122B-26
122B-37
122B-38
122B-30
122B-28
122B-5
122B-46
122B-29
122B-34
122B-48
122A-4
122B-11
122B-12
122B-33
122B-42
122B-1
122B-35
122B-28
122B-27
122B-18
123A-26
122B-4
122B-39
122B-8
122B-9

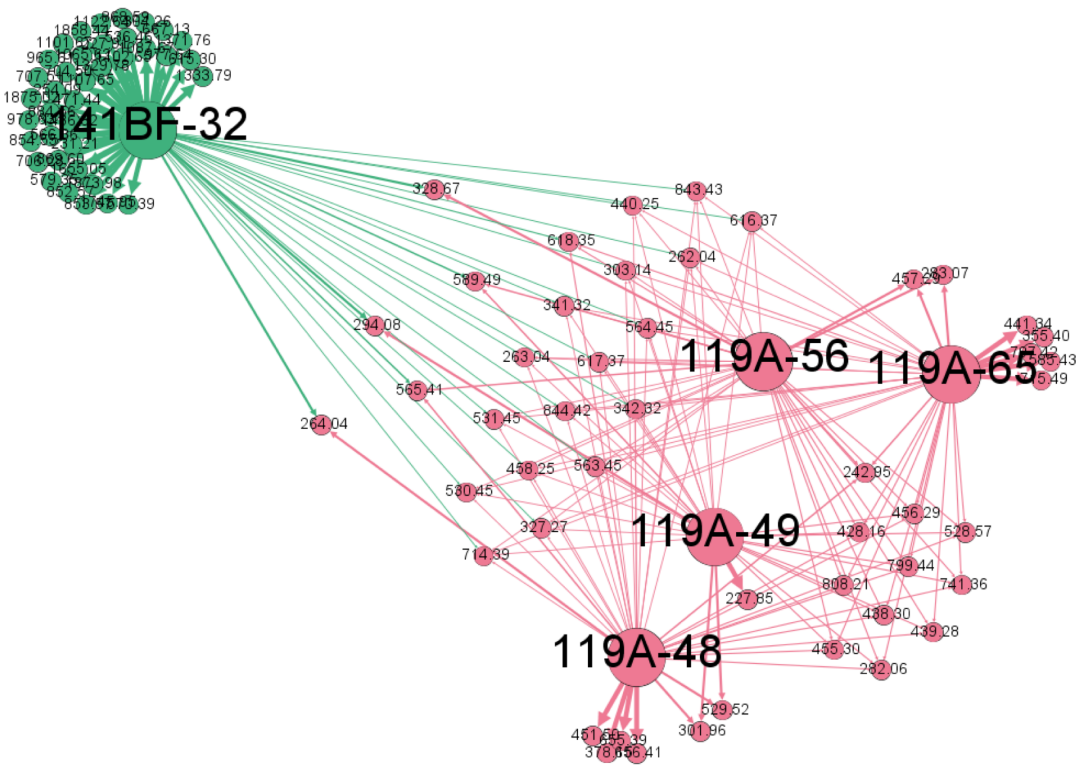
Group 28


142BF-20
142BF-17
142BF-18
142BF-28
142BF-25
142BF-21


142BF-25
142BF-17
142BF-20
142BF-21
142BF-18
142BF-28
458.28
1101.60
973.53
1359.71
352.21
1057.58
715.43
438.29
440.29
327.18
456.30
282.11
283.13
455.31
616.36
1230.66
451.84
439.32
1102.57
972.55
264.07
348.17
1103.55
844.48
412.28
586.36
428.18
585.39
714.44
457.32
1281.64
617.36
974.51
542.25
563.67
379.17
527.82
262.08
1232.56
459.28
263.08
557.35
843.49
1139.53
618.33

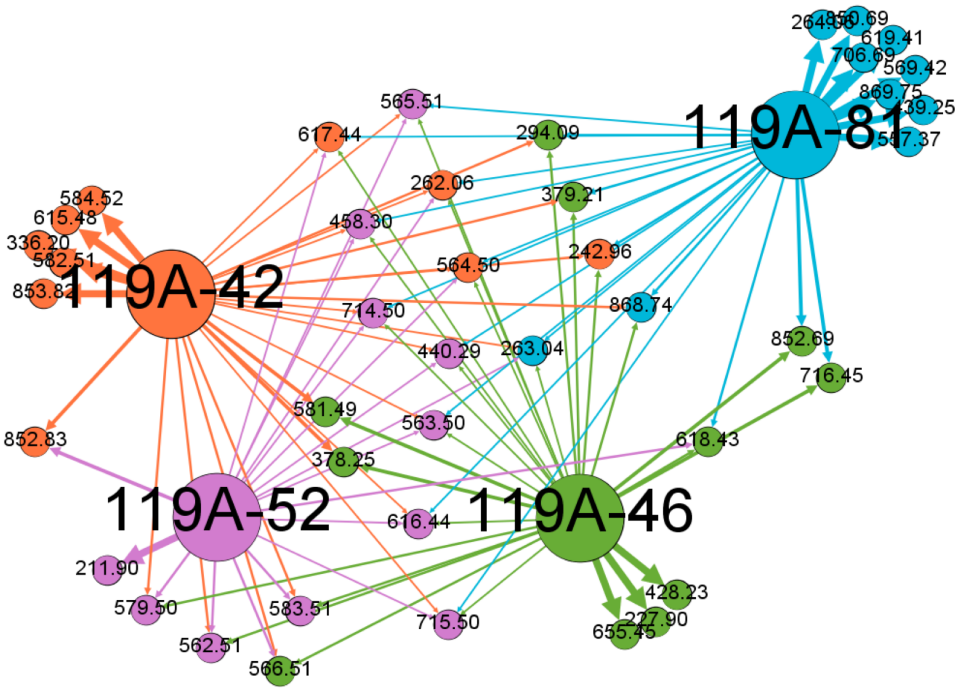
Group 29


119A-65
119A-48
119A-56
119A-49
141BF-32

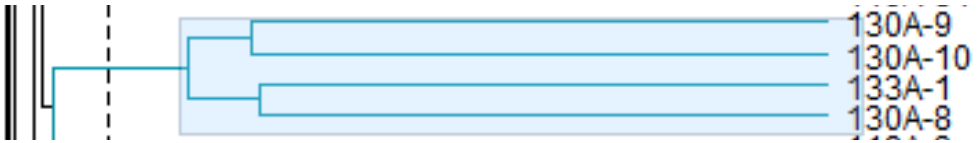

141BF-32
119A-56
119A-65
119A-49
119A-48
328.67
440.25
843.43
616.87
618.35
262.04
303.14
589.49
341.32
564.45
294.08
263.04
617.37
565.41
531.45
844.42
342.32
264.04
458.25
563.45
530.45
242.95
327.27
456.29
428.16
528.57
714.39
799.44
808.21
741.86
227.85
438.30
439.28
455.30
282.06
529.52
301.96

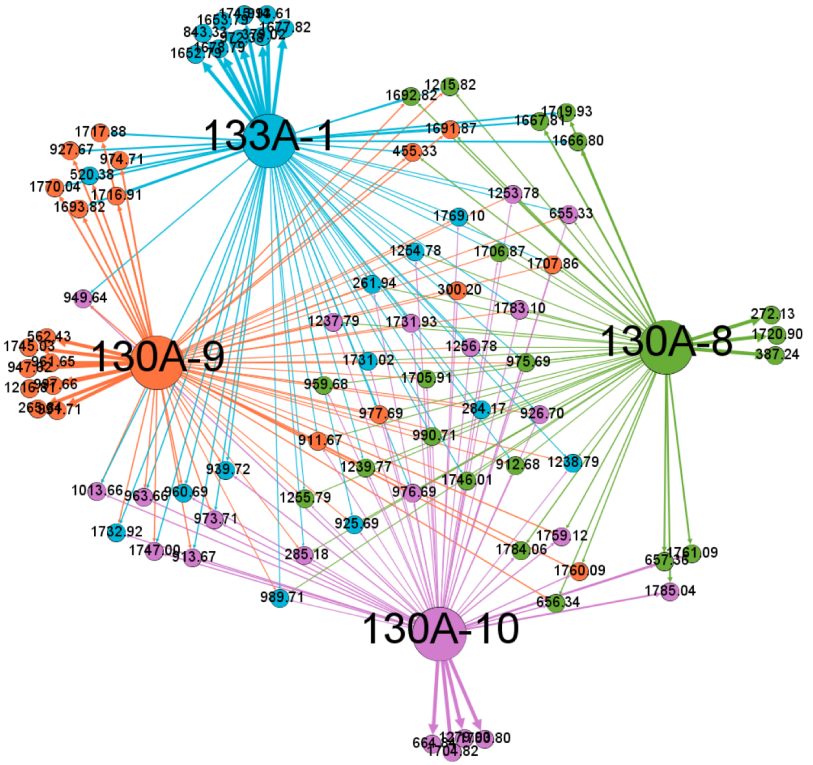
Group 30


119A-46
119A-42
119A-52
119A-81

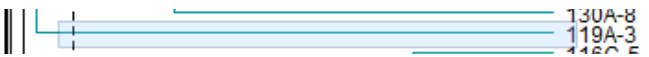

119A-81
119A-42
119A-52
119A-46
565.51
617.44
294.09
262.06
379.21
458.30
564.50
242.96
714.50
868.74
440.29
263.04
581.49
563.50
378.25
616.44
715.50
852.69
716.45
618.43
852.83
211.90
579.50
583.51
562.51
566.51
428.23
227.90
655.45
584.52
615.48
336.20
562.51
853.82
619.41
706.69
569.42
869.75
547.37

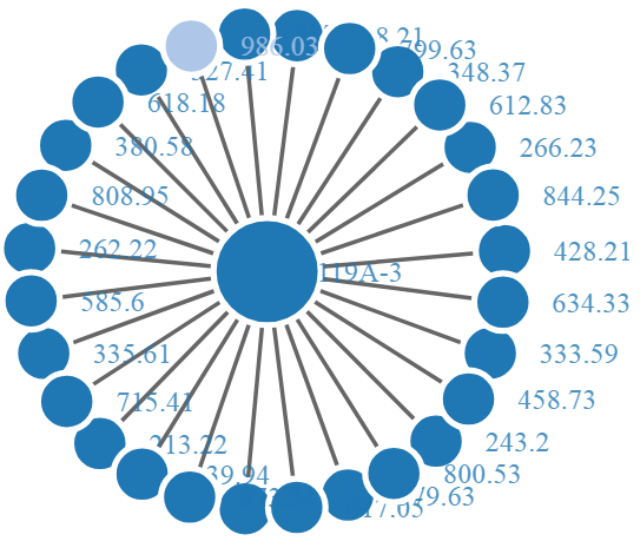
Group 31



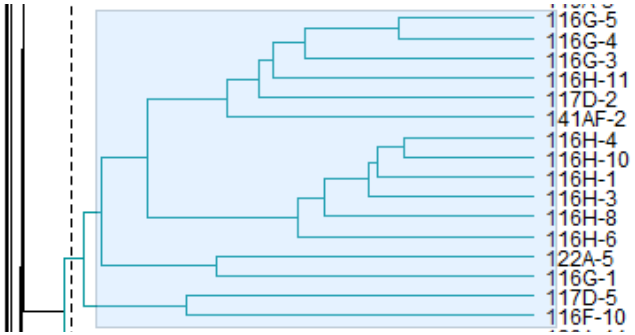


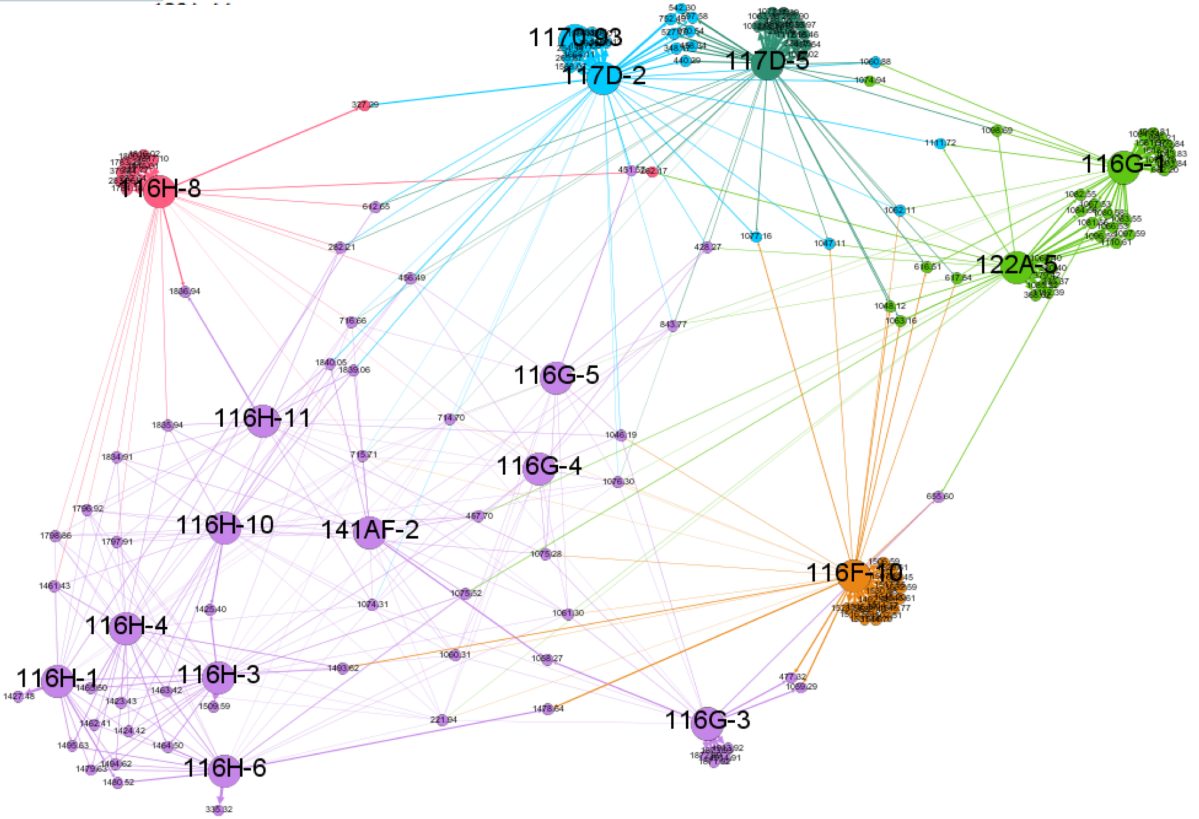
Group 32


130A-8
119A-3

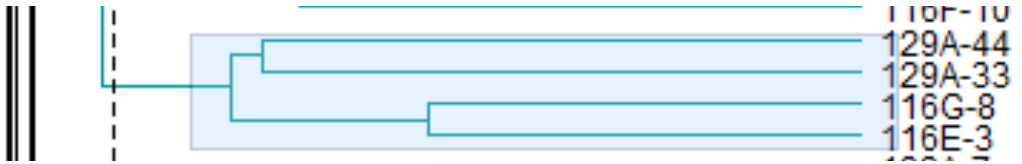

986.03
8.21
99.63
348.37
612.83
266.23
844.25
428.21
634.33
333.59
458.73
243.2
800.53
49.63
17.05
39.94
713.22
715.41
335.61
585.6
262.22
808.95
380.58
618.18
27.41
119A-3

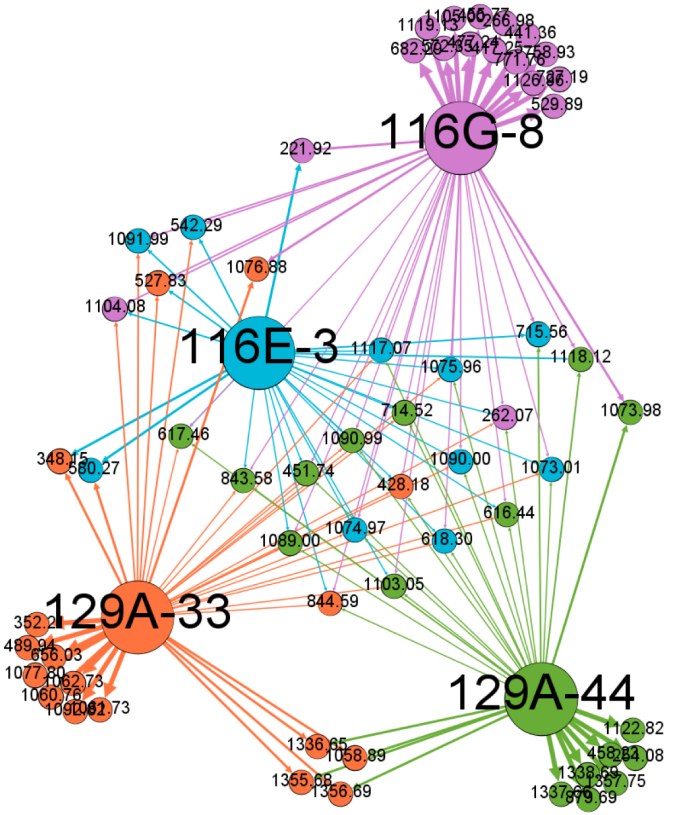
Group 33


116G-5
116G-4
116G-3
116H-11
117D-2
141AF-2
116H-4
116H-10
116H-1
116H-3
116H-8
116H-6
122A-5
116G-1
117D-5
116F-10

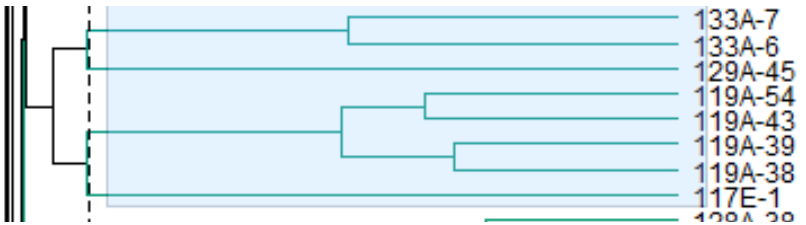

117D-3
117D-2
117D-5
116H-8
116G-1
122A-5
116G-5
116H-11
116G-4
116H-10
141AF-2
116F-10
116H-4
116H-1
116H-3
116G-3
116H-6

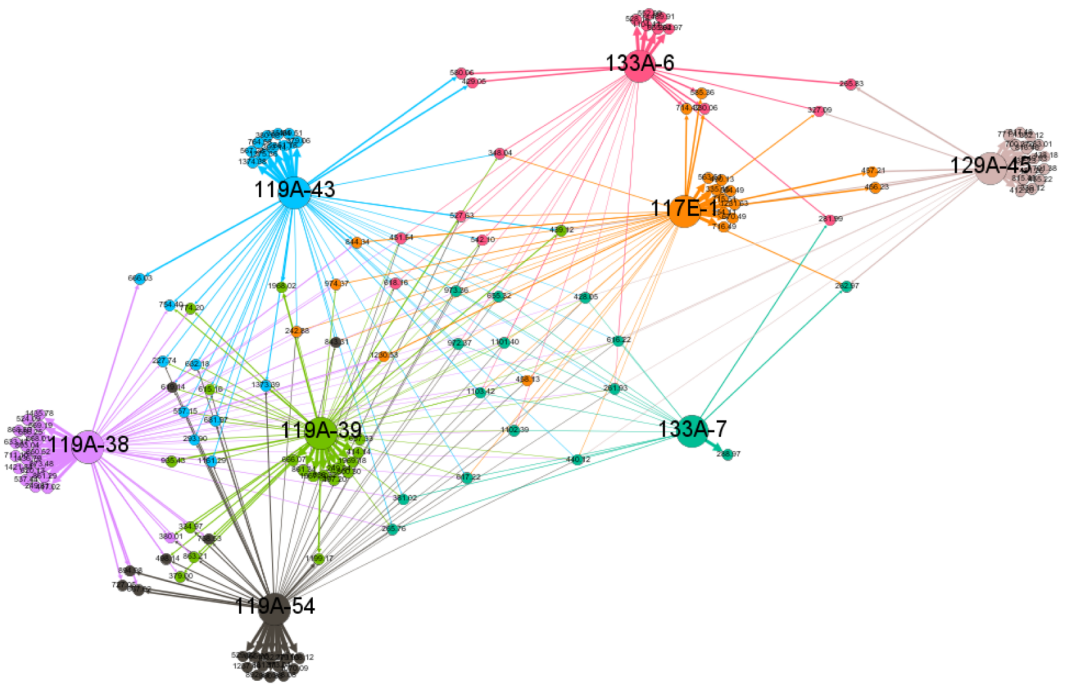
Group 34


129A-44
129A-33
116G-8
116E-3


116G-8
116E-3
129A-33
129A-44
221.92
542.29
1091.99
1076.88
527.83
1104.08
1117.07
715.56
1118.12
1075.96
714.52
262.07
1073.98
617.46
1090.99
1090.00
1073.01
348.15
580.27
843.58
451.74
428.18
616.44
1074.97
1089.00
618.30
1103.05
844.59
352.21
489.94
656.03
1077.80
1060.76
1122.82
1336.65
1058.89
1355.68
1356.69
1338.55
1357.75

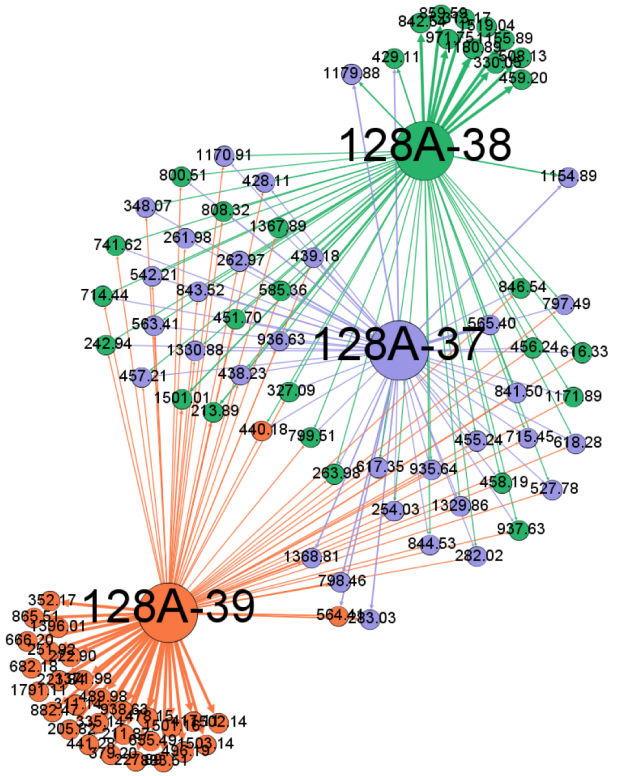
Group 35


133A-7
133A-6
129A-45
119A-54
119A-43
119A-39
119A-38
117E-1

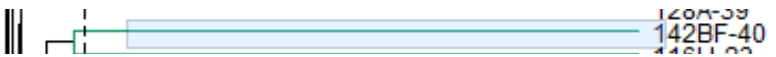

133A-6
119A-43
129A-45
117E-1
119A-38
119A-39
133A-7
119A-54

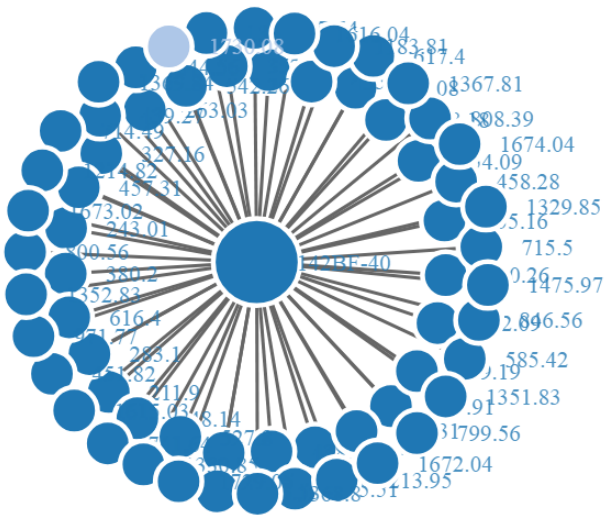
Group 36





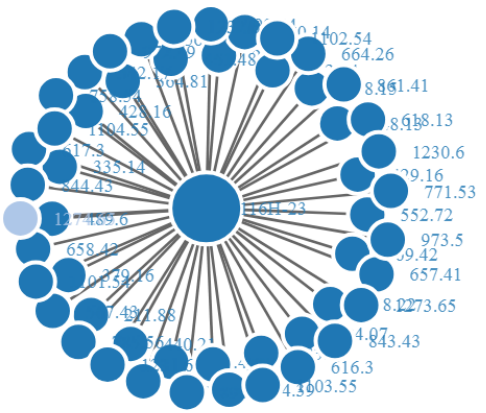
Group 37


142BF-40

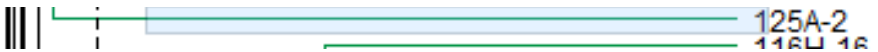

1367.81
1674.04
458.28
1329.85
715.5
1475.97
585.42
1351.83
1799.56
1672.04

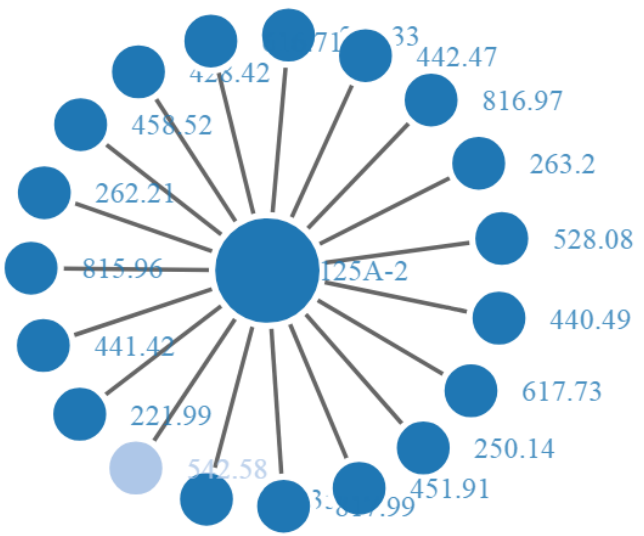
Group 38

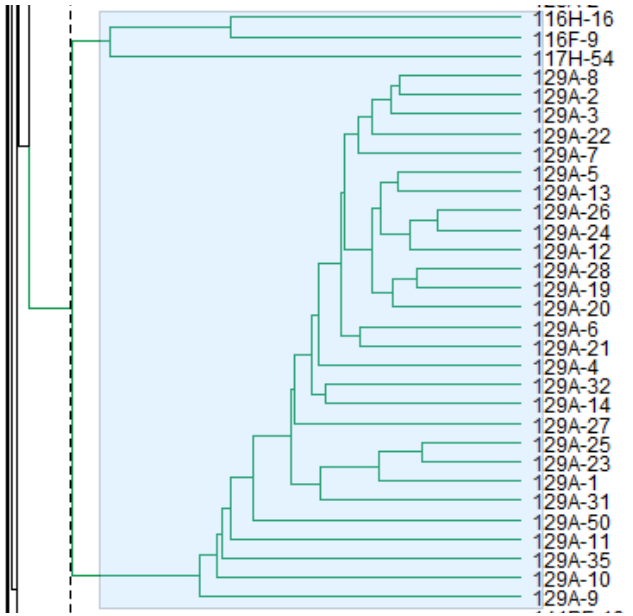

116H-23

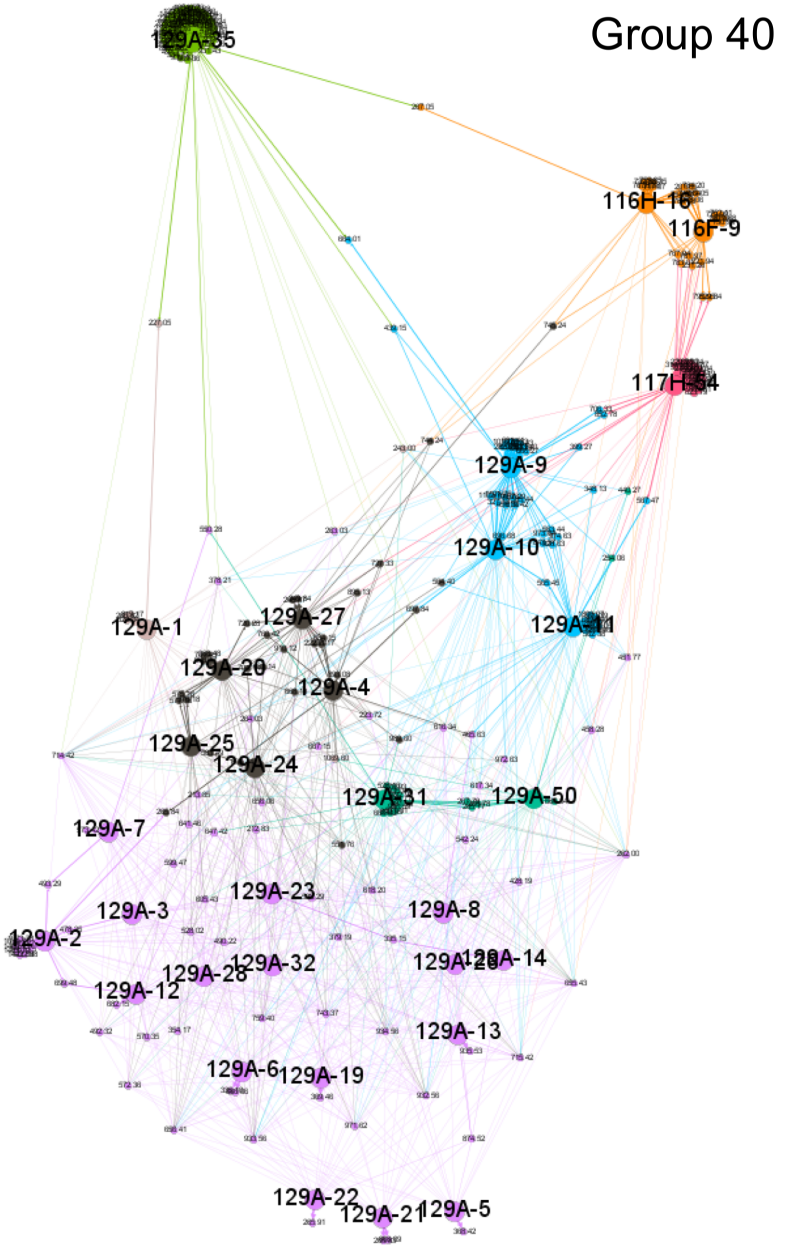

664.26
1230.6
771.53
552.72
973.5
657.41
843.43
616.3
658.43
116H-23

Group 39

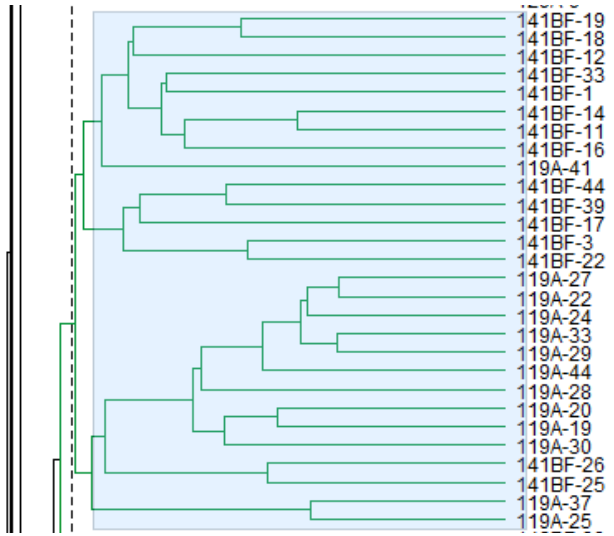

125A-2

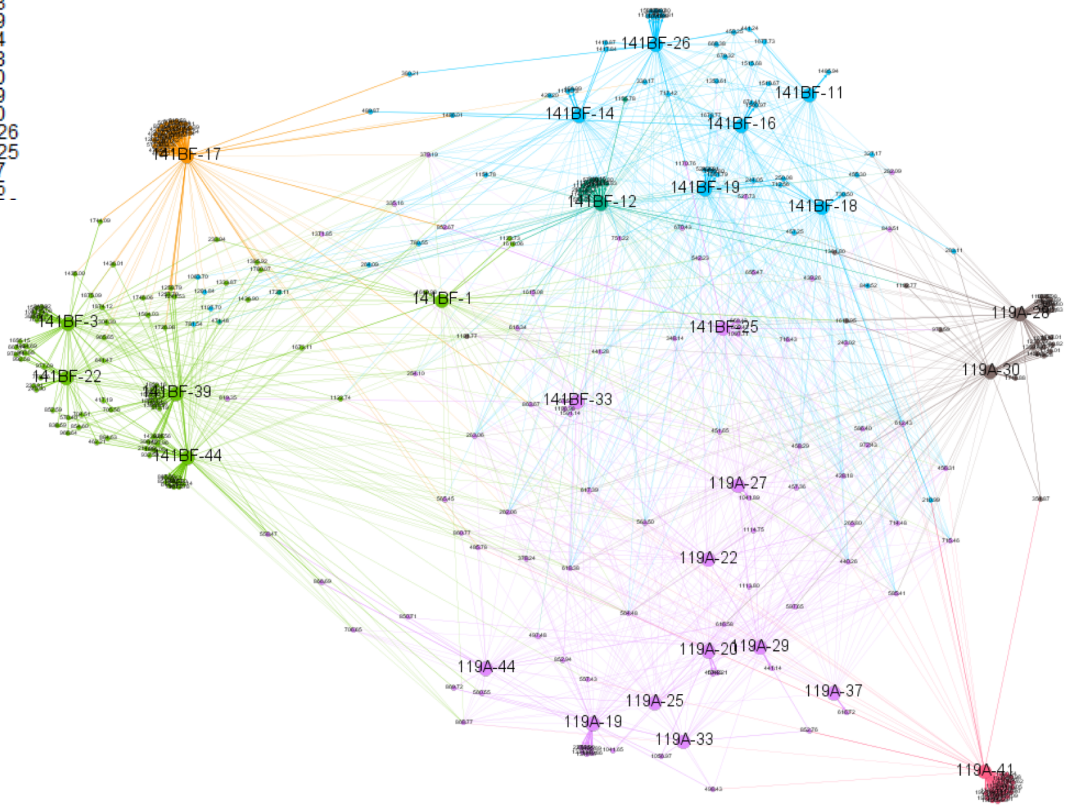

442.47
816.97
263.2
528.08
440.49
617.73
250.14
451.91
441.42
262.21
458.52
815.96


116H-16
116F-9
117H-54
129A-8
129A-2
129A-3
129A-22
129A-7
129A-5
129A-13
129A-26
129A-24
129A-12
129A-28
129A-19
129A-20
129A-6
129A-21
129A-4
129A-32
129A-14
129A-27
129A-25
129A-23
129A-1
129A-31
129A-50
129A-11
129A-35
129A-10
129A-9

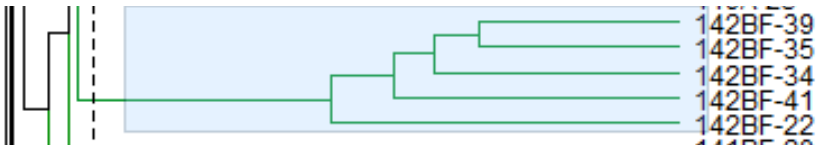

Group 40
129A-35
116H-16
116F-9
117H-54
129A-9
129A-10
129A-1
129A-27
129A-11
129A-20
129A-4
129A-25
129A-24
129A-31
129A-50
129A-7
129A-23
129A-3
129A-8
129A-2
129A-32
129A-14
129A-28
129A-12
129A-13
129A-6
129A-19
129A-22
129A-21
129A-5

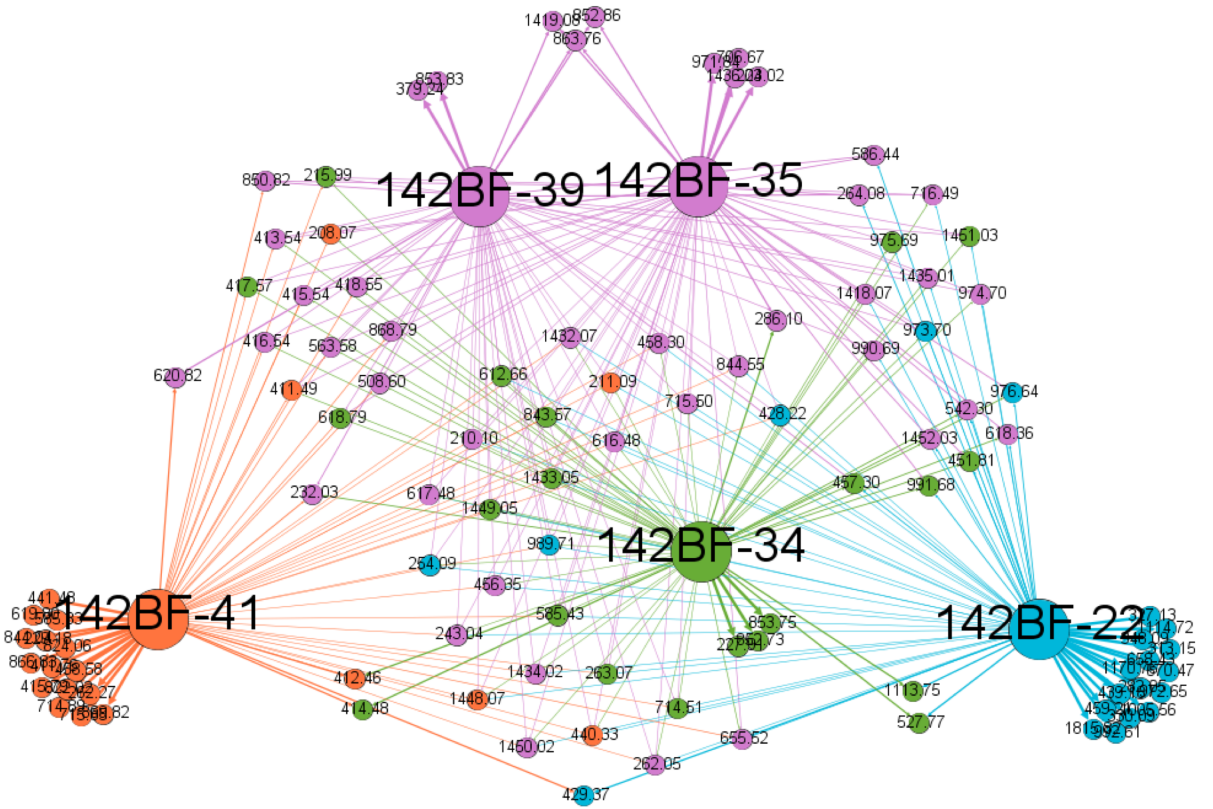
Group 41


141BF-19
141BF-18
141BF-12
141BF-33
141BF-1
141BF-14
141BF-11
141BF-16
119A-41
141BF-44
141BF-39
141BF-17
141BF-3
141BF-22
119A-27
119A-22
119A-24
119A-33
119A-29
119A-44
119A-28
119A-20
119A-19
119A-30
141BF-26
141BF-25
119A-37
119A-25

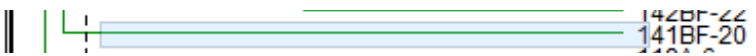

141BF-26
141BF-11
141BF-14
141BF-16
141BF-17
141BF-19
141BF-12
141BF-18
141BF-1
141BF-3
141BF-25
119A-28
141BF-22
141BF-39
119A-30
141BF-33
141BF-44
119A-27
119A-22
119A-20
119A-29
119A-44
119A-37
119A-25
119A-19
119A-33
119A-41

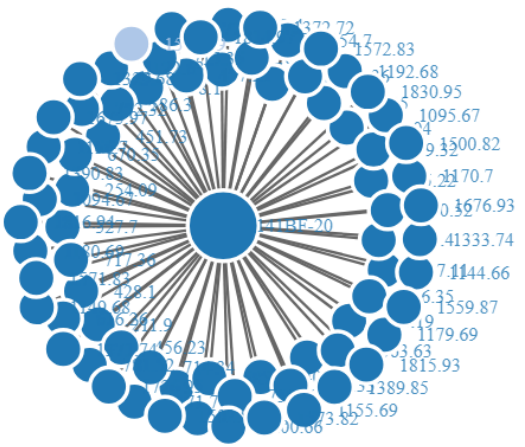
Group 42


142BF-39
142BF-35
142BF-34
142BF-41
142BF-22

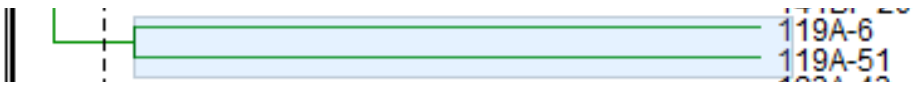

142BF-39
142BF-35
142BF-34
142BF-41
142BF-22

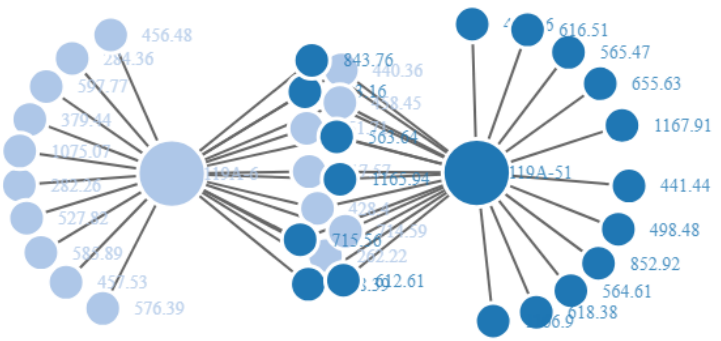
Group 43


141BF-20

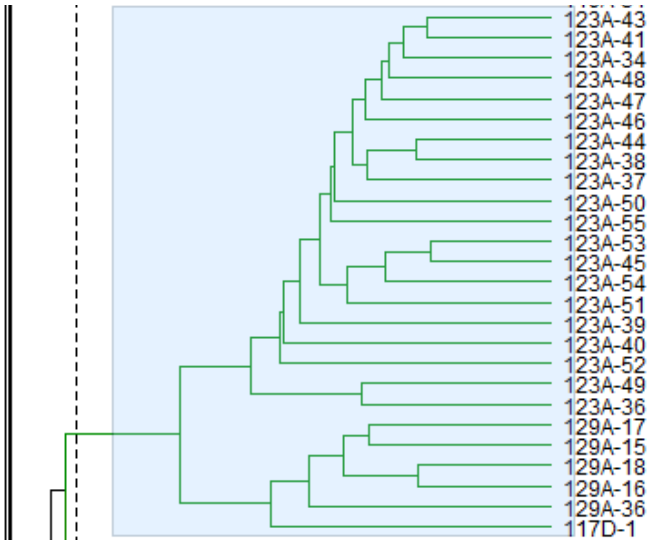

1572.83
1830.95
1095.67
1500.82
1170.7
1676.93
41333.74
1559.87
1179.69
1815.93
1389.85
1155.69

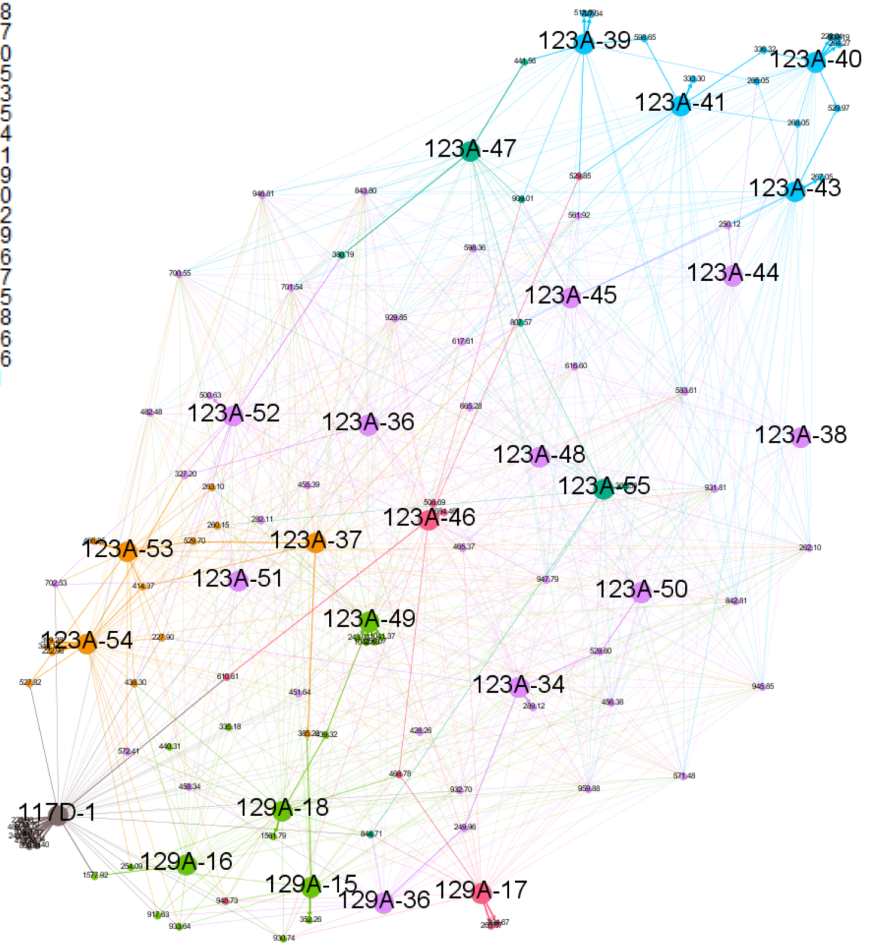
Group 44


119A-6
119A-51

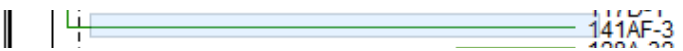

456.48
379.44
1075.07
576.39
843.76
715.56
612.61
616.51
565.47
655.63
1167.91
441.44
498.48
852.92
564.61
518.38

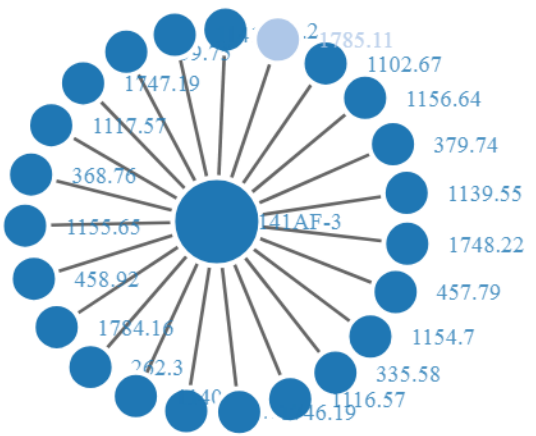
Group 45


123A-43
123A-41
123A-34
123A-48
123A-47
123A-46
123A-44
123A-38
123A-37
123A-50
123A-55
123A-53
123A-45
123A-54
123A-51
123A-39
123A-40
123A-52
123A-49
123A-36
129A-17
129A-15
129A-18
129A-16
129A-36
117D-1

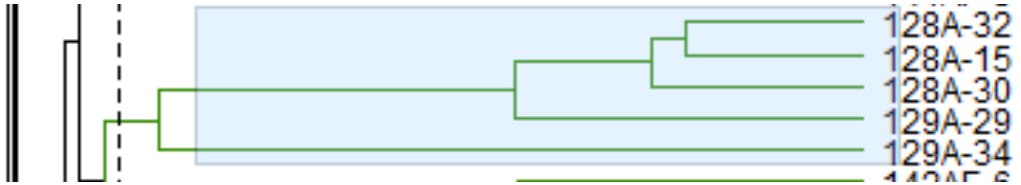

123A-39
123A-40
123A-41
123A-47
123A-43
123A-44
123A-45
123A-52
123A-36
123A-38
123A-48
123A-55
123A-46
123A-53
123A-37
123A-51
123A-50
123A-49
123A-54
123A-34
117D-1
129A-18
129A-16
129A-15
129A-36
129A-17

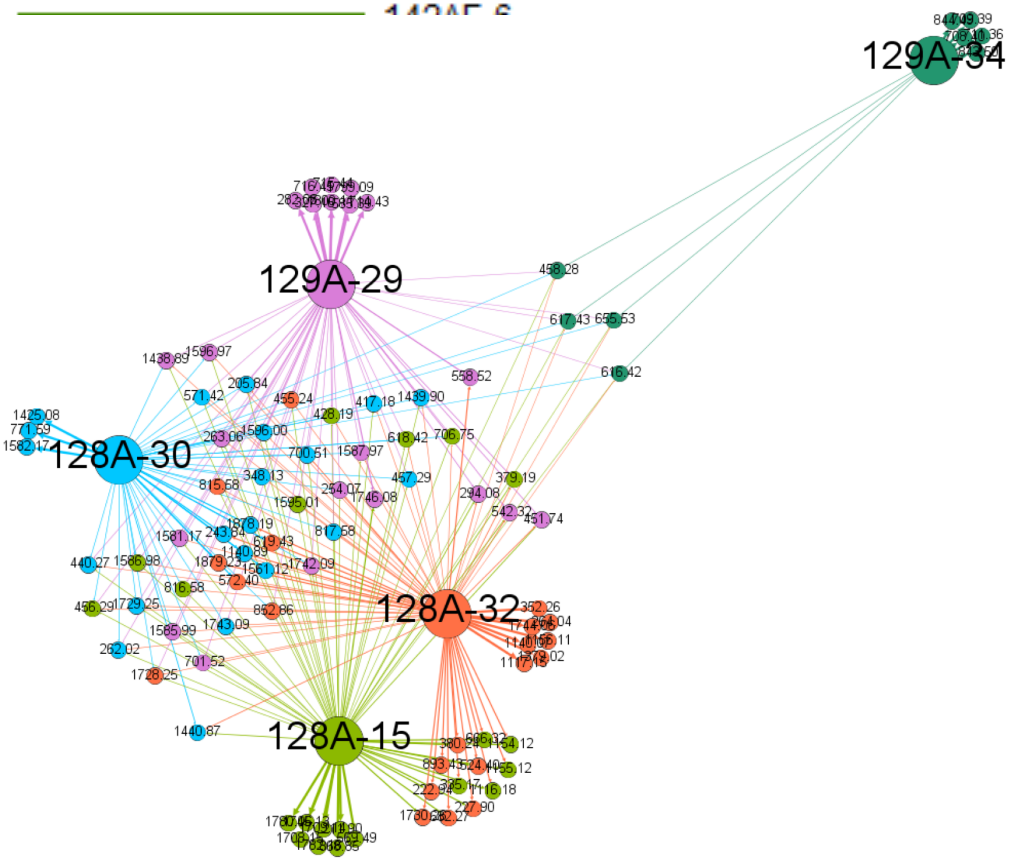
Group 46


141AF-3


1785.11
1102.67
1156.64
379.74
1139.55
1748.22
457.79
1154.7
335.58
1116.57
1747.19
1155.65
458.92
141AF-3

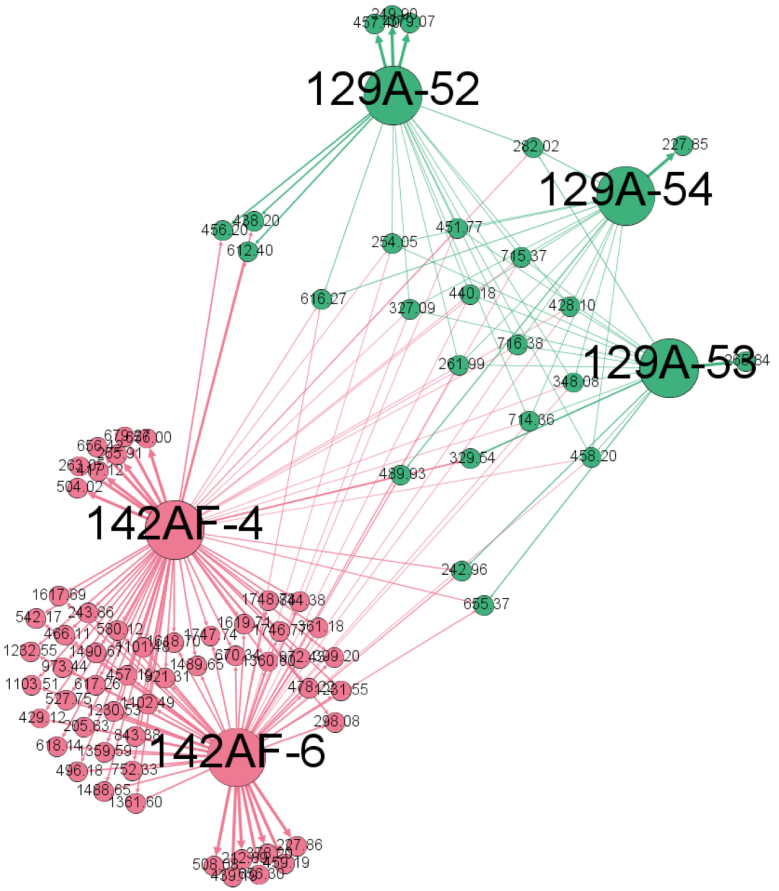
Group 47


128A-32
128A-15
128A-30
129A-29
129A-34

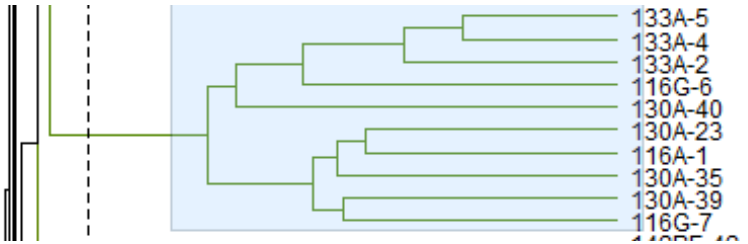

129A-34
129A-29
128A-30
128A-32
128A-15

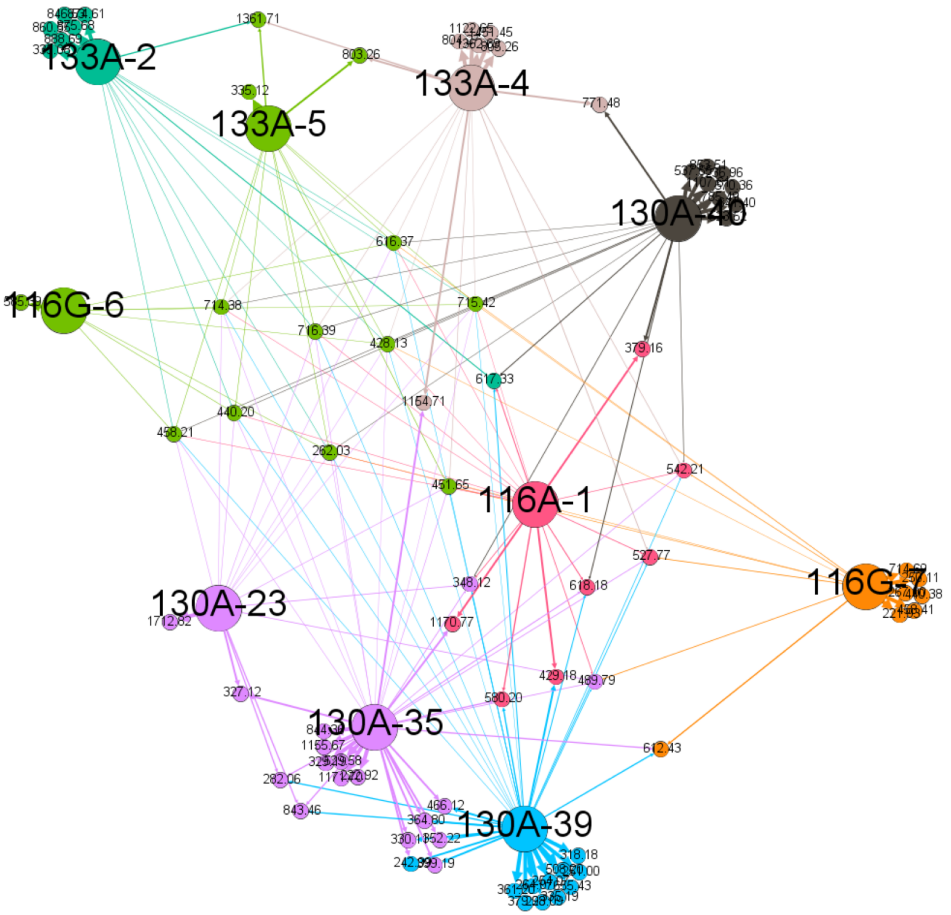
Group 48


142AF-6
142AF-4
129A-54
129A-53
129A-52

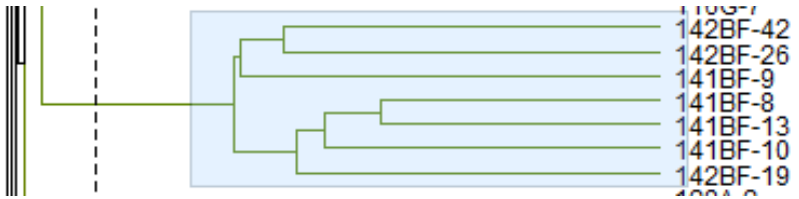

129A-52
129A-54
129A-53
142AF-4
142AF-6

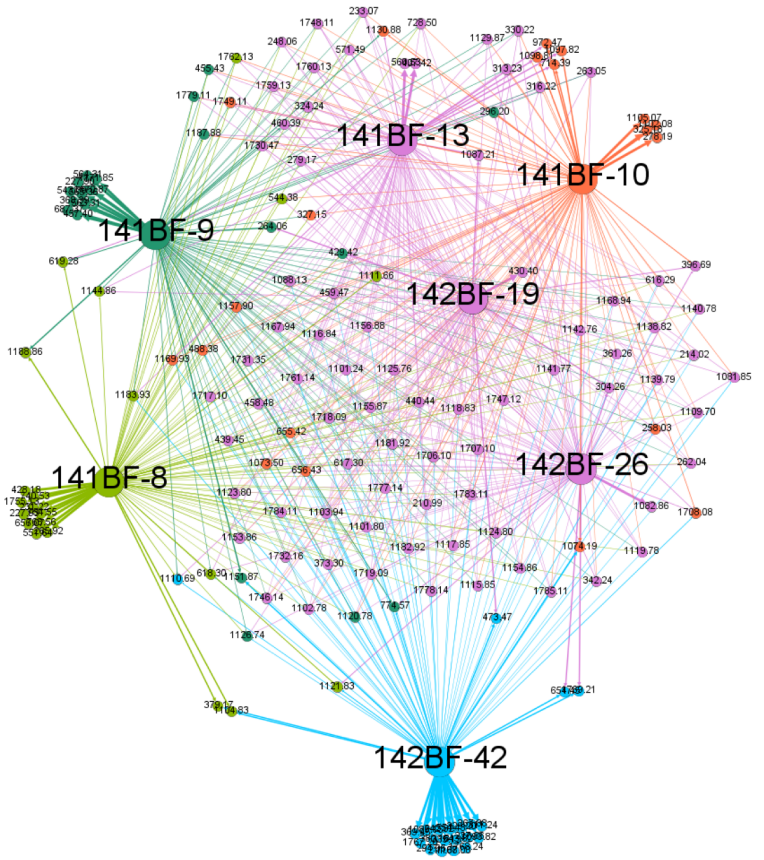
Group 49


133A-5
133A-4
133A-2
116G-6
130A-40
130A-23
116A-1
130A-35
130A-39
116G-7

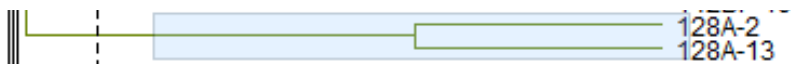

133A-2
133A-5
133A-4
130A-40
116G-6
116A-1
130A-23
116G-7
130A-35
130A-39

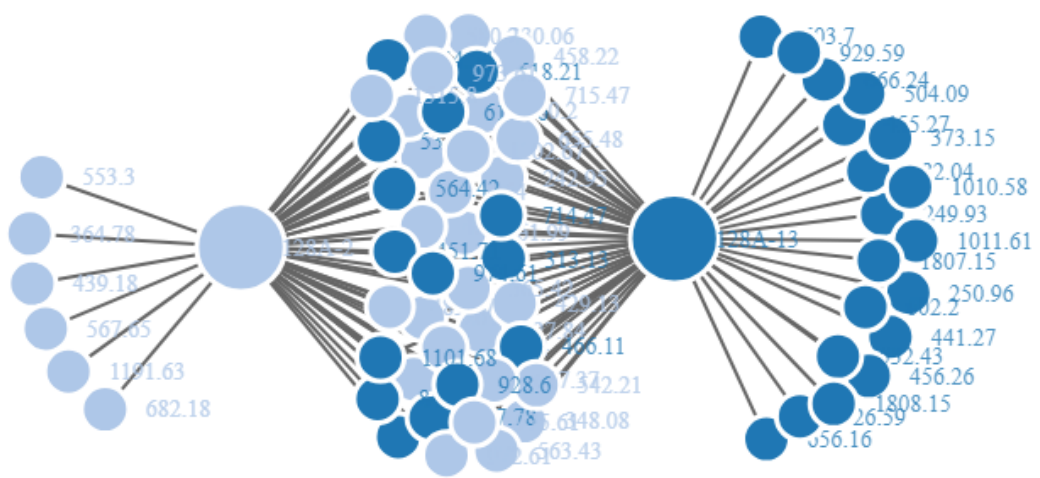
Group 50


142BF-42
142BF-26
141BF-9
141BF-8
141BF-13
141BF-10
142BF-19

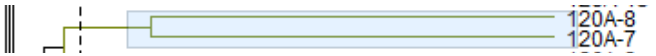

141BF-13
141BF-10
141BF-9
142BF-19
141BF-8
142BF-26
142BF-42

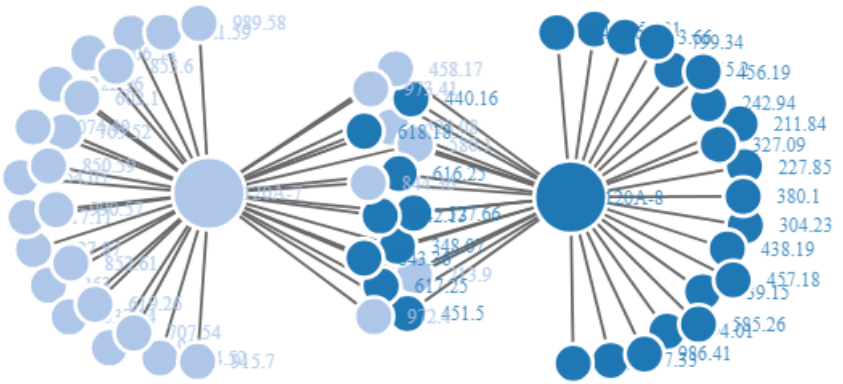
Group 51


128A-2
128A-13

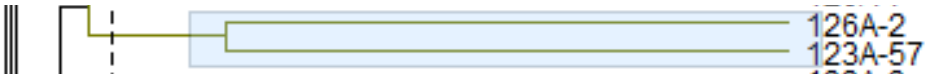

553.3
364.78
439.18
567.65
1101.63
682.18
458.22
715.47
564.42
1101.68
928.6
348.08
563.43
929.59
504.09
373.15
1010.58
1011.61
1807.15
250.96
441.27
456.26
1808.15
656.16

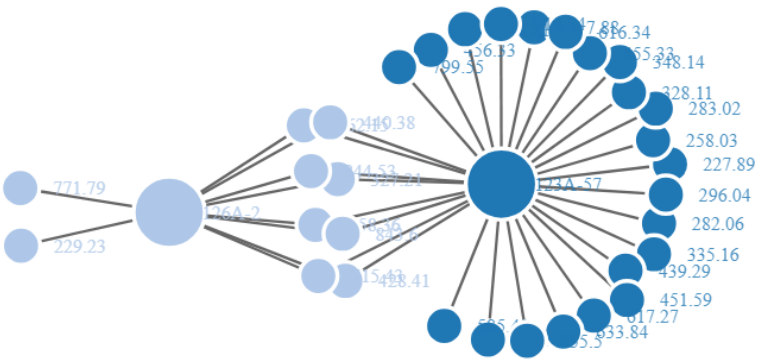
Group 52


120A-8
120A-7

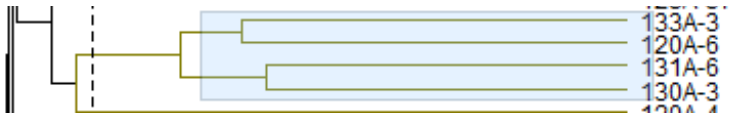

989.58
458.17
440.16
618.18
616.25
451.5
707.54
456.19
242.94
211.84
327.09
227.85
380.1
304.23
438.19
457.18
585.26
986.41

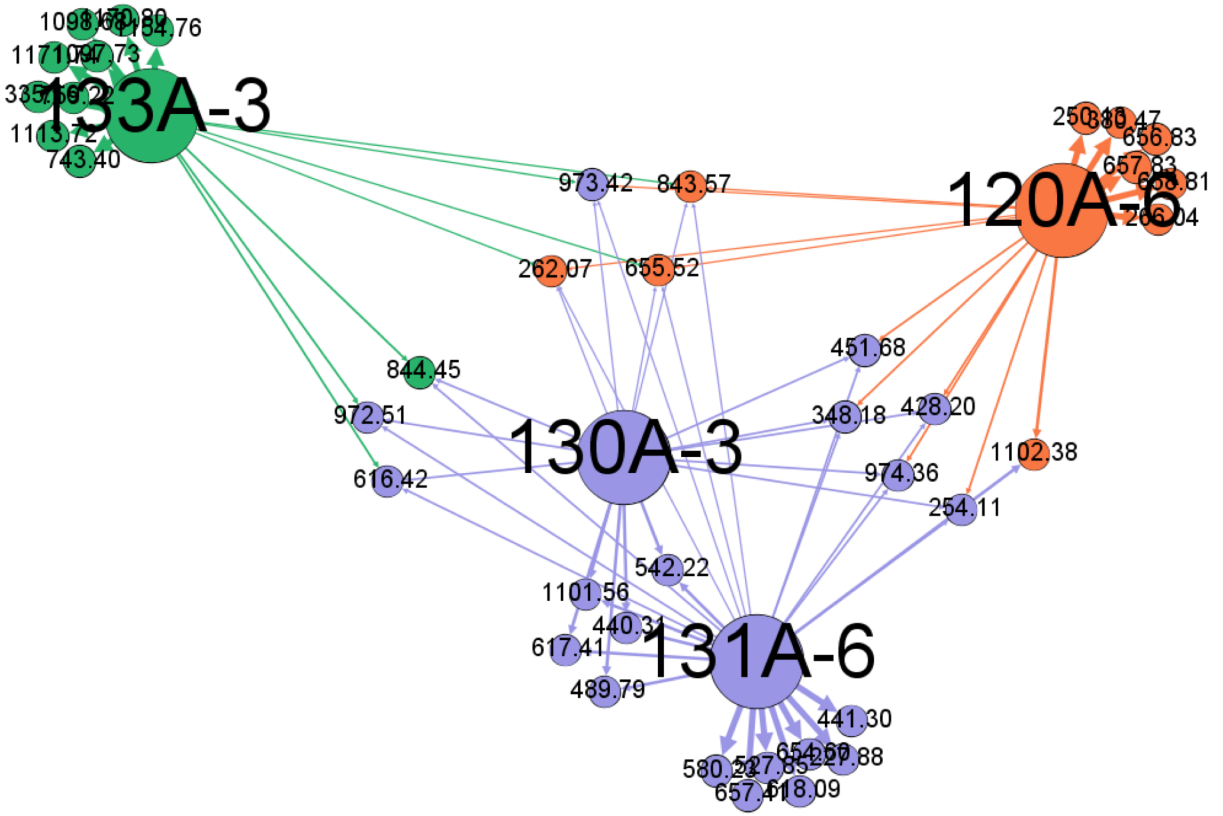
Group 53


126A-2
123A-57


771.79
229.23
348.14
328.11
283.02
258.03
227.89
296.04
282.06
335.16
439.29
451.59

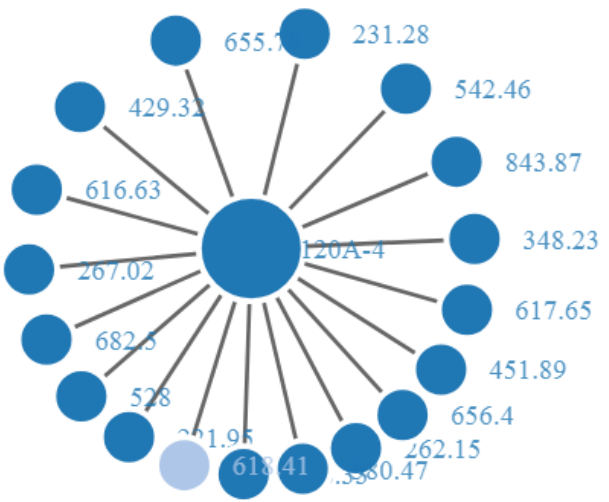
Group 54


133A-3
120A-6
131A-6
130A-3


133A-3
1113.72
743.40
973.42
843.57
120A-6
656.83
657.83
262.07
655.52
844.45
972.51
616.42
130A-3
451.68
348.18
428.20
1102.38
974.36
254.11
542.22
1101.56
440.31
617.41
489.79
131A-6
441.30
657.41
618.09

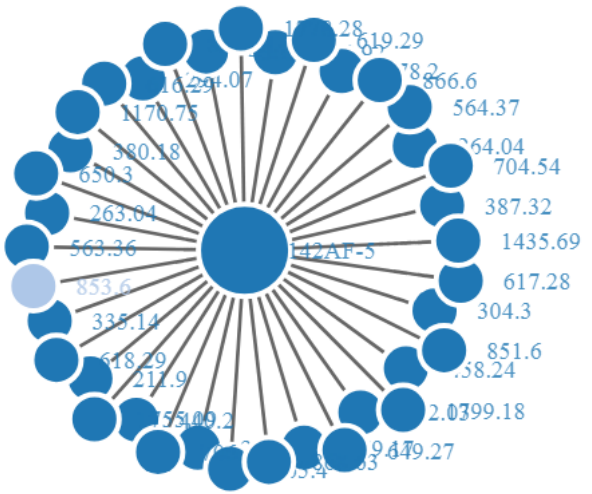
Group 55


120A-4

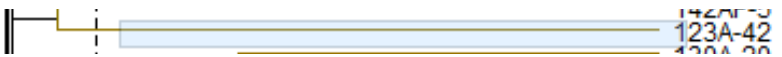

231.28
542.46
843.87
348.23
617.65
451.89
656.4
262.15
429.32
616.63
267.02
120A-4

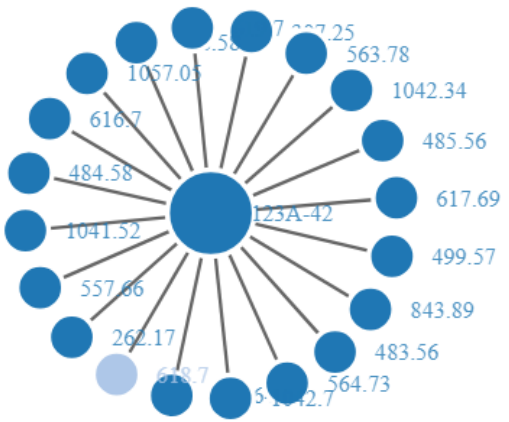
Group 56


120A-4
142AF-5


619.29
866.6
564.37
704.54
387.32
1435.69
617.28
304.3
851.6
1399.18
649.27
1170.75
380.18
650.3
263.64
563.36
853.6
335.14
618.29
212.9

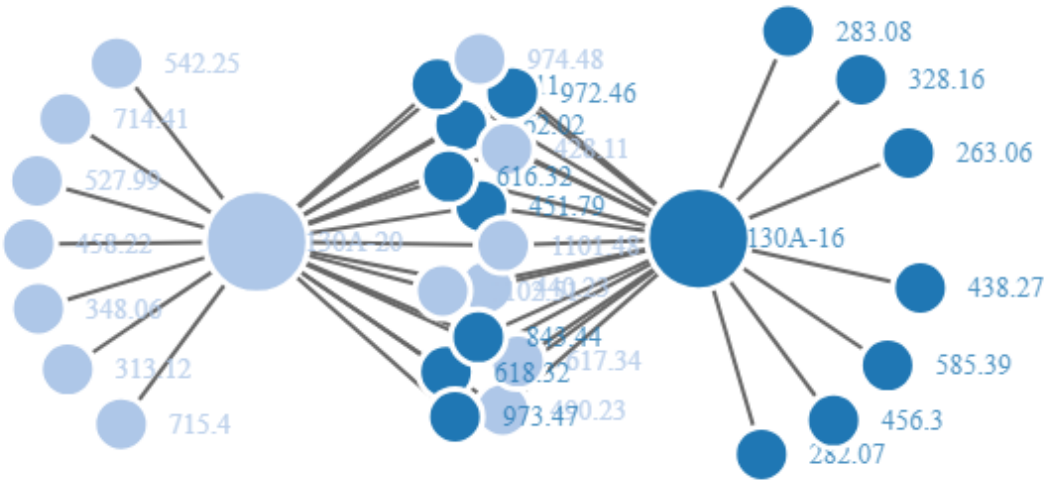
Group 57


123A-42


563.78
1042.34
485.56
617.69
499.57
843.89
483.56
564.73
1041.52
557.66
262.17
484.58
616.7
1057.05
123A-42

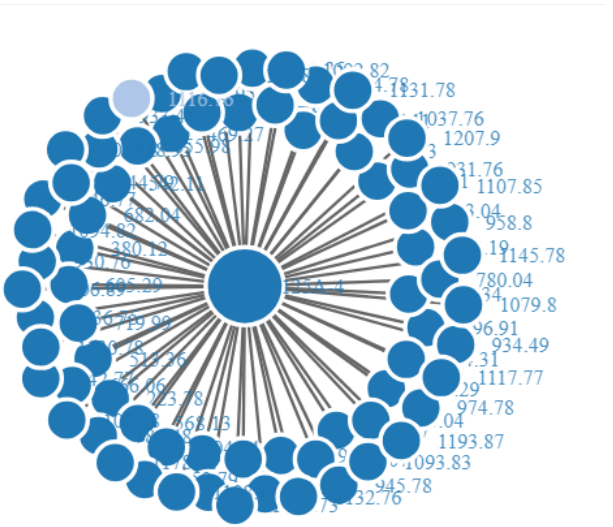
Group 58


130A-20
130A-16


542.25
714.41
527.99
458.22
348.06
313.12
715.4
974.48
972.46
616.32
451.79
618.32
973.47
283.08
328.16
263.06
130A-16
438.27
585.39
456.3
282.07

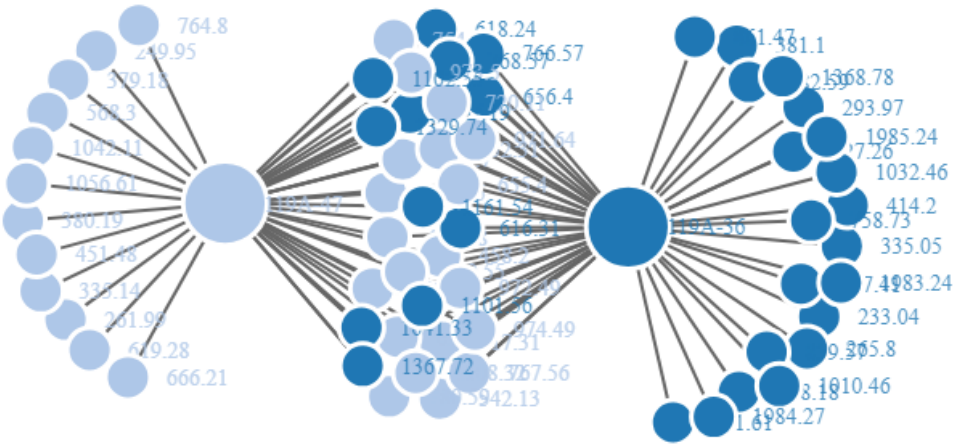
Group 59


125A-4


1131.78
1037.76
1207.9
1107.85
958.8
1145.78
780.04
1079.8
934.49
1117.77
974.78
1193.87
1093.83

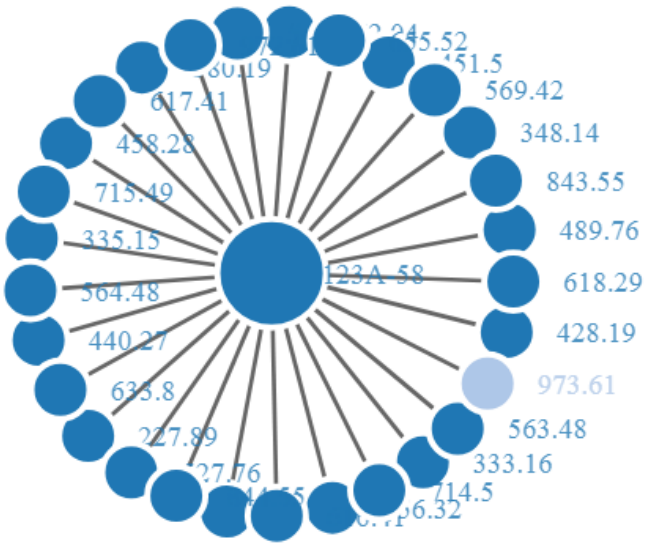
Group 60

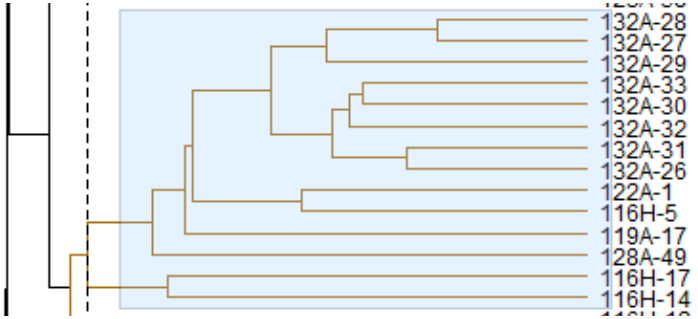

119A-47
119A-36


764.8
379.18
1042.11
1056.61
380.19
451.48
666.21
618.24
766.57
656.4
1329.74
1367.72
1368.78
293.97
1985.24
1032.46
414.2
335.05
1983.24
233.04
1010.46
1984.27

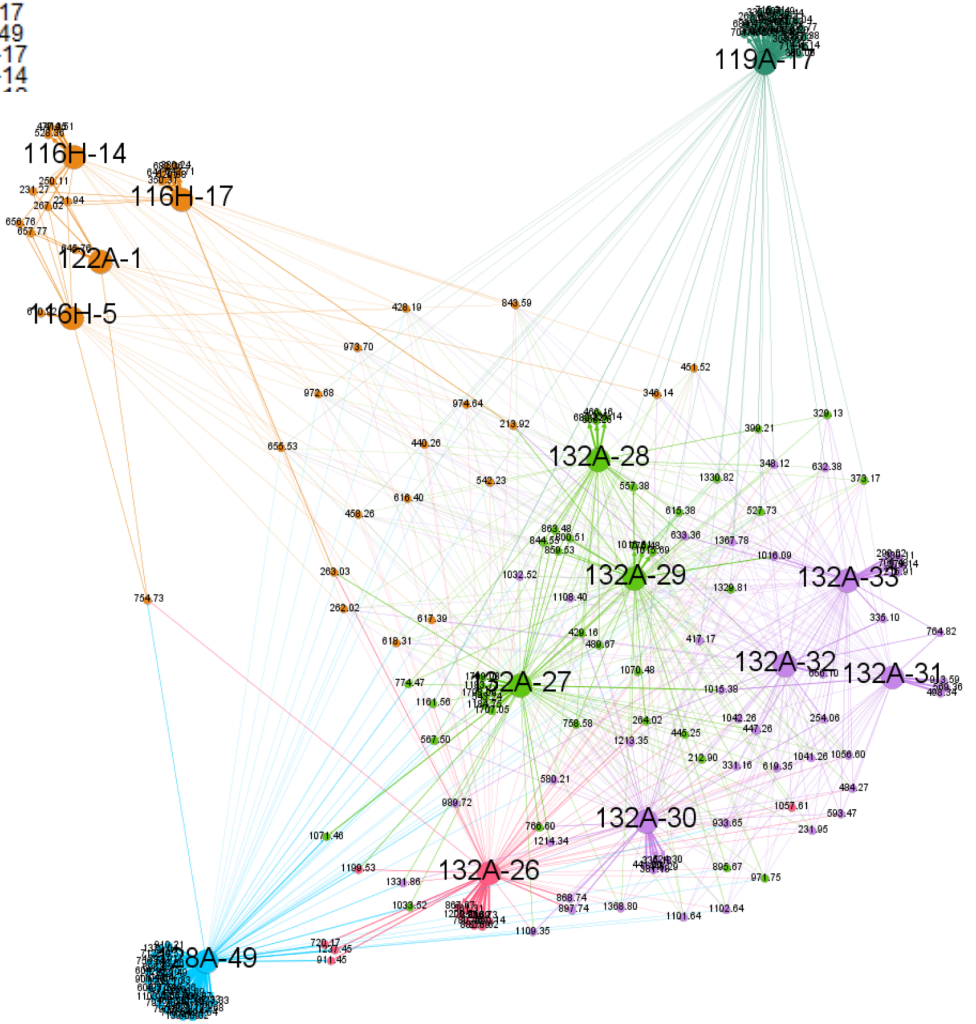
Group 61


123A-58


569.42
348.14
843.55
489.76
618.29
428.19
973.61
563.48
333.16
617.41
458.28
715.49
335.15
564.48
440.27

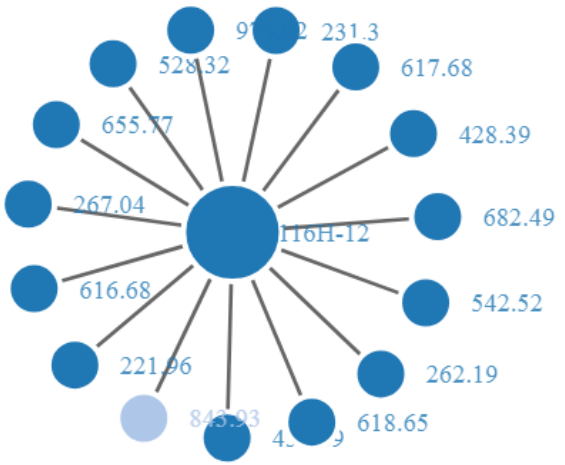

132A-28
132A-27
132A-29
132A-33
132A-30
132A-32
132A-31
132A-26
122A-1
116H-5
119A-17
128A-49
116H-17
116H-14

Group 62


119A-17
116H-14
116H-17
122A-1
116H-5
132A-28
132A-29
132A-33
132A-27
132A-32
132A-31
132A-30
132A-26
128A-49

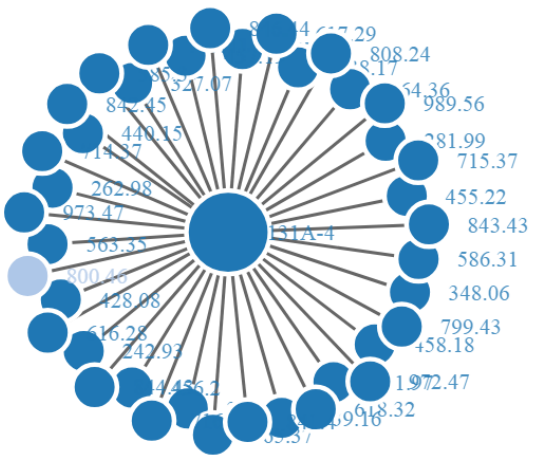
Group 63


116H-12


231.3
617.68
428.39
682.49
542.52
262.19
618.65
616.68
267.04
655.77

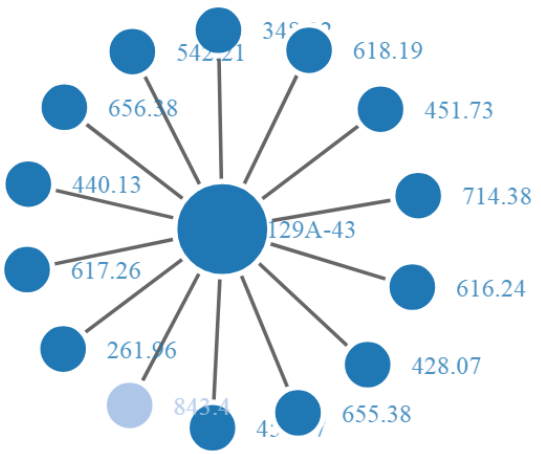
Group 64


131A-4


808.24
989.56
715.37
455.22
843.43
586.31
348.06
799.43
458.18
131A-4

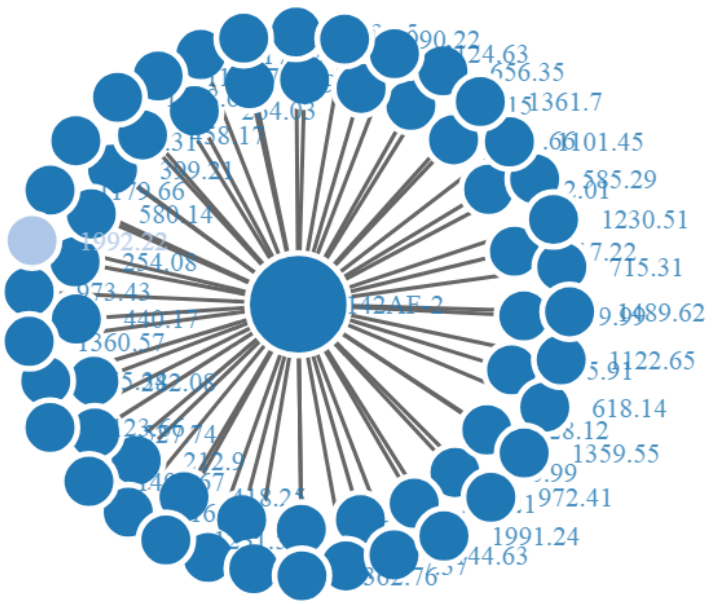
Group 65


129A-43

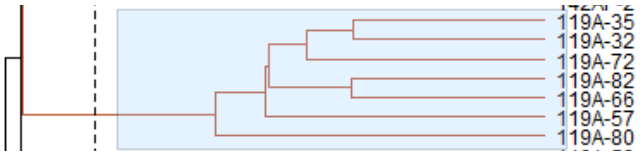

618.19
656.38
451.73
440.13
714.38
129A-43
617.26
616.24
428.07
655.38

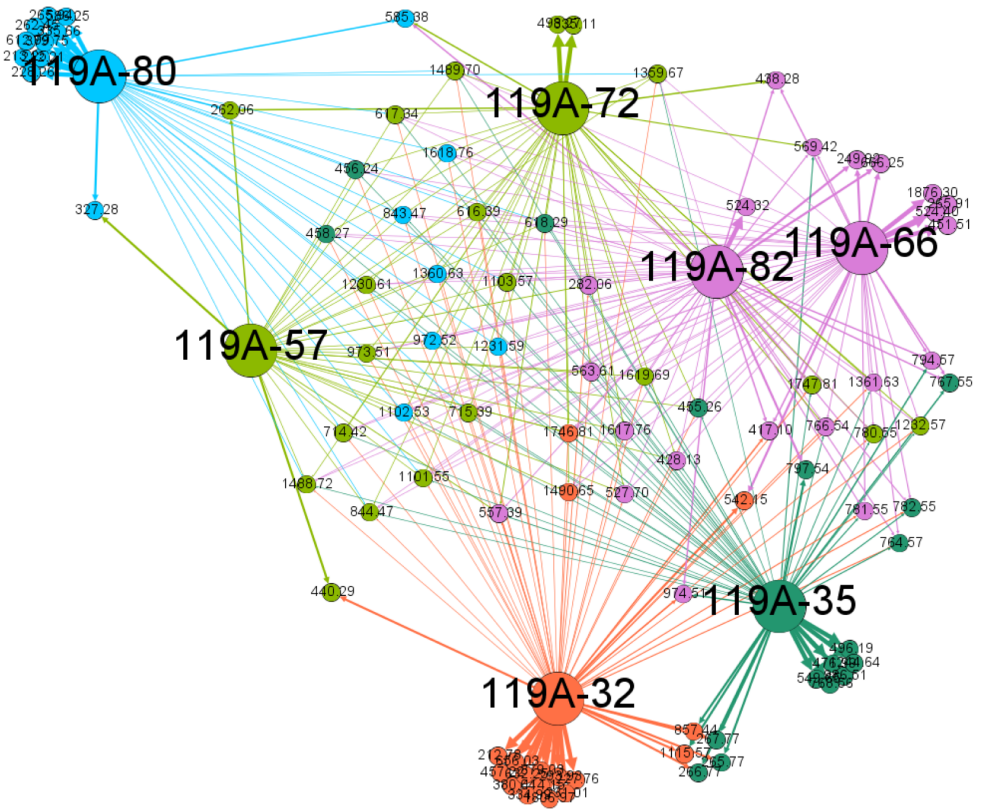
Group 66


142AF-2


142AF-2
1361.7
1230.51
715.31
1122.65
618.14
1359.55
1991.24

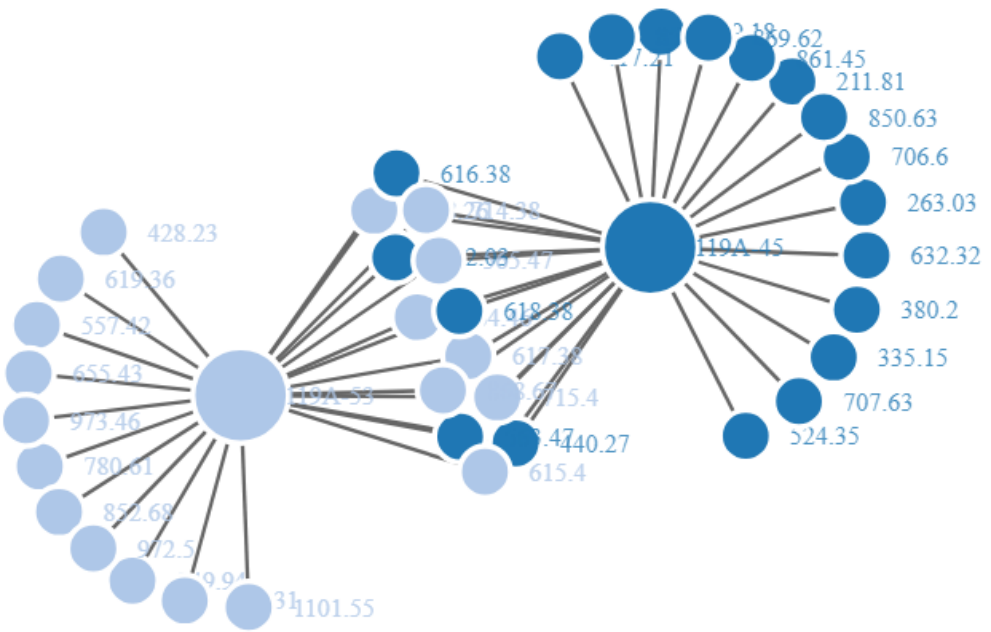
Group 67


119A-35
119A-32
119A-72
119A-82
119A-66
119A-57
119A-80


119A-80
119A-72
119A-66
119A-82
119A-57
119A-35
119A-32

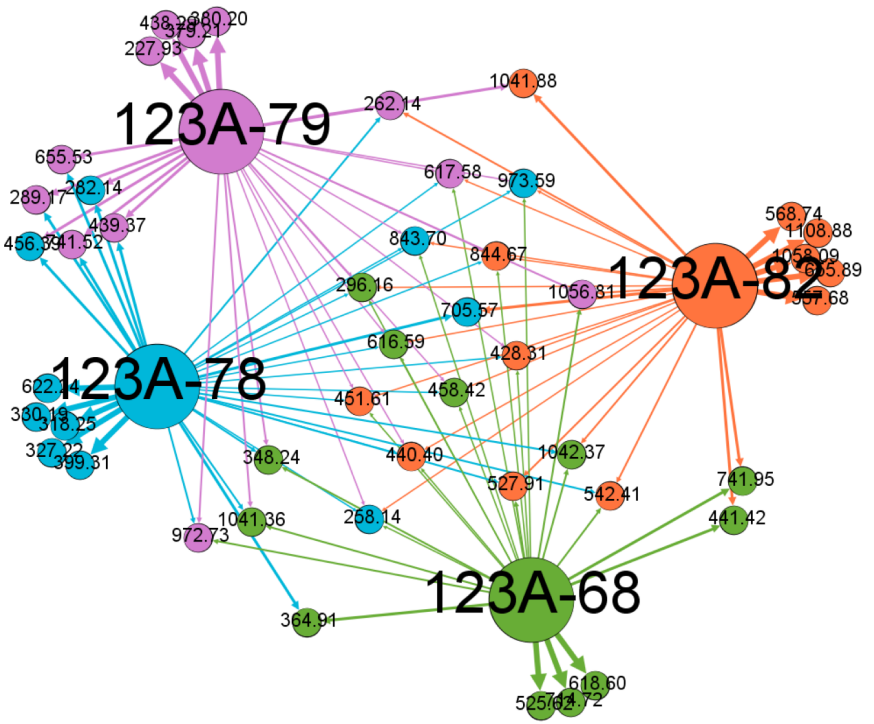
Group 68


119A-53
119A-45

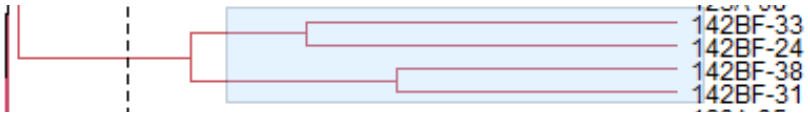

616.38
428.23
619.36
557.42
655.43
973.46
780.61
852.68
972.5
1101.55
861.45
211.81
850.63
706.6
263.03
632.32
380.2
335.15
707.63
440.27
615.4

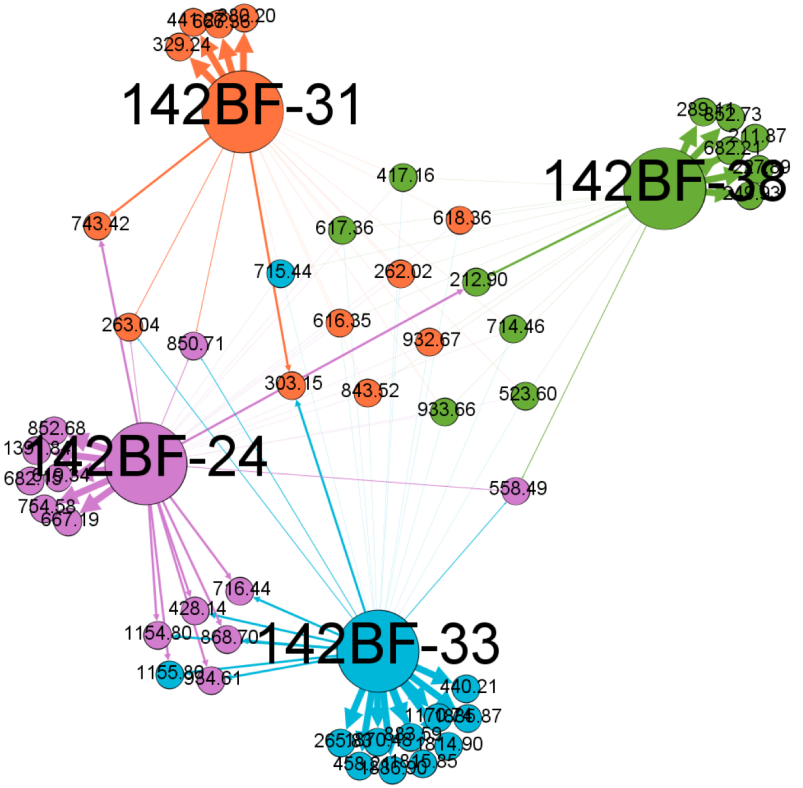
Group 69


123A-82
123A-79
123A-78
123A-68

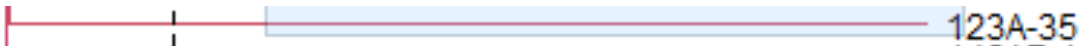

123A-79
123A-82
123A-78
123A-68
227.93
262.14
1041.88
655.53
289.17
617.58
973.59
568.74
1108.88
843.70
844.67
296.16
1056.81
705.57
616.59
428.31
622.14
451.61
458.42
348.24
440.40
1042.37
527.91
542.41
741.95
441.42
258.14
1041.36
972.73
364.91
618.60

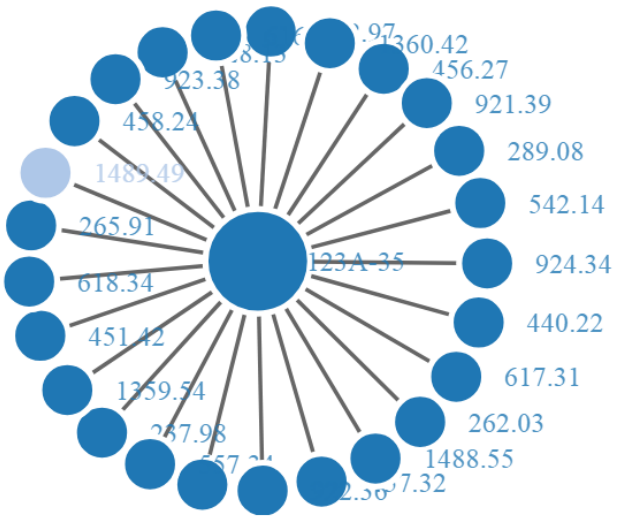
Group 70


142BF-33
142BF-24
142BF-38
142BF-31

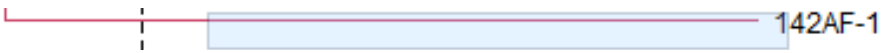

142BF-31
329.24
743.42
417.16
617.36
618.36
715.44
262.02
212.90
263.04
850.71
616.85
932.67
714.46
303.15
843.52
933.66
523.60
142BF-38
142BF-24
852.68
754.58
667.19
558.49
716.44
428.14
1154.80
868.70
142BF-33
440.21
714.90

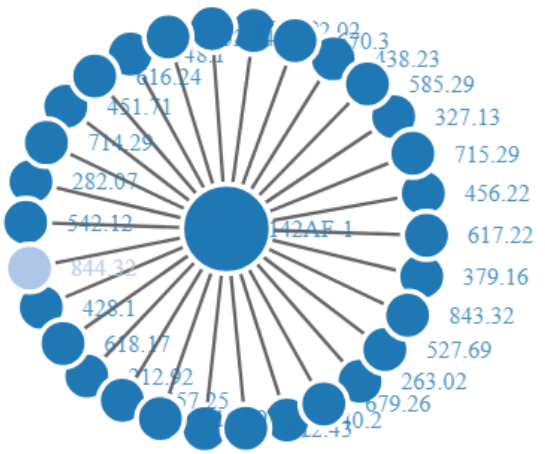
Group 71

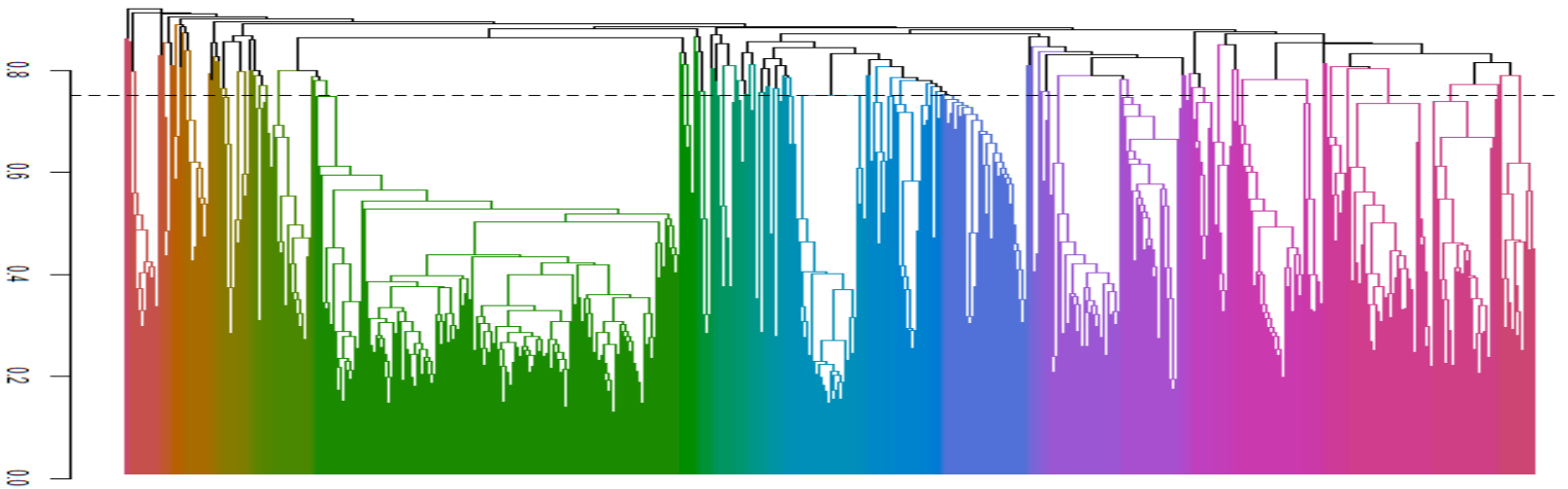

123A-35


360.42
456.27
921.39
289.08
542.14
924.34
440.22
617.31
262.03
1488.55
923.38
458.24
265.91
618.34
451.42
123A-35

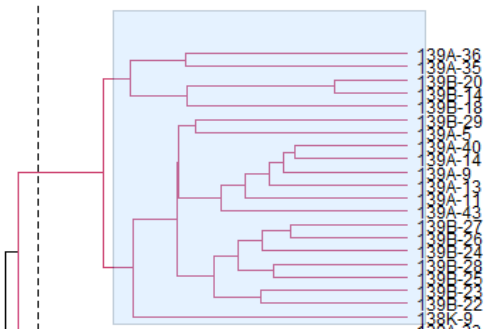
Group 72


142AF-1

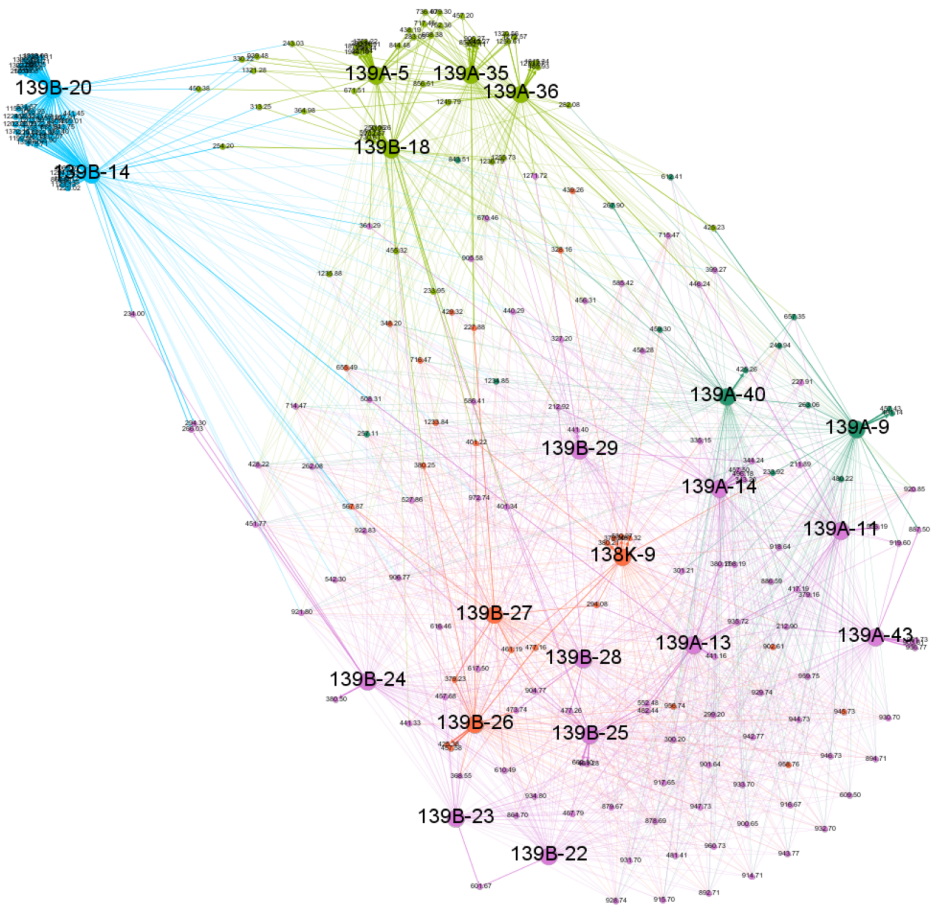

616.24
451.71
714.29
282.07
542.12
844.32
428.1
618.17
585.29
327.13
715.29
456.22
617.22
379.16
843.32
527.69
263.02
679.26



Appendix 3: Group1 to Group 101. Dendrogram resulting from hierarchical clustering of MALDI-TOF MS protein spectra of 883 marine isolates using IDBac. The dendrogram was “cut” at a threshold of 0.75 (dashed lines) to create groups of similar isolates. 101 distinct groups were created and the respective MANs are presented in the following figures.

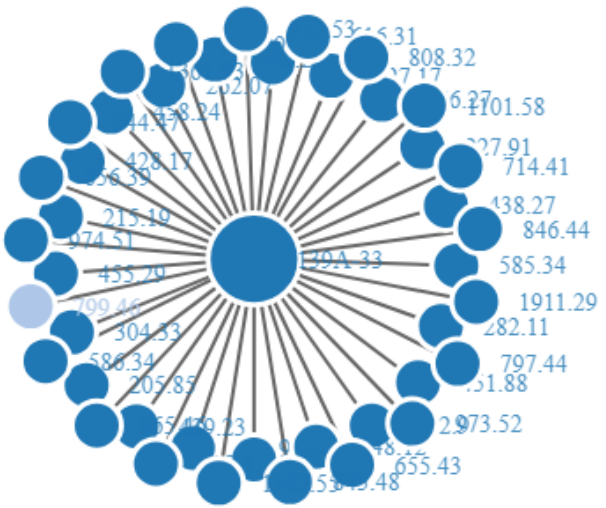

139A-36
139A-35
139B-20
139B-14
139B-18
139B-29
139A-5
139A-40
139A-14
139A-9
139A-11
139A-43
139A-13
139B-27
139B-28
139B-24
139B-26
139B-25
139B-23
139B-22
138K-9

Group 1

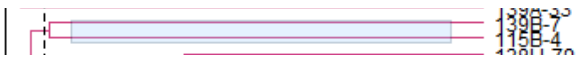

139B-20
139A-5
139A-35
139A-36
139B-18
139B-14
139A-40
139A-9
139B-29
139A-14
139A-11
138K-9
139B-27
139A-43
139A-13
139B-28
139B-24
139B-26
139B-25
139B-23
139B-22

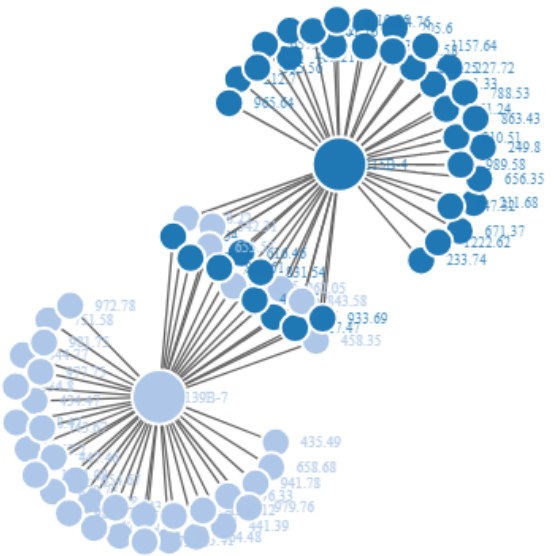
Group 2


138K-9
139A-33
139B-7

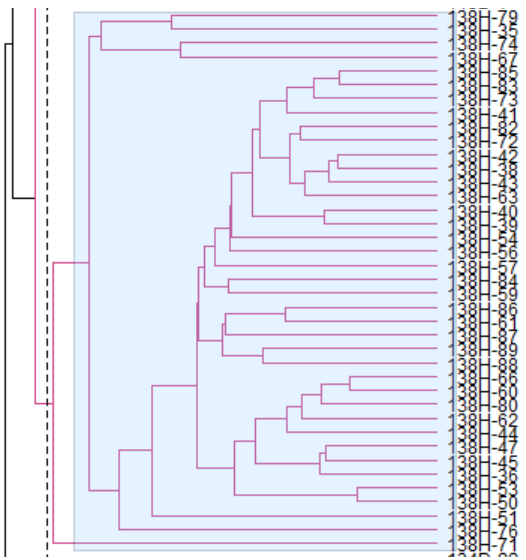

808.32
1101.58
714.41
438.27
846.44
585.34
1911.29
282.11
797.44
655.43

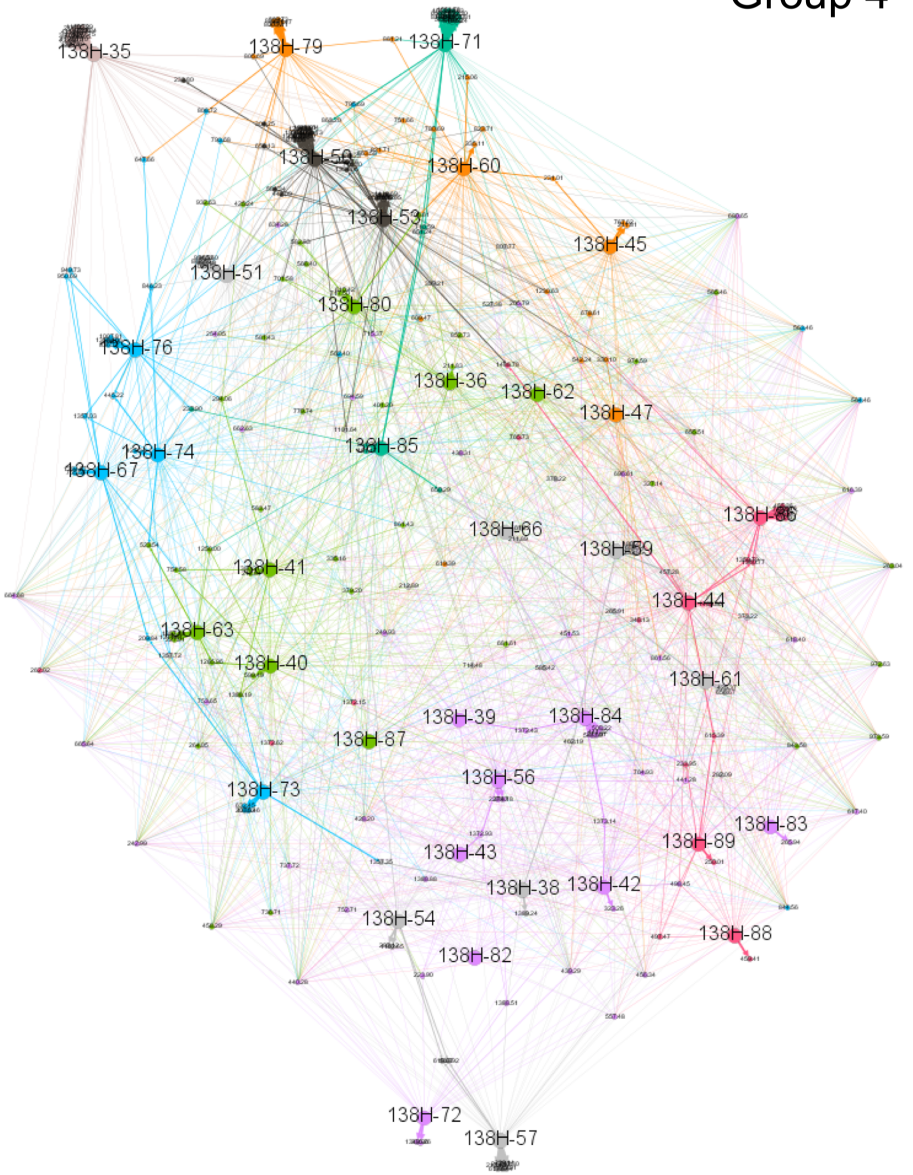
Group 3


139B-7
115B-4

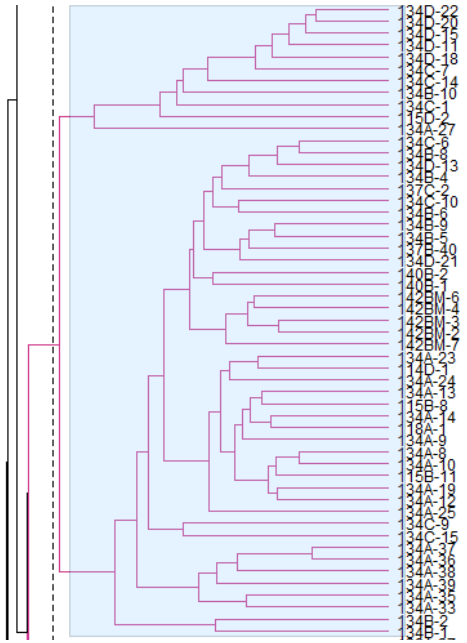

1157.64
788.53
863.43
249.8
989.58
656.35
671.37
233.74
972.78
933.69
458.35
435.49
658.68
941.78
441.39

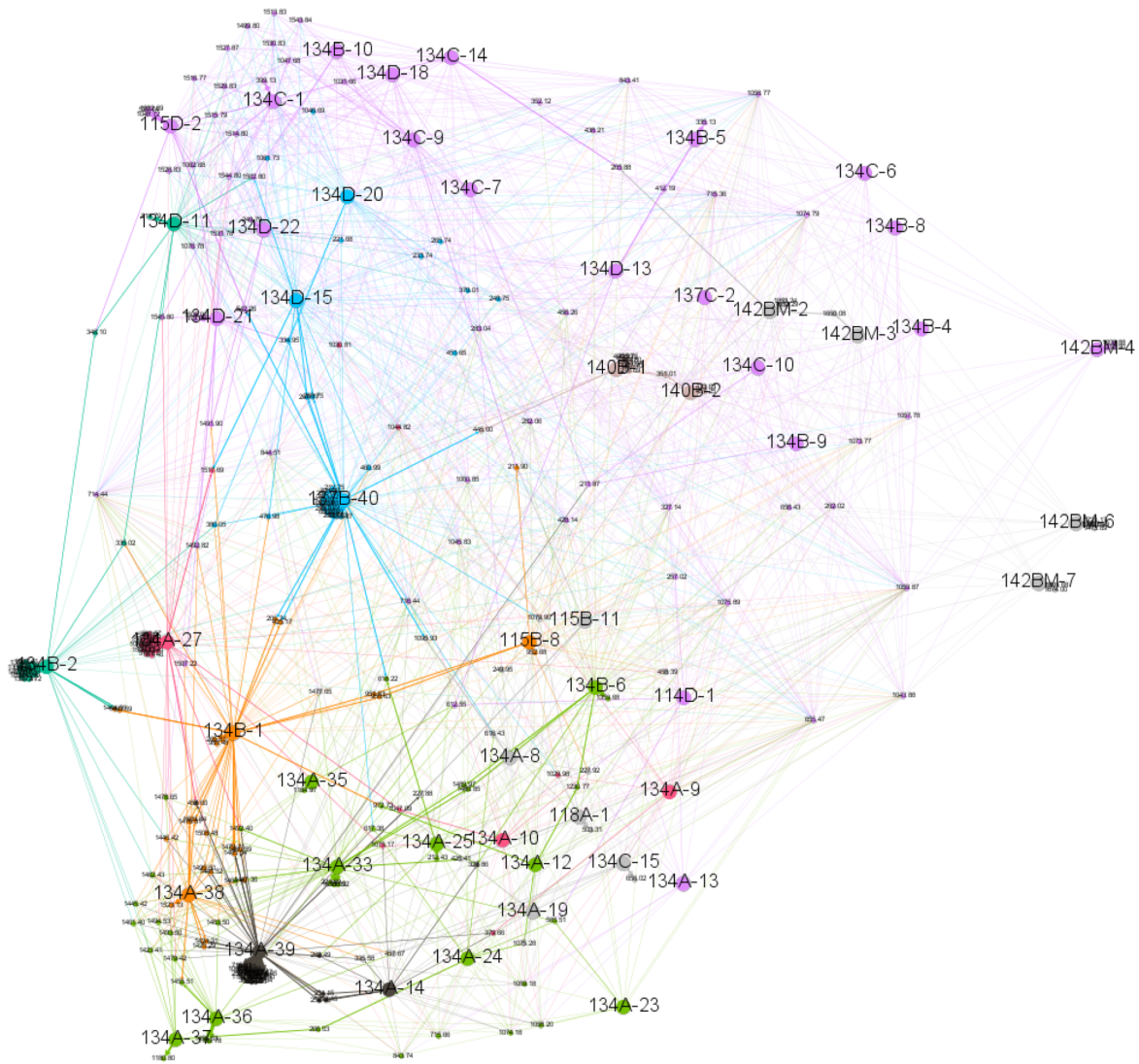
Group 4



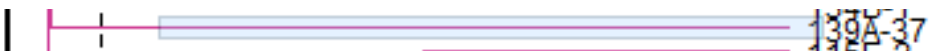

138H-35
138H-79
138H-71
138H-50
138H-60
138H-53
138H-45
138H-51
138H-80
138H-76
138H-36
138H-62
138H-47
138H-74
138H-85
138H-67
138H-66
138H-59
138H-41
138H-44
138H-63
138H-40
138H-61
138H-39
138H-84
138H-87
138H-56
138H-73
138H-83
138H-89
138H-43
138H-38
138H-42
138H-54
138H-88
138H-82
138H-72
138H-57

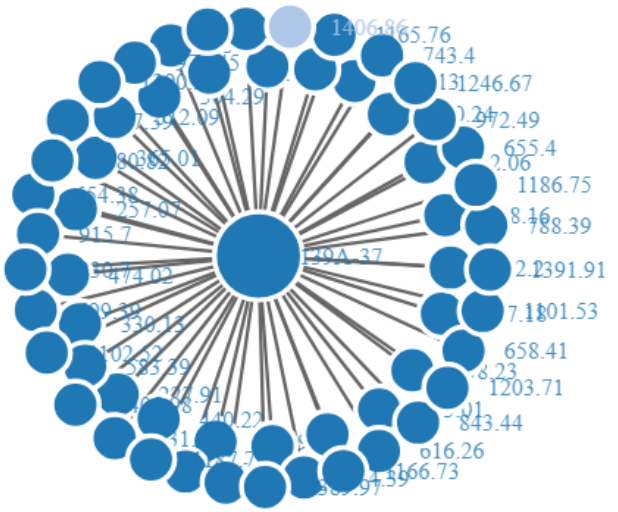
Group 5


134D-22
134D-20
134D-15
134D-11
134D-18
134C-7
134C-14
134B-10
134C-1
115D-2
134A-27
134C-6
134B-8
134D-13
134B-4
137C-2
134C-10
134B-6
134B-9
134B-5
137B-40
134D-21
140B-2
140B-1
142BM-6
142BM-4
142BM-3
142BM-2
142BM-7
134A-23
14D-1
134A-24
134A-13
115B-8
134A-14
118A-1
134A-9
134A-8
134A-10
115B-11
134A-19
134A-12
134A-25
134C-9
134C-15
134A-37
134A-36
134A-38
134A-39
134A-35
134A-33
134B-2
134B-1

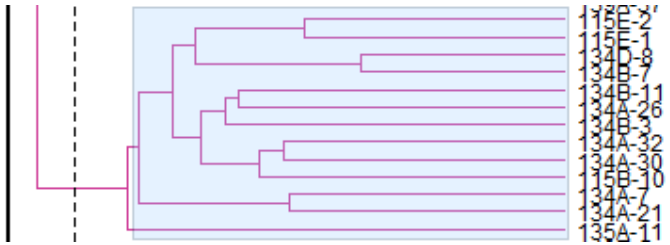

134B-10
134C-14
134D-18
134C-1
115D-2
134C-9
134B-5
134C-6
134D-20
134C-7
134D-11
134D-22
134B-8
134D-13
134D-15
137C-2
142BM-2
134D-21
142BM-3
134B-4
142BM-4
140B-1
134C-10
140B-2
134B-9
137B-40
142BM-6
142BM-7
115B-11
134A-27
115B-8
134B-2
134B-6
114D-1
134B-1
134A-8
134A-35
134A-9
118A-1
134A-25
134A-10
134A-33
134A-12
134C-15
134A-13
134A-38
134A-19
134A-39
134A-24
134A-14
134A-23
134A-36
134A-37

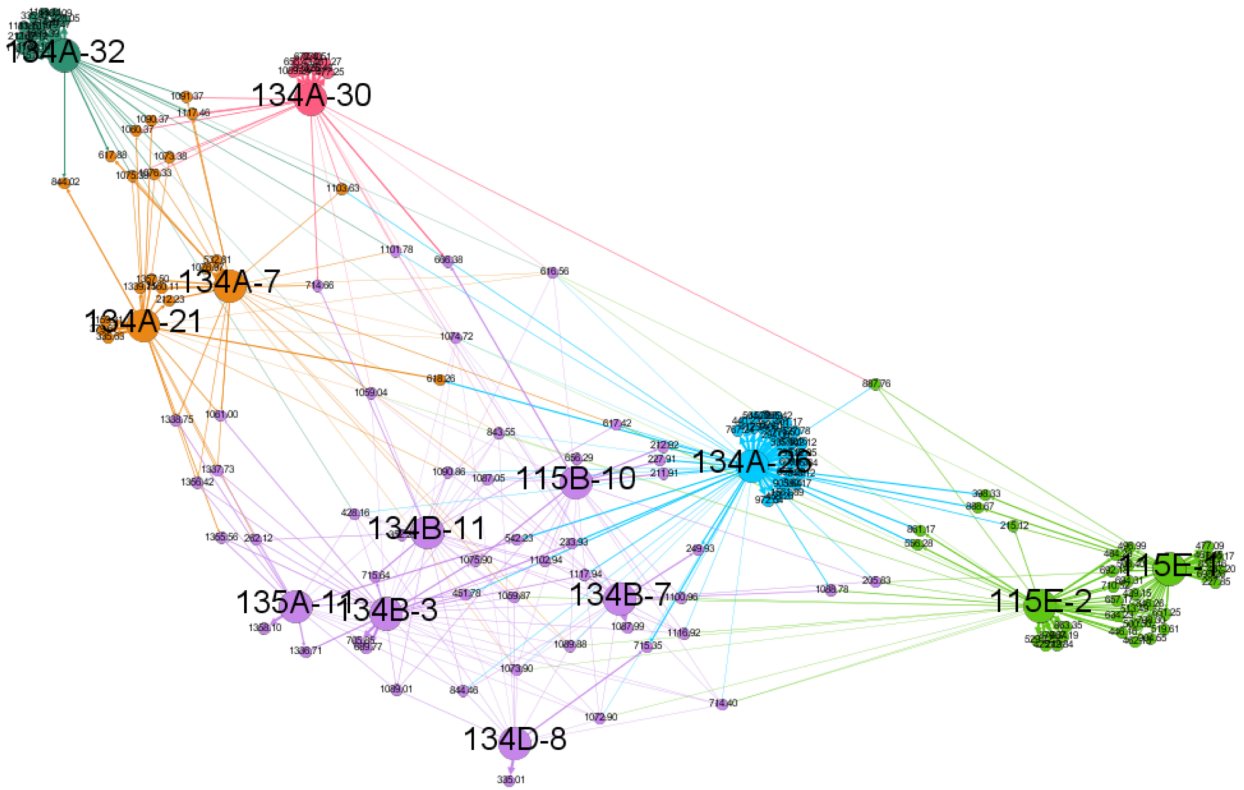
Group 6


139A-37

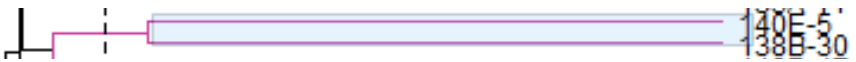

1406.86
65.76
743.4
1246.67
972.49
655.4
1186.75
788.39
2391.91
1101.53
658.41
1203.71
843.44
616.26
166.73
139A-37
915.7
257.07
330.13
583.39
440.22

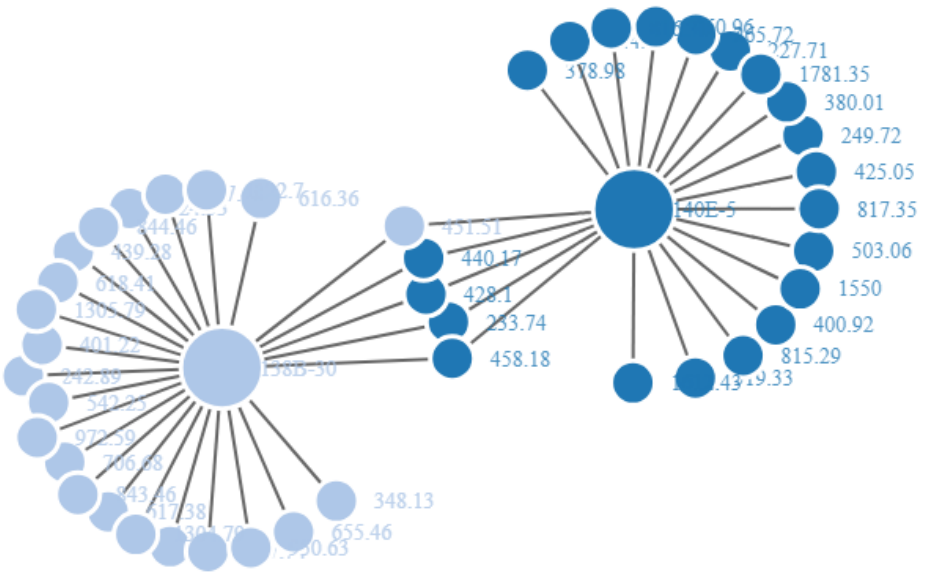
Group 7


115E-2
115E-1
134D-8
134B-7
134B-11
134A-26
134B-3
134A-32
134A-30
115B-10
134A-7
134A-21
135A-11


134A-32
134A-30
134A-7
134A-21
134A-26
115B-10
134B-11
135A-11
134B-3
134B-7
115E-2
115E-1
134D-8

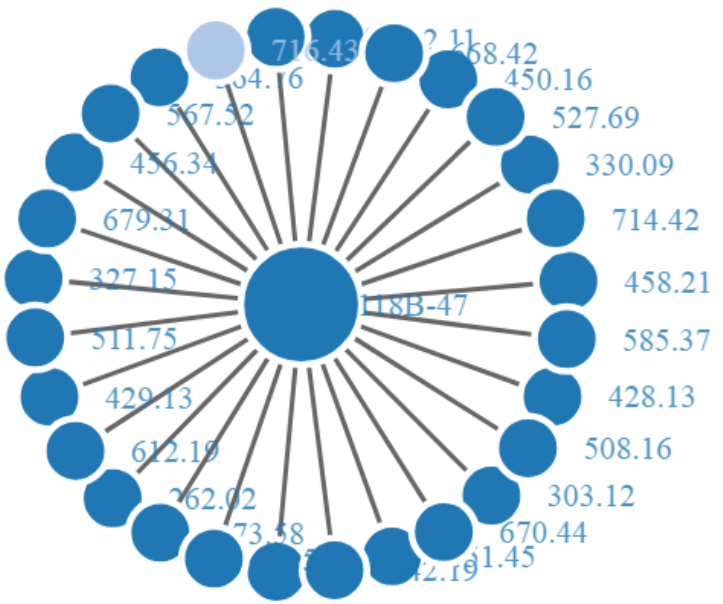
Group 8


140E-5
138B-30

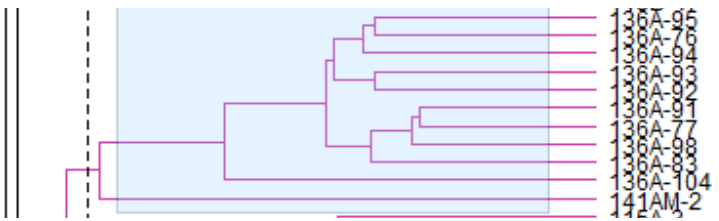

1781.35
380.01
249.72
425.05
817.35
503.06
1550
400.92
815.29
140E-5
451.51
440.17
428.1
253.74
458.18
616.36
348.13
655.46

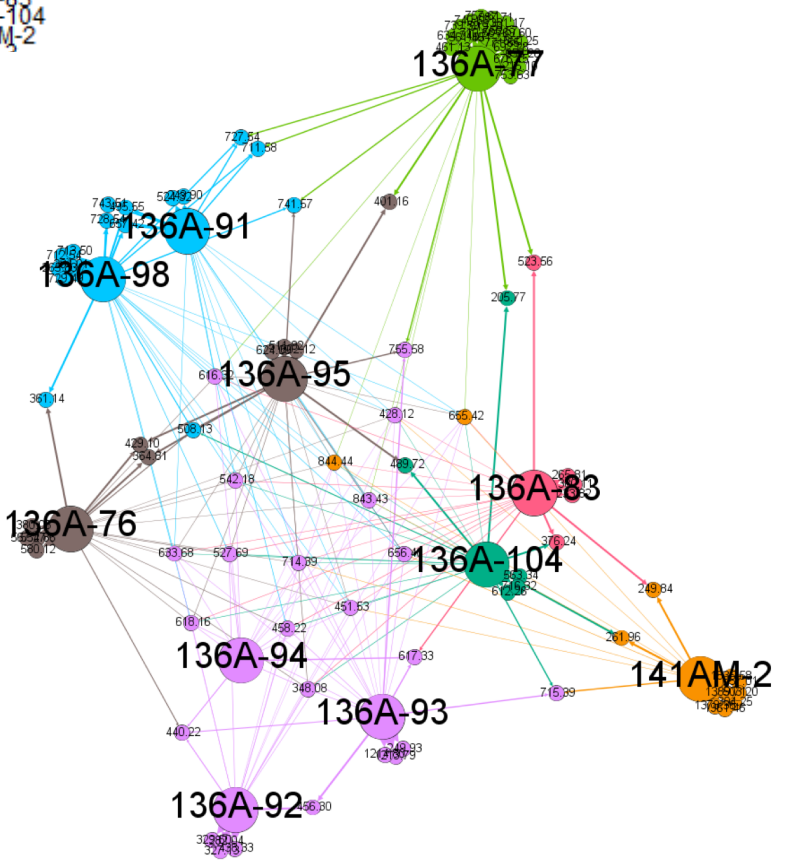
Group 9


118B-47

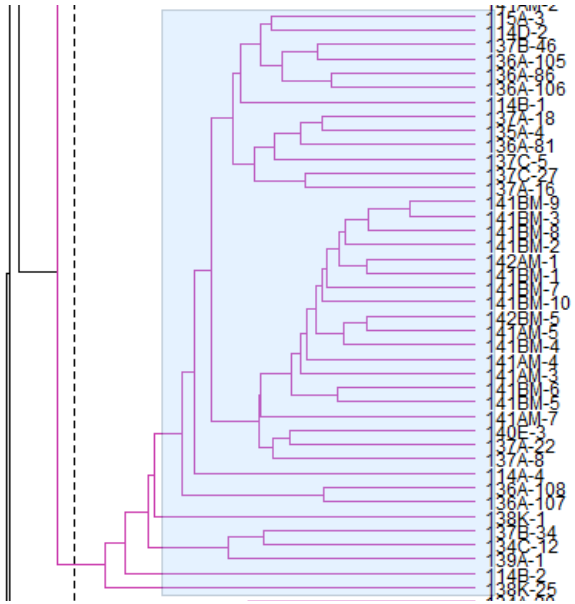

716.43
450.16
527.69
330.09
714.42
458.21
585.37
428.13
508.16
303.12
670.44
567.52
456.34
679.31
327.15
511.75
429.13
612.19
118B-47

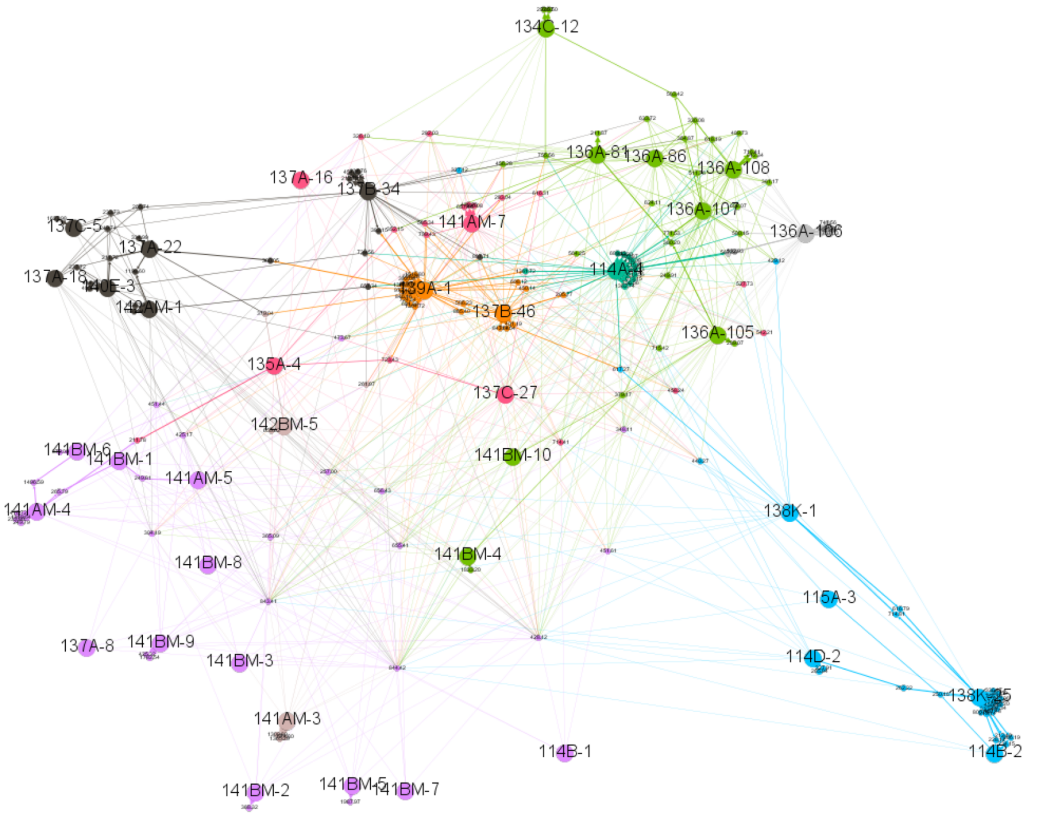
Group 10


136A-95
136A-76
136A-94
136A-93
136A-92
136A-91
136A-77
136A-98
136A-83
136A-104
141AM-2

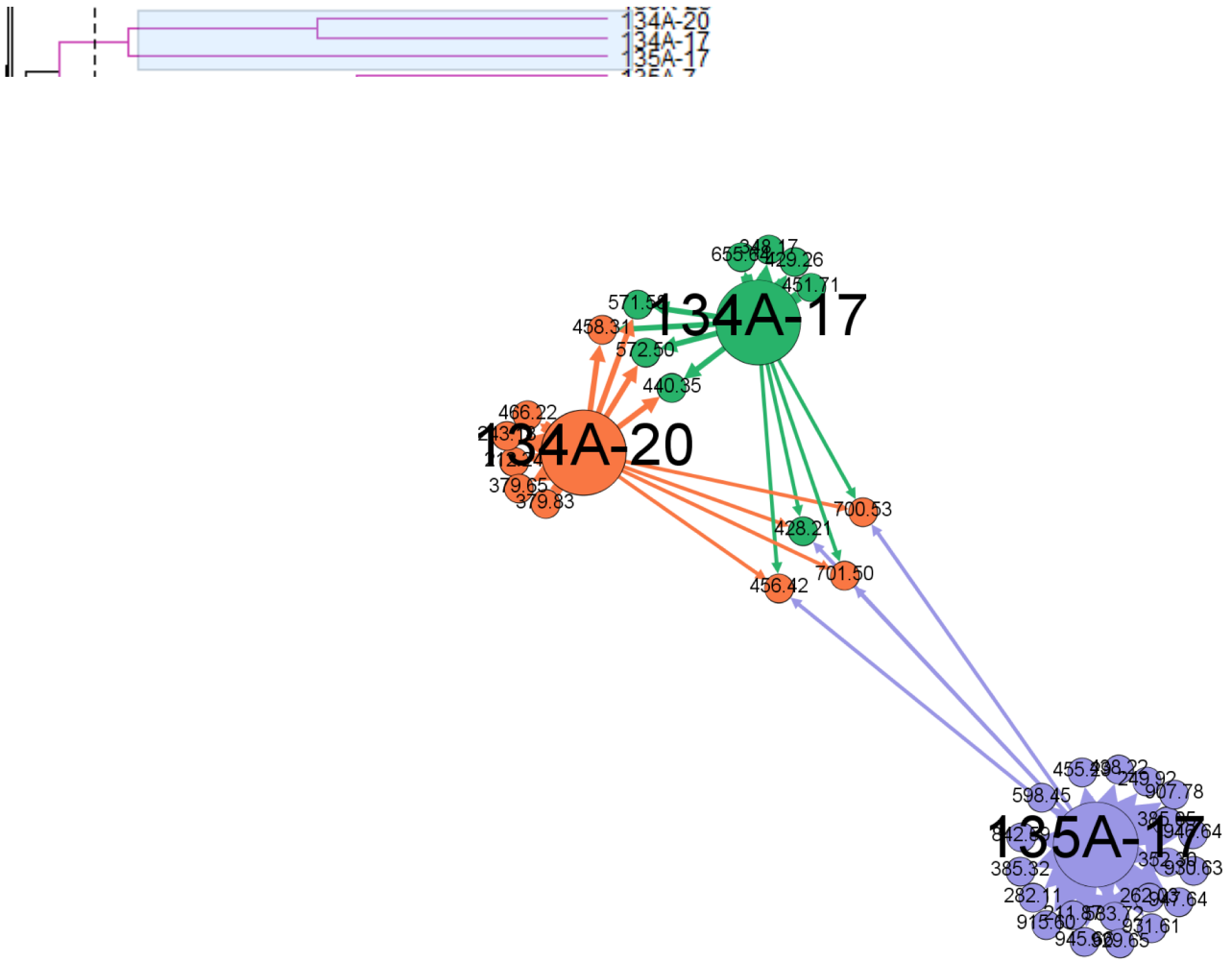

136A-77
136A-91
136A-98
136A-95
136A-83
136A-76
136A-104
136A-94
141AM-2
136A-93
136A-92

Group 11


15A-3
14D-2
37B-46
36A-105
36A-86
36A-106
14B-1
37A-18
35A-4
36A-81
37C-5
37C-27
37A-16
41BM-9
41BM-3
41BM-8
41BM-2
42AM-1
41BM-1
41BM-7
41BM-10
42BM-5
41AM-5
41BM-4
41AM-4
41AM-3
41BM-6
41BM-5
41AM-7
40E-3
37A-22
37A-8
14A-4
36A-108
36A-107
38K-1
37B-34
34C-12
39A-1
14B-2
38K-25

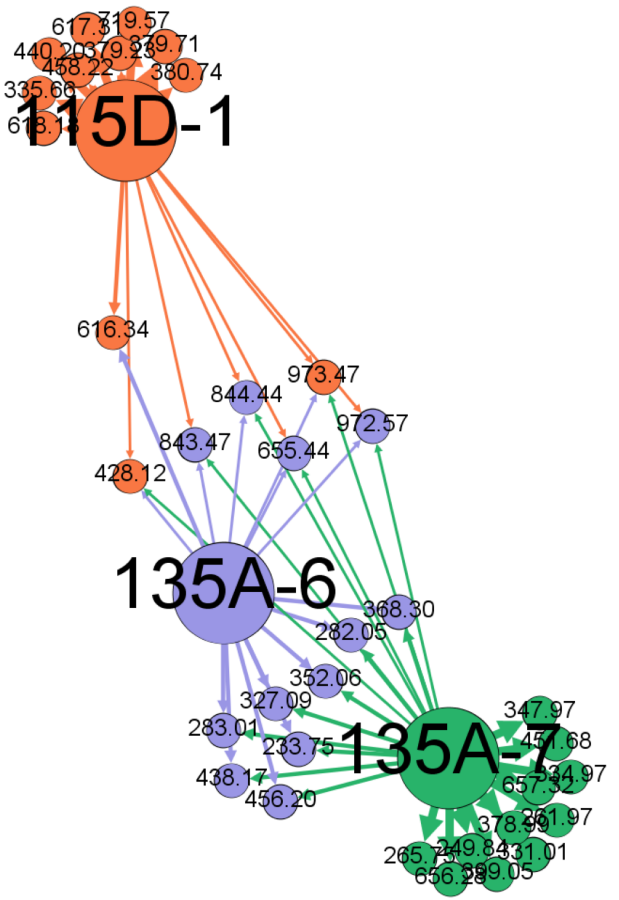

134C-12
136A-81
136A-86
136A-108
137A-16
137B-34
136A-107
137C-5
141AM-7
136A-106
137C-22
114A-4
137C-18
140E-3
139A-1
142AM-1
137B-46
136A-105
135A-4
137C-27
142BM-5
141BM-6
141BM-1
141BM-10
141AM-5
141AM-4
138K-1
141BM-4
141BM-8
115A-3
137A-8
141BM-9
114D-2
141BM-3
138K-25
141AM-3
114B-1
114B-2
141BM-2
141BM-5
141BM-7

Group 12

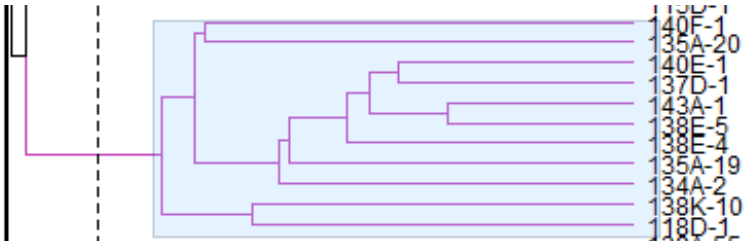

134A-20
134A-17
135A-17
134A-17
134A-20
135A-17
655.64
429.26
451.71
571.54
458.31
572.50
440.35
466.22
379.65
379.83
700.53
428.21
701.50
456.42
598.45
907.78
385.32
282.11
915.60
945.92
931.61

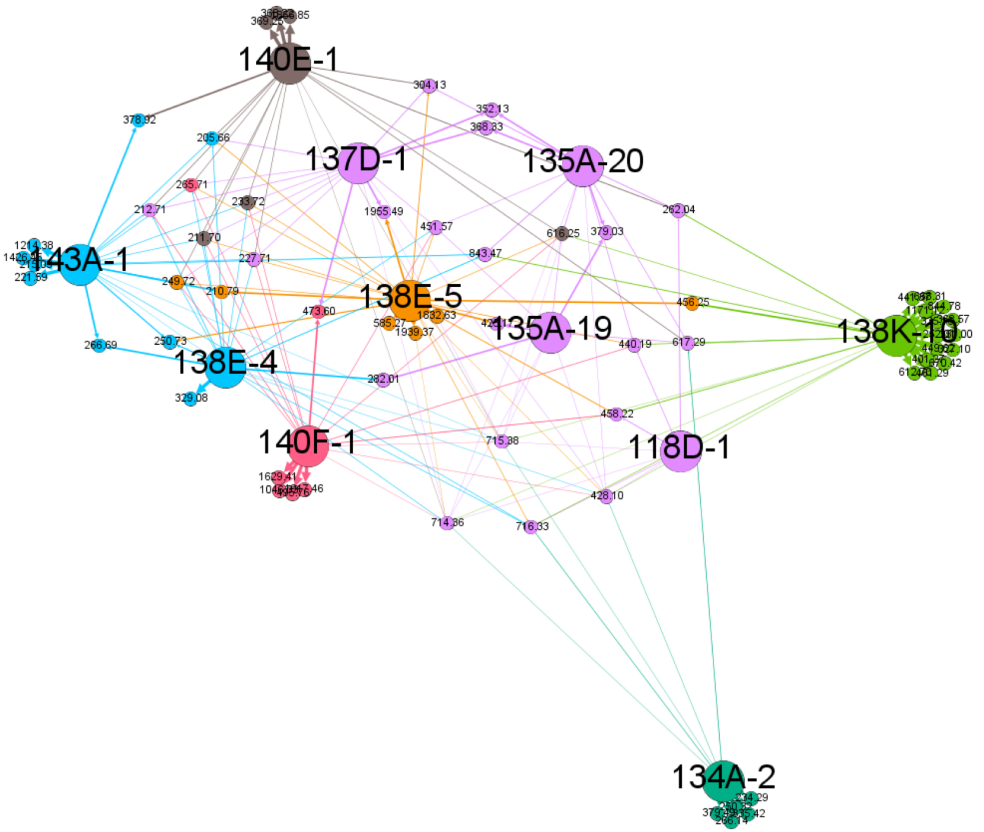
Group 13


135A-7
135A-6
115D-1

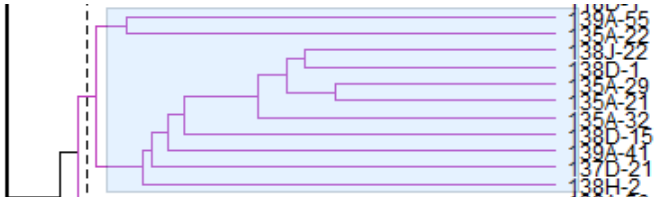

115D-1
135A-6
135A-7
616.34
973.47
844.44
972.57
843.47
655.44
428.12
368.30
282.05
352.06
327.09
283.01
233.75
438.17
456.20
347.97
451.68
534.97
657.32
361.97
378.99
249.84
331.01
265.75
656.28
699.05
380.74
335.66
458.22
379.23
379.71
617.31
719.57

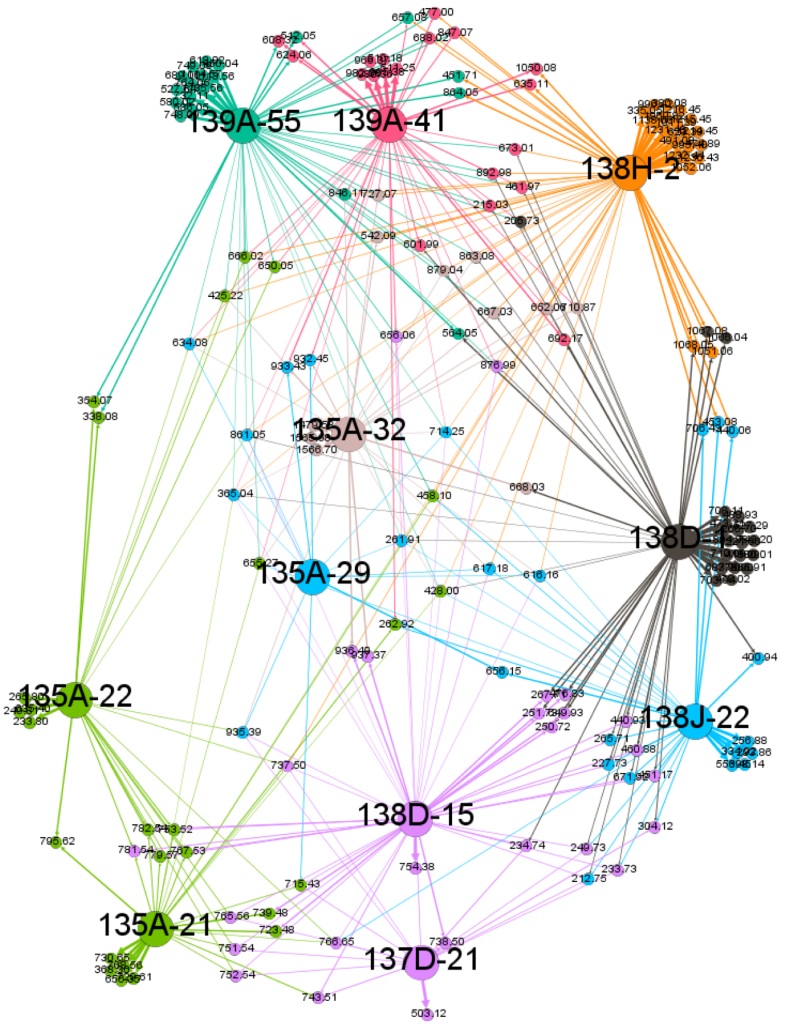
Group 14


140F-1
135A-20
140E-1
137D-1
143A-1
138E-5
138E-4
135A-19
134A-2
138K-10
118D-1

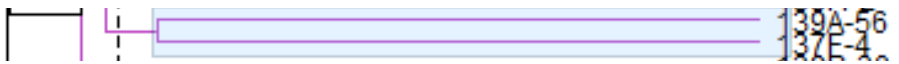

140E-1
137D-1
135A-20
143A-1
138E-5
135A-19
138K-10
138E-4
140F-1
118D-1
134A-2

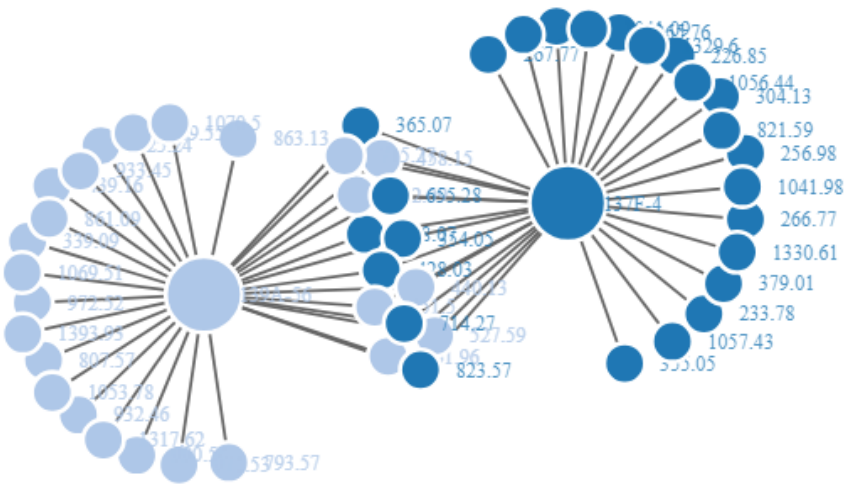
Group 15


139A-55
135A-22
138J-22
138D-1
135A-29
135A-21
135A-32
138D-15
139A-41
137D-21
138H-2

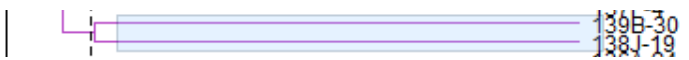

139A-55
139A-41
138H-2
135A-32
138D-1
135A-29
135A-22
138J-22
138D-15
135A-21
137D-21

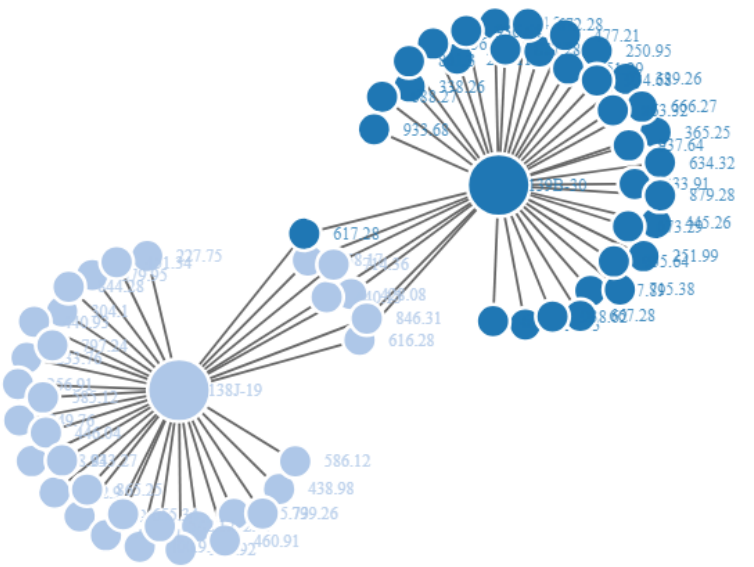
Group 16


139A-56
137E-4

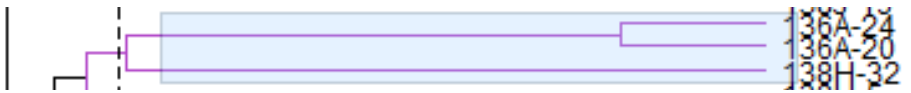

365.07
863.13
823.57
1057.43
233.78
379.01
1330.61
266.77
1041.98
256.98
821.59
304.13
1056.44
226.85

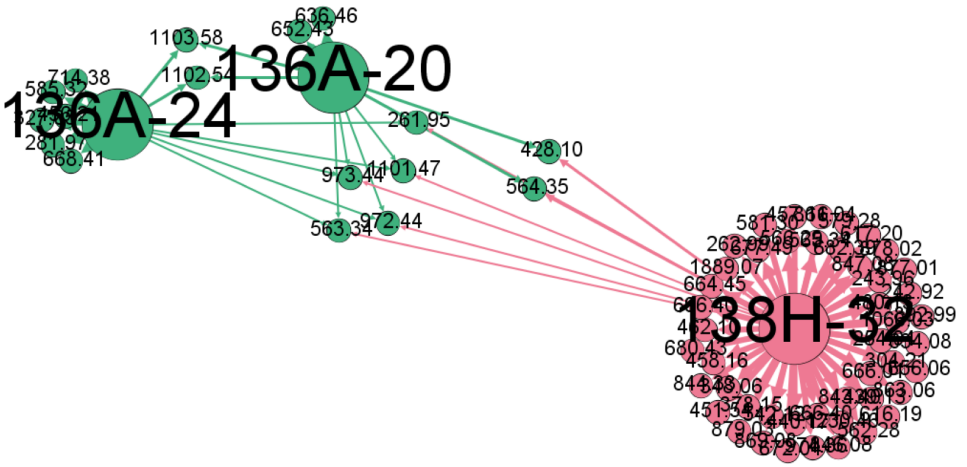
Group 17


139B-30
138J-19

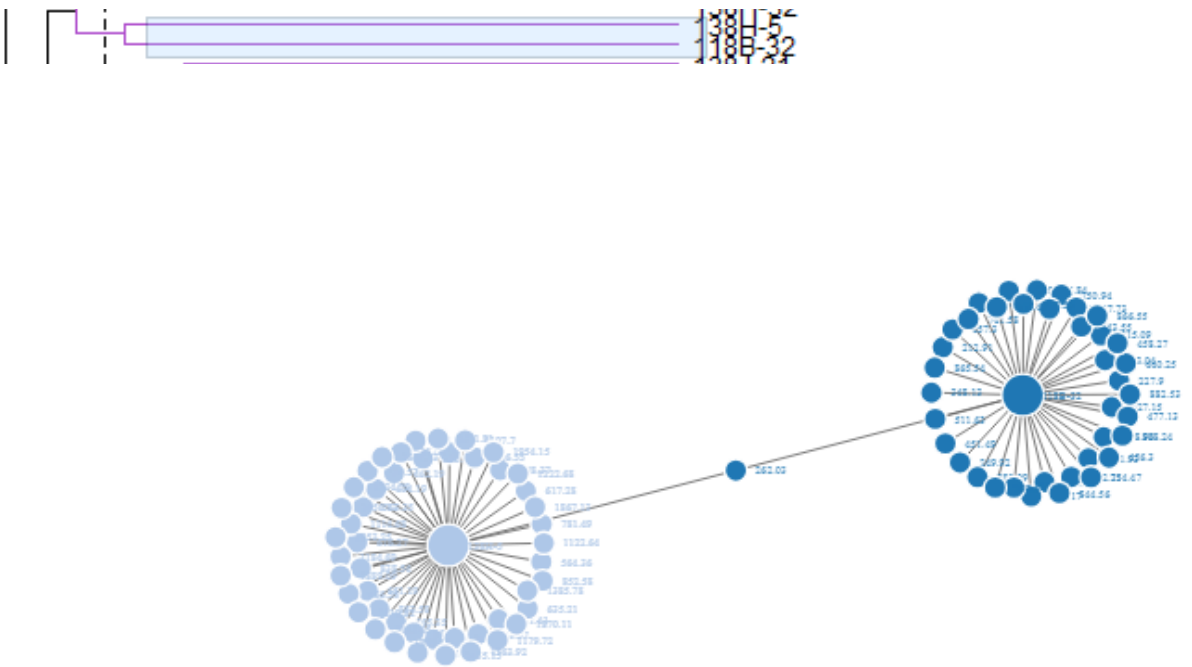

250.95
666.27
365.25
634.32
879.28
445.26
251.99
617.28
227.75
846.31
616.28
586.12
438.98
460.91

Group 18

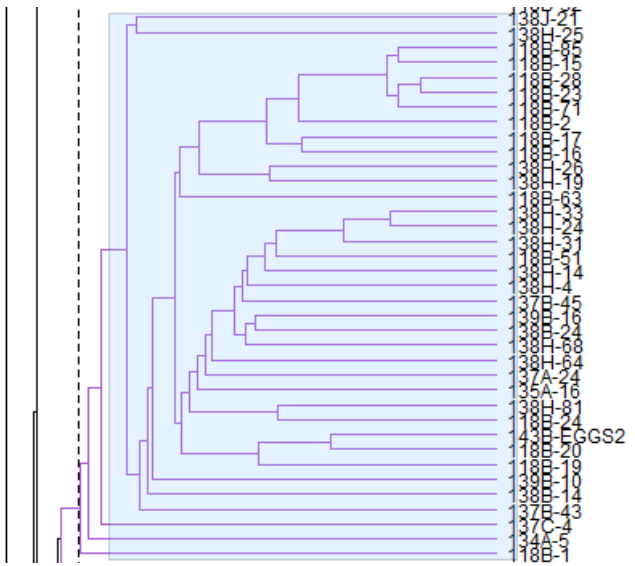

136A-24
136A-20
138H-32


136A-20
136A-24
138H-32
1103.58
1102.54
261.95
428.10
564.35
973.44
1101.47
563.34
972.44
668.41
281.97
1889.07
664.45
680.45
458.16

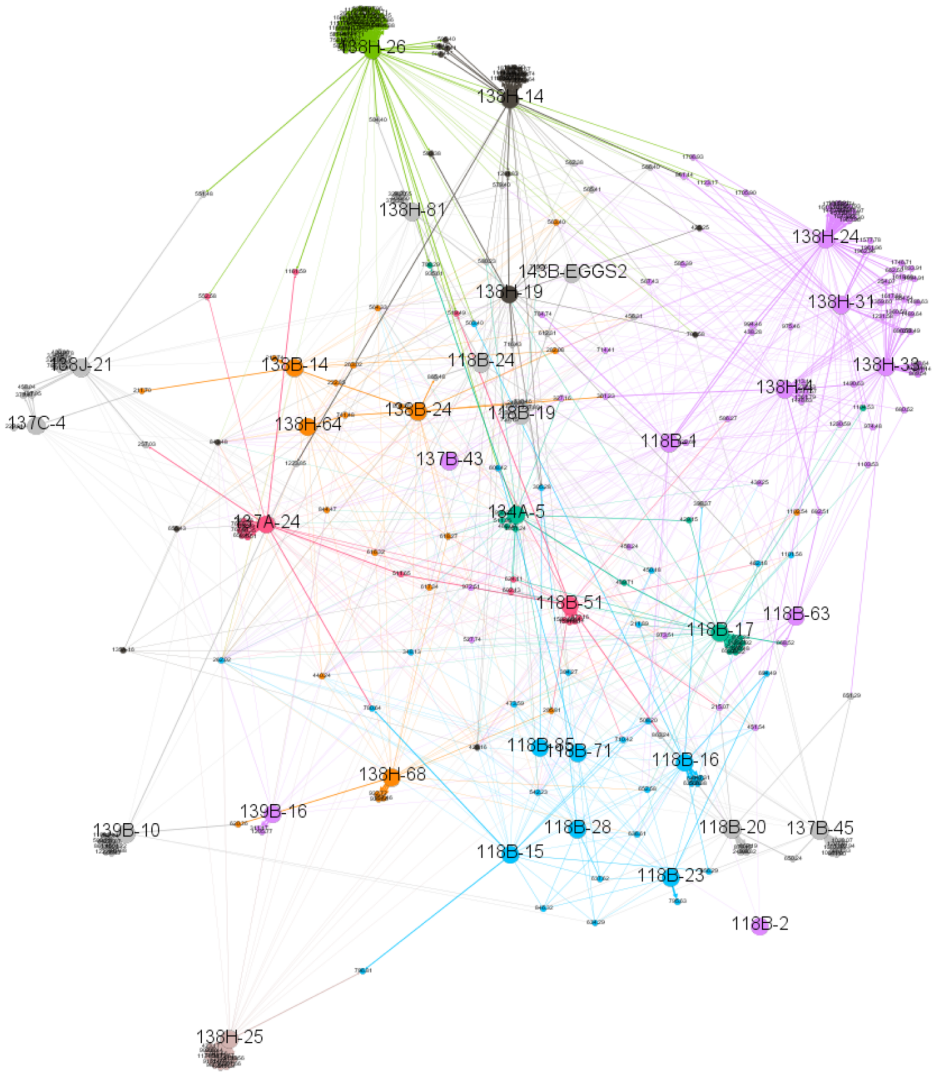
Group 19



Group 20

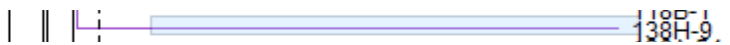


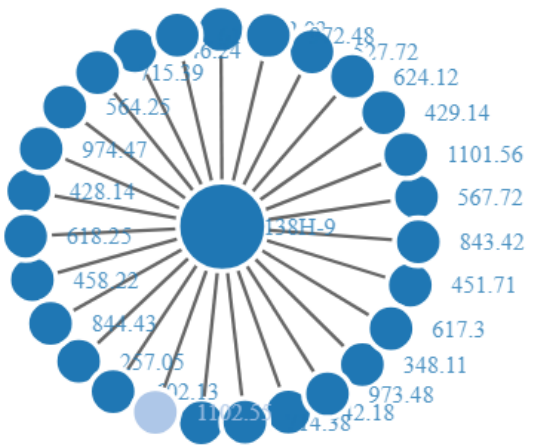
138J-21
138H-25
118B-85
118B-15
118B-28
118B-23
118B-71
118B-2
118B-17
118B-16
138H-26
138H-19
118B-63
138H-33
138H-24
138H-31
118B-51
138H-14
138H-4
137B-45
139B-16
138B-24
138H-68
138H-64
137A-24
135A-16
138H-81
118B-24
143B-EGGS2
118B-20
118B-19
139B-10
138B-14
137B-43
137C-4
134A-5
118B-1



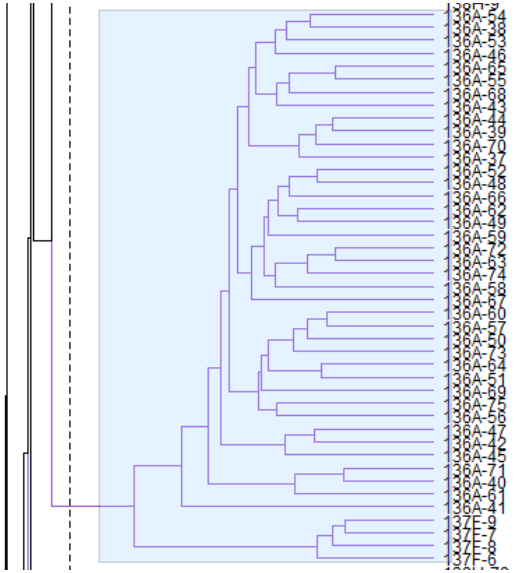
138H-26
138H-14
138H-81
138H-24
143B-EGGS2
138H-19
138H-31
138J-21
138B-14
118B-24
138H-33
138H-4
137C-4
138B-24
118B-19
138H-64
118B-1
137B-43
134A-5
137A-24
118B-51
118B-63
118B-17
118B-85
118B-71
118B-16
138H-68
139B-16
139B-10
118B-28
118B-20
137B-45
118B-15
118B-23
118B-2
138H-25

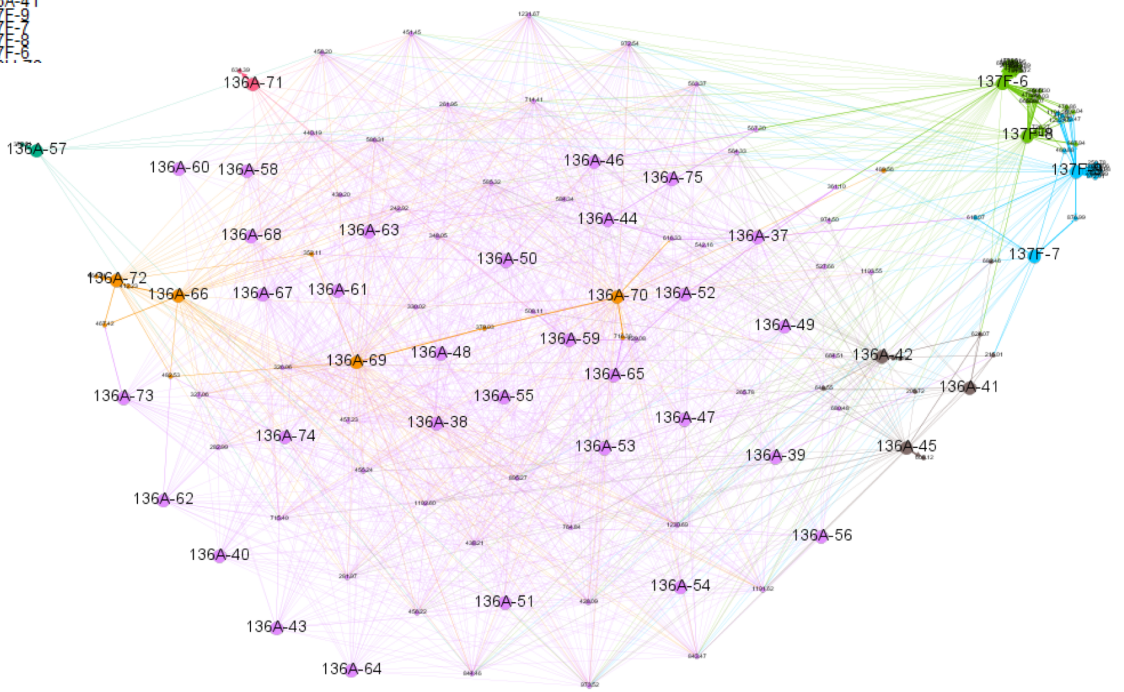
Group 21


138H-9

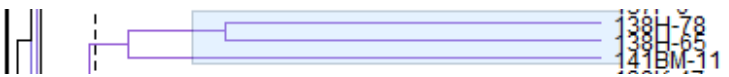

624.12
429.14
1101.56
567.72
843.42
451.71
617.3
348.11
973.48
974.47
428.14
618.25
458.22
844.43
138H-9

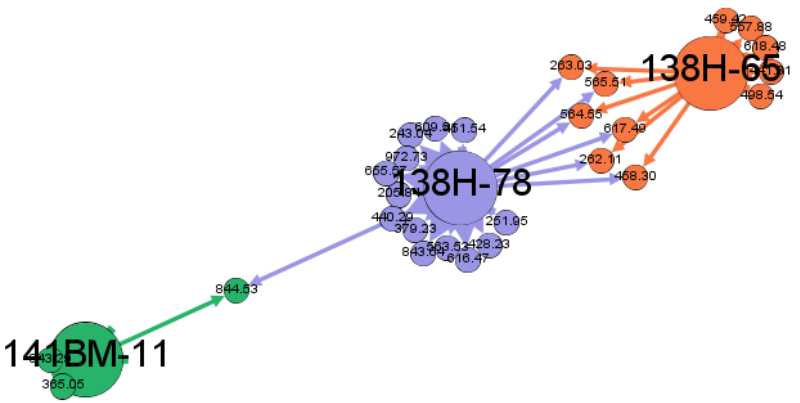
Group 22




136A-71
137F-6
136A-57
136A-60
136A-58
136A-46
136A-75
137F-8
136A-44
136A-68
136A-63
136A-37
137F-7
136A-50
136A-72
136A-66
136A-67
136A-61
136A-70
136A-52
136A-49
136A-59
136A-69
136A-48
136A-42
136A-65
136A-41
136A-73
136A-55
136A-47
136A-38
136A-74
136A-53
136A-39
136A-45
136A-62
136A-56
136A-40
136A-54
136A-51
136A-43
136A-64

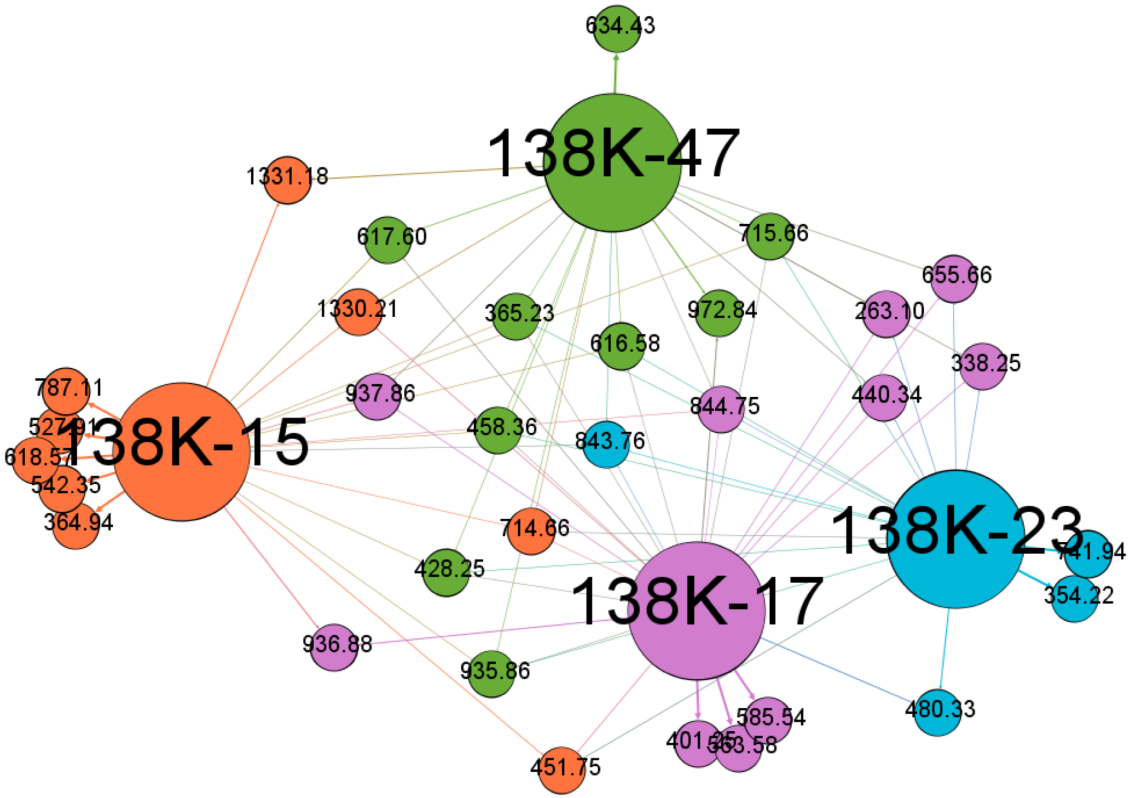
Group 23


138H-78
138H-65
141BM-11

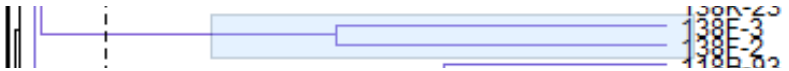

138H-65
138H-78
141BM-11
263.03
565.51
564.55
617.49
262.11
458.30
498.54
972.73
251.95
428.23
844.53
365.05

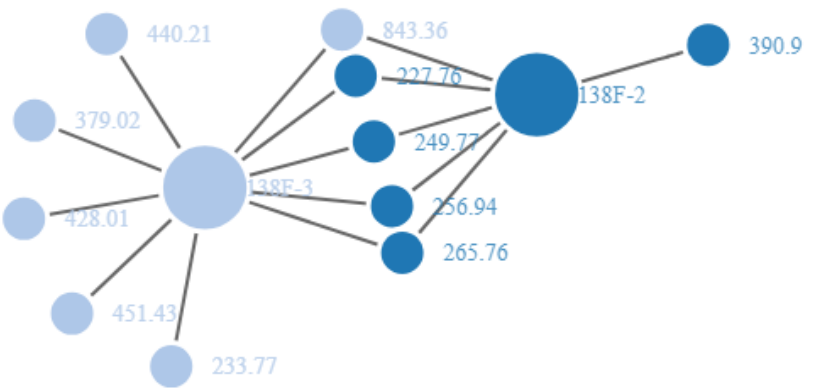
Group 24

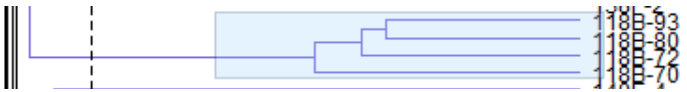

138K-47
138K-15
138K-17
138K-23

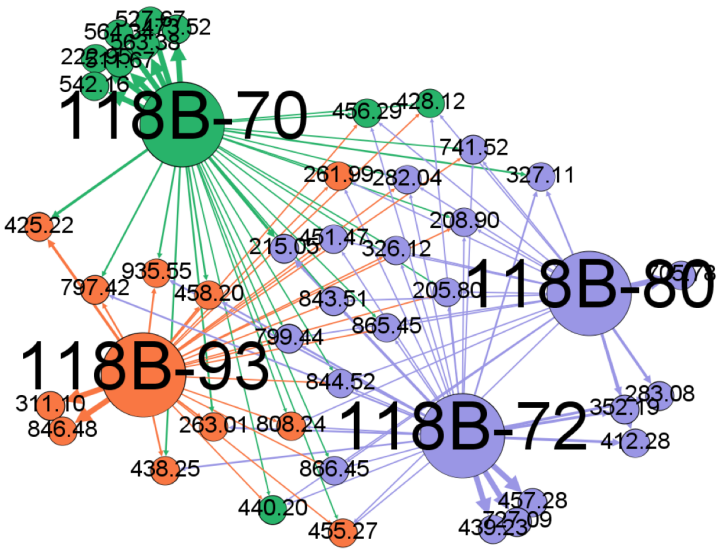

634.43
138K-47
1331.18
617.60
715.66
655.66
1330.21
365.23
972.84
263.10
616.58
338.25
787.11
937.86
844.75
440.34
458.36
843.76
618.57
542.35
364.94
714.66
741.94
428.25
138K-17
138K-23
354.22
936.88
935.86
585.54
480.33
451.75

Group 25

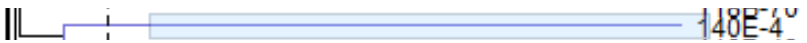


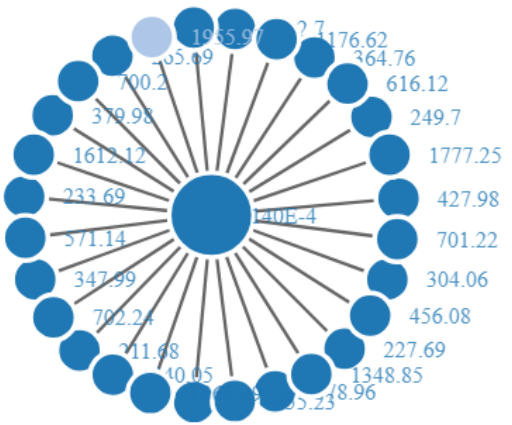

440.21
843.36
390.9
227.76
379.02
138F-2
249.77
138F-3
428.01
256.94
265.76
451.43
233.77


118B-93
118B-80
118B-72
118B-70

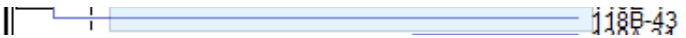

118B-70
118B-93
118B-80
118B-72
542.16
456.29
428.12
741.52
327.11
261.99
282.04
425.22
208.90
215.05
451.47
326.12
935.55
797.42
458.20
205.80
843.51
865.45
799.44
844.52
311.10
846.48
263.01
808.24
283.08
352.19
412.28
438.25
866.45
457.28
440.20
455.27
439.23

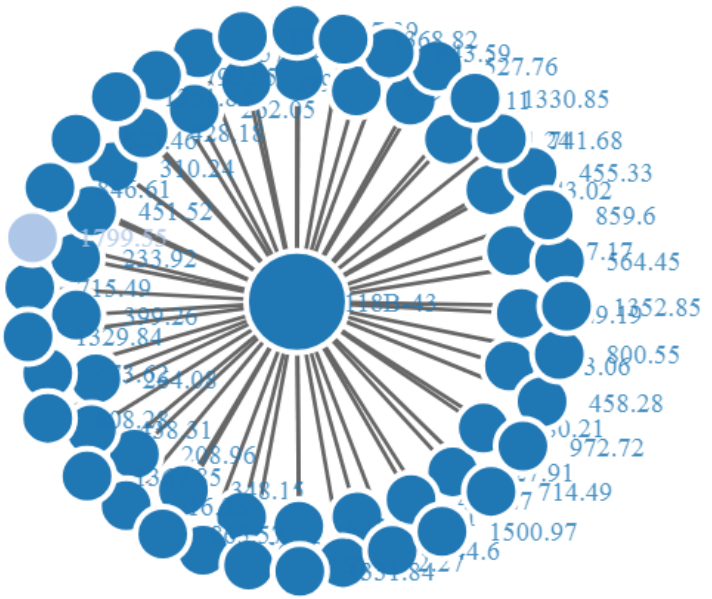
Group 27


140E-4

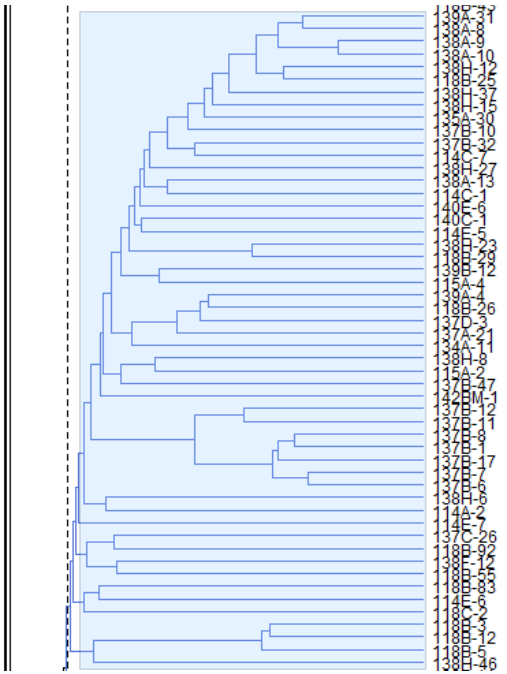

1176.62
364.76
616.12
249.7
1777.25
427.98
701.22
304.06
456.08
227.69
1348.85
140E-4

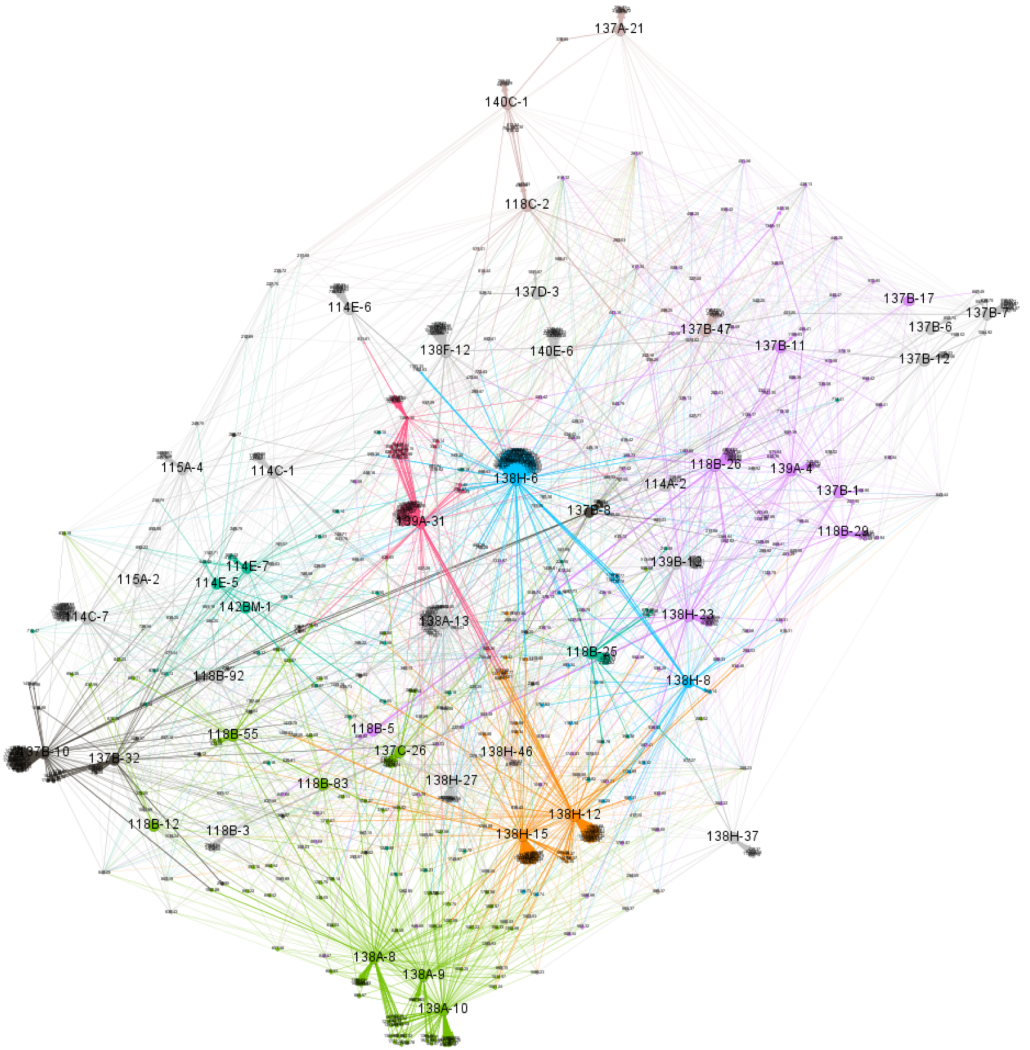
Group 28

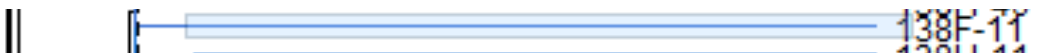

118B-43

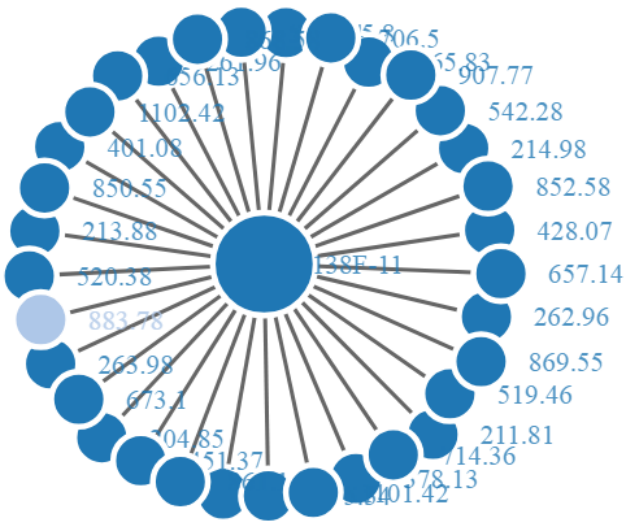

1330.85
455.33
859.6
564.45
800.55
458.28
972.72
714.49
1500.97
451.52
1329.84
118B-43

Group 29

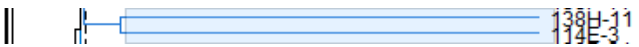

138A-31
138A-8
138A-9
138A-10
138H-12
118B-25
138H-37
138H-15
135A-30
137B-10
137B-32
114C-7
138H-27
138A-13
114C-1
140E-6
140C-1
114E-5
138H-23
118B-29
139B-12
115A-4
139A-4
118B-26
137D-3
137A-21
114A-11
138H-8
115A-2
137B-47
142BM-1
137B-12
137B-11
137B-8
137B-1
137B-17
137B-7
137B-6
138H-6
114A-2
114E-7
137C-26
118B-92
138F-12
118B-55
118B-83
114E-6
118C-2
118B-12
118B-3
118B-5
138H-46

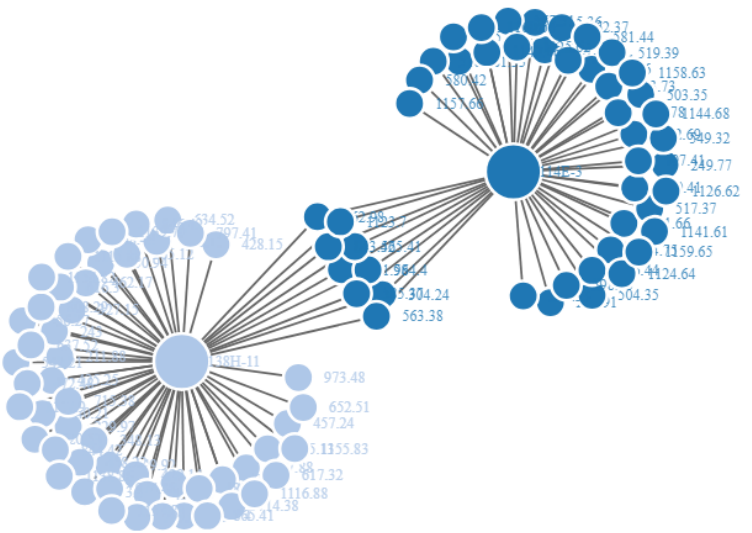

137A-21
140C-1
118C-2
137D-3
114E-6
137B-17
137B-7
137B-6
137B-47
138F-12
140E-6
137B-11
137B-12
115A-4
114C-1
138H-6
118B-26
139A-4
114A-2
137B-1
139A-31
137B-8
118B-29
139B-12
114E-7
115A-2
114E-5
142BM-1
114C-7
138A-13
138H-23
118B-25
118B-92
138H-8
118B-55
118B-5
137B-10
137B-32
137C-26
138H-46
118B-83
138H-27
138H-12
118B-12
118B-3
138H-15
138H-37
138A-8
138A-9
138A-10


138F-11

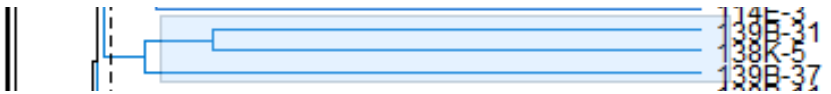

907.77
542.28
214.98
852.58
428.07
657.14
262.96
869.55
519.46
211.81
1102.42
401.08
850.55
213.88
520.38
263.98
673.1
138F-11

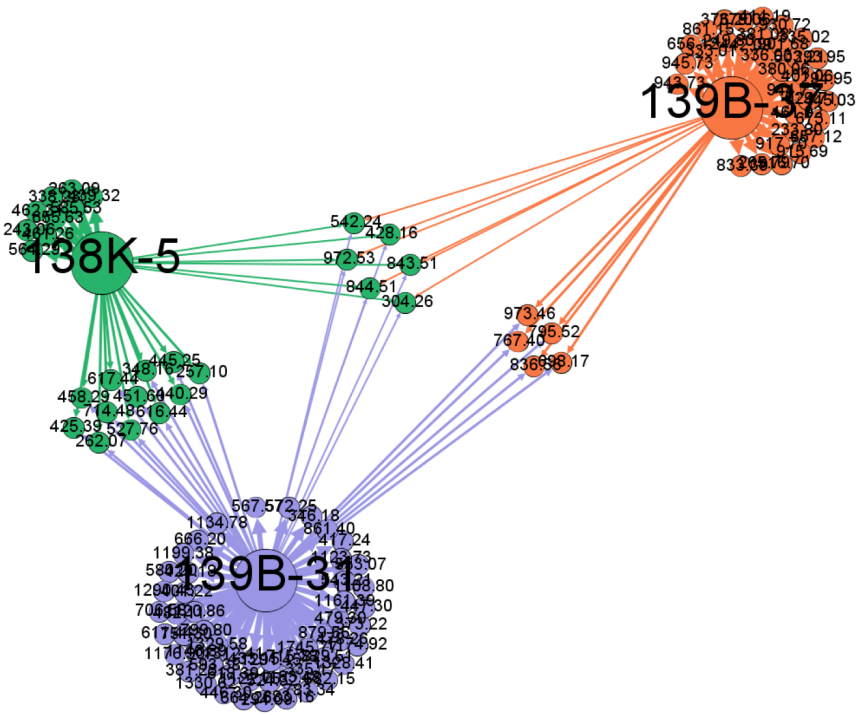
Group 31




1158.63
1144.68
1126.62
517.37
1141.61
1124.64
563.38
973.48
652.51
457.24
817.32
1116.88

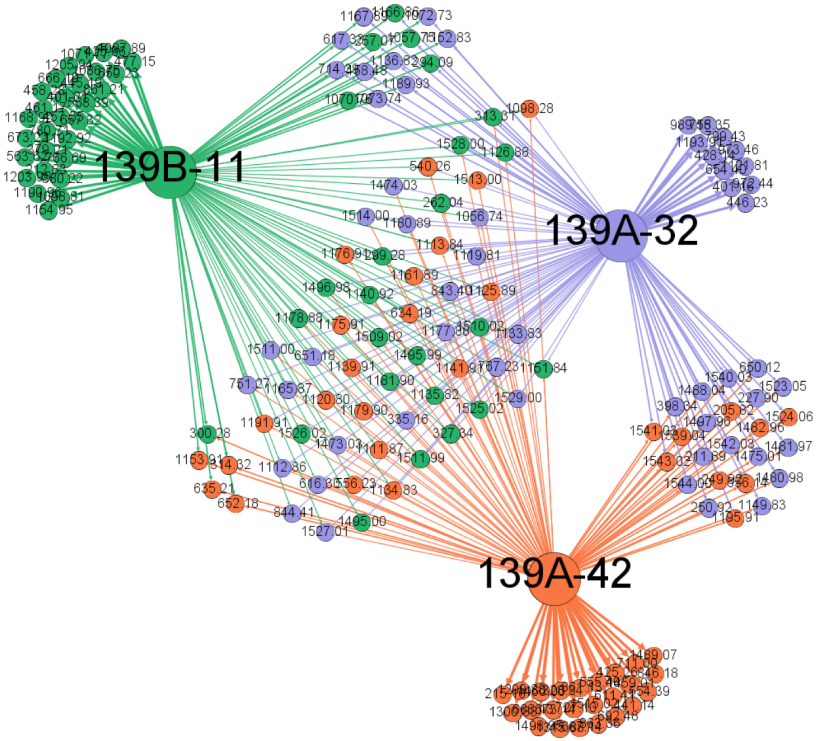
Group 32


139B-31
138K-5
139B-37

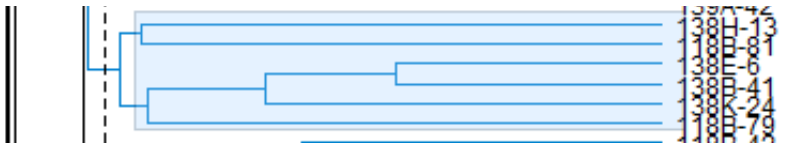

139B-37
138K-5
139B-31
542.24
428.16
972.53
843.51
844.51
304.26
973.46
785.52
767.40
836.17
567.57
2.25
346.18
861.40
1134.78
666.20
1199.38
417.24
1160.80
1161.39
479.20
879.85
1745.74
1174.92
617.44
458.29
451.64
640.29
714.48
616.44
425.39
527.76
262.07

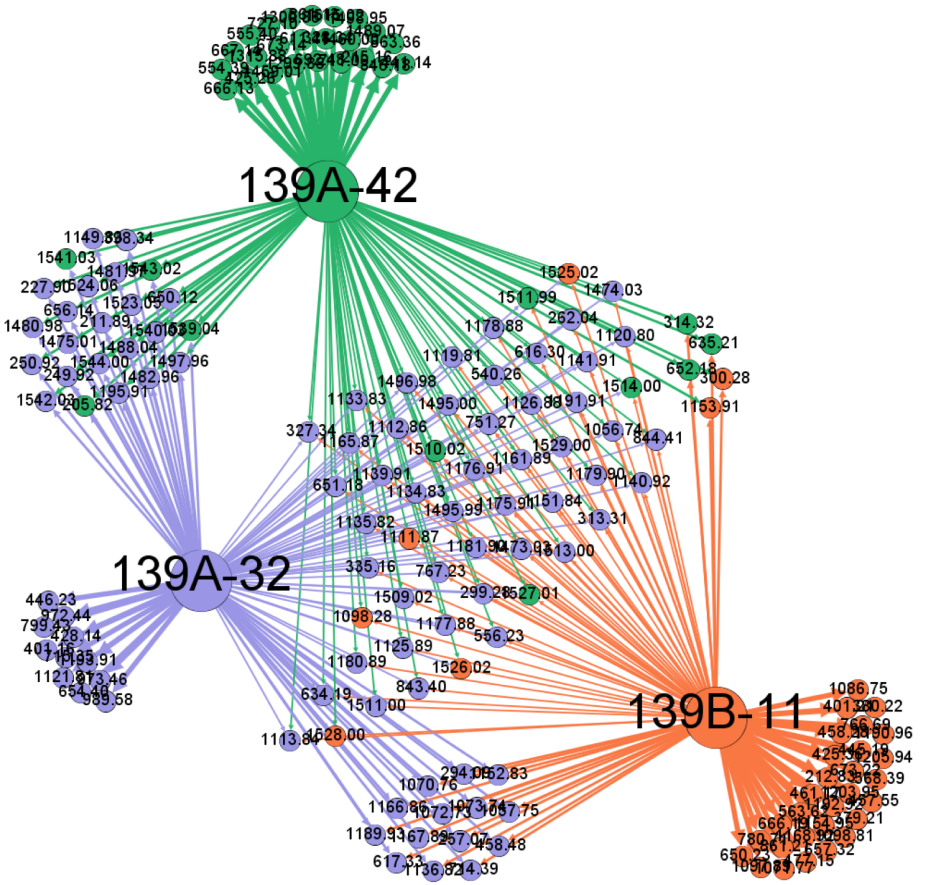
Group 33



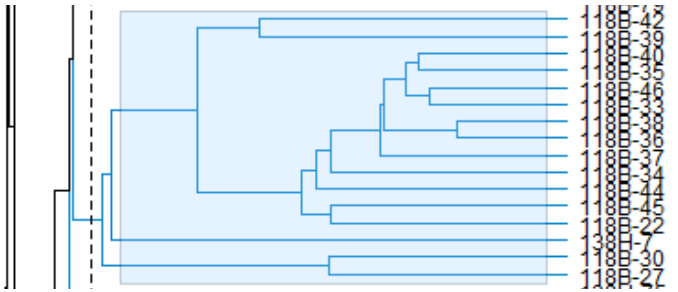


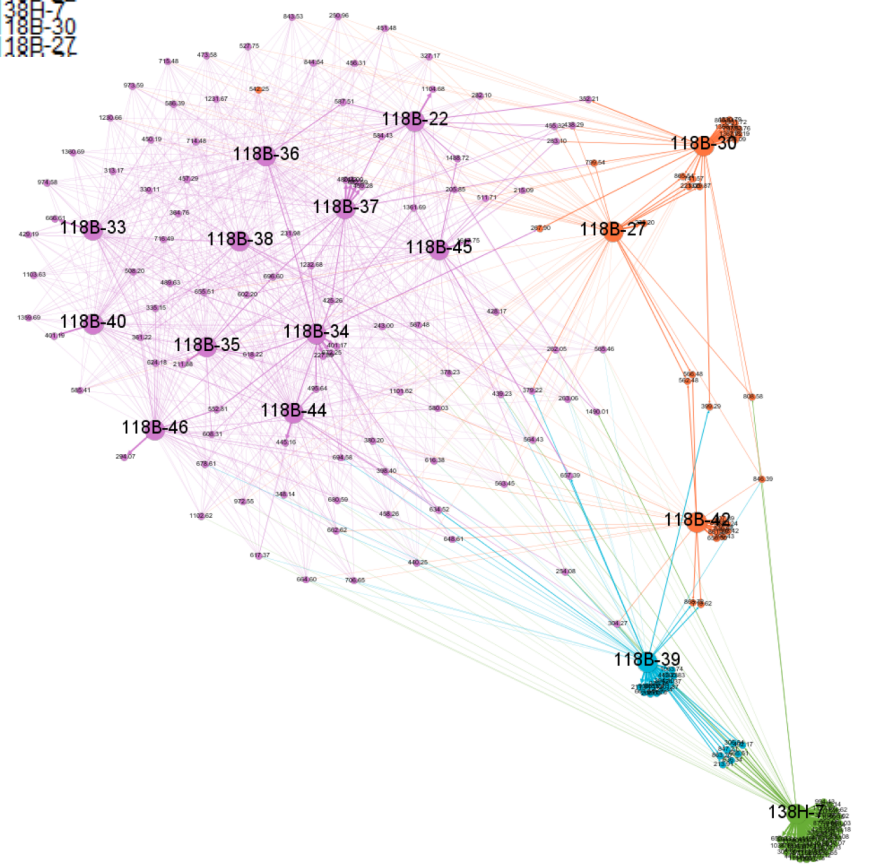
Group 34



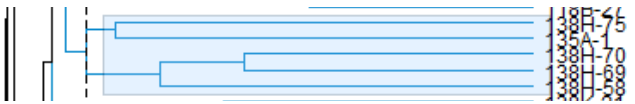


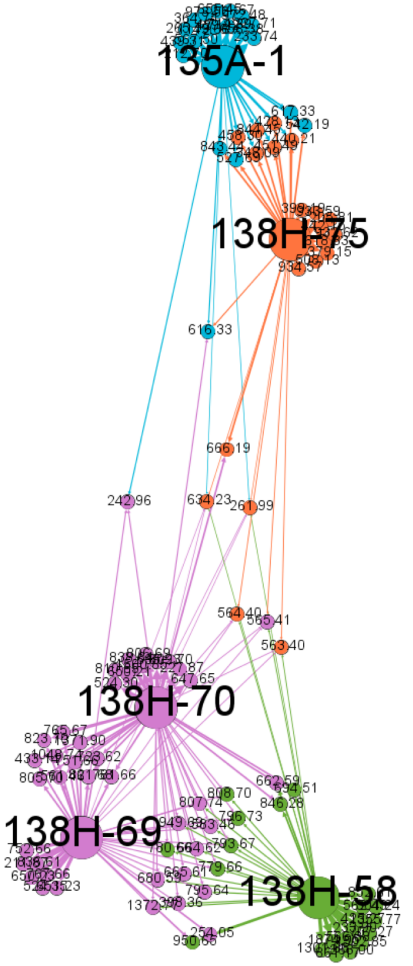
Group 35


118B-42
118B-39
118B-40
118B-35
118B-46
118B-33
118B-38
118B-36
118B-37
118B-34
118B-44
118B-45
118B-22
118B-30
118B-27
138H-7


118B-22
118B-30
118B-36
118B-37
118B-33
118B-38
118B-45
118B-27
118B-40
118B-34
118B-35
118B-44
118B-46
118B-42
118B-39
138H-7

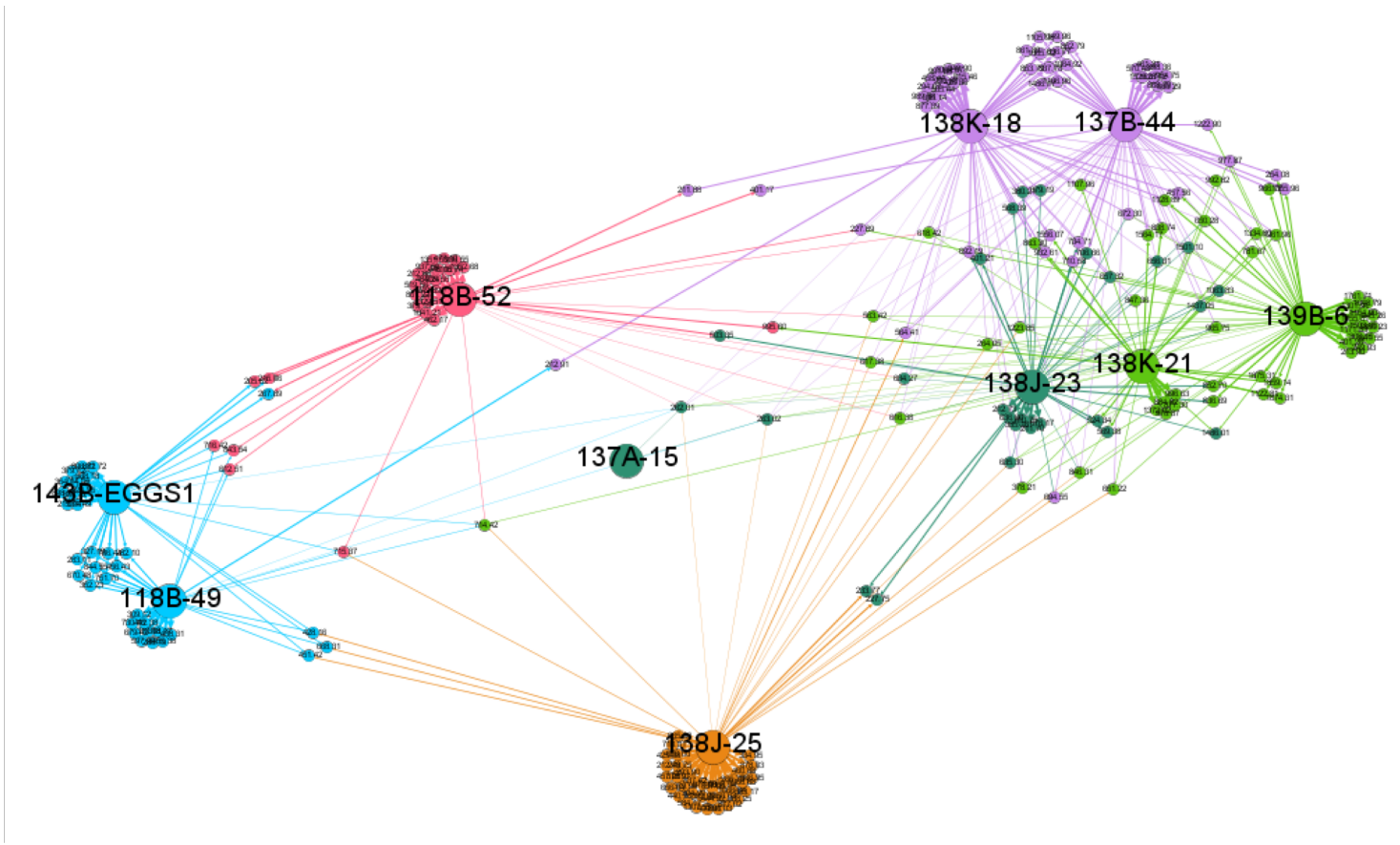
Group 36

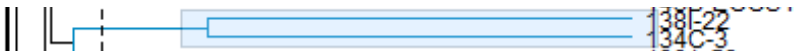

138H-75
135A-1
138H-70
138H-69

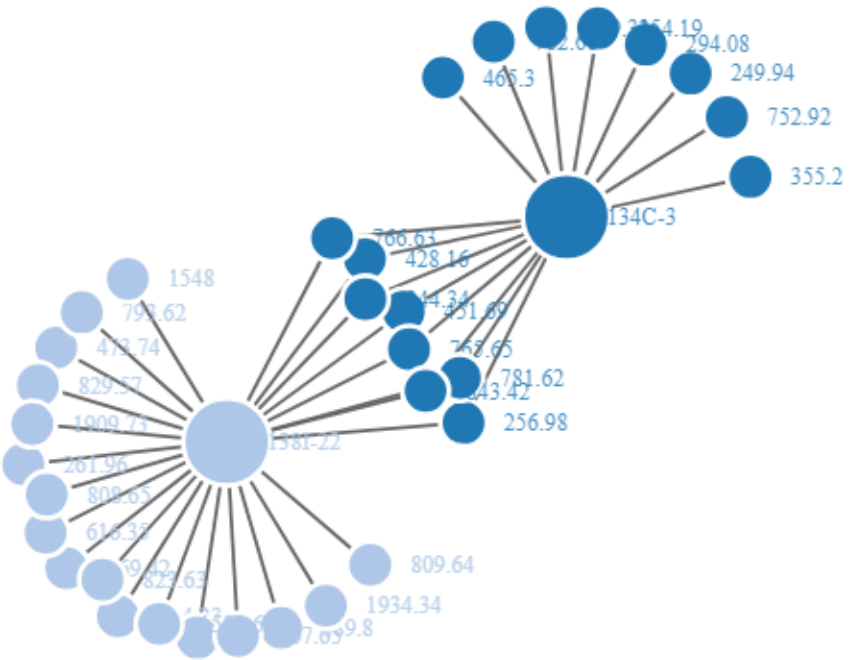

135A-1
138H-75
138H-70
138H-69
138H-58

Group 37

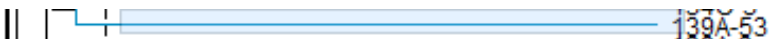

138K-21
138K-18
138J-23
137B-44
139B-6
138J-25
137A-15
118B-52
118B-49
143B-EGGS1

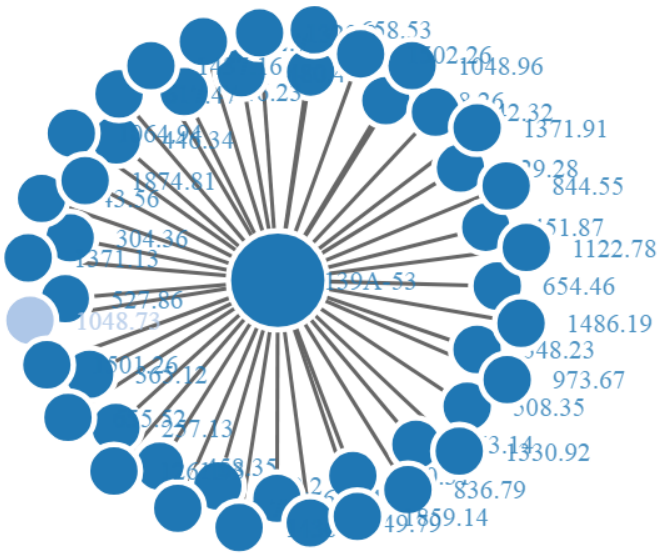

138K-18
137B-44
118B-52
139B-6
138K-21
138J-23
137A-15
143B-EGGS1
118B-49
138J-25


138I-22
134C-3

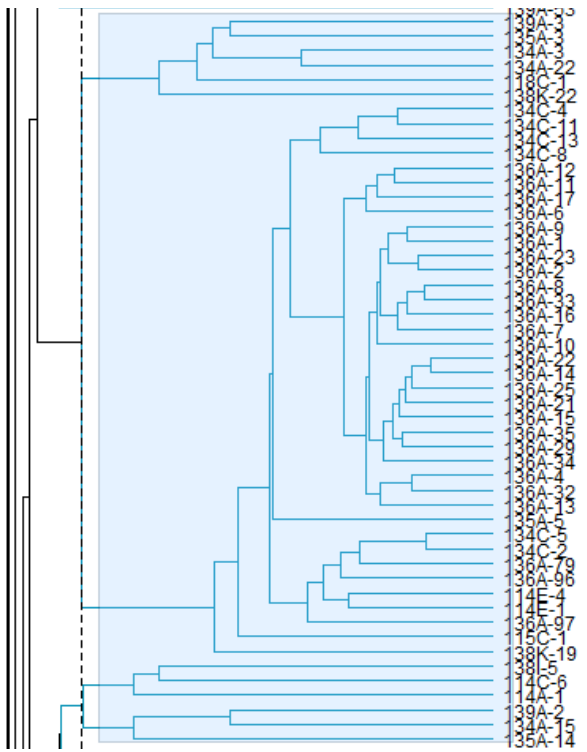

294.08
249.94
752.92
355.2
134C-3
428.16
451.69
781.62
256.98
1548
809.64
1934.34

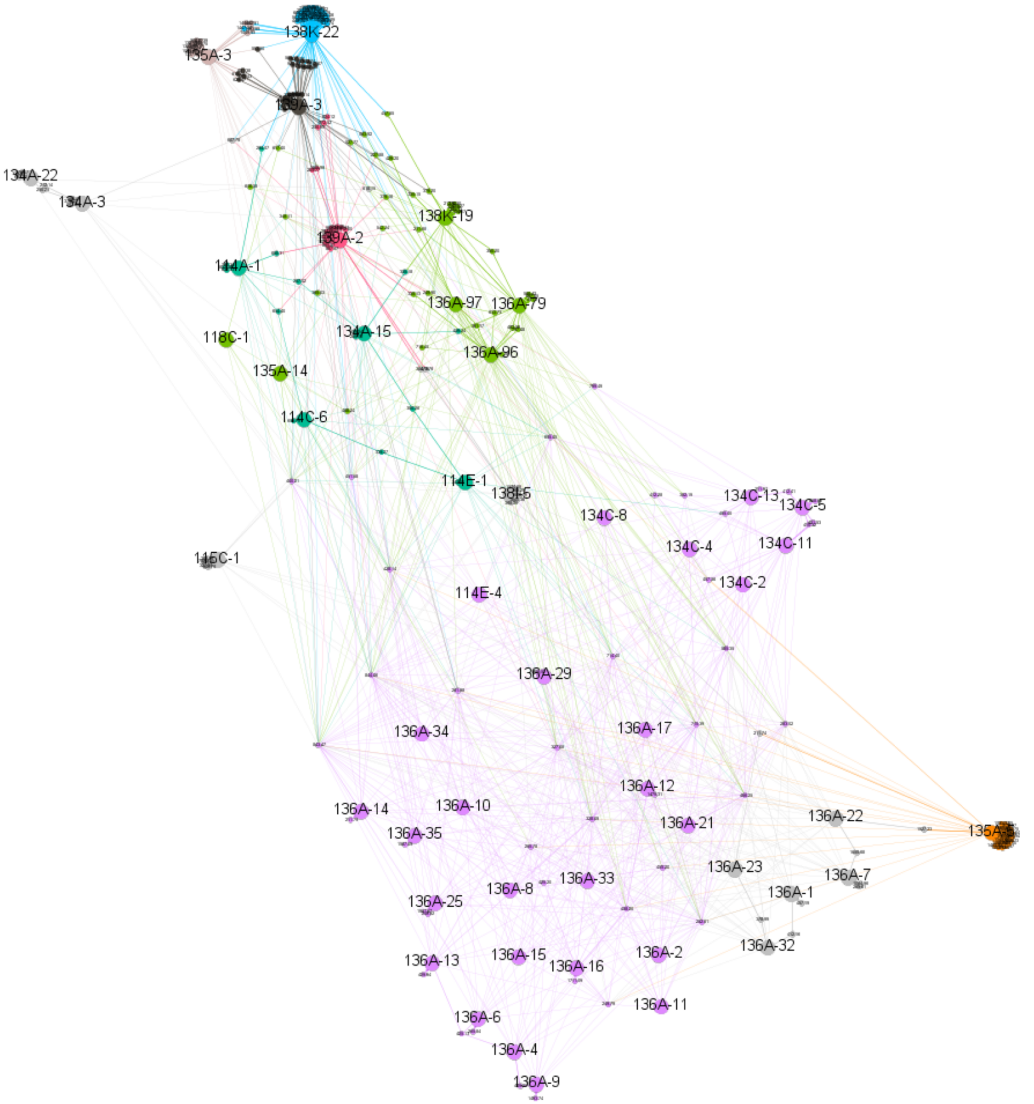
Group 39


139A-53

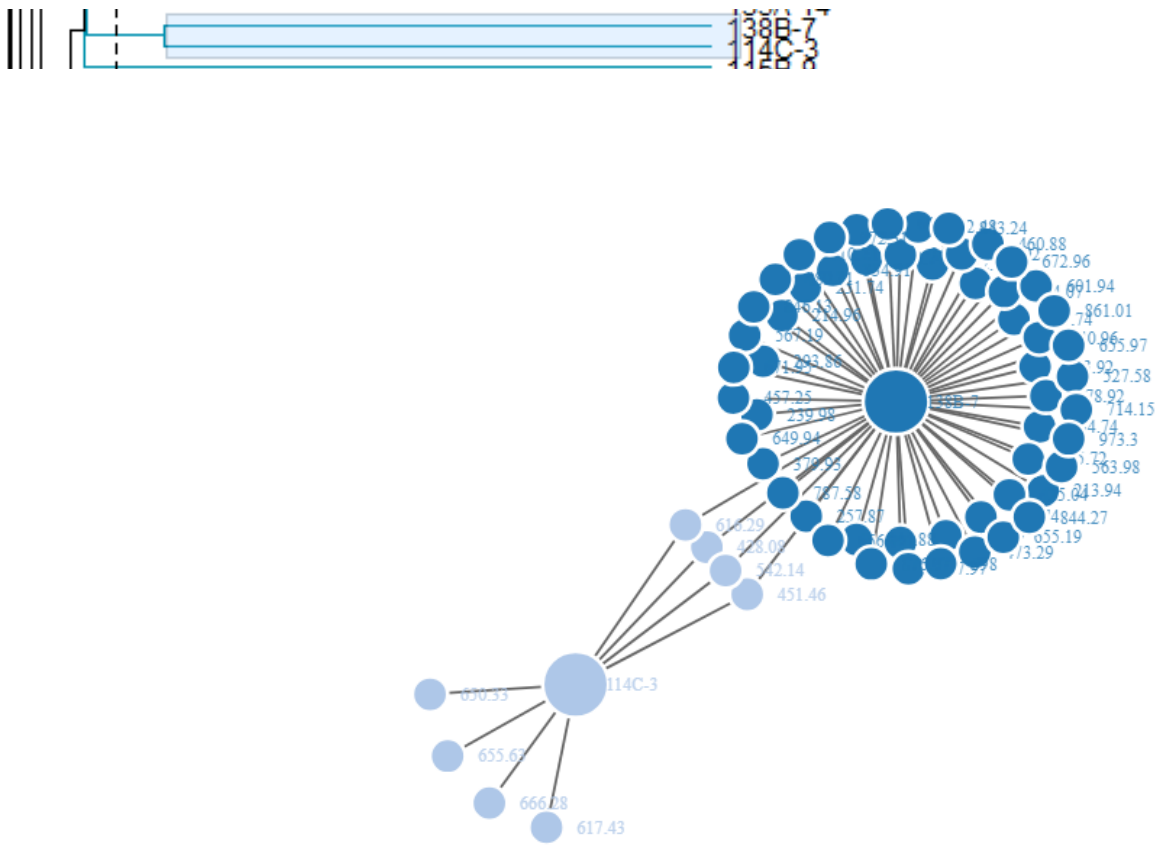

1048.96
1371.91
844.55
1122.78
654.46
1486.19
973.67
1330.92
836.79
1874.81
304.36
1048.73
1371.13
139A-53

Group 40

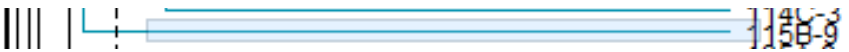


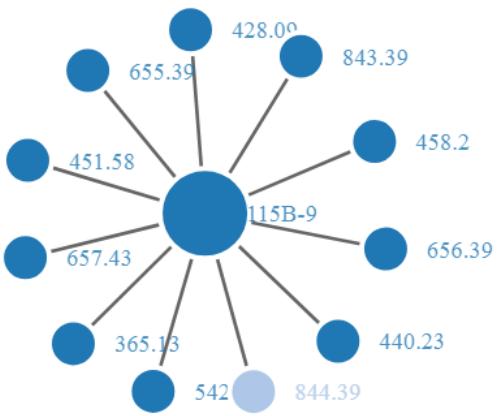

138K-22
135A-3
139A-3
134A-22
134A-3
138K-19
139A-2
144A-1
136A-97
136A-79
118C-1
134A-15
136A-96
135A-14
144C-6
114E-1
138I-5
134C-13
134C-5
134C-8
134C-4
134C-11
145C-1
134C-2
114E-4
136A-29
136A-34
136A-17
136A-12
136A-14
136A-10
136A-21
136A-22
135A-5
136A-35
136A-23
136A-7
136A-33
136A-8
136A-1
136A-25
136A-32
136A-13
136A-15
136A-16
136A-2
136A-11
136A-6
136A-4
136A-9

Group 41

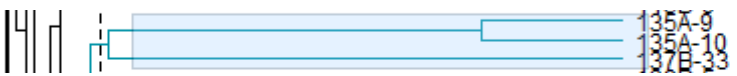


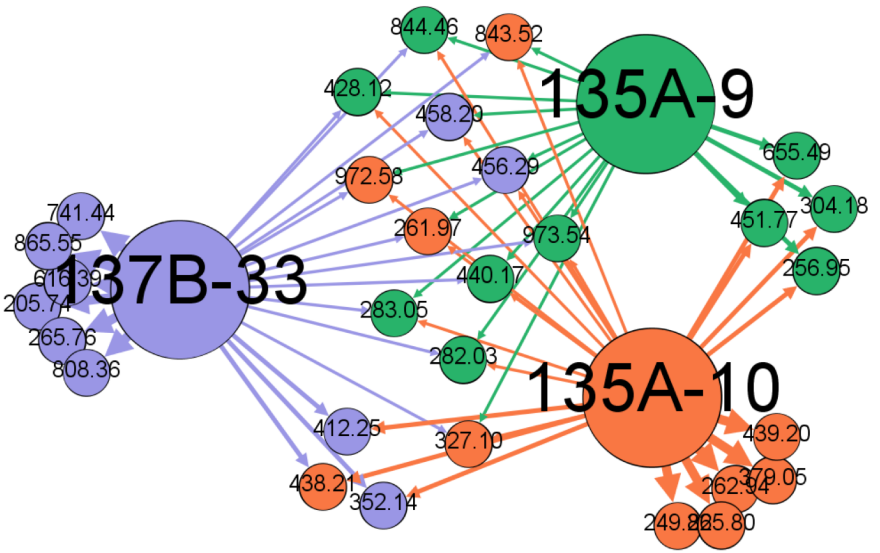
Group 42



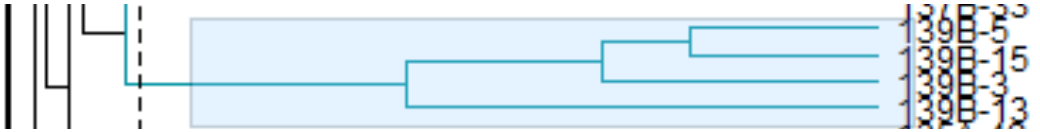

428.09
843.39
655.39
458.2
451.58
115B-9
656.39
657.43
440.23
365.13
542
844.39

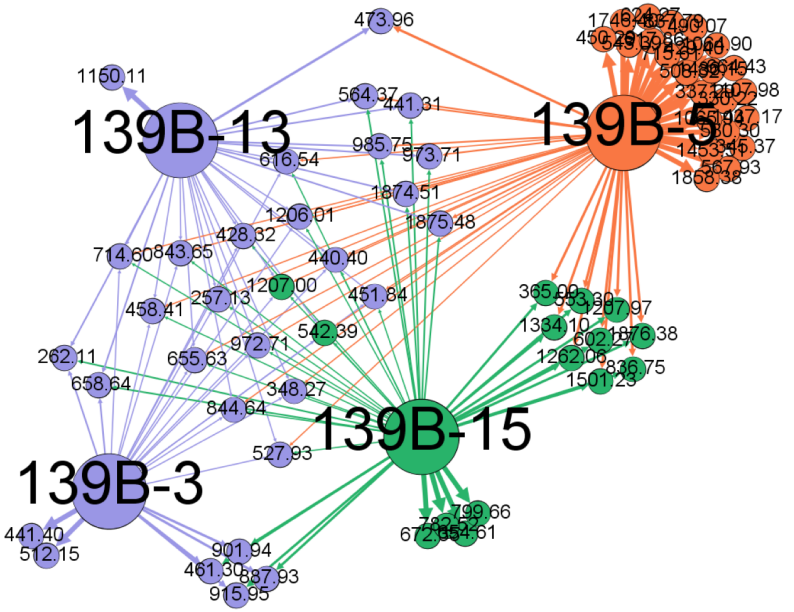
Group 43


135A-9
135A-10
137B-33


844.46
843.52
135A-9
428.12
458.20
655.49
972.58
456.29
304.18
741.44
261.97
973.54
451.77
865.55
137B-33
440.17
256.95
616.39
205.74
283.05
265.76
282.03
808.36
135A-10
412.25
439.20
327.10
438.21
352.14
379.05
262.94
249.82
285.80

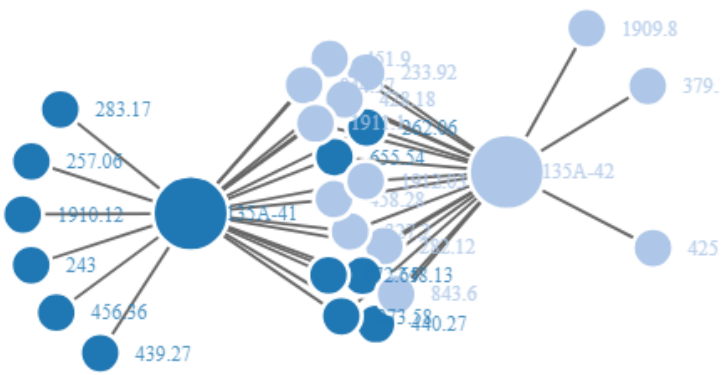
Group 44

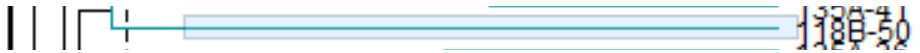

139B-5
139B-15
139B-3
139B-13

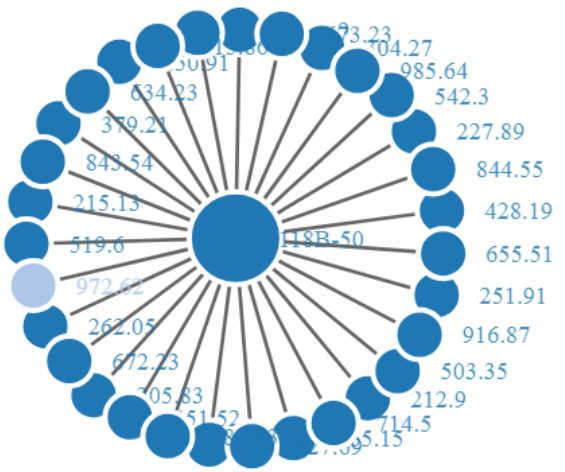

139B-13
139B-5
139B-15
139B-3
1150.11
473.96
564.37
441.81
985.75
973.71
616.54
1874.51
1206.01
1875.48
428.82
714.60
843.65
440.40
1207.00
257.13
451.84
458.41
542.39
972.71
262.11
655.63
658.64
348.27
844.64
527.93
441.40
512.15
901.94
461.30
887.93
915.95
799.66
365.00
1207.97
1334.10
1876.38
1262.06
1501.23
1858.38

Group 45


135A-42
135A-41

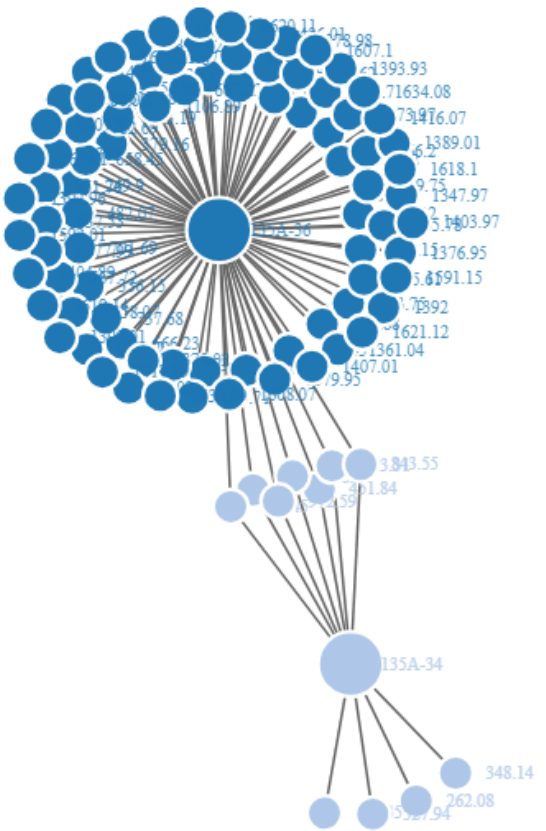

283.17
257.06
1910.12
243
456.36
439.27
135A-41
233.92
655.54
135A-42
1909.8
379.
425
843.6
440.27



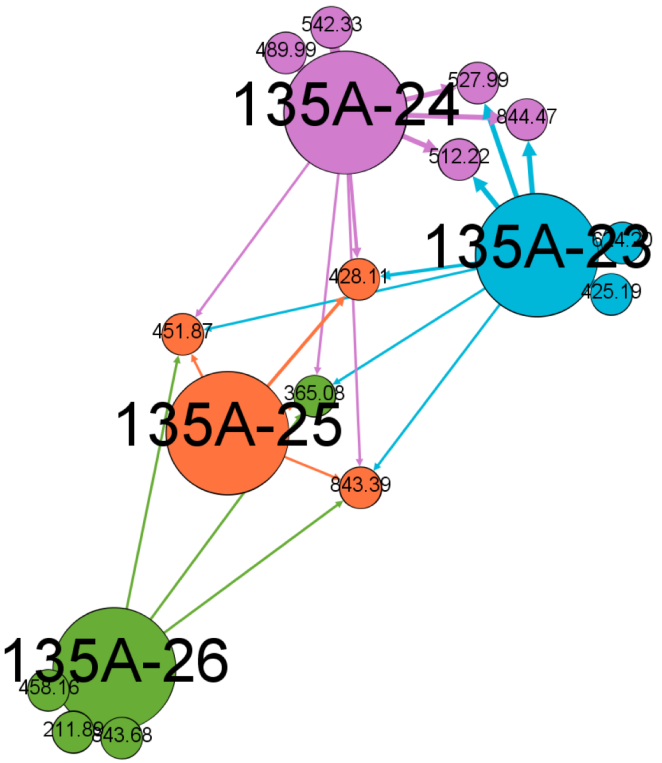

704.27
985.64
542.3
227.89
844.55
428.19
655.51
251.91
916.87
503.35
212.9
714.5
519.6
262.05
672.23

Group 47


135A-36
135A-34

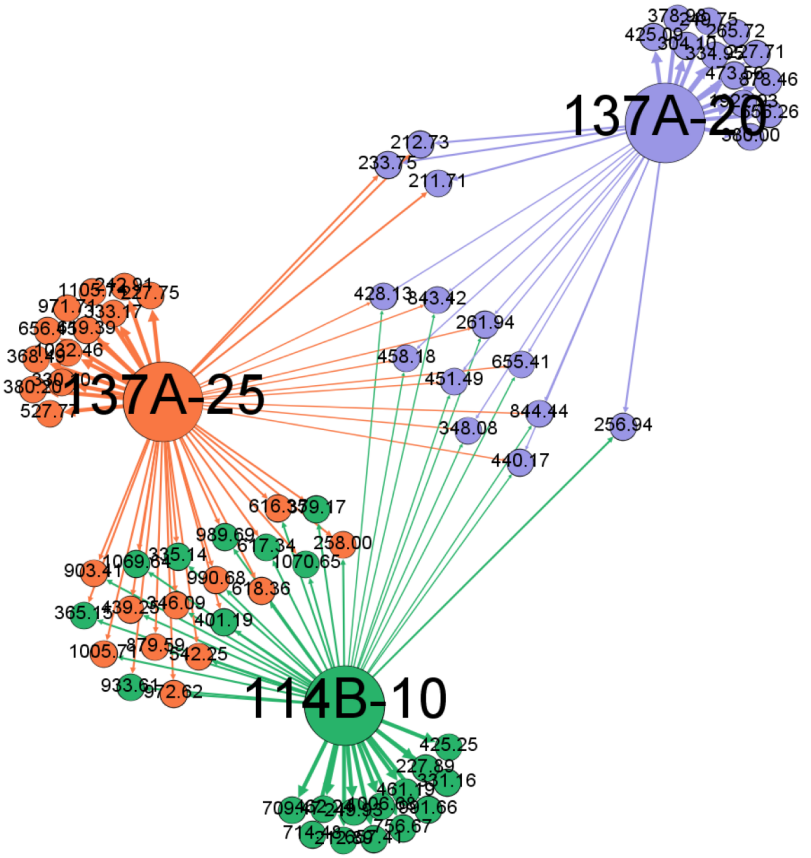

1607.1
1393.93
1634.08
1389.01
1618.1
1347.97
1376.95
1621.12
1361.04
1407.01
135A-34
348.14
262.08


135A-25
135A-23
135A-24
135A-26

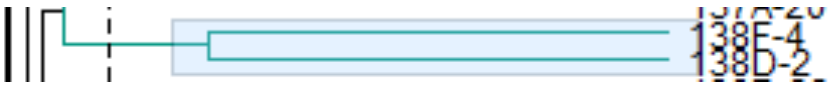

542.33
489.99
135A-24
527.99
844.47
512.22
135A-23
624.23
428.11
425.19
451.87
365.08
135A-25
843.39
135A-26
458.16
211.89
843.68

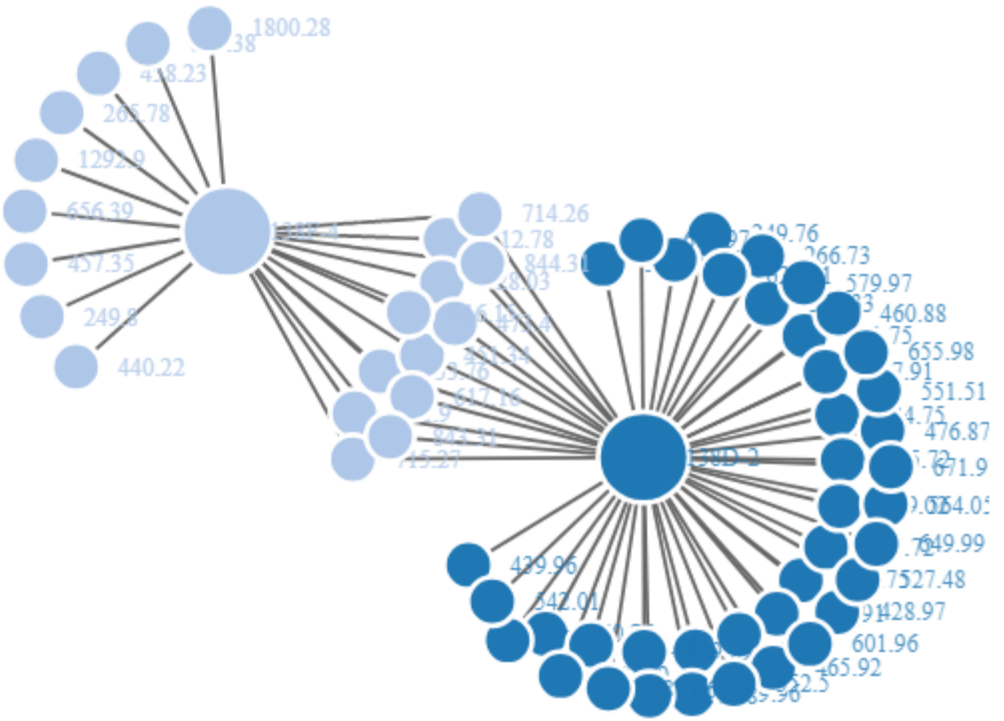
Group 49


137A-25
114B-10
137A-20

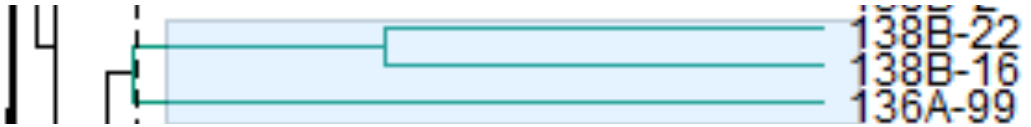

137A-20
137A-25
114B-10
212.73
233.75
211.71
261.94
458.18
655.41
451.49
844.44
348.08
256.94
440.17
258.00
1070.65
618.36
401.19
542.25
425.25
227.89
331.16
756.67

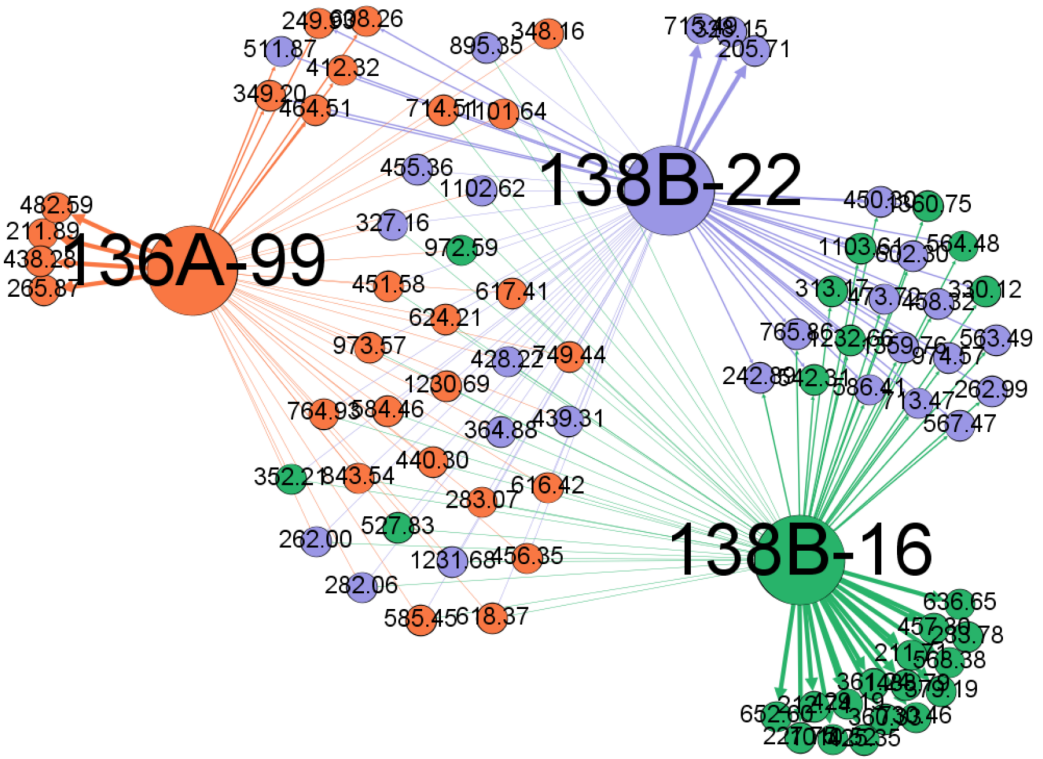
Group 50


138E-4
138D-2

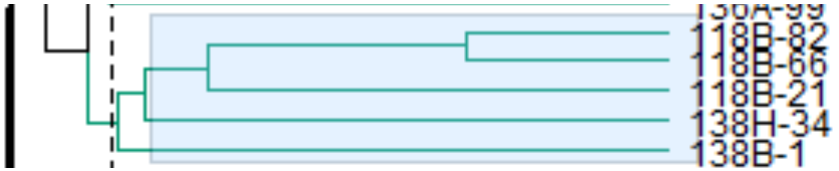

1800.28
438.23
265.78
1292.9
656.39
457.35
249.8
440.22
714.26
844.30
266.73
579.97
460.88
655.98
551.51
476.87
649.99
527.48
428.97
601.96
465.92
439.96
542.01

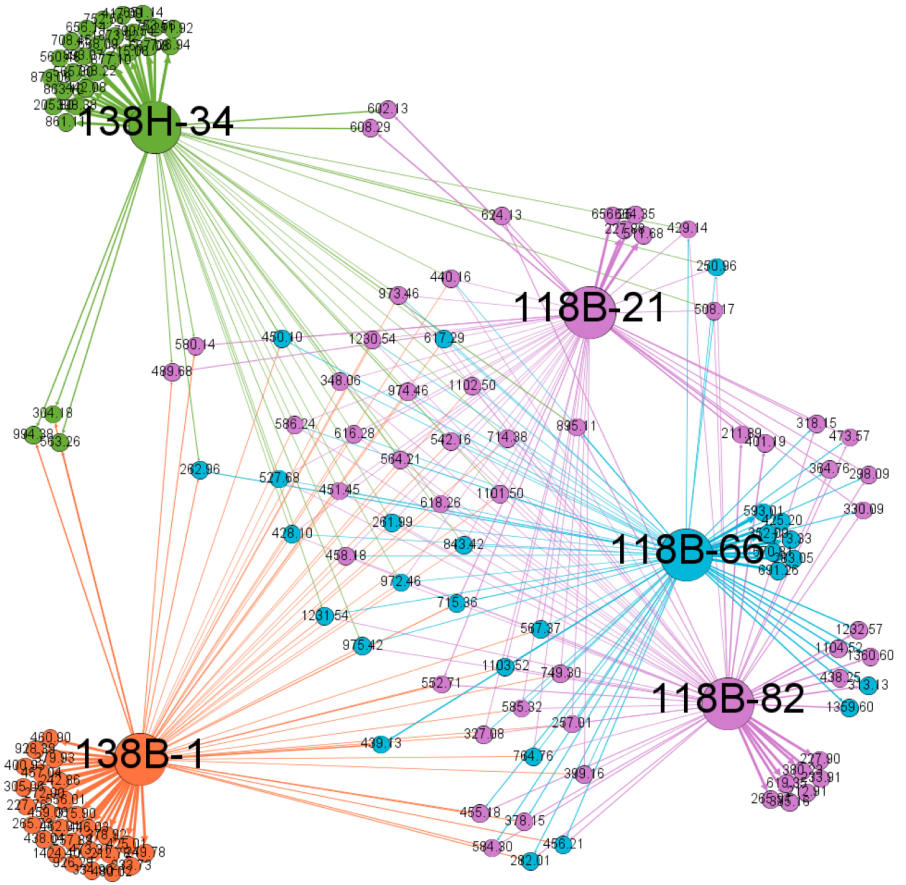
Group 51


138B-22
138B-16
136A-99

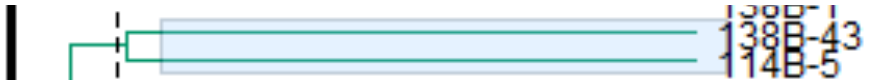

138B-22
136A-99
138B-16
482.59
211.89
438.28
265.87
511.87
412.32
349.20
464.51
895.35
348.16
205.71
455.36
1102.62
327.16
972.59
451.58
617.41
624.21
973.57
749.44
1230.69
764.93
584.46
364.88
439.31
440.30
616.42
283.07
527.83
262.00
1231.68
456.35
282.06
618.37
564.48
330.12
563.49
262.99
567.47
636.65
568.38
379.19

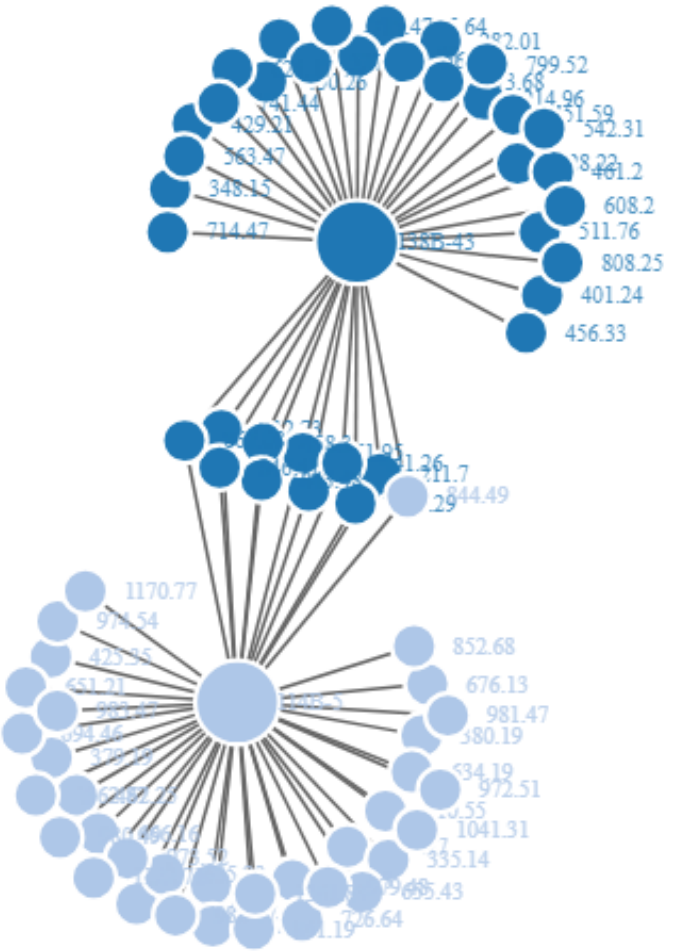
Group 52

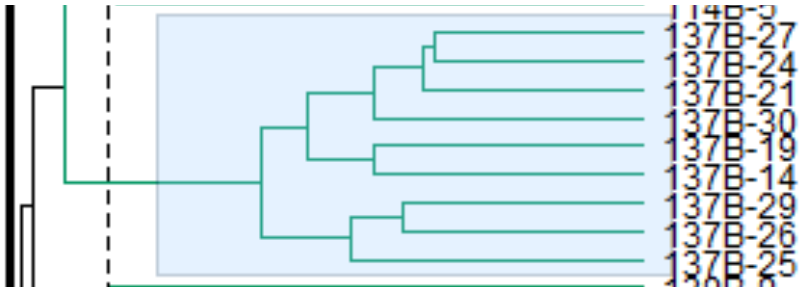

138A-99
118B-82
118B-66
118B-21
138H-34
138B-1


138H-34
118B-21
118B-66
118B-82
138B-1

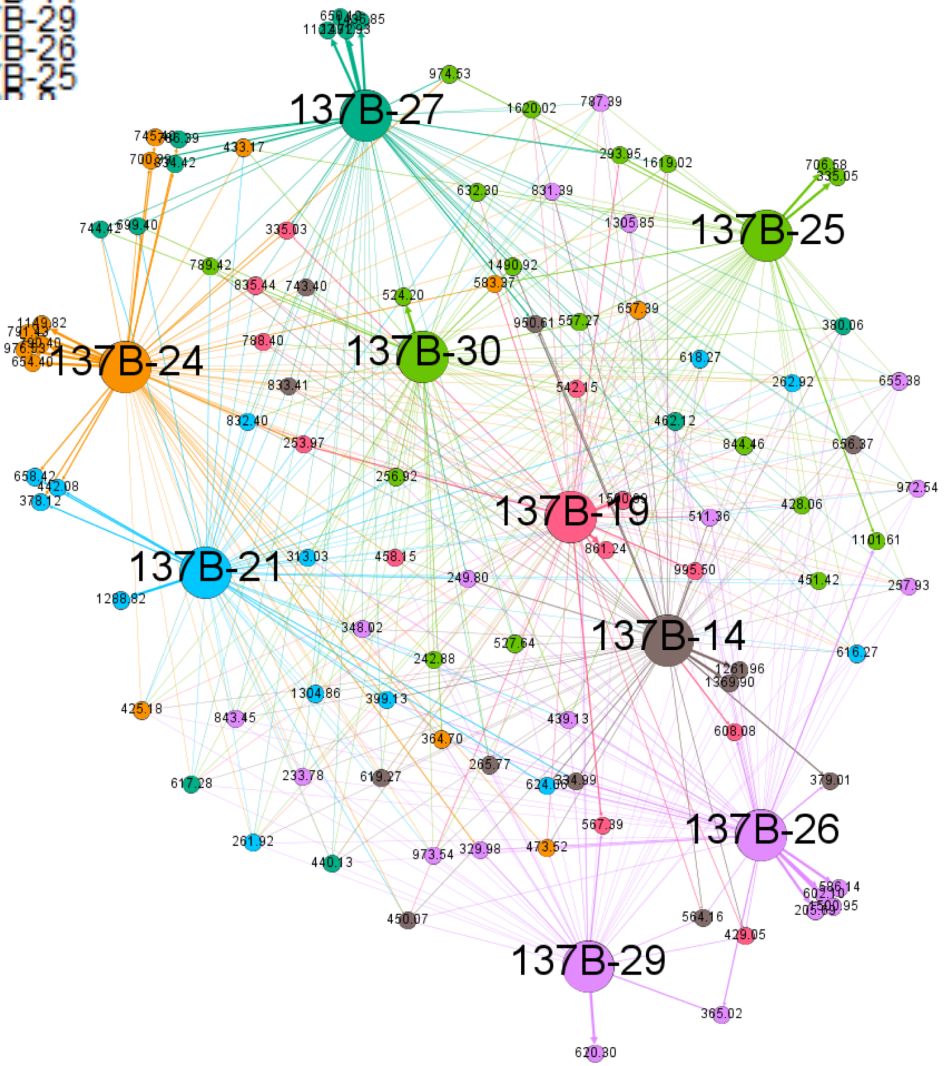
Group 53


138B-43
114B-5

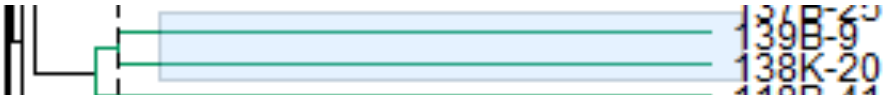

799.52
542.31
461.2
608.2
511.76
808.25
401.24
456.33
348.15
714.47
1170.77
852.68
676.13
981.47
972.51
1041.31
335.14

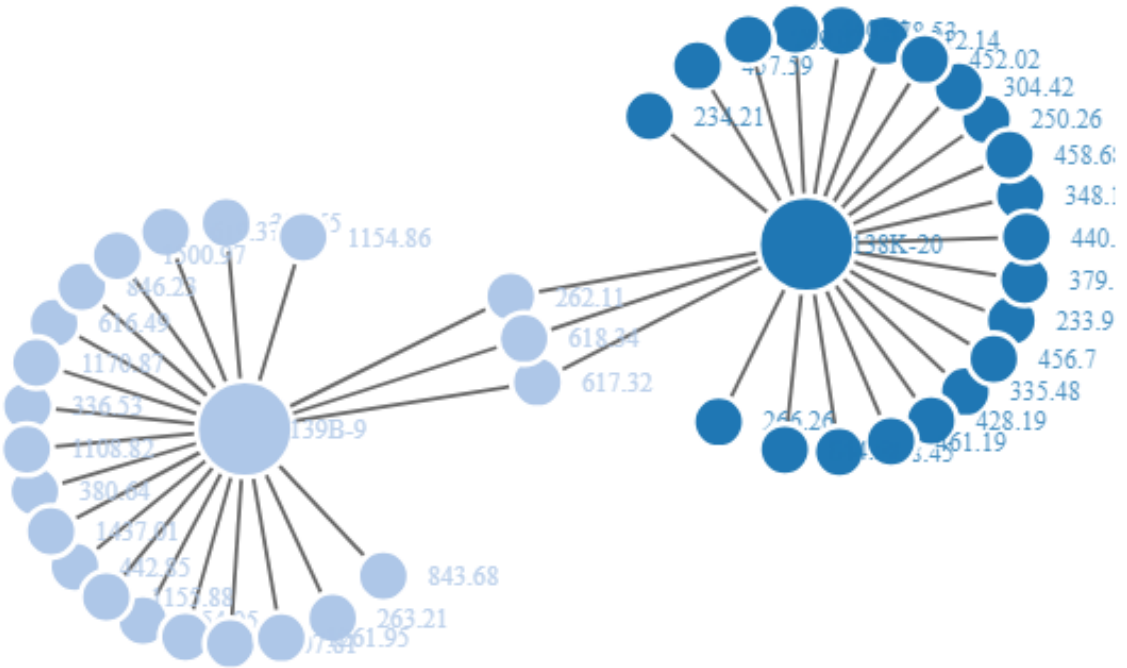

137B-27
137B-24
137B-21
137B-30
137B-19
137B-14
137B-29
137B-26
137B-25

Group 54

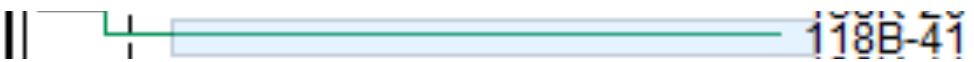

137B-27
137B-24
137B-30
137B-25
137B-19
137B-21
137B-14
137B-26
137B-29

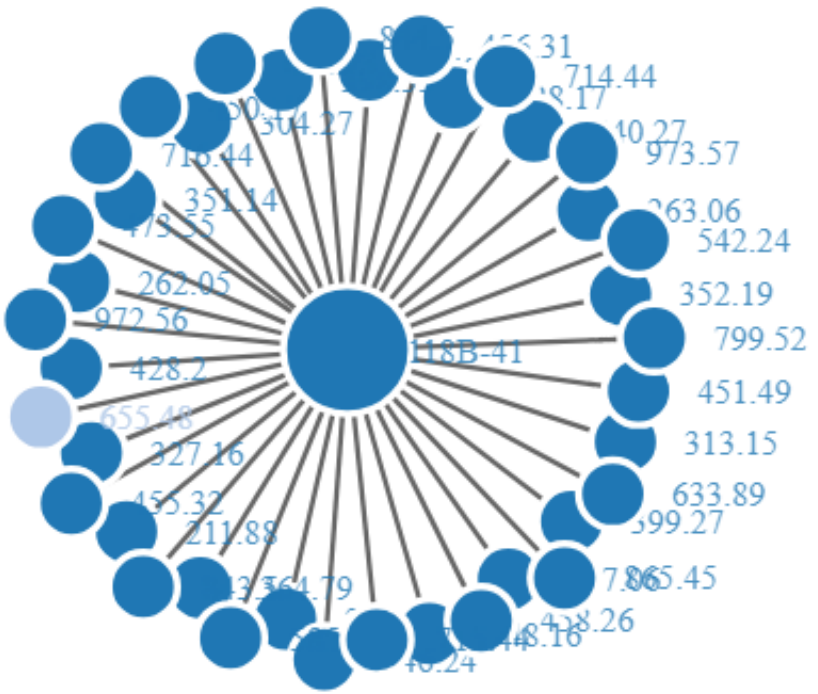
Group 55


139B-9
138K-20

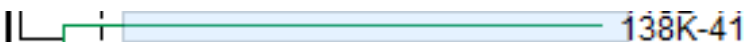

1154.86
262.11
618.24
617.32
843.68
263.21
234.21
452.02
304.42
250.26
458.6
348.
440.
379.
233.9
456.7
335.48
428.19
461.19

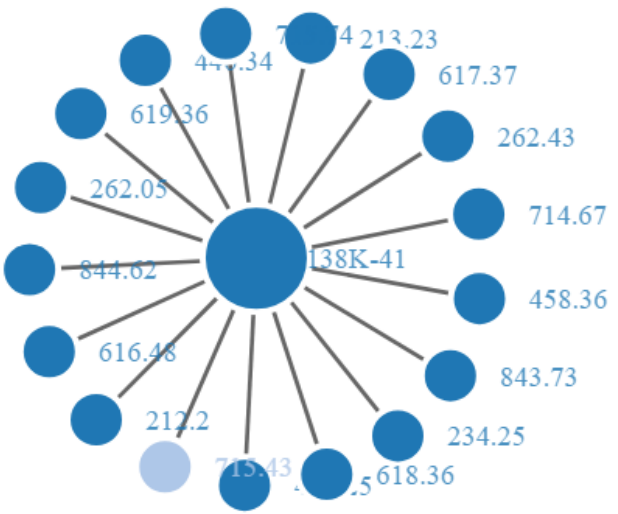
Group 56

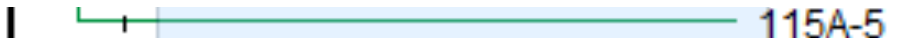

118B-41

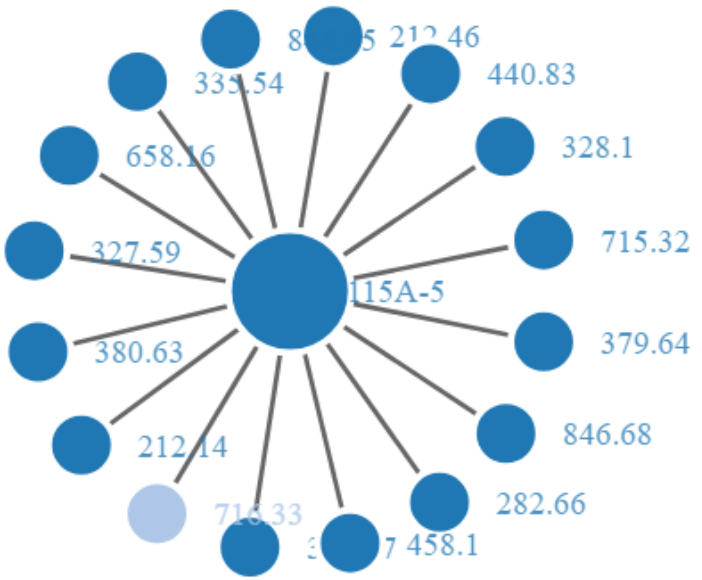

714.44
973.57
542.24
352.19
799.52
451.49
313.15
633.89
262.05
972.56
428.2
655.48
327.16
211.88

Group 57

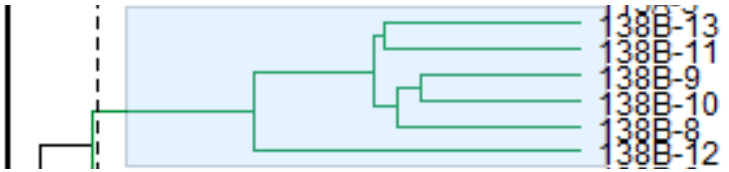

138K-41

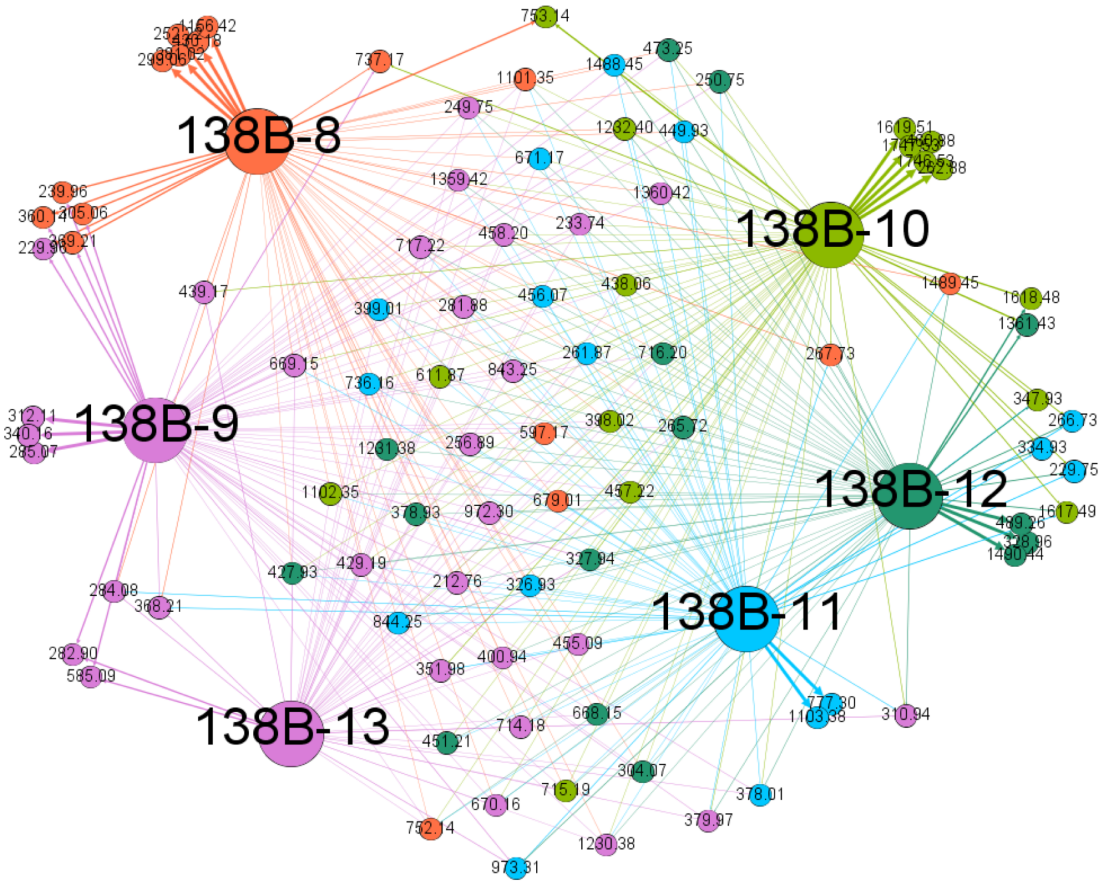

213.23
617.37
262.43
714.67
458.36
843.73
234.25
618.36
262.05
844.62
616.48
212.2
138K-41


115A-5


440.83
328.1
715.32
379.64
846.68
282.66
458.1
327.59
380.63
658.16

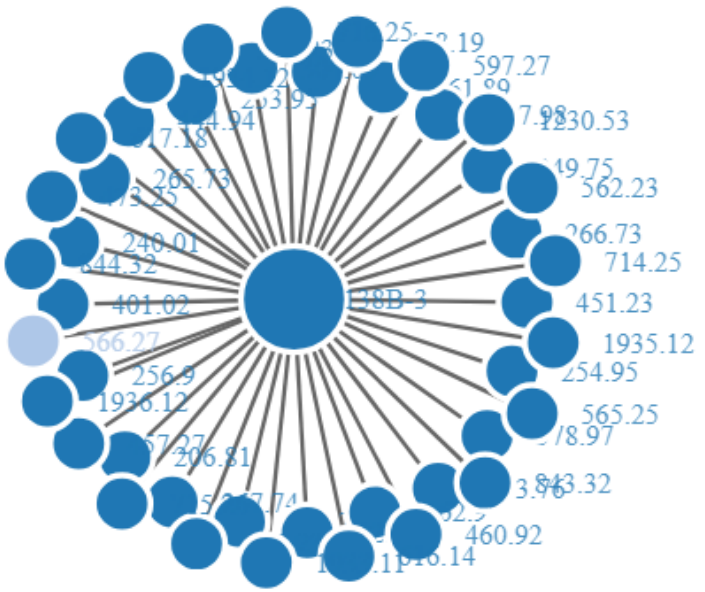
Group 59


138B-13
138B-11
138B-9
138B-10
138B-8
138B-12

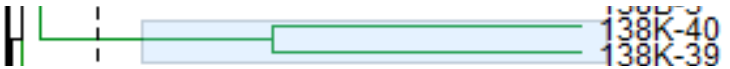

138B-8
138B-10
138B-9
138B-12
138B-11
138B-13

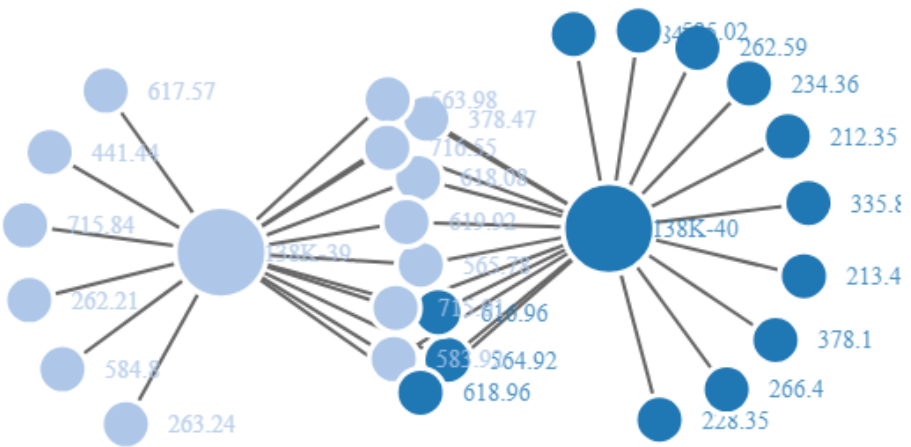
Group 60


138B-3

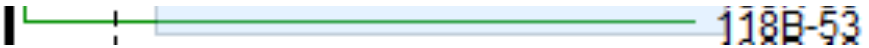

597.27
1230.53
562.23
266.73
714.25
451.23
1935.12
254.95
565.25
460.92
240.01
401.02
256.9
206.81

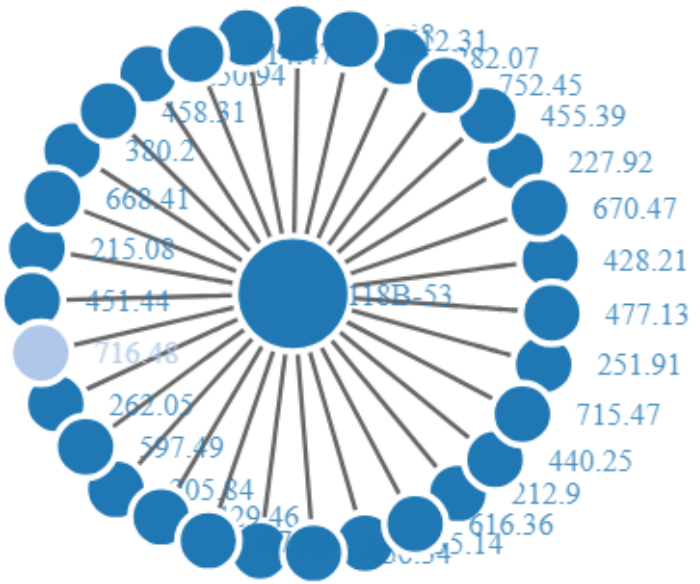
Group 61

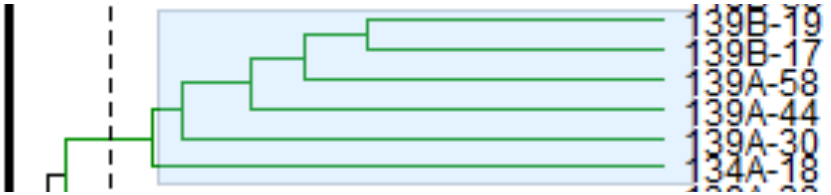

138K-40
138K-39


617.57
441.44
715.84
262.21
584.8
263.24
563.98
378.47
618.08
619.92
565.78
618.96
564.92
618.96
262.59
234.36
212.35
335.8
213.4
378.1
266.4
228.35
138K-40

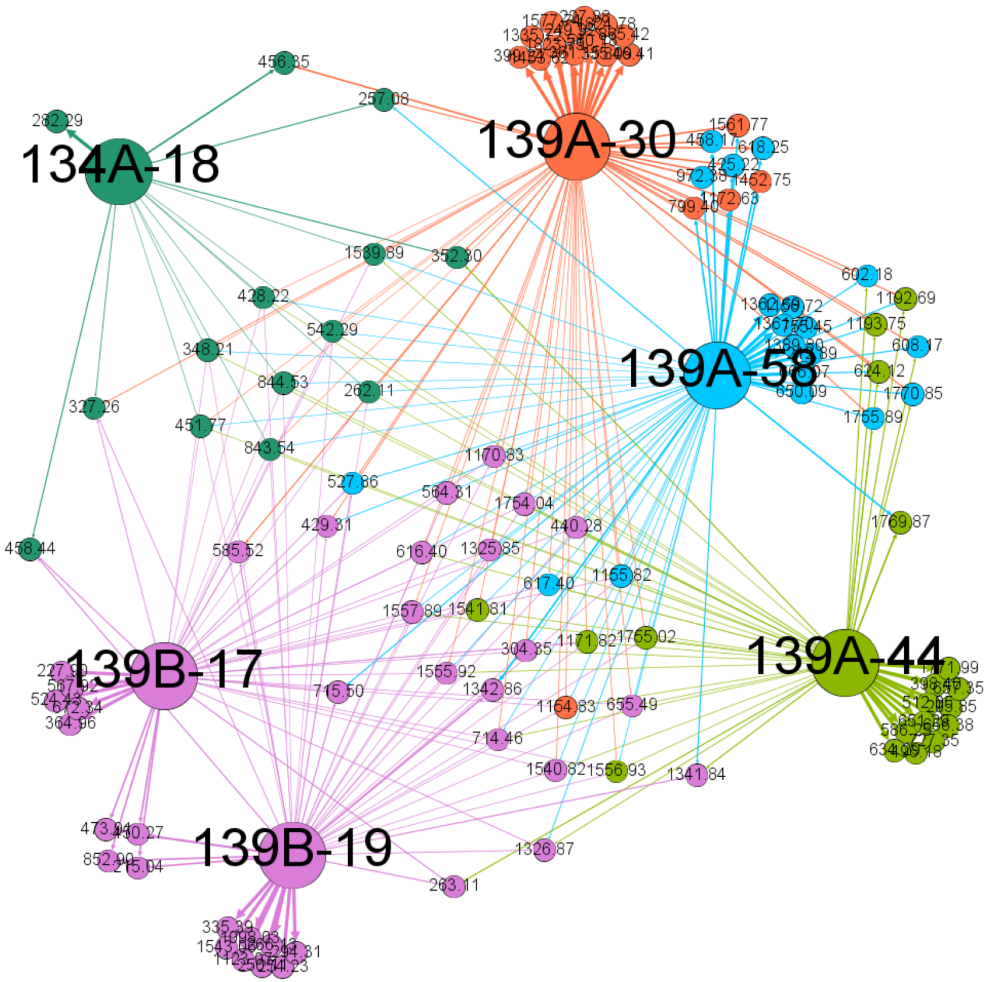
Group 62


118B-53

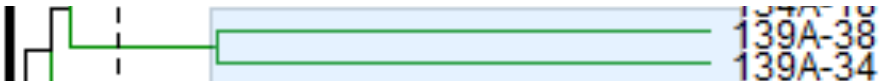

752.45
455.39
227.92
670.47
428.21
477.13
251.91
715.47
440.25
212.9
616.36
458.31
380.2
668.41
215.08
451.44
716.48
262.05
597.49

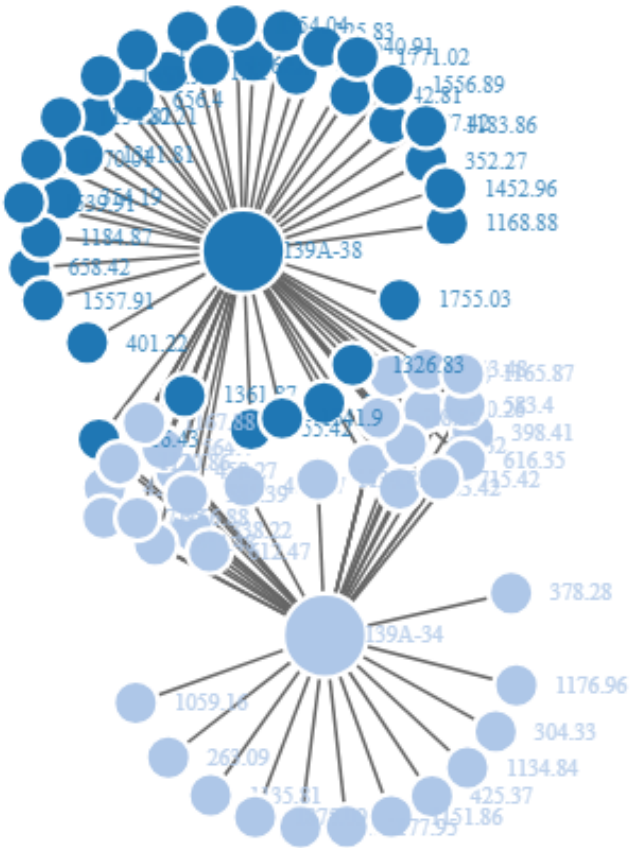

139B-19
139B-17
139A-58
139A-44
139A-30
134A-18

Group 63

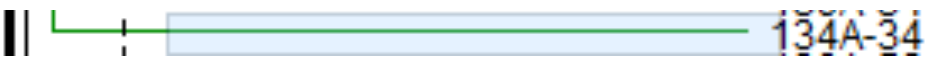

139A-30
134A-18
139A-58
139A-44
139B-17
139B-19

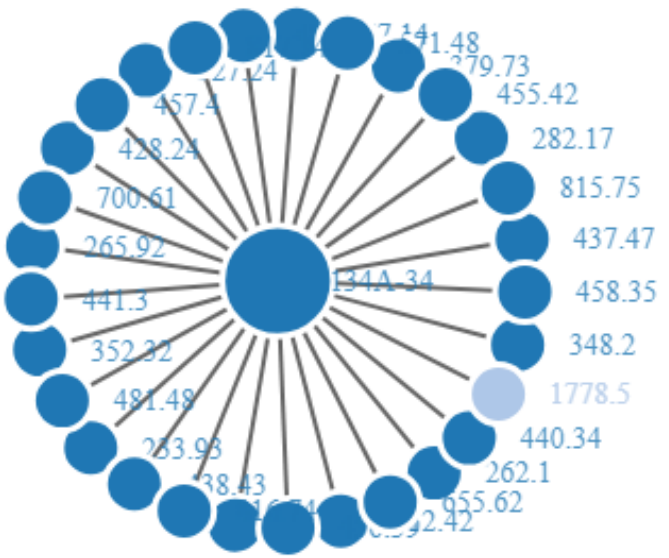
Group 64


139A-38
139A-34

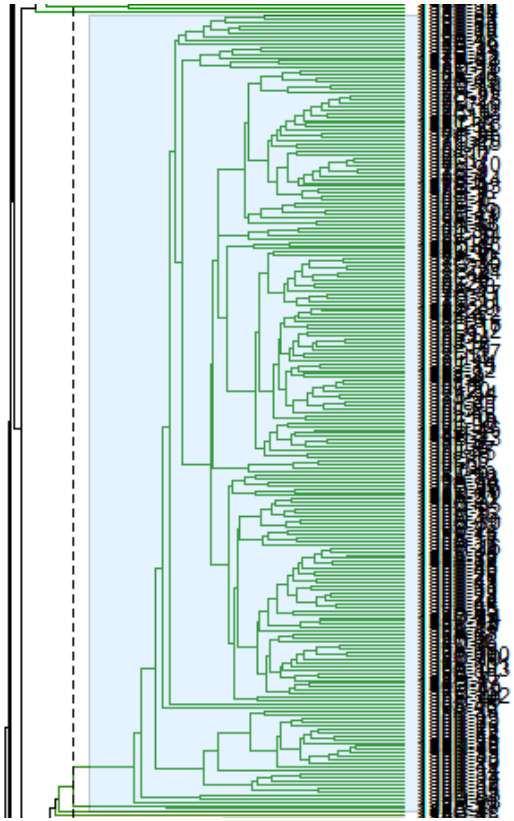

1771.02
1556.89
352.27
1452.96
1168.88
139A-38
1755.03
1326.83
1165.87
398.41
616.35
378.28
139A-34
1176.96
304.33
1134.84
425.37

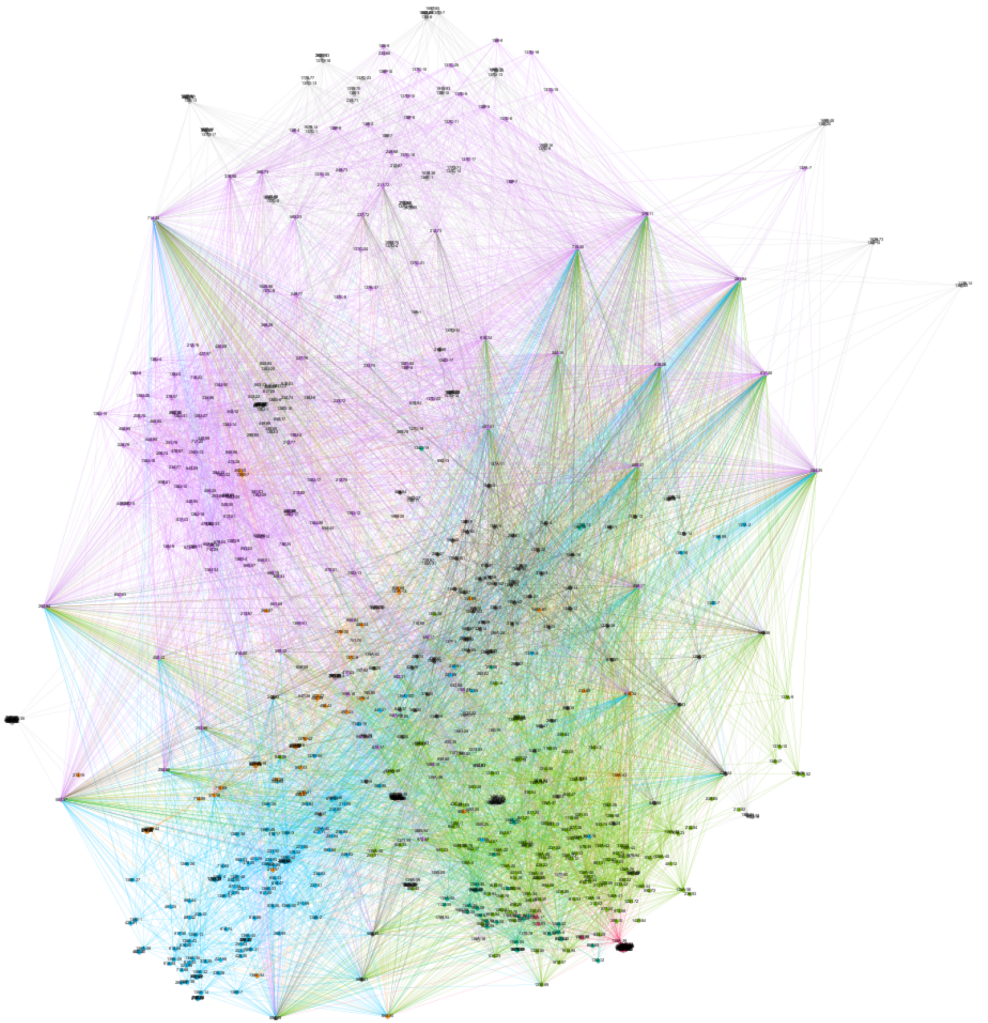
Group 65


134A-34

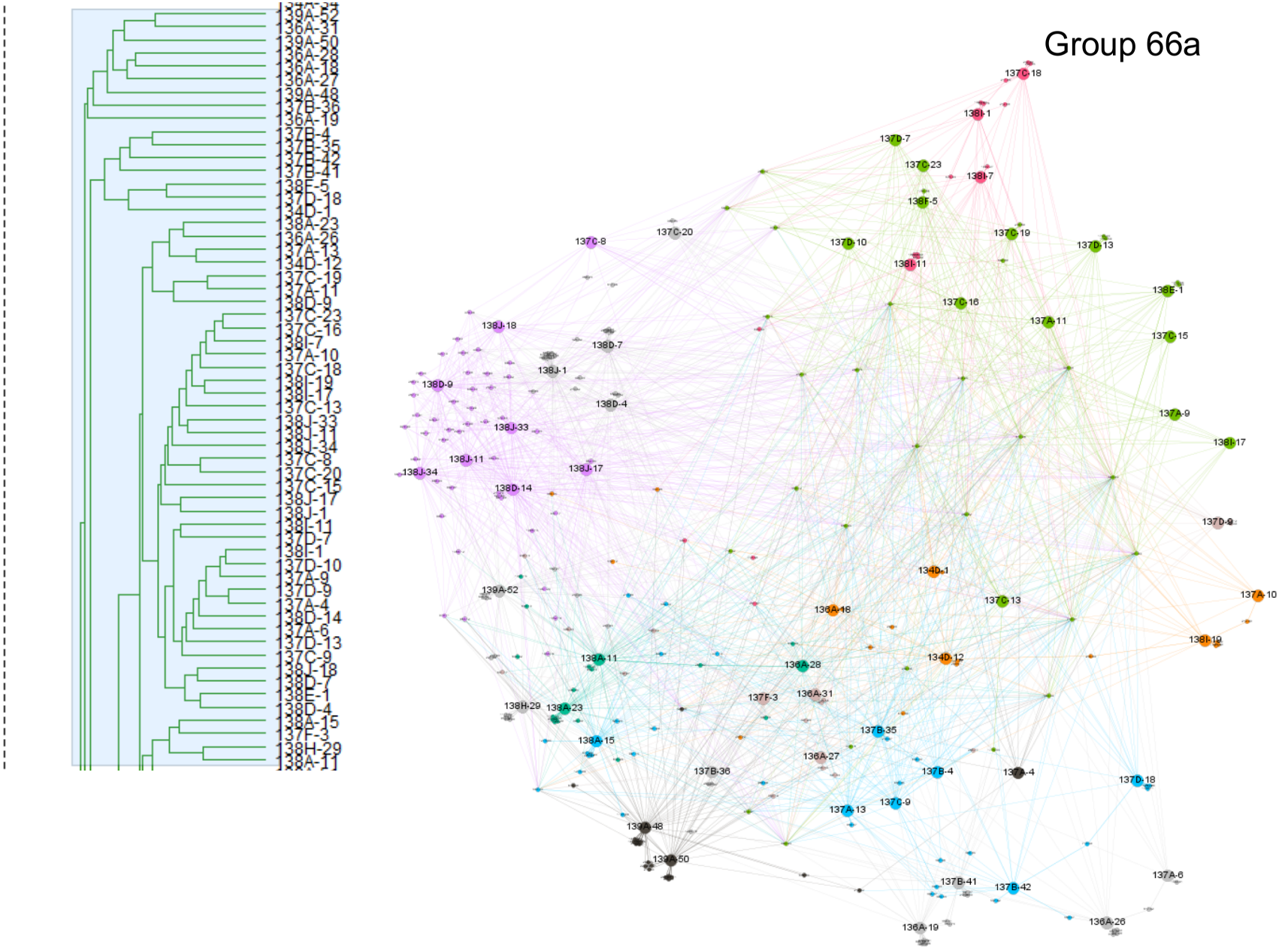

455.42
282.17
815.75
437.47
458.35
348.2
1778.5
440.34
262.1
700.61
265.92
441.3
352.32
481.48

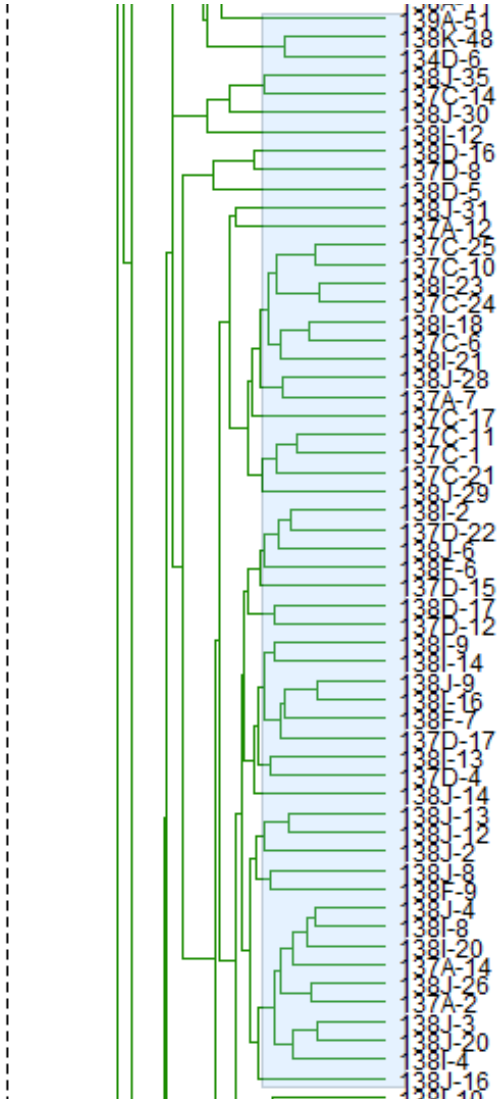
Group 66

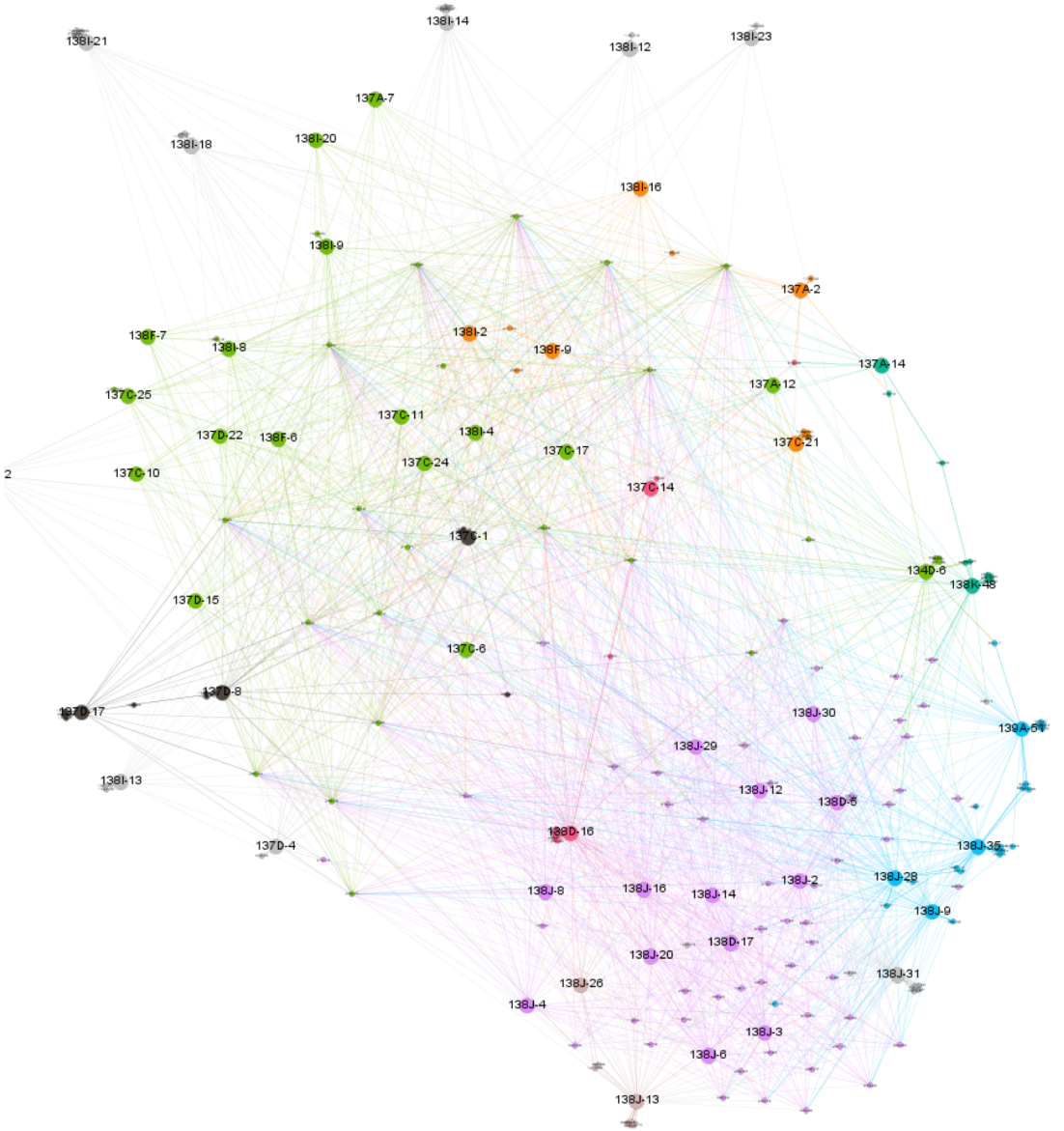


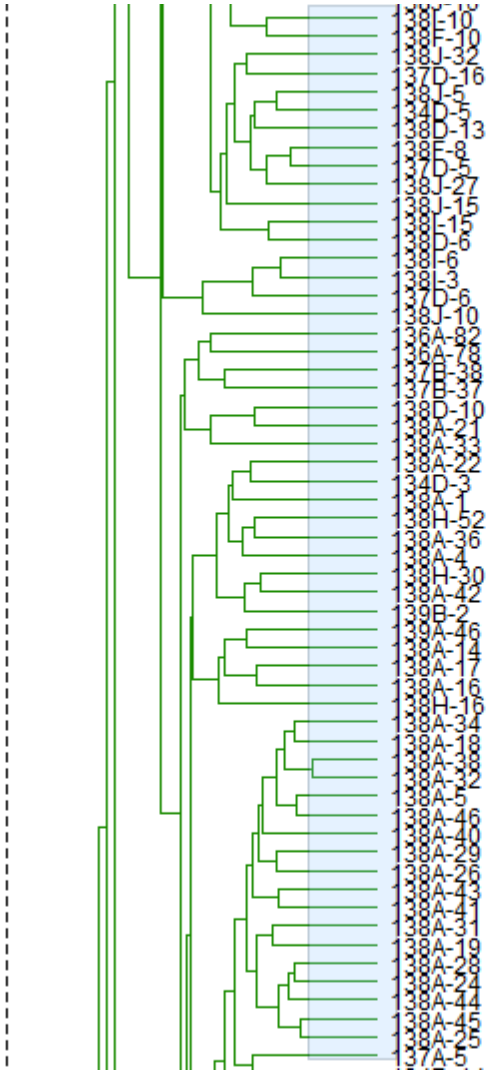


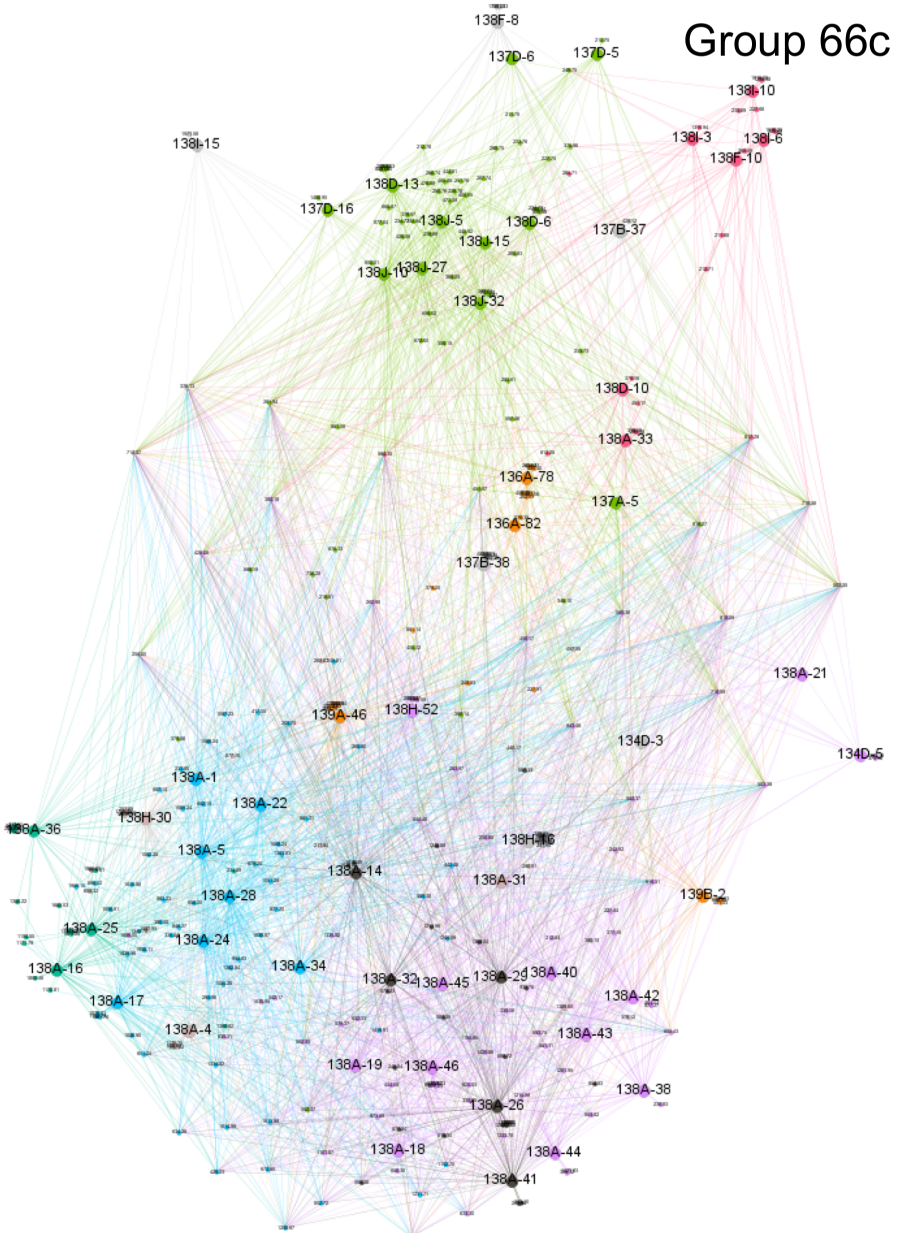
*Note: This group was subdivided into 4 groups, since it contained greater than 200 strains.

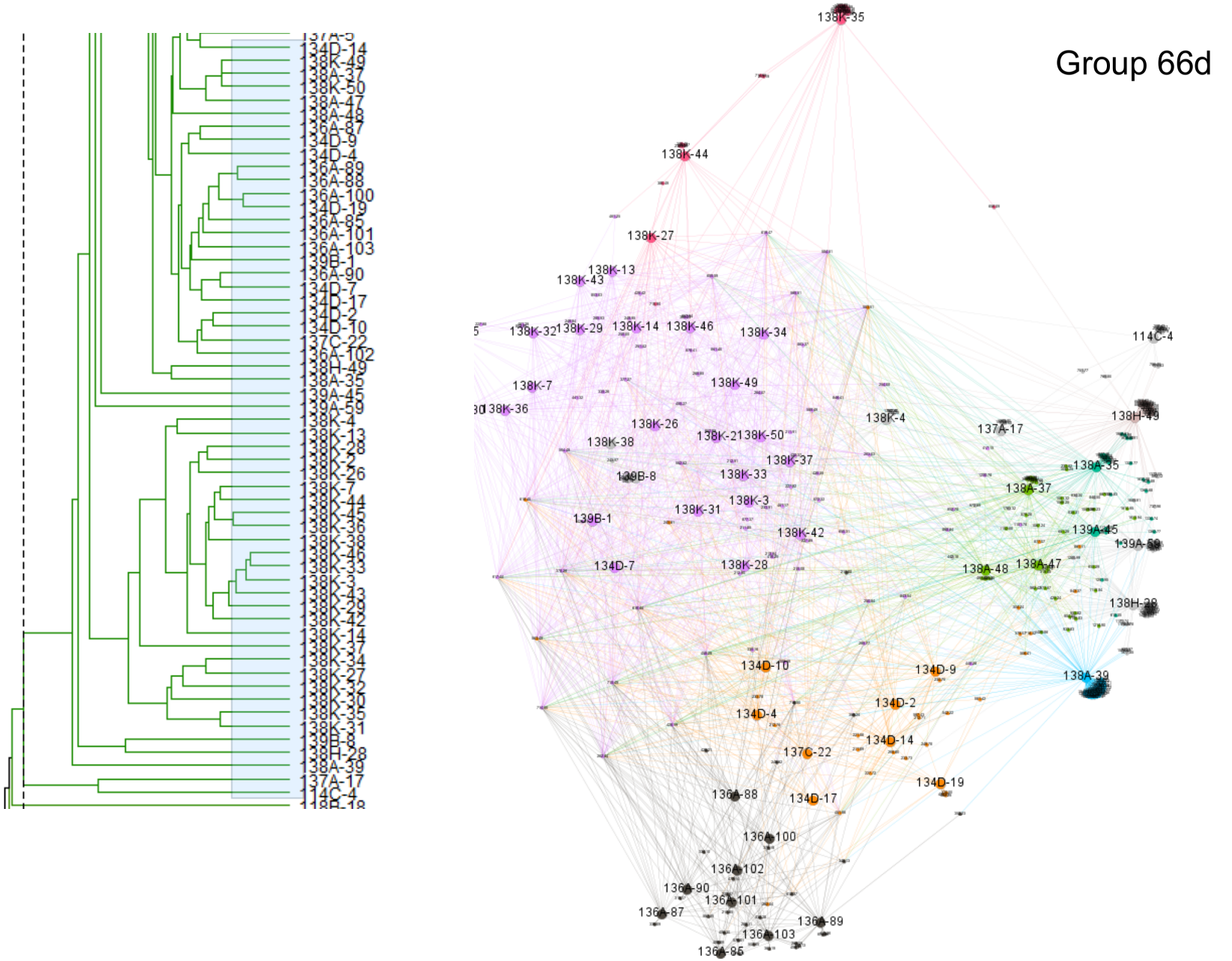

Group 66a
137C-18
138I-1
137D-7
137C-23
138I-7
138F-5
137C-19
137D-13
137C-20
137C-8
137D-10
138I-11
138E-1
137C-16
137A-11
137C-15
138J-18
138D-7
138J-1
138D-9
138D-4
137A-9
138J-33
138I-17
138J-11
138J-34
138J-17
138D-14
137D-9
134D-1
139A-52
137C-13
137A-10
136A-18
138I-19
138A-11
136A-28
134D-12
137F-3
136A-31
138H-29
138A-23
137B-35
138A-15
136A-27
137B-36
137B-4
137A-4
137D-18
137C-9
137A-13
139A-48
139A-50
137A-6
137B-41
137B-42
136A-26
136A-19



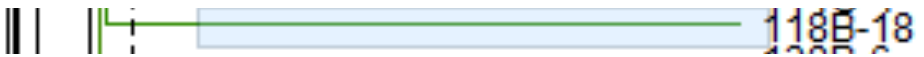

Group 66b
138I-14
138I-21
138I-12
138I-23
137A-7
138I-18
138I-20
138I-16
138I-9
137A-2
138F-7
138I-8
138I-2
138F-9
137A-14
137A-12
137C-25
137C-11
137D-22
138F-6
138I-4
137C-21
137C-17
137C-24
137C-10
137C-14
137C-1
138D-6
138I-48
137D-15
137C-6
137D-8
137D-17
138J-30
139A-51
138J-29
138I-13
138J-12
138D-5
138D-16
137D-4
138J-35
138J-28
138J-8
138J-16
138J-14
138J-2
138J-9
138D-17
138J-20
138J-31
138J-26
138J-4
138J-3
138J-6
138J-13

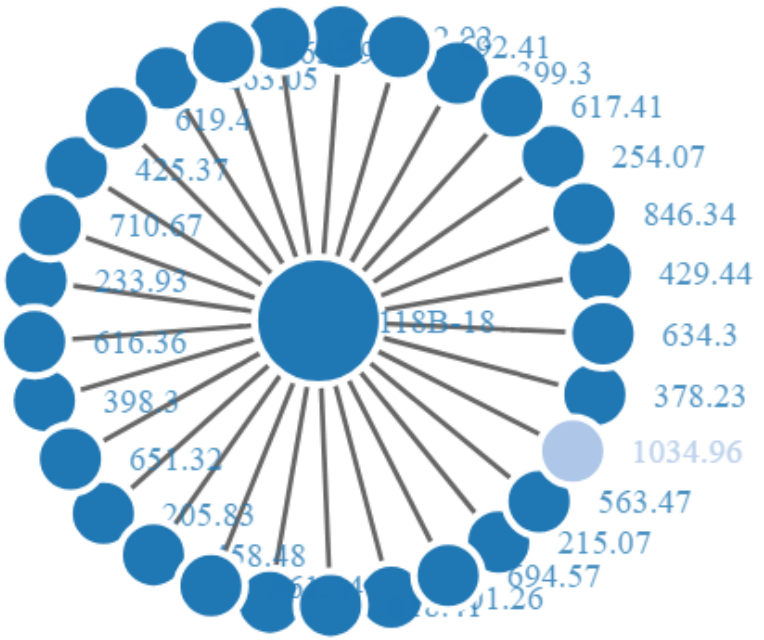

138I-10
138F-10
138J-32
137D-16
138J-5
134D-5
138D-13
138F-8
137D-5
138J-27
138J-15
138I-15
138D-6
138I-6
138I-3
137D-6
138J-10
136A-82
136A-78
137B-38
137B-37
138D-10
138A-21
138A-33
138A-22
134D-3
138A-1
138H-52
138A-36
138A-4
138H-30
138A-42
139B-2
139A-46
138A-14
138A-17
138A-16
138H-16
138A-34
138A-18
138A-38
138A-32
138A-5
138A-46
138A-40
138A-29
138A-26
138A-43
138A-31
138A-19
138A-28
138A-24
138A-44
138A-45
138A-25
137A-5

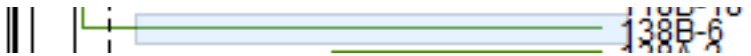

Group 66c
138F-8
137D-6
137D-5
138I-10
138I-15
138I-3
138I-6
138F-10
138D-13
137D-16
138J-5
138D-6
137B-37
138J-15
138J-10
138J-27
138J-32
138D-10
138A-33
136A-78
137A-5
136A-82
137B-38
138A-21
139A-46
138H-52
134D-3
134D-5
138A-1
138A-22
138H-30
138A-36
138A-5
138H-16
138A-14
138A-31
139B-2
138A-28
138A-25
138A-24
138A-16
138A-34
138A-32
138A-45
138A-29
138A-40
138A-42
138A-17
138A-4
138A-43
138A-19
138A-46
138A-38
138A-26
138A-18
138A-44
138A-41

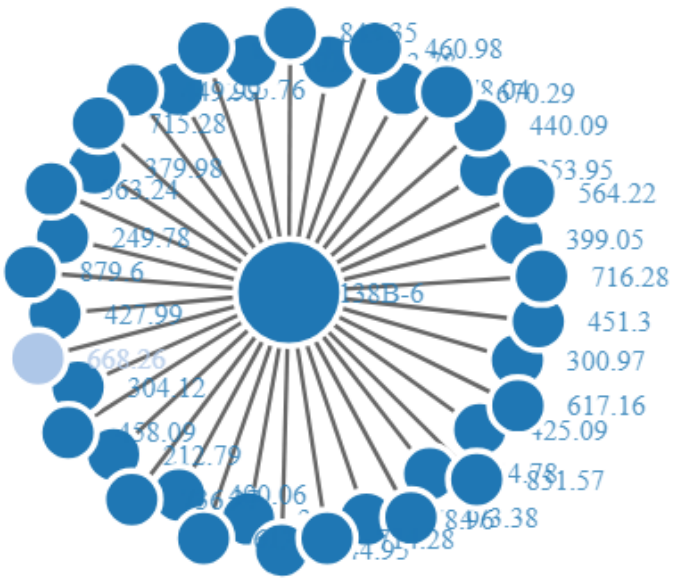

Group 66d
138K-35
138K-44
138K-27
138K-13
138K-43
138K-32
138K-29
138K-14
138K-46
138K-34
138K-7
138K-49
138K-36
138K-26
138K-38
138K-21
138K-50
138K-37
139B-8
138K-33
138K-3
138K-31
139B-1
138K-42
134D-7
138K-28
138K-4
114C-4
138H-49
137A-17
138A-35
138A-37
139A-45
139A-59
138A-48
138A-47
138H-28
138A-39
134D-10
134D-9
134D-2
134D-4
134D-14
137C-22
134D-19
136A-88
134D-17
136A-100
136A-102
136A-90
136A-101
136A-87
136A-89
136A-103
136A-85

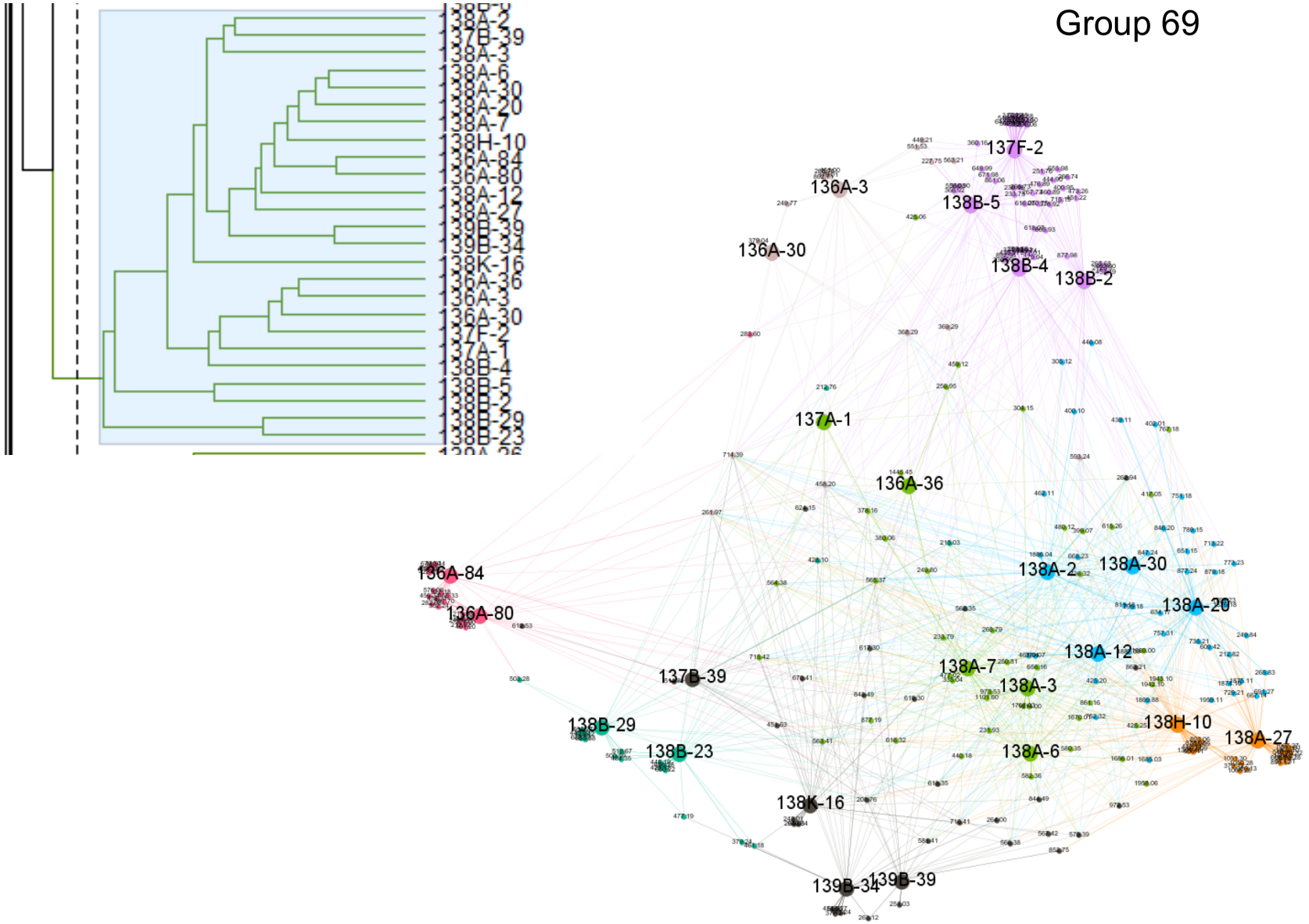
Group 67


118B-18

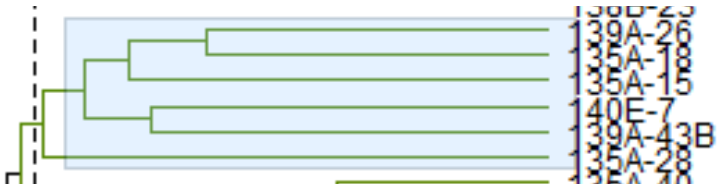

617.41
254.07
846.34
429.44
634.3
378.23
1034.96
563.47
215.07
694.57
619.4
425.37
710.67
233.93
616.36
398.3
651.32
118B-18


138B-6

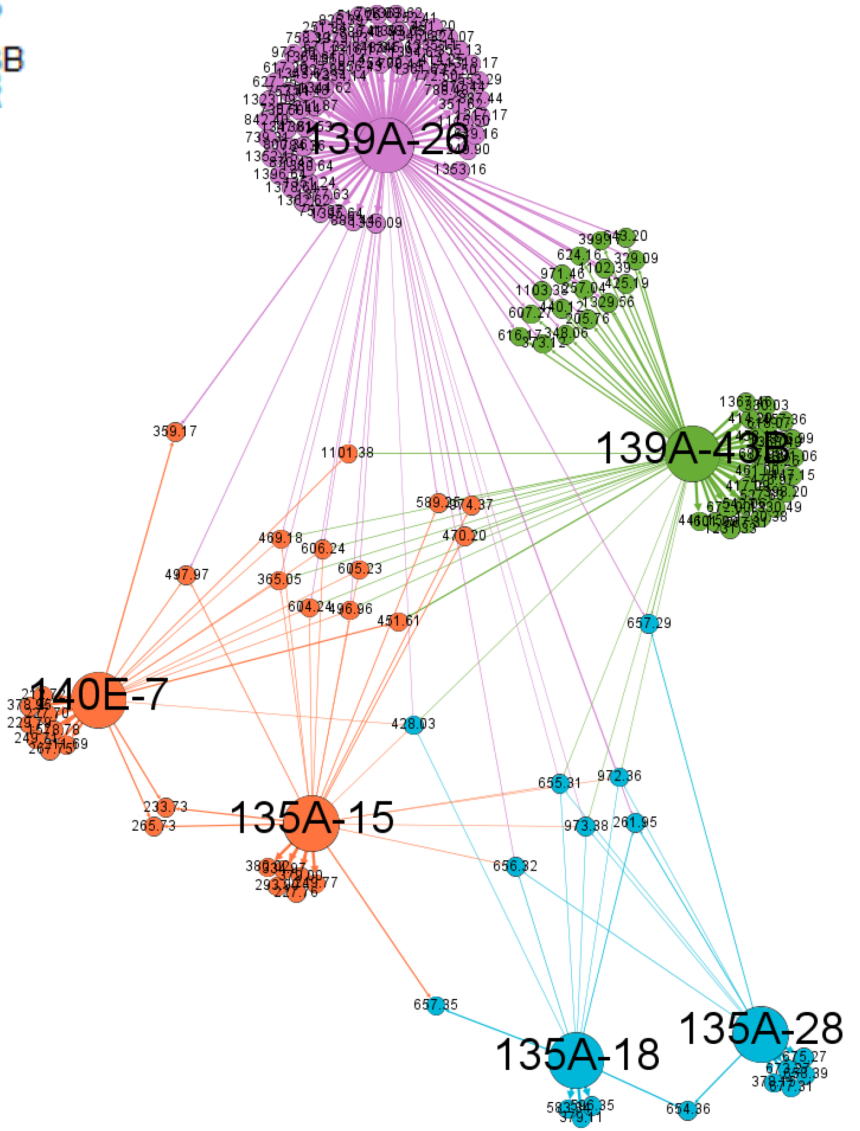

460.98
440.09
564.22
399.05
716.28
451.3
300.97
617.16
138B-6


Group 69
137F-2
136A-3
138B-5
136A-30
138B-4
138B-2
137A-1
136A-36
136A-84
136A-80
138A-2
138A-30
138A-20
138A-12
138A-7
137B-39
138A-3
138B-29
138H-10
138A-27
138B-23
138A-6
138K-16
139B-34
139B-39

Group 70

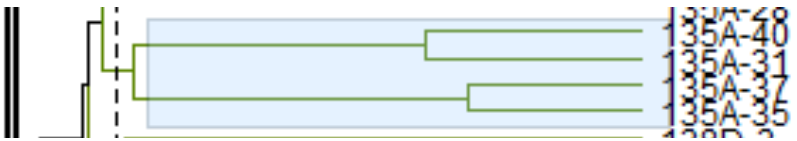


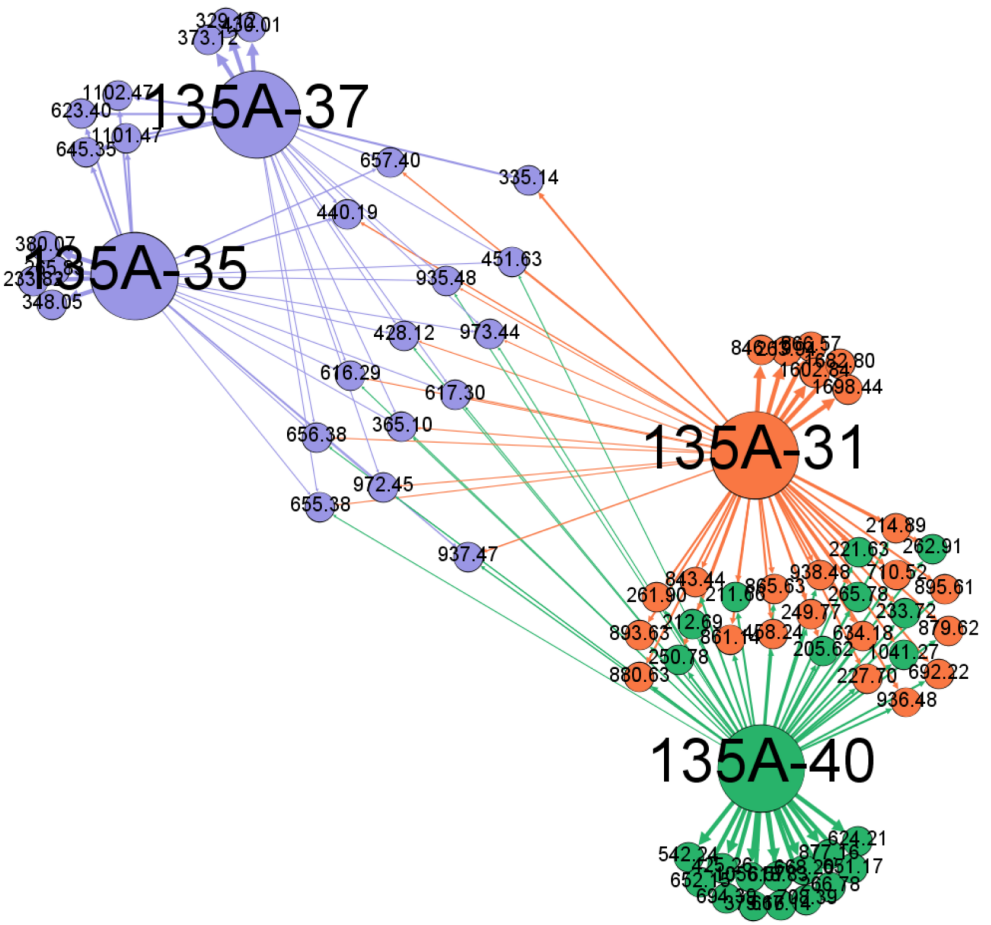
139A-26
135A-18
135A-15
140E-7
139A-43B
135A-28



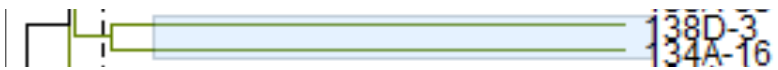
139A-26
139A-43B
140E-7
135A-15
135A-18
135A-28

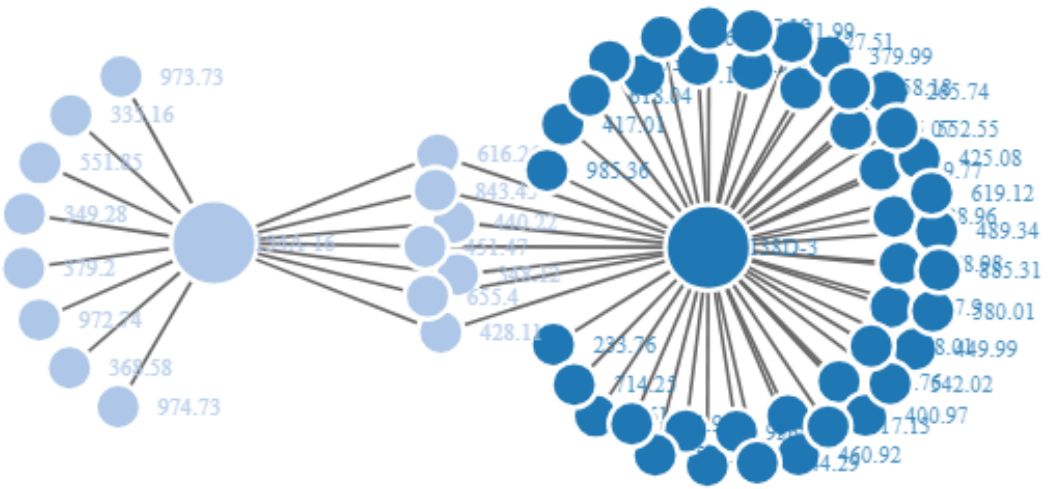
Group 71


135A-40
135A-31
135A-37
135A-35

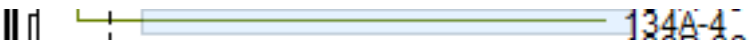

135A-37
135A-35
135A-31
135A-40
373.12
1102.47
623.40
1101.47
645.35
657.40
335.14
440.19
380.07
451.63
935.48
348.05
428.12
973.44
616.29
617.30
365.10
656.38
972.45
655.38
937.47
1698.44
214.89
221.63
262.91
710.52
895.61
843.44
261.90
265.78
233.72
249.77
879.62
893.63
212.69
634.18
205.62
1041.27
250.78
880.63
227.70
692.22
936.48
624.21
542.24
1051.17

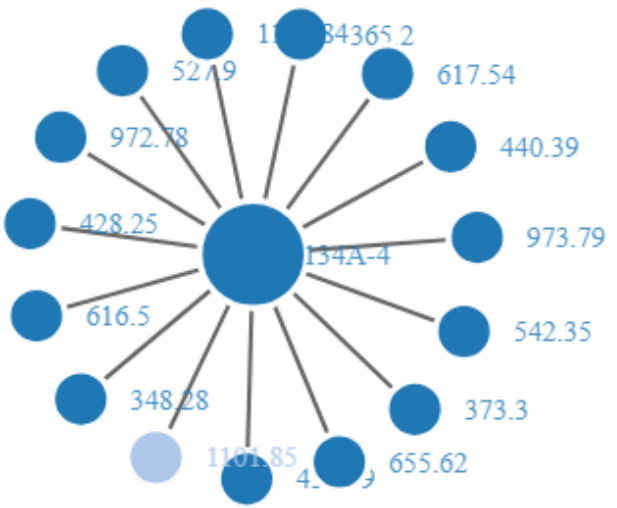
Group 72


138D-3
134A-16

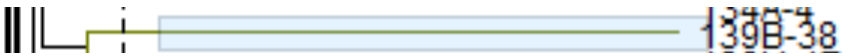

973.73
333.16
551.85
349.28
379.2
972.74
368.58
974.73
616.2
843.4
440.22
451.47
655.4
428.1
985.36
417.01
233.76
714.25
379.99
285.74
552.55
425.08
619.12
489.34
885.31
380.01
449.99
542.02
400.97
460.92

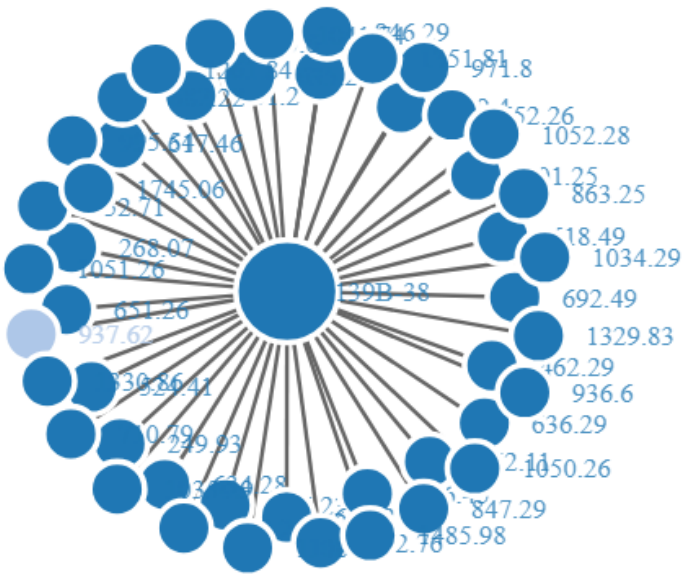
Group 73



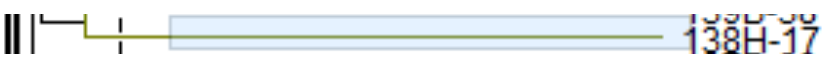


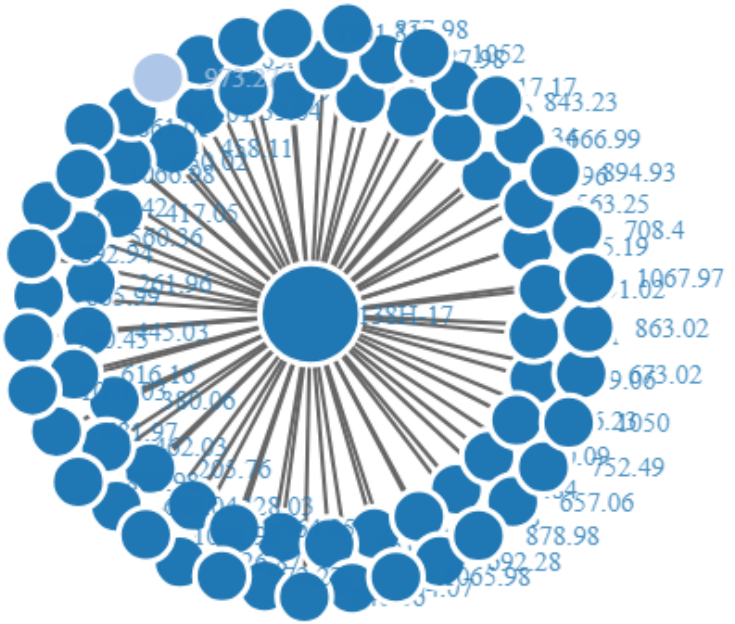
Group 74


139B-38

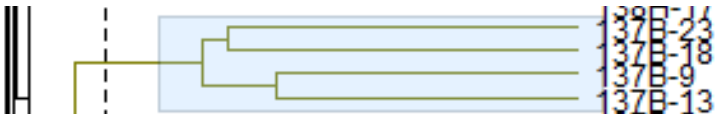

971.8
1052.28
863.25
1034.29
692.49
1329.83
936.6
1050.26
847.29

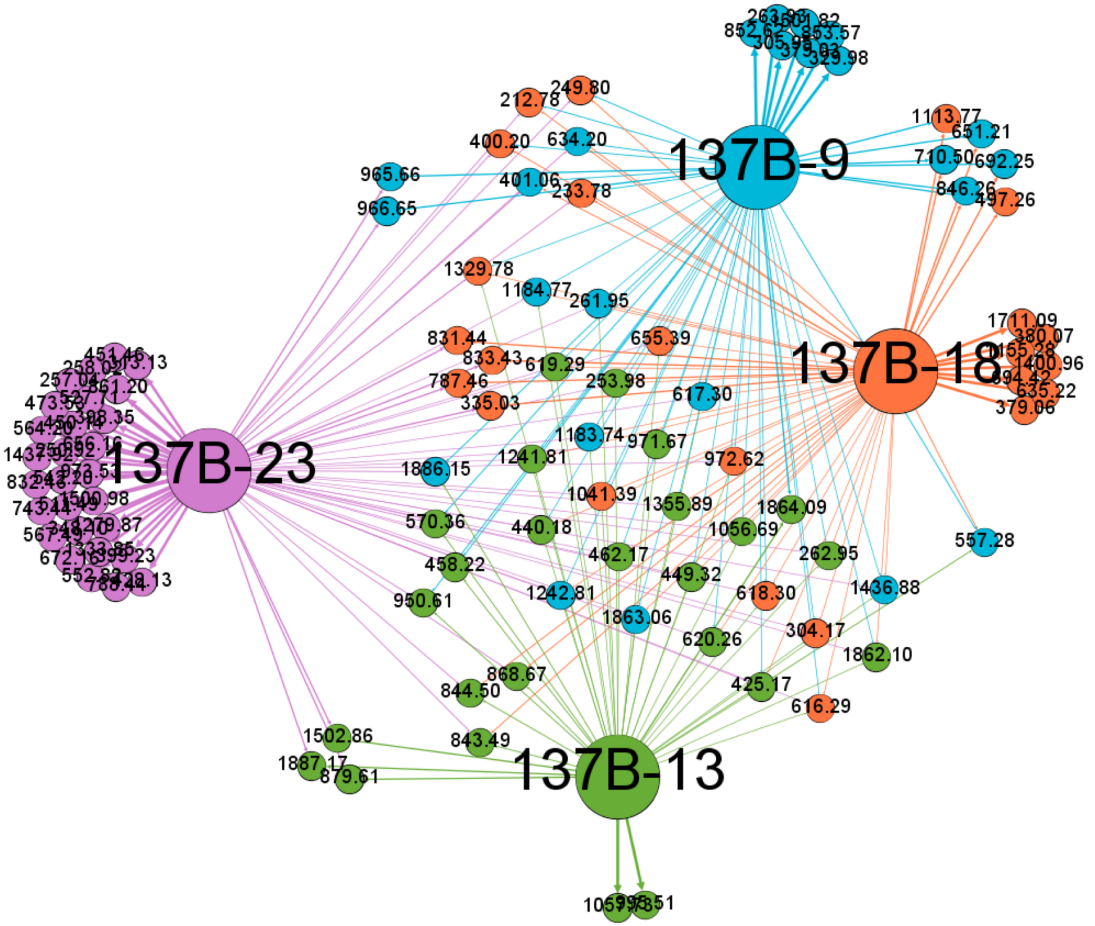
Group 75



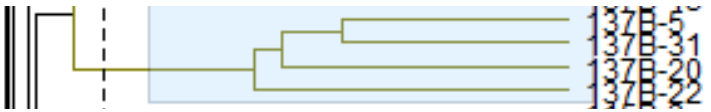


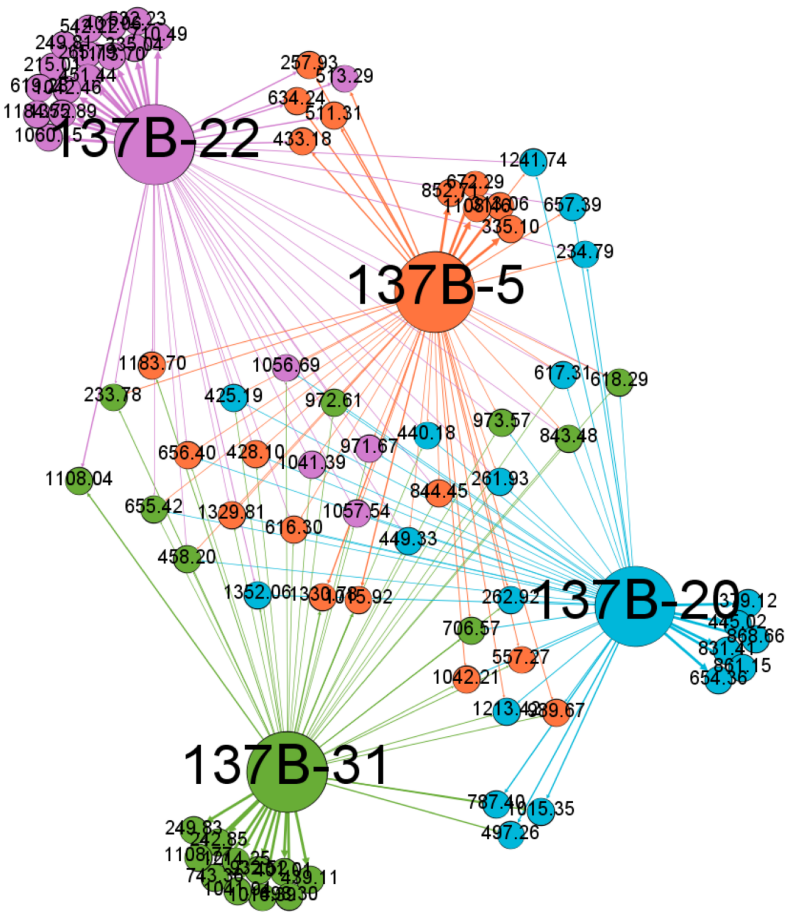
Group 76


137B-23
137B-18
137B-9
137B-13

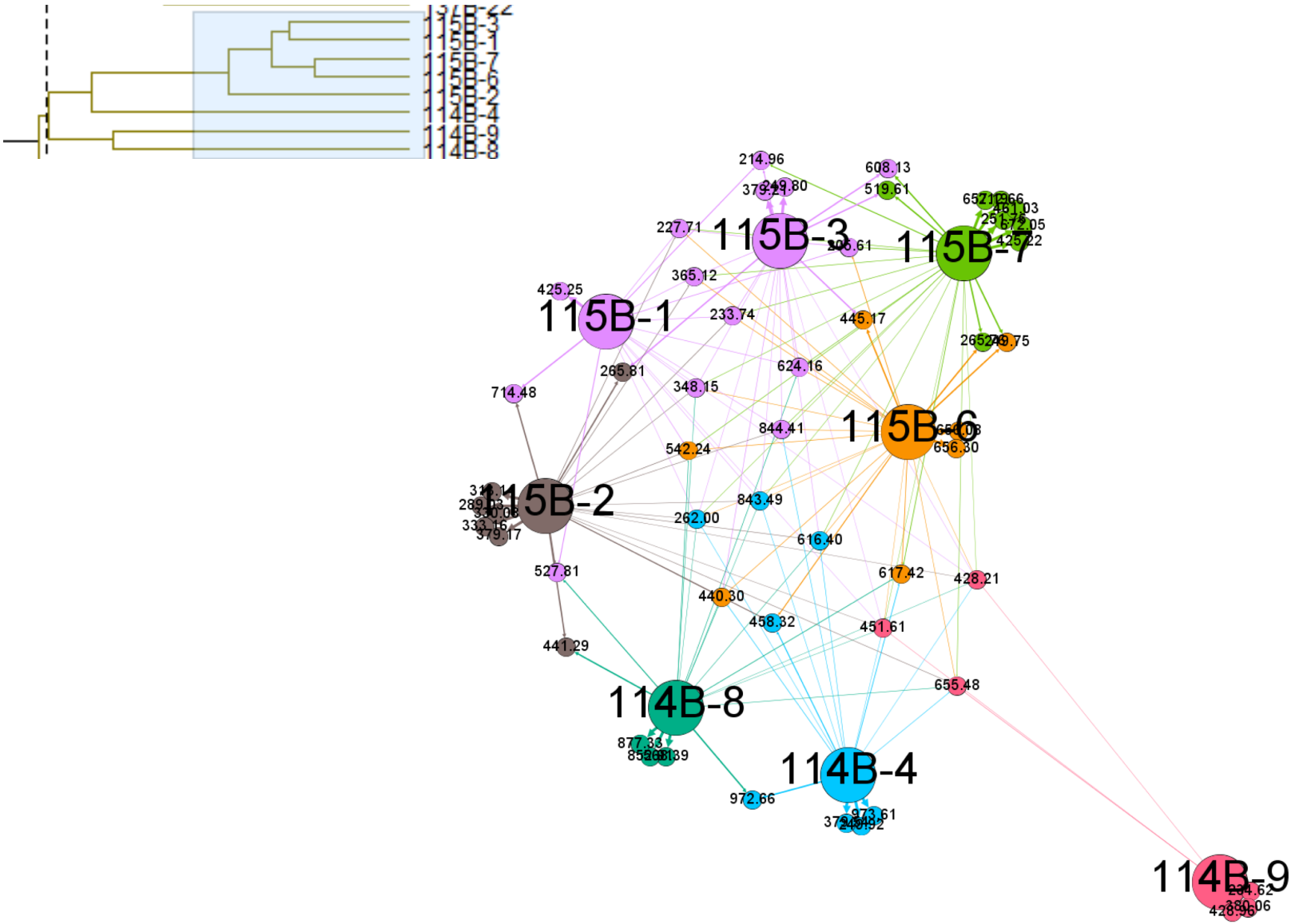

137B-9
137B-18
137B-23
137B-13
249.80
212.78
400.20
634.20
965.66
401.06
233.78
966.65
1329.78
1184.77
261.95
831.44
655.39
833.43
619.29
787.46
253.98
335.03
617.30
1183.74
971.67
1241.81
1886.15
972.62
1041.39
1355.89
1864.09
570.36
440.18
1056.69
462.17
262.95
458.22
449.32
950.61
1242.81
618.30
1436.88
1863.06
620.26
304.17
1862.10
868.67
844.50
425.17
616.29
1502.86
843.49
1887.17
879.61
557.28
1113.77
651.21
710.50
692.25
846.26
497.26
1711.09
380.07
1400.96
535.22
379.06

Group 77


137B-5
137B-31
137B-20
137B-22

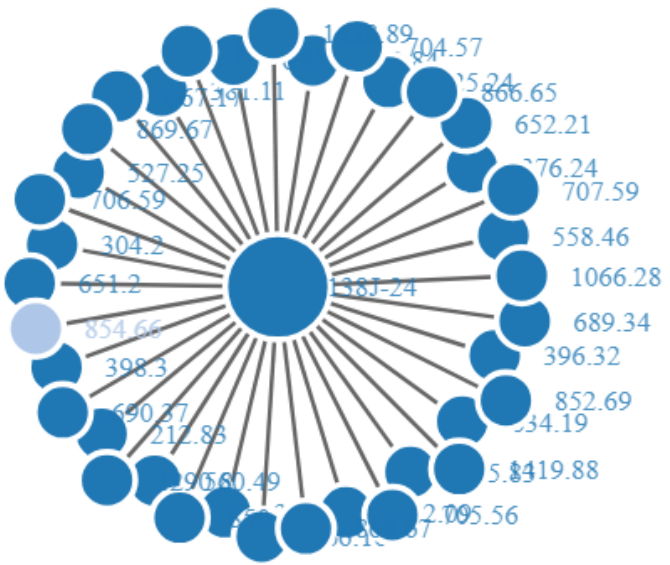

137B-22
137B-5
137B-20
137B-31
257.93
513.29
634.24
511.31
433.18
1241.74
657.39
335.10
234.79
1163.70
1056.69
233.78
425.19
972.61
617.31
618.29
973.57
843.48
440.18
656.40
428.10
1041.39
971.67
1108.04
844.45
261.93
655.42
1329.81
1057.54
616.30
449.33
458.20
262.92
706.57
557.27
1042.21
787.40
1015.35
497.26
249.83
439.11

Group 78

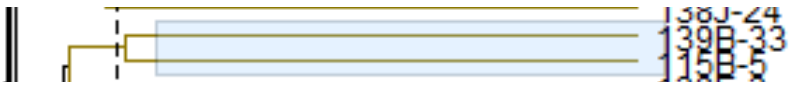

115B-3
115B-7
115B-1
115B-6
115B-2
114B-8
114B-4
114B-9
214.96
608.13
519.61
227.71
425.25
365.12
233.74
445.17
624.16
265.81
714.48
348.15
844.41
542.24
656.30
843.49
262.00
616.40
527.81
617.42
428.21
440.30
458.32
451.61
441.29
655.48
877.33
972.66
973.61

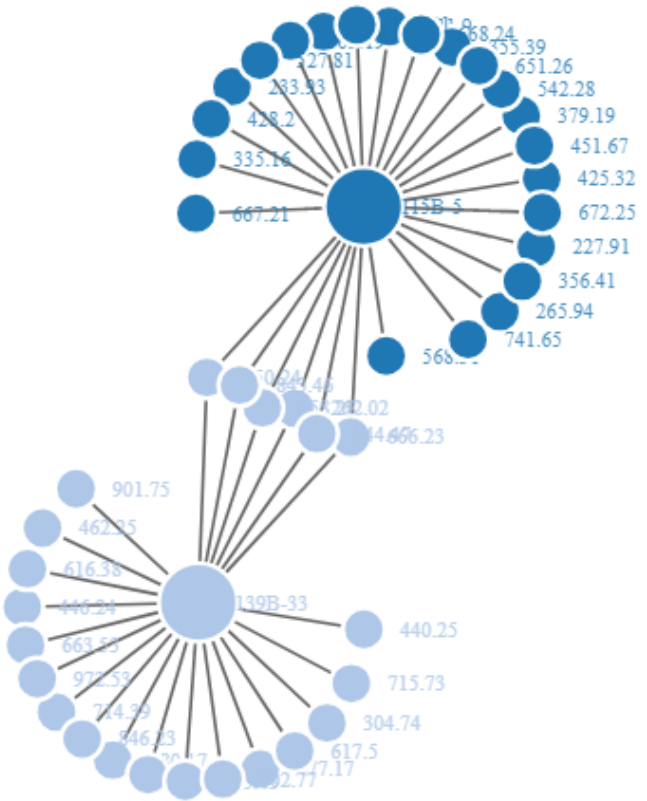
Group 79


138J-24


704.57
866.65
652.21
707.59
558.46
1066.28
689.34
396.32
852.69
869.67
527.25
706.59
304.2
651.2
854.66
398.3
212.83
138J-24

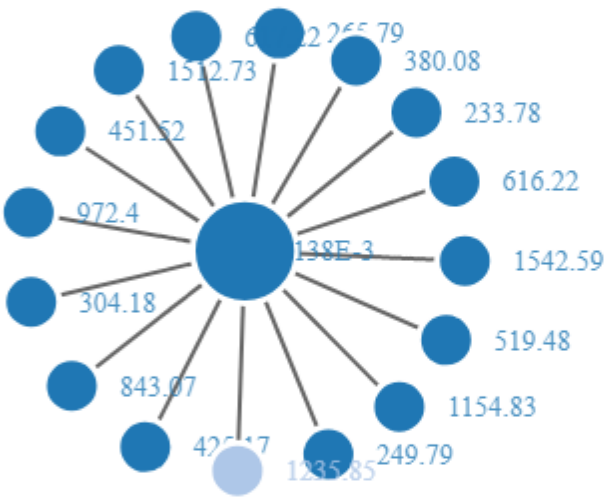
Group 80


138J-24
139B-33
115B-5

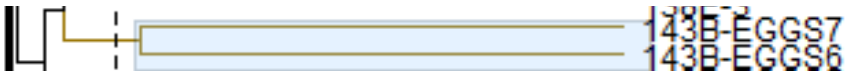

651.26
542.28
379.19
451.67
425.32
672.25
227.91
356.41
265.94
741.65
667.21
115B-5
901.75
616.38
139B-33
440.25
715.73
304.74
617.5

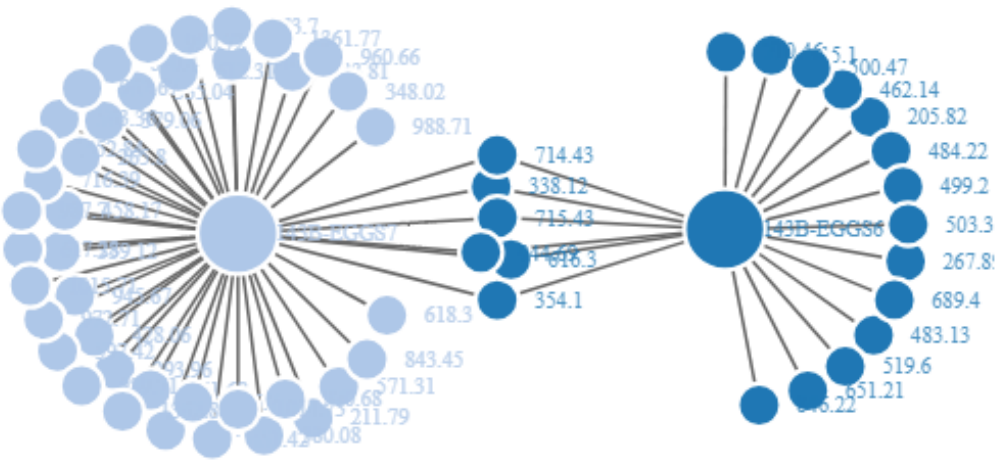
Group 81




380.08
233.78
616.22
1542.59
519.48
1154.83
249.79
304.18
972.4
451.52

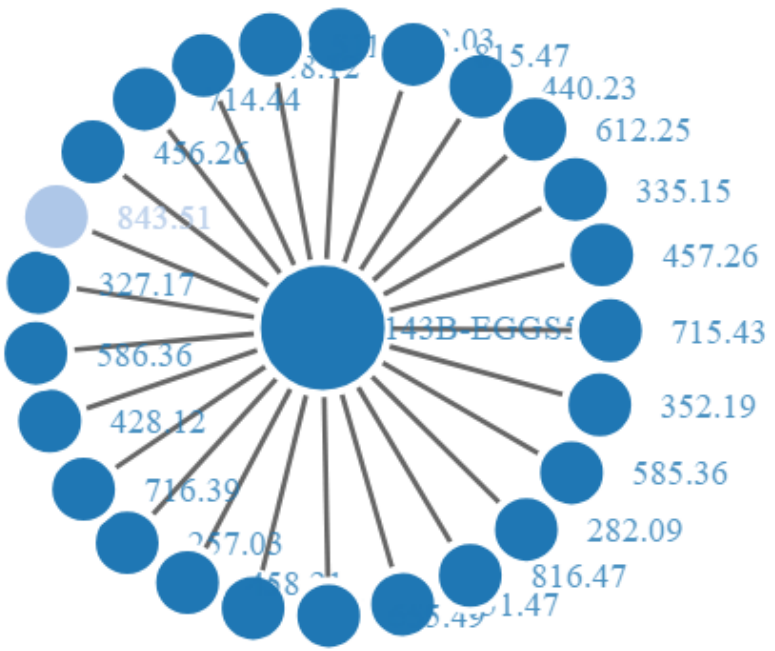
Group 82


143B-EGGS7
143B-EGGS6

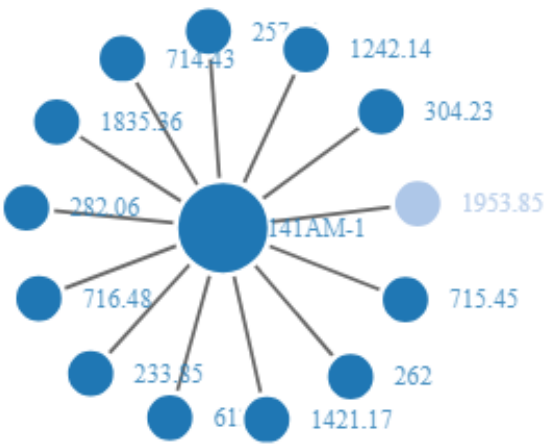

714.43
338.12
715.43
616.3
354.1
988.71
348.02
618.3
843.45
571.31
211.79
462.14
205.82
484.22
499.2
503.3
689.4
483.13
519.6

Group 83

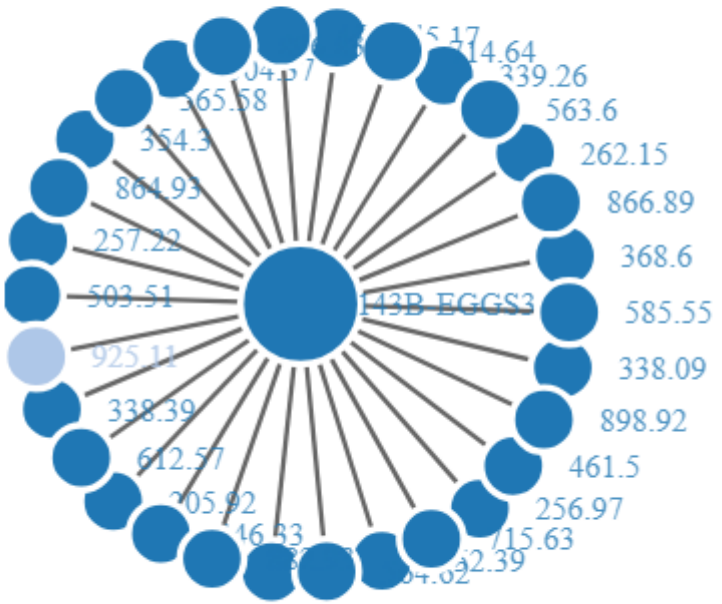

143B-EGGS5


440.23
612.25
335.15
457.26
715.43
352.19
585.36
282.09
816.47
456.26
843.51
327.17
586.36
428.12
716.39

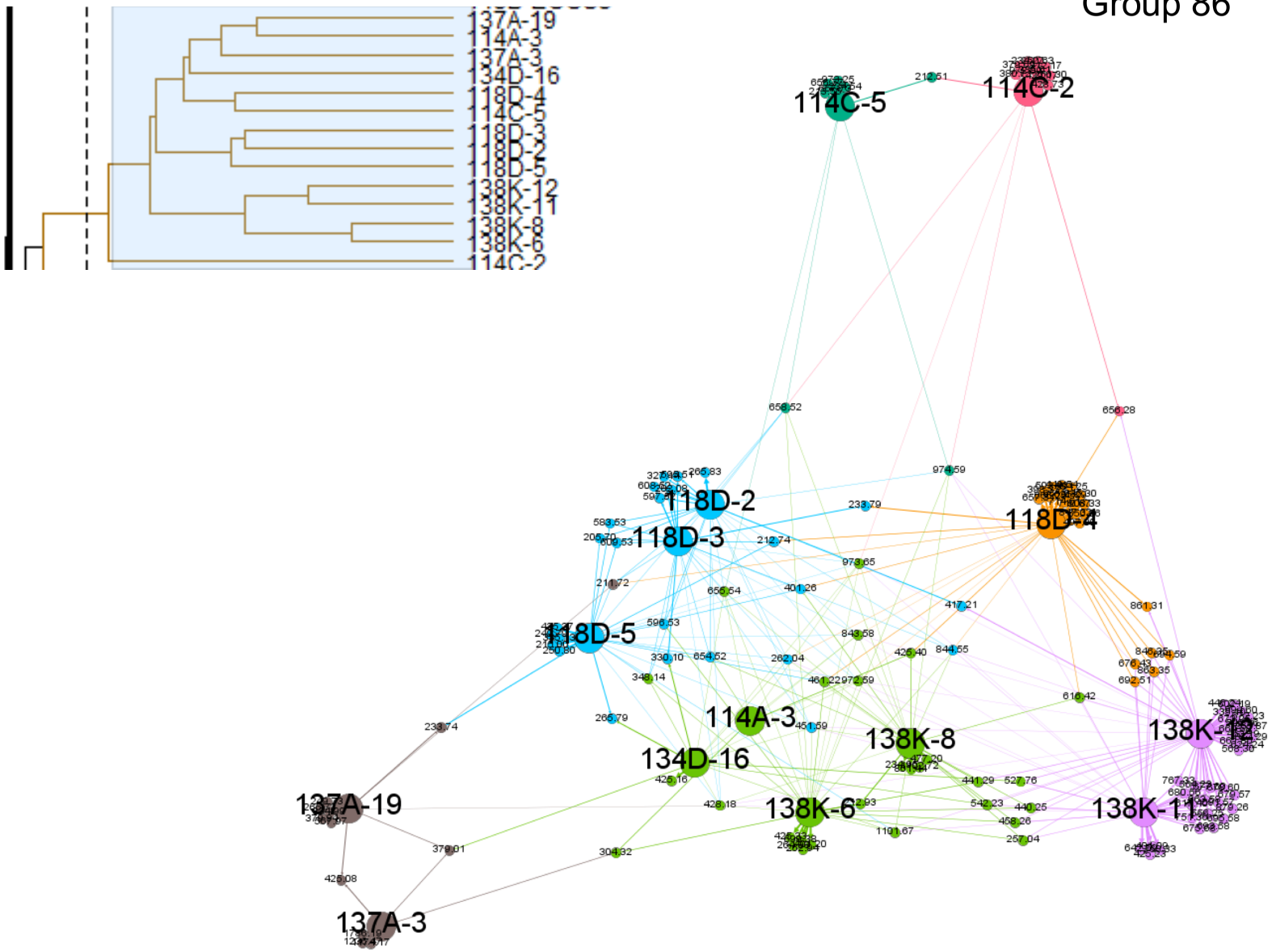

141AM-1


257
1242.14
304.23
1953.85
715.45
262
1421.17
61
233.35
716.48
282.06
1835.36
714.43
141AM-1


143B-EGGS3

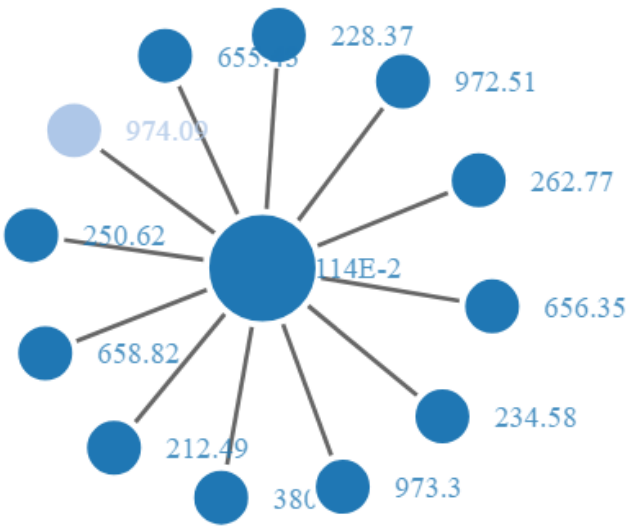

339.26
563.6
262.15
866.89
368.6
585.55
338.09
898.92
461.5
256.97
715.63
365.58
354.3
864.93
257.22
503.51
925.11
338.39
612.57
143B-EGGS3

Group 86

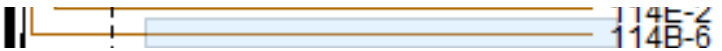

137A-19
114A-3
137A-3
134D-16
118D-4
114C-5
118D-3
118D-2
118D-5
138K-12
138K-11
138K-8
138K-6
114C-2
114C-5
118D-2
118D-3
118D-4
118D-5
114A-3
138K-8
134D-16
138K-6
138K-11
137A-19
137A-3

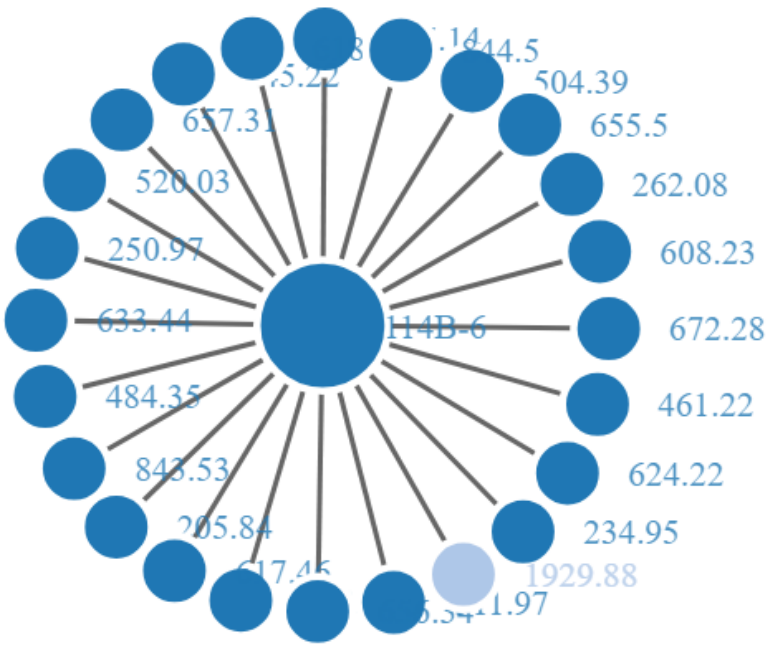
Group 87


114E-2

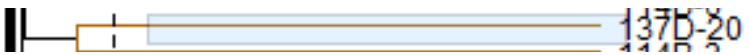

228.37
972.51
974.09
262.77
250.62
114E-2
656.35
658.82
234.58
212.49
973.3

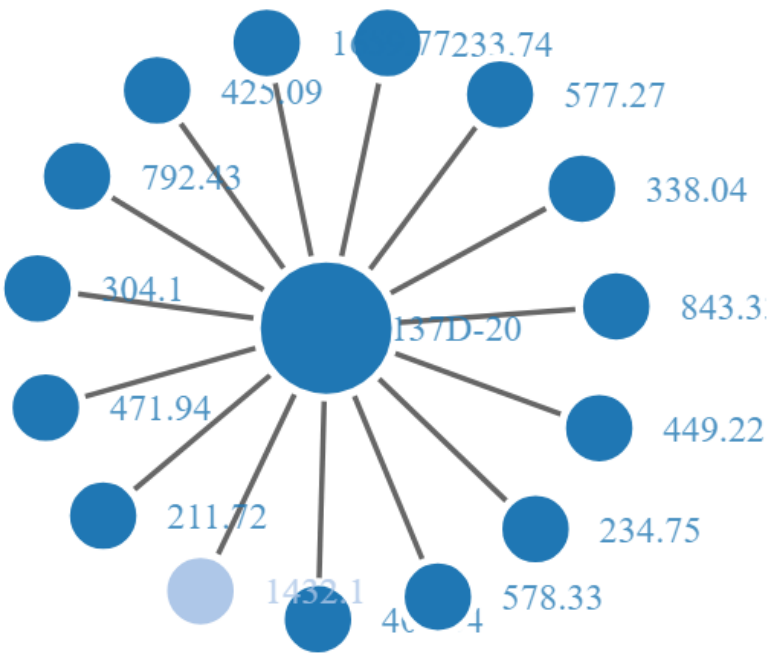
Group 88

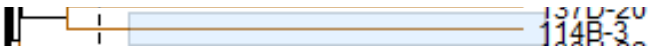

114E-2
114B-6

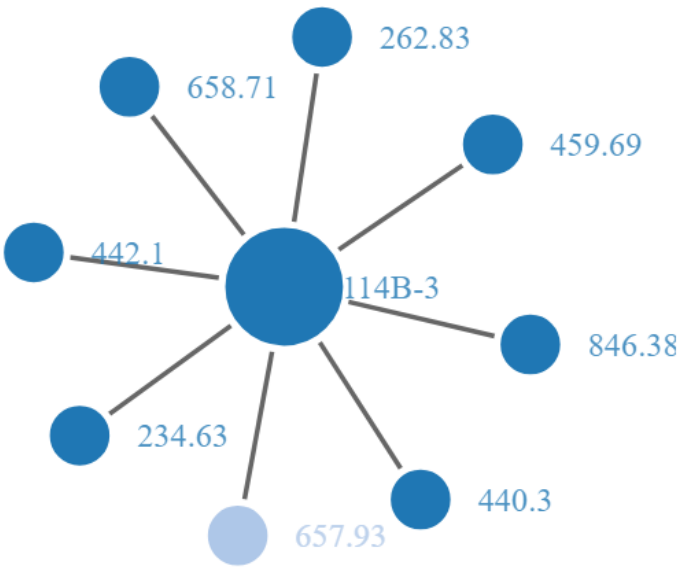

504.39
655.5
262.08
608.23
672.28
461.22
624.22
234.95
1929.88
520.03
250.97
633.44
484.35
843.53
114B-6

Group 89

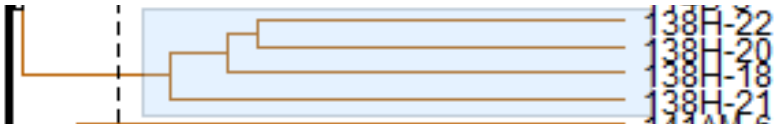

137D-20

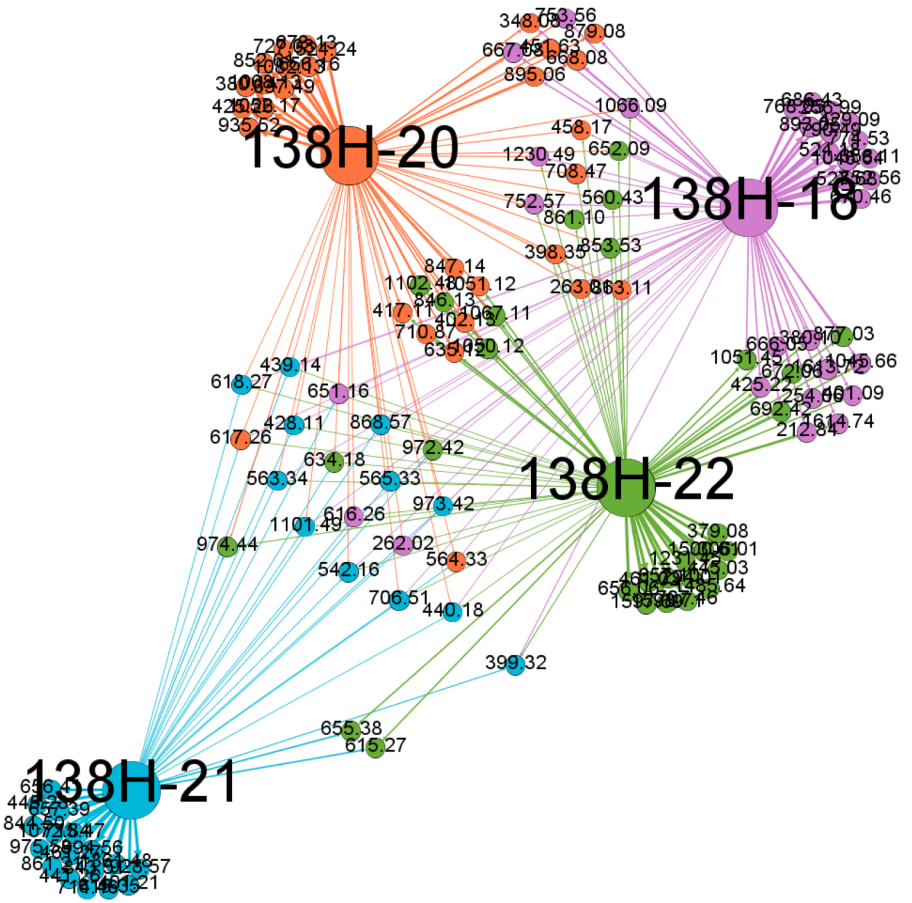

77233.74
577.27
338.04
843.3
449.22
234.75
578.33
792.43
304.1
471.94
211.72
137D-20


114B-3


262.83
658.71
459.69
442.1
114B-3
846.38
234.63
440.3
657.93

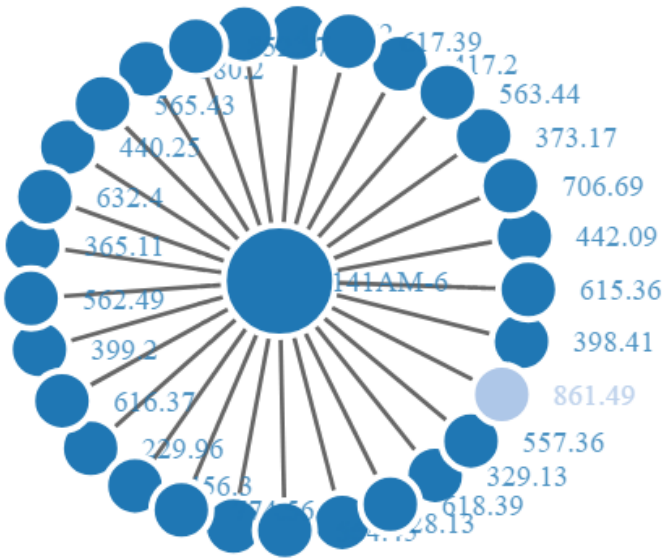
Group 91





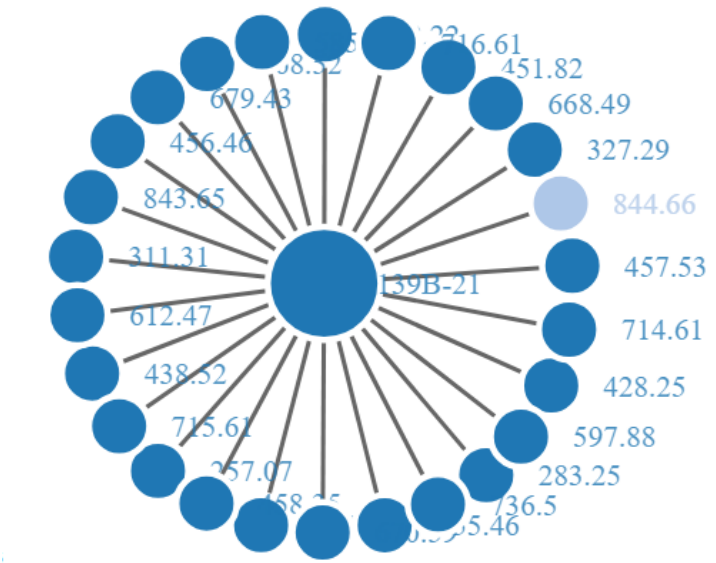
Group 92


141AM-6

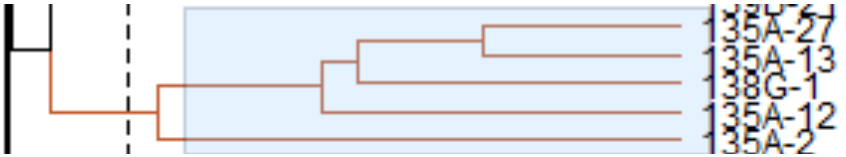

617.39
417.2
563.44
373.17
706.69
442.09
615.36
398.41
861.49
557.36
329.13
618.39
565.43
440.25
632.4
365.11
562.49
399.2
616.37
229.96
56.8
141AM-6

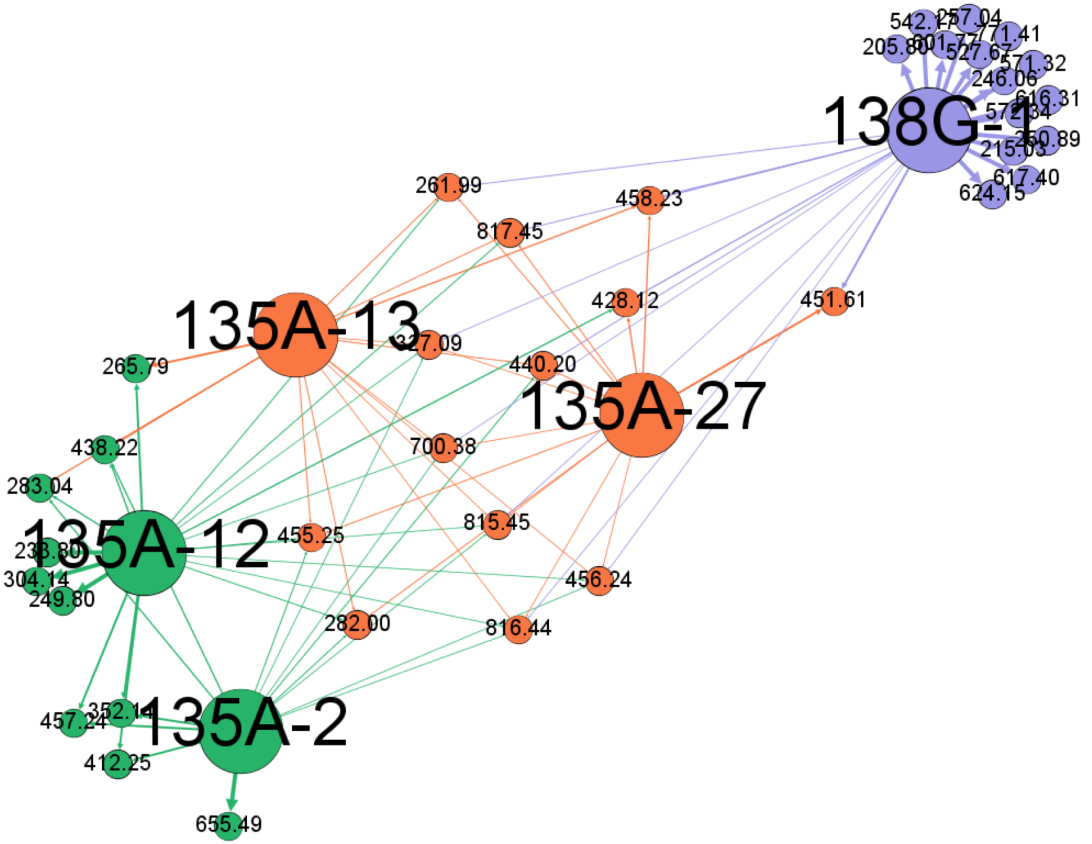
Group 93


139B-21


451.82
668.49
327.29
844.66
457.53
139B-21
714.61
428.25
597.88
283.25
679.43
456.46
843.65
311.31
612.47
438.52

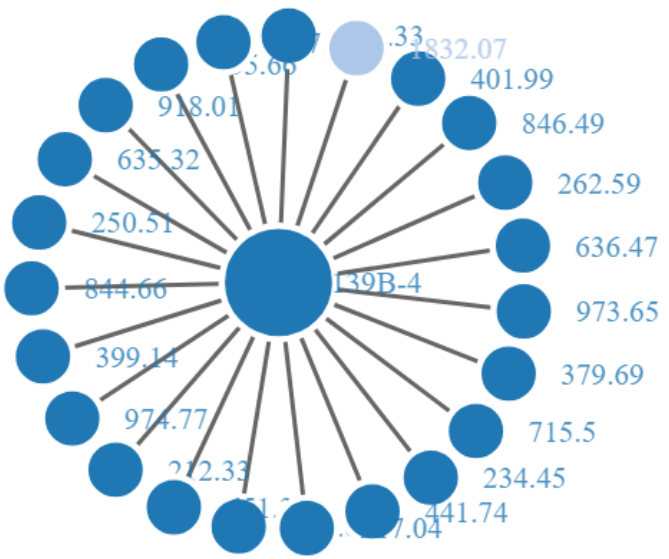
Group 94

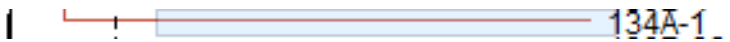

135A-27
135A-13
138G-1
135A-12
135A-2

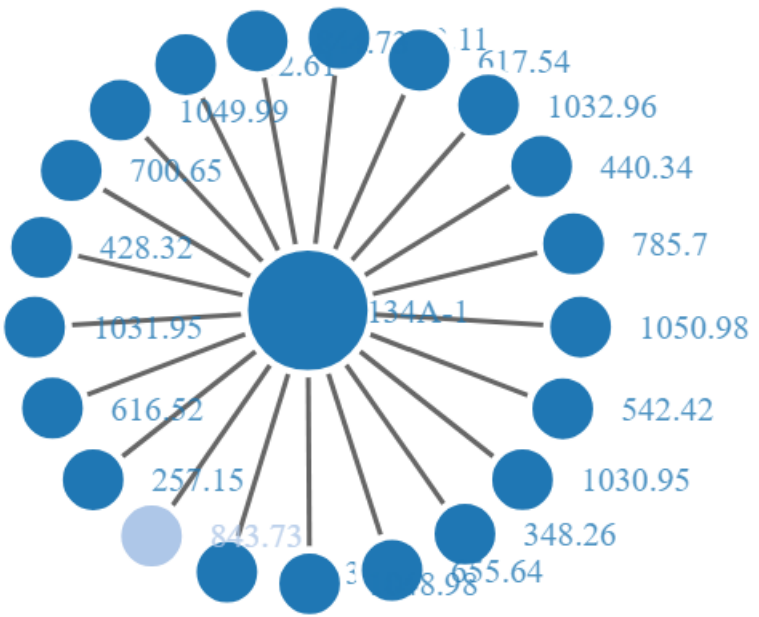

138G-1
135A-13
135A-27
135A-12
135A-2
261.99
458.23
817.45
428.12
451.61
265.79
440.20
438.22
700.38
283.04
815.45
455.25
304.14
249.80
456.24
282.00
816.44
412.25
655.49
246.06
571.32
617.40
624.15

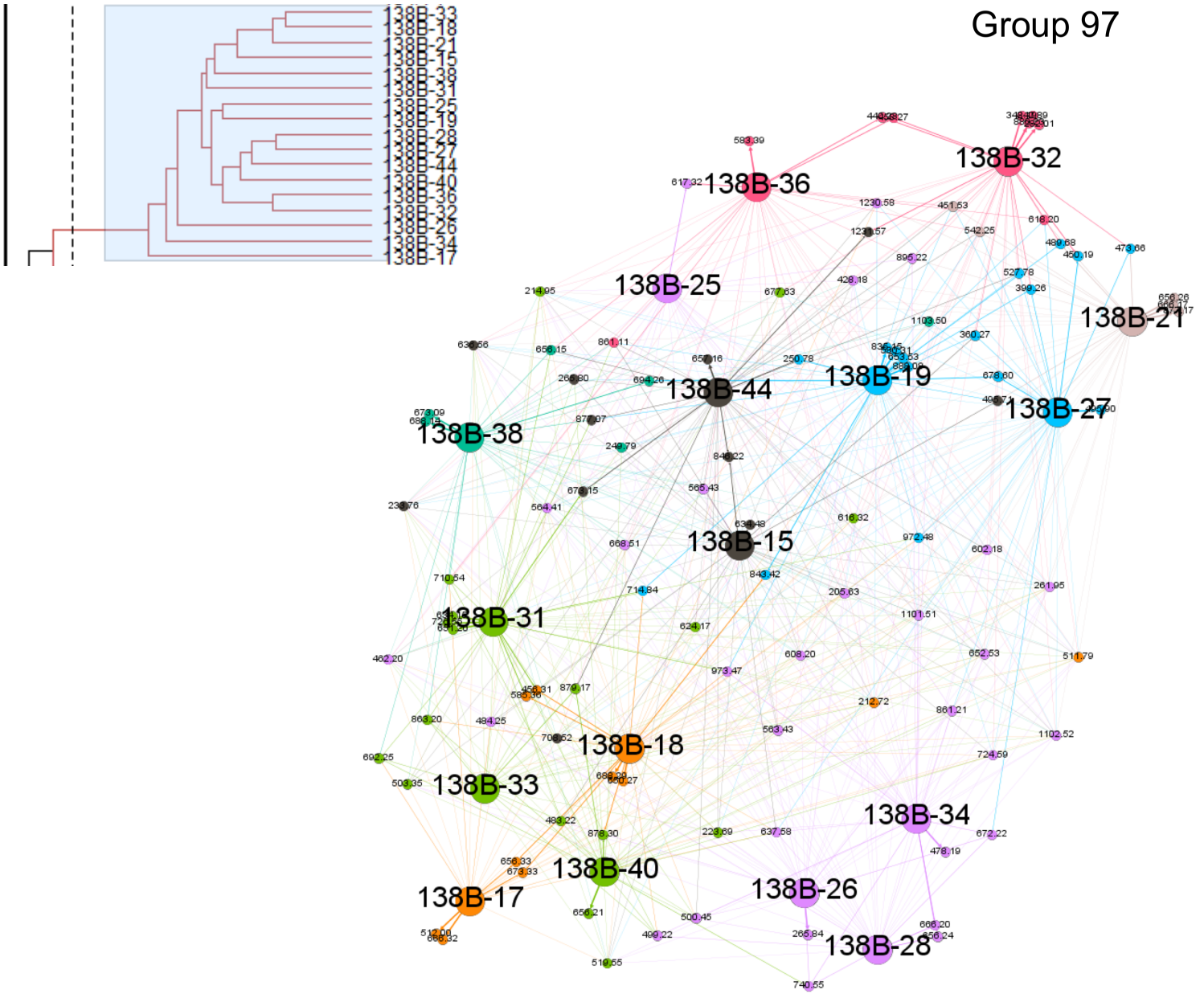
Group 95

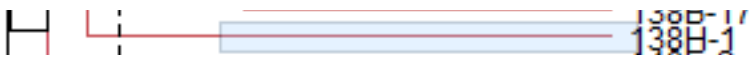

139B-4

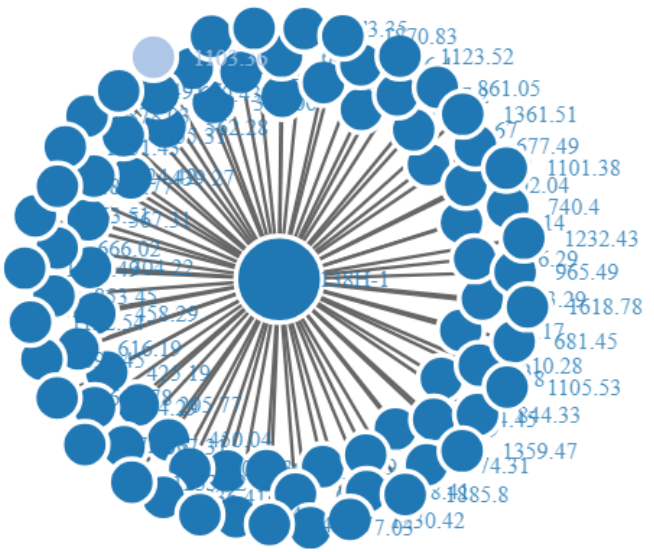

1832.07
401.99
846.49
262.59
636.47
973.65
379.69
715.5
234.45
441.74
918.01
635.32
250.51
844.66
399.14
974.77


134A-1

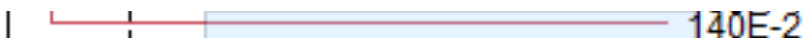

617.54
1032.96
440.34
785.7
1050.98
542.42
1030.95
348.26
1049.99
700.65
428.32
1031.95
616.52
134A-1

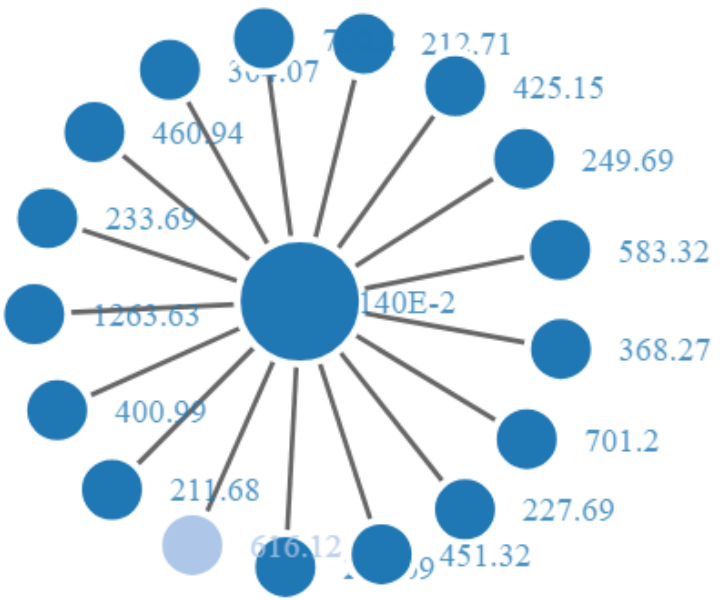

Group 97
138B-33
138B-18
138B-21
138B-15
138B-38
138B-31
138B-25
138B-32
138B-36
138B-44
138B-19
138B-27
138B-18
138B-33
138B-34
138B-40
138B-17
138B-26
138B-28


138B-17
138B-1

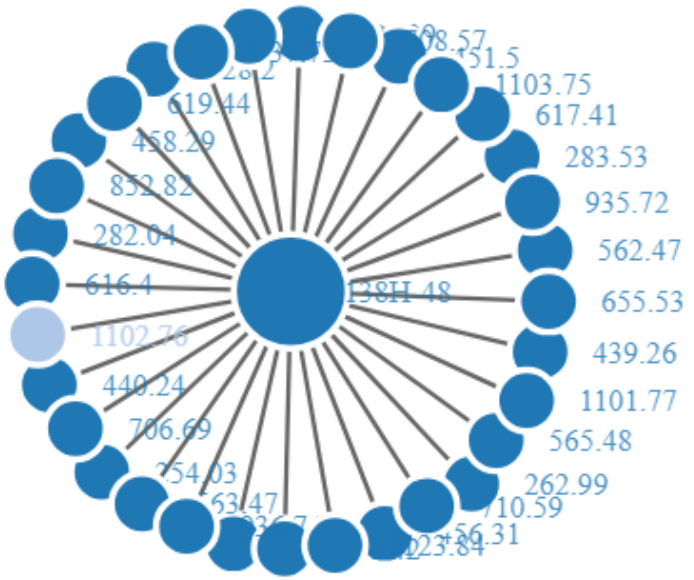

1123.52
861.05
1361.51
577.49
1101.38
740.4
1232.43
965.49
1618.78
681.45
1105.53
1359.47
1385.8

Group 99

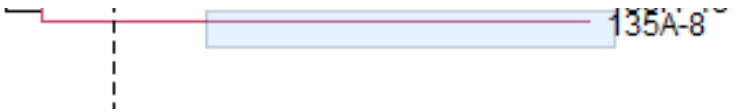

140E-2

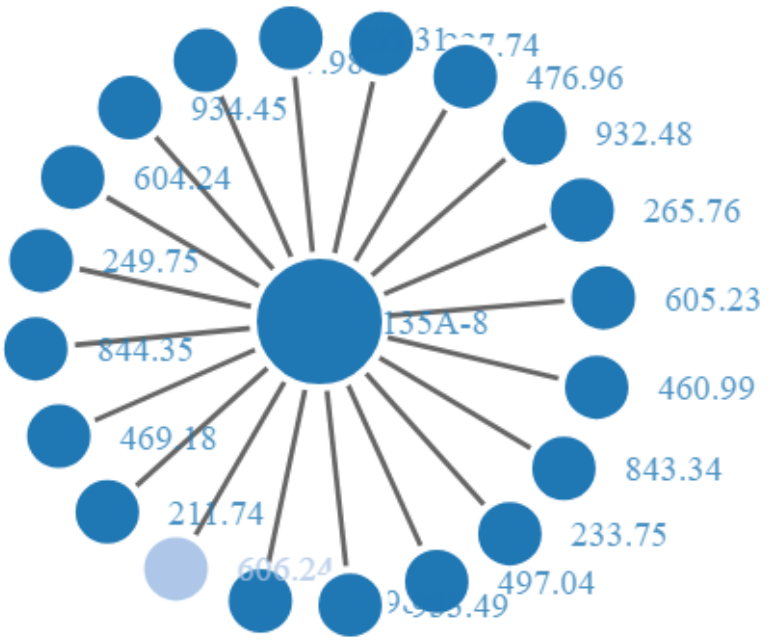

212.71
425.15
249.69
583.32
140E-2
368.27
701.2
227.69
451.32
233.69
1263.63
400.99


140E-2
138H-48


1103.75
617.41
283.53
935.72
562.47
655.53
439.26
1101.77
565.48
262.99
619.44
458.29
852.82
282.04
616.4
1102.76
440.24
706.69
254.03
138H-48

Group 101


135A-8


476.96
932.48
265.76
605.23
460.99
843.34
233.75
497.04
606.24
844.35
469.18
249.75
604.24
934.45