Genome-based natural product biosynthetic gene cluster discovery: from sequencing to mining

Hao Wang

Division of Microbiology and Biotechnology
Department of Food and Environmental Sciences
Faculty of Agriculture and Forestry and
Viikki Doctoral Programme in Molecular Biosciences
University of Helsinki

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Agriculture and Forestry of the University of Helsinki, for public examination in auditorium 1041 at Viikki Biocenter 2, Viikinkaari 5, on March 14th at 12 o'clock noon.

HELSINKI 2014

Supervisor:
Professor Kaarina Sivonen
Department of Food and Environmental Sciences
University of Helsinki, Finland

Reviewers:
Docent Petri Auvinen
Institute of Biotechnology
University of Helsinki, Finland

Professor Elke Dittmann
Institute of Biochemistry and Biology
University of Potsdam, Germany

Opponent:
Dr. Björn Voß
AG Genetics & Experimental Bioinformatics
Institute of Biology III
University of Freiburg, Germany

Front cover: Illustration of *Anabaena* sp. 90 chromosome I and the distribution of nonribosomal peptide synthetases and modular polyketide synthases across the three domains of life

ISSN 1799-7372
ISBN 978-952-10-9780-5 (Paperback)
ISBN 978-952-10-9781-2 (PDF)

Unigrafia
Helsinki 2014

Table of Contents

List of original publications

I **Wang H**, Sivonen K, Rouhiainen L, Fewer DP, Lyra C, Rantala-Ylinen A, Vestola J, Jokela J, Rantasärkkä K, Li Z, Liu B. (2012) Genome-derived insights into the biology of the hepatotoxic bloom-forming cyanobacterium *Anabaena* sp. strain 90. *BMC Genomics*. 13: 613.

II **Wang H**, Fewer DP, Sivonen K. (2011) Genome mining demonstrates the widespread occurrence of gene clusters encoding bacteriocins in cyanobacteria. *PLoS ONE* 6(7): e22384.

III **Wang H**, Fewer DP, Rouhiainen L and Sivonen K. An atlas of nonribosomal peptide and modular polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. *Submitted manuscript.*

The publications are referred to in the text by their roman numerals.

The author's contribution

I Hao Wang participated the design of the study, conducted the bioinformatic analysis, contributed to the result interpretation and wrote the first draft of the manuscript.

II Hao Wang designed the study, conducted the bioinformatic analysis, interpreted the results and wrote the first draft of the manuscript.

III Hao Wang designed the study, conducted the bioinformatic analysis, interpreted the results and wrote the first draft of the manuscript.

Abbreviations

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
cDNA	Complementary DNA
rRNA	Ribosomal RNA
SNP	Single-nucleotide polymorphism
SMRT	Single molecule real time
MPSS	Massively parallel signature sequencing
Mb	Mega bases
Gb	Giga bases
Kb	Kilo bases
Bp	Base pair
BLAST	Basic Local Alignment and Search Tool
NCBI	National Center for Biotechnology Information
EMBL	European Molecular Biology Laboratory
DDBJ	DNA Data Bank of Japan
WGA	Whole genome assembly
WGS	Whole genome shotgun
ORF	Open reading frame
IS	Insertion sequence
MITE	Miniature inverted-repeat transposable element;
RM	Restriction-modification
NRPS	Nonribosomal peptide synthetase
NRP	Nonribosomal peptide
PKS	Polyketide synthase
PK	Polyketide
FAS	Fatty acid synthase
RiPP	Ribosomaly synthesized and post-translationally modified peptide
A	Adenylation
PCP	Peptidyl carrier domain
C	Condensation
TE	Thioesterase
E	Epimerization
H	Heterocyclization
AT	Acyltransferase
ACP	Acyl carrier domain
KS	Ketosynthase
KR	Ketoreductase
DH	Dehydratase
ER	Enoylreductase
FTP	File transfer protocol

Abstract

Natural products are small molecules produced by a range of living organisms. They may be toxic or have pharmaceutical applications as antibiotics, anticancer, antiparasitic and anti-fungal agents. Natural products are commonly synthesized by nonribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs), such as microcystins. Ribosomal pathways in cyanobacteria are also known for the synthesis of bacteriocins, lantibiotics, cyanobactins and microviridins. Genes encoding biosynthetic enzymes of these systems are often found together and form gene clusters.

The filamentous cyanobacterium *Anabaena* sp. strain 90, a hepatotoxin producer isolated from a bloom of a Finnish lake, was selected for genome sequencing, in order to explore its full capacity of bioactive compound production. The 5.3-Mb *Anabaena* sp. 90 genome displays a multichromosome composition with five circular replicons: two chromosomes and three plasmids. A total of four nonribosomal biosynthetic gene clusters, which are responsible for the production of anabaenopeptilides, anabaenopeptins, microcystins and the novel glycolipopeptides hassallidins, were identified in chromosome I. Genome annotation revealed that *Anabaena* sp. 90 genome also harbors an anacyclamide-encoding cyanobactin gene cluster and seven putative bacteriocin gene clusters, which belong to the ribosomal pathways. These biosynthetic gene clusters amount to a total of ~250 kb, and 5% of the genome.

Analysis of the *Anabaena* sp. 90 genome suggested that cyanobacteria might produce bacteriocins. A thorough genome mining at the phylum level was conducted targeting the discovery of cyanobacterial bacteriocin biosynthetic pathways. The results demonstrated the common presence of bacteriocin gene clusters in cyanobacteria. A total of 145 bacteriocin gene clusters were discovered, the majority of them were previously unknown. Based on their gene organization and domain composition, these gene clusters were classified into seven groups. This classification is supported by the phylogenetic analysis, which also indicates independent evolutionary trajectories of the gene clusters in different groups. By scrutinizing the surrounding regions of these gene clusters, a total of 290 putative precursors were located. They showed diverse structures and very little sequence conservation of the core peptide.

To explore the distribution of NRPSs and PKSs, a comprehensive genome-mining study was carried out and demonstrated their widespread occurrence across the three domains of life, with the discovery of 3,339 gene clusters from 991 organisms, by examining a total of 2,699 genomes. The majority of these gene clusters were found in bacteria, in which high correlation between bacterial genome size and the capacity of NRPS and PKS biosynthetic pathways was observed. Currently, PKSs are classified into three types. Type I PKSs and NRPSs are known to share a modular scheme with a multidomain structure. Surprisingly, a large number (8,906) of enzymes encoding a single NRPS or type I PKS functional domain were found. These monodomain enzymes have a similar genetic organization to type II PKSs, which are nonmodular enzymes. The finding of common occurrence of nonmodular NRPSs and type I PKSs substantially differs from the current knowledge. Furthermore, a total of 314 gene clusters comprised mostly of monodomain enzymes were found. In addition, sequence analysis suggested that the evolution of NRPS machineries was a combination of common descent and horizontal gene transfer.

1. Introduction

1.1 Natural product biosynthesis

Natural products are small compounds produced by living organisms. They are often biologically active with pharmaceutical applications, since a substantial number of drugs are derived from these natural products (Newman and Cragg 2007). Natural products include a wide range of bioactive compounds, such as nonribosomal peptides, polyketides, fatty acids, bacteriocins, cyanobactins and lantibiotics. They are biosynthesized by corresponding enzyme systems.

Significant portions of known natural products are derived from nonribosomal peptides and polyketides (Donadio et al. 2005). They were biosynthesized by nonribosomal peptide synthetase (NRPSs) and polyketide synthases (PKSs), respectively. Both enzyme systems function as production lines that assembly substrate monomers into end products (Fischbach and Walsh 2006). More recently, an increasing number of natural products have been discovered that are ribosomally synthesized and post-translationally modified peptides (RiPPs) (Arnison et al. 2013). These peptides usually are crafted from short ribosomally produced precursors by a number of proteases and then modified by certain tailoring enzymes (**Fig. 1**). Ribosomal peptides usually are classified according to the different set of peptidases and modification enzymes they utilized.

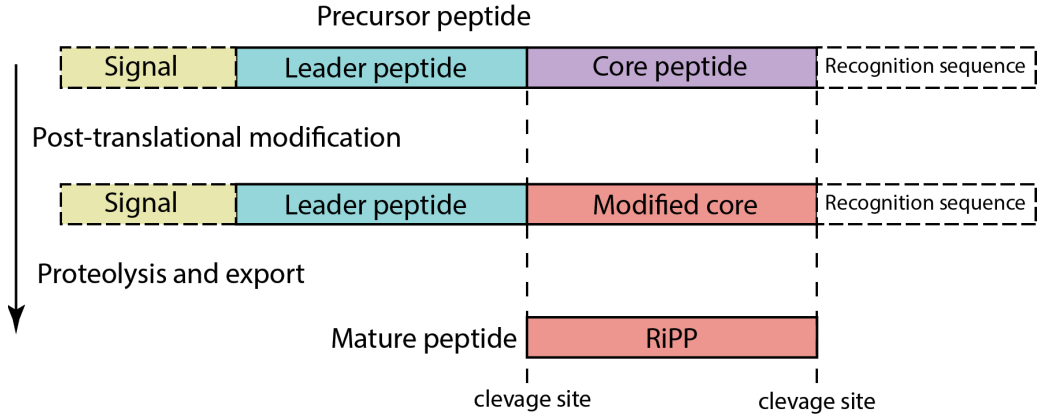

Fig. 1. Illustration of biosynthesis for RiPPs from the precursor peptide (modified from Arnison et al. 2013). The precursor peptide usually contains a leader peptide and a core region that is transformed into the mature product. N-terminal signal peptide and C-terminal recognition sequence may present in some precursors. Many of the post-translational modifications are guided by leader peptide and recognition sequence.

Genes encoding the enzymes involved into the biosynthesis, modification and regulation of these natural products often form gene clusters. Based on present knowledge of natural product biosynthesis, most of the natural product biosynthetic machineries for an organism could be determined using bioinformatic analysis given that the genome is available (Walsh and Fischbach 2010).

1.1.1 Nonribosomal peptide biosynthesis

Nonribosomal peptides form a diverse group of natural products that are synthesized by large multifunctional nonribosomal peptide synthetases (NRPSs) (Finking and Marahiel 2004). NRPSs often possess multiple domains that organize into modules. Each module, except the starter one, usually has three core biosynthetic domains: adenylation (A), peptidyl carrier (PCP) and condensation (C) domains, which coordinately incorporate one amino acid or hydroxyl acid to the product, in addition to optional auxiliary domains of thioesterase (TE), epimerization (E), N-methylation, heterocyclization (H), formylation and oxidation (**Fig. 2**). Nonribosomal peptide biosynthetic gene clusters encode proteins with one or multiple modules, which form complexes operating as assembly lines (Fischbach and Walsh 2006), and often follow the colinearity rule (Marahiel et al. 1997). Currently, NRPSs have been known only in modular scheme with three subgroups: linear, iterative and nonlinear (Mootz et al. 2002).

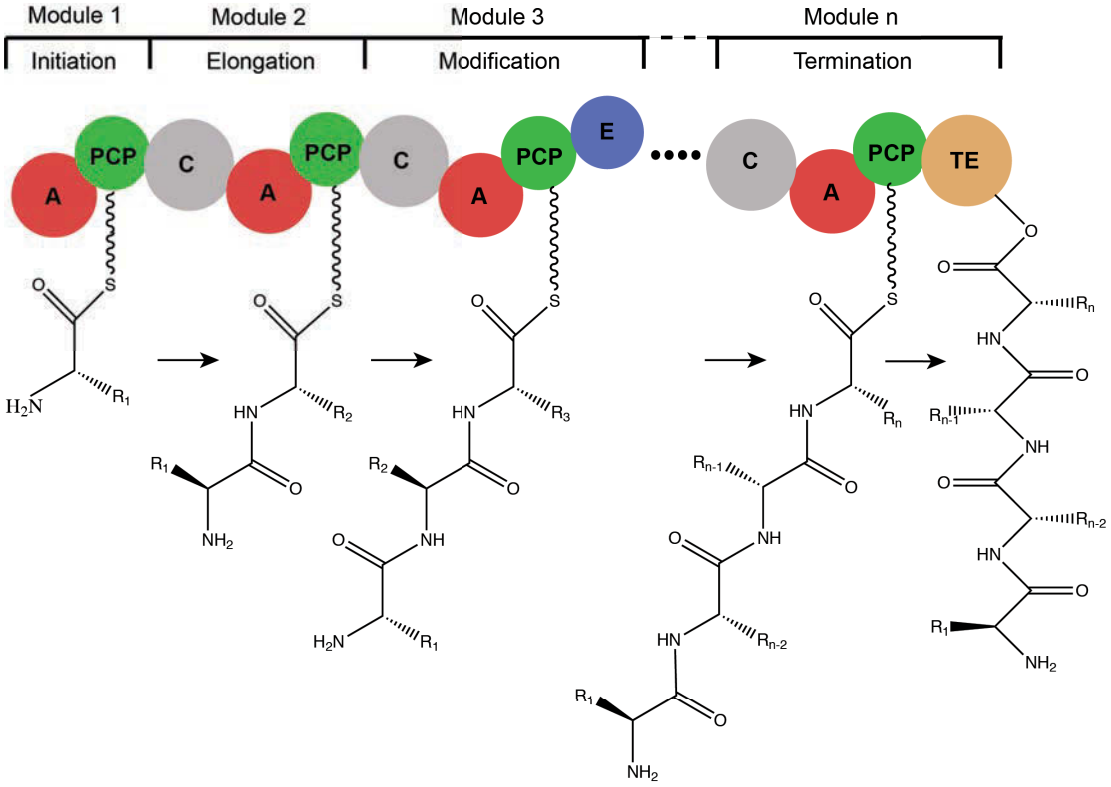

Fig. 2. Illustration of NRPS biosynthesis assembly line (modified from Strieker et al., 2010). The amino acids are activated by the adenylation (A) domain and transferred by the peptidyl carrier (PCP) domain. Peptide bond is formed by the condensation (C) domain. Amino acids may be modified, for example by epimerization (E) domain. The release of final product is usually catalyzed by thioesterase (TE) domain through either hydrolysis or macrocyclization. The numbers of modules and modification domains can be very variable.

1.1.2 Polyketide biosynthesis

Polyketides are a group of secondary metabolites synthesized from carboxylic acids by polyketide synthases (PKSs), which are not only structurally and enzymatically similar but also evolutionarily related to the fatty acid synthases (FASs) (Jenke-Kodama et al. 2005). Currently, PKSs are classified into three types. This classification is adapted from the predated field of FASs (Staunton and Weissman 2001).

Type I PKSs and FASs are large megasynthases and share a module-based biosynthetic strategy (Meier and Burkart 2009), which is also employed by NRPSs. However, their modularity levels are different in that types I FASs have only one module that iteratively stacks the same monomer to a polymer (Smith and Tsai 2007), whereas NRPSs and type I PKSs have varied modules and function as assembly lines (Fischbach and Walsh 2006). The occurrence of hybrid systems between NRPS and type I PKS can be attributed to their similar biosynthetic mechanism (Du et al. 2001). The type I PKS modules, which are responsible for the monomer carboxylic acid incorporation, also constitute three core domains: acyltransferase (AT), acyl carrier domain (ACP) and ketosynthase (KS), in addition to additional modification domains of ketoreductase (KR), dehydratase (DH), enoylreductase (ER) and aminotransferase.

In contrast to type I PKSs, type II and III PKSs are short proteins with a single functional domain (Jenke-Kodama et al. 2005). Like type II FASs, type II PKSs have a minimal set of individual enzymes: two KS and one ACP proteins, which are clustered together and iteratively used for biosynthesis of aromatic polyketides (Hertweck et al. 2007). Type III PKSs are also monofunctional proteins, which lack ACP protein and form homodimers with a single active site for chain elongation (Shen 2003, Yu et al. 2012).

However, the occurrence of many transition state PKSs has diminished credibility of the present classification (Muller 2004, Wenzel and Muller 2005). For instance, some iterative PKSs are almost indistinguishable from type I FASs (Moss et al. 2004). In addition, a transition state of type II and III PKSs was found to be independent of the ACP domain (Kwon et al. 2001). A transition state between type I and II PKSs was also discovered in the trans-AT PKS clusters, in which a distinct AT protein iteratively loads the monomers *in trans* for other modular PKSs that lack AT domain (Cheng et al. 2003). Therefore, it would be expected that the discovery of novel PKS architectures alongside the rapid expansion of genomic data would lead to reclassification of PKSs (Muller 2004).

1.1.3 Bacteriocin biosynthesis

Bacteriocins, a class of ribosomal peptides with diverse structures, are emerging as alternatives to traditional antibiotics due to their significant inhibition against other bacteria (Cotter et al. 2013), in addition to other applications, such as food preservatives (Galvez et al. 2008). They have been found in all major lineages of bacteria (Riley and Wertz 2002). However, bacteriocin research has focused mostly on lactic acid bacteria (Nes et al. 2007). Bacteriocins are grouped into four classes (Cotter et al. 2005, Oman 2011, Stepper 2011). The bacteriocin precursors commonly contain a conserved N-terminal leader sequence with double glycine motif (Oman and van der Donk 2010), which is recognized and cleaved by a C39 peptidase domain (Håvarstein et al. 1995). The double glycine motif was later refined to M(R/K)ELX$_3$E(I/L)X$_2$(I/V)XG(G/A) (Dirix et al. 2004). Associated tailoring enzymes may provide post-translational modifications such as lanthionine formation (Willey and van der Donk 2007), macrocyclization (Willey and van der Donk 2007), dehydration (Onaka et al. 2005), or heterocyclization (Li et al. 1996, Milne et al. 1999). A previous genome mining study for bacteriocin gene clusters in Gram-negative bacteria revealed a number of ABC transporter domain-containing proteins (Dirix et al. 2004). They were shown to be responsible for the export of bacteriocins (Håvarstein et al. 1995). A three-gene transport cluster was shown to be associated with

the bacteriocin production (Haft et al. 2010). Two subgroups of double-glycine-type precursors (NHLP and N11P) were also discovered (Haft et al. 2010). Lantibiotics are a class of bacteriocins containing the amino acid Lanthionine, which are modified by the bifunctional lanthionine synthetase (Willey and van der Donk 2007). Genome-mining tools (de Jong et al. 2006, 2010) and databases (Hammami 2007, Hammami et al. 2010) are now available for bacteriocin analysis.

1.2 Genome sequencing

The aim of genome sequencing is to determine the entirety of an organism's genetic information, which includes both the coding sequences (genes) and non-coding sequences of the DNA or RNA (Ridley 2000). Genome sequencing had been started since seventies of last century (Sanger et al. 1977a, Staden 1979), soon after the invention of the Sanger sequencing method (Sanger et al. 1977b). Ever since, this research field has boomed and led to the interdisciplinary scientific field of bioinformatics (Hagen 2000). Enabled by the high efficiency obtained from technical innovations (Smith et al. 1986, Shizuya et al. 1992), a large number of high-throughput sequencing projects were achieved, from the first bacterial genome (Fleischmann et al. 1995) to the human genomes (Venter et al. 2001, Lander et al. 2001). At the meantime, the whole genome shotgun (WGS) sequencing (**Fig. 3**) (Venter et al. 1998, Myers et al. 2000) was shown to be more cost-effective than the hierarchical shotgun (HS) approach (McPherson et al. 2001, Waterston et al. 2002), and thus commonly accepted as standard.

Starting from this century, genome sequencing has been greatly accelerated as the emerging of many ultra-high-throughput sequencing techniques and methods, which include Massively Parallel Signature Sequencing (MPSS) (Brenner et al. 2000), 454 pyrosequencing (Margulies et al. 2005), Polony sequencing (Shendure et al. 2005), Solexa (Illumina) (Bentley et al. 2008), SOLiD (Valouev et al. 2008), and Ion Torrent (Rusk 2011). These systems are collectively termed as the next-generation sequencing (NGS) platforms and extensively reviewed (Mardis 2008, Shendure and Ji 2008, Mardis 2013), in addition to the upcoming 3rd generation methods, like Helioscope single molecule sequencing (Harris et al. 2008), Nanopore (Clarke et al. 2009), Single Molecule real time SMRT sequencing (PacBio) (Eid et al. 2009), DNA nanoball sequencing (Drmanac et al. 2010) and Tunnelling currents (Di Ventra 2013). Compared to the capillary Sanger sequencing, NGS platforms obtained dramatically magnified yield in data volume at a much lower cost (**Table 1**). The quickly accumulated sequences delivered by NGS instruments profoundly impacted biological research by contributing significant amount of genomes (Mardis 2013), and provoked other more ambitious projects, such as the 1000 Genome Project for human (1000 Genomes Project Consortium et al. 2010) and the Genome 10K Project to cover almost every vertebrate genus (Genome 10K Community of Scientists 2009).

1.2.1 Assembly, scaffolding and gap closure

The controversy between very long stretch of genome sequences and limited read lengths of current technology is solved by over amount of random sequencing and reconstruction (or assembly) using computer programs (Myers et al. 2000). Genome assembly requires intensive computational resources to build up continuous genomic pieces (contigs) by merging sequencing reads having overlaps. A growing number of assemblers has been

developed and utilized in assembling prokaryotic and eukaryotic genomes, such as Phrap (Green 1994), Celera Assembler (Myers et al. 2000), ARACHNE (Batzoglou et al. 2002), AMOScmp (Pop et al. 2004), as well as Velvet (Zerbino and Birney 2008) and SOAP (Li et al. 2010a) that were designed for short-read NGS data. Previously in the Sanger sequencing, a *de novo* genome project usually generates sequence data to 8-10X genome coverage for assembly, according to the Lander-Waterman model (Lander and Waterman 1988). Genome assemblies of NGS projects now often accumulate data amount to 20X coverage or more (Lim et al. 2012), and therefore have exerted more challenges on bioinformatic tools for genome assembly (Dolled-Filhart et al. 2013).

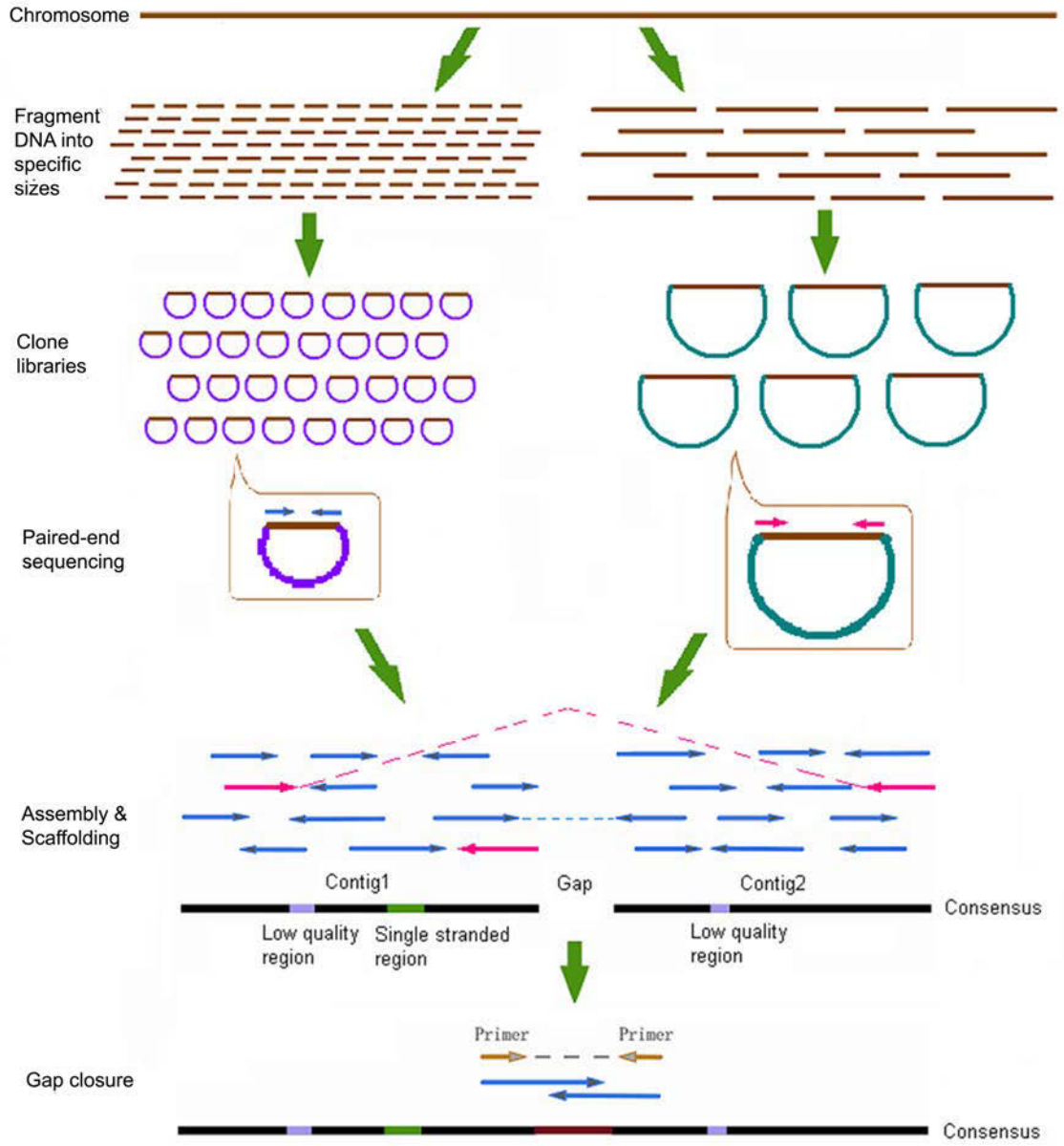

Fig. 3. Schematic presentation of whole genome shotgun (WGS) strategy used in the Sanger sequencing.

Table 1. Comparison of the next-generation, 3rd generation and the Sanger sequencing platforms (adapted from Liu et al. 2012, Quail et al. 2012).

Platform and model	Method and mechanism	Read length (Bp)	Data output per run	Run time	Cost/Gb ($)	Paired reads	Max insert size	Accuracy
454 GS FLX (Roche)	Pyrosequencing	700	0.7 Gb	24 hours	10,000	Yes	1000 Bp	99.9%
HiSeq 2000 (Illumina)	Sequencing by synthesis	20-250	600 Gb	1 - 11 days	50 - 150	Yes	700 Bp	98%
SOLiD 4 (Life Tech)	Sequencing by ligation	50+50	120 Gb	1 - 2 weeks	130	Yes	300 Bp	99.94%
Ion Torrent PGM	pH-mediated semiconductor sequencing	200	1 Gb	2 hours	1,000	Yes	400 Bp	98%
PacBio RS	Single-molecule real-time sequencing	~1,500	100 Mb	0.5 - 2 hours	750 - 1,500	No	10-30 Kb	87%
Sanger 3730xl (ABI)	Dideoxy chain termination	400 - 900	1.9 - 84 Kb	0.3 - 3 hours	2,400,000	Yes	500-1000 Bp	99.99%

Repetitive genomic regions larger than read length often cause misassemblies and lead to gaps (Salzberg and Yorke 2005). Scaffolding is the linking and ordering of contigs into clusters (scaffolds) with regard to paired-end reads located in different contigs (**Fig. 3**). Gaps within scaffolds can be resolved by targeted sequencing, such as primer walking over the PCR products of the gap regions (**Fig. 3**). Due to their short read lengths, however, NGS genome projects are usually fragmented, which is difficult to be improved by simply increasing the genome coverage (Treangen and Salzberg 2011). These obstacles restricted the application of NGS in *de novo* sequencing and derived a growing number of partial genomes, which conflict with the original aim of genome sequencing (Alkan et al. 2011). A method of adding tuned-size paired-end data according to repeat structure of a genome was shown to be effective in improving assembly quality of prokaryotic genomes (Wetzel et al. 2011). Another more promising method using long SMRT read has been recently developed to obtain complete genome (English et al. 2012).

1.2.2 Genome annotation

Genome annotation is an essential task of interpreting the biological information from assembled sequences, since all the downstream biological researches are based on this information (Stein 2001). There are two main tasks in genome annotation: gene finding and protein function assignment. Different gene finding programs were developed for the determination of open reading frames in prokaryotic (Delcher et al. 2007) and eukaryotic genomes (Burge and Karlin 1997, Majoros et al. 2004, Keller et al. 2011), since additional efforts are required in recognizing intron-exon boundaries for eukaryotic genes. In addition, homology (Altschul et al. 1997) and motif (Lowe and Eddy 1997) search programs were used for identifying non-coding RNAs. According to the principle that shared sequence implies shared function, protein functions are derived from the most similar hits among various protein databases, such as the non-redundant database, Uniprot, COG (Tatusov et al. 2003), TIGRFAMs (Haft et al. 2003) and Pfam (Finn et al. 2014). InterPro is a very useful annotation package by gathering many protein signature databases (Hunter et al. 2009), it also includes a search engine, InterProScan, for whole proteome analysis (Mulder and Apweiler 2007). Nowadays, annotation is a routine work after genome sequencing and mostly processed by automatic pipelines compiled by different tools. However, the involvement of manual curation can significantly improve the annotation quality.

1.2.3 Genome databases

Since the completion of the first bacterial genome *Haemophilus influenza* (Fleischmann et al. 1995), there has been an exponential growth of sequenced genomes (**Fig. 4**). To date, various genome databases are available, such as the Joint Genome Institute genome portal (Grigoriev et al. 2012), the Comprehensive Microbial Resource (Peterson et al. 2001) and the Saccharomyces Genome Database (Cherry et al. 2012). These specialized genome databases usually are manually curated and with detail annotation, but often targeting for specific organisms and lineages. The genomes of these databases were also deposited into the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/) of United States of America. NCBI houses the most comprehensive biological information databases in the world.

There are two genome databases in NCBI: GenBank and RefSeq (Pruitt et al. 2009). Both contain complete and partial genomes, and can be accessed through the file transfer protocol (FTP) services. GenBank, together with the European Molecular Biology Laboratory (EMBL) and the DNA Data Bank of Japan (DDBJ), are the three major portals for international biological sequence submission. Sequences produced in laboratories, genome centers and companies throughout the world should be submitted to one of the three portals, in order to obtain the accession numbers for publications. Submitted sequences are synchronized between the three databases in a daily basis so as to maintain the data consistency. The GenBank genome database (ftp.ncbi.nih.gov/genbank/genomes/) deposits the submitted genomes, which are required to meet certain annotation criteria before acceptance. However, GenBank does not curate the genomes and just act as an archive of sequence data. In contrast, the RefSeq genome database (ftp.ncbi.nlm.nih.gov/genomes/) is curated (Pruitt et al. 2009). The genome records of RefSeq are mostly from GenBank, in addition to other curated sources. However, they will be reviewed by NCBI staff and may be re-annotated to have more literature references and complete annotation. Therefore, these standardized RefSeq genomes are amenable to data mining studies.

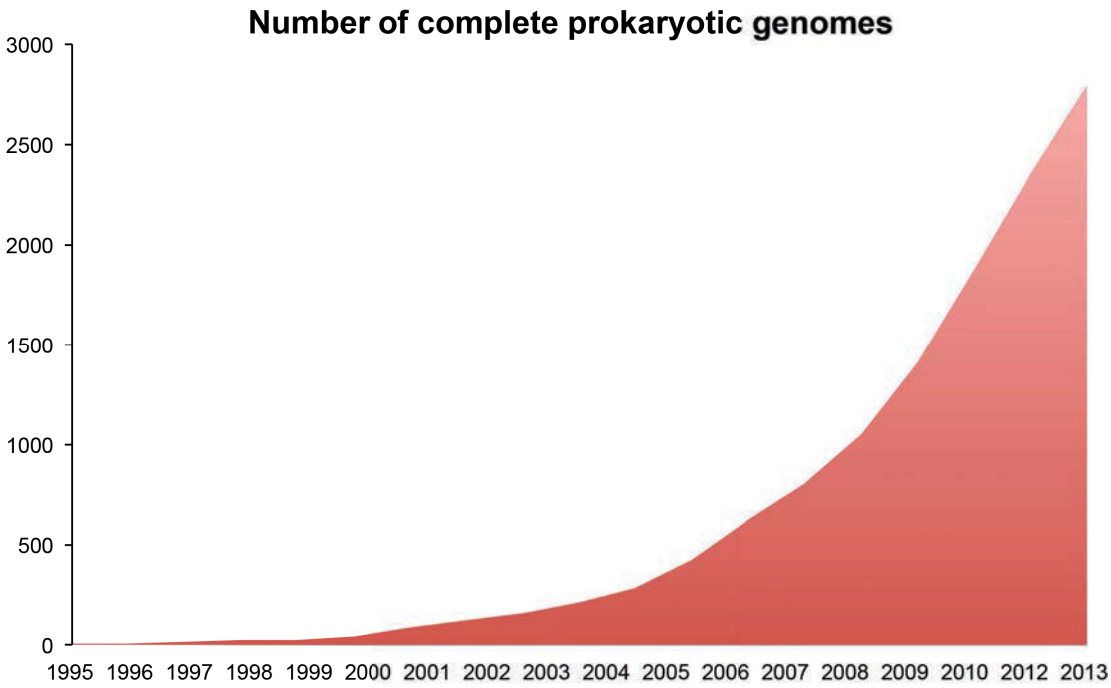

Fig. 4. The exponential growth curve of complete prokaryotic genomes deposited in NCBI by years starting from 1995, when the first complete bacterial genome was published (Fleischmann et al. 1995).

1.3 Cyanobacteria

Cyanobacteria are known as one of the earliest lineages on earth according to the fossil records that could be dated back to at least 2.8 billion years ago (Olson 2006). Ever since, as the only prokaryotic lineage capable of oxygenic photosynthesis, they have been significantly contributing to the atmospheric composition in biosphere (Bryant 1994), which have had impact on all life on earth. In the future, photosynthetic activities in

cyanobacteria may be utilized as revolutionary biotechnological applications in reducing carbon emissions by directing CO_2 into fuels (Atsumi et al. 2009).

Cyanobacteria were considered as a rich source of vast array of secondary metabolites produced by nonribosomal pathways (Welker and von Döhren 2006, Sivonen and Börner 2008, Dittmann et al. 2013). Recent studies demonstrated that cyanobacteria also are emerging as a prolific source of ribosomal natural products (Sivonen et al. 2010, Velasquez and van der Donk 2011, Arnison et al. 2013). Biotechnological and biomedical applications have been derived from these highly diverse natural products (Rastogi and Sinha 2009, Nunnery et al. 2010). Therefore, cyanobacterial natural product research has been attracting scientific research interests.

1.3.1 *Anabaena*

Anabaena is a genus of filamentous nitrogen-fixing cyanobacteria (Rippka et al. 2001). They are common in fresh and brackish water systems (Sivonen et al. 1990, Halinen et al. 2008). *Anabaena* can develop heterocyst cells for nitrogen fixation under nitrogen limitation (Flores and Herrero 2010). *Anabaenas* also produce a diverse range of natural products both from nonribosomal (Rouhiainen et al. 2000, Rouhiainen et al. 2004, Rouhiainen et al. 2010, Rantala-Ylinen et al. 2011) and ribosomal biosynthetic pathways (Leikoski et al. 2010). Many of these are toxins and toxic blooms of *Anabaena* cause serious health risk to human and livestocks (Sivonen 2009).

1.3.2 Cyanobacterial genomics

To date, a total of 73 complete genomes of cyanobacteria have been released (**Table 2**). The previously biased taxonomic distribution toward marine *Prochlorococcus* and *Synechococcus* strains in this phylum had been recently improved (Shih et al. 2013). These cyanobacteria have different lifestyles and habitats, from marine and freshwater aquatic environments to terrestrial soil and deserts, and display diverse genomic organizations (Hess 2011). Comparative genomic analysis were applied to address the issues of evolution (Swingley et al. 2008a), physiology and ecology of cyanobacteria (Hess 2008). In addition, genome sequencing revealed that a number of cyanobacteria were shown to be potent producers of natural products (Kaneko et al. 2007, Frangeul et al. 2008, Rounge et al. 2009, Voss et al. 2013).

1.4 Comparative genomic analysis for natural product biosynthetic gene clusters

The purpose of genome sequencing is to determine nucleotide sequence of genetic materials (chromosomes and/or plasmids) and identify the locations of genes, indicated by the blue solid boxes and horizontal lines in **Fig. 5**. The trend of genomic research is switching from sequencing to mining, given the huge amount of genomic data sequenced (**Fig. 4**). In this study, a comparative genomic analysis approach was developed for the characterization and comparison of natural product biosynthetic gene clusters, provided by the growing number of software tools and databases (Rausch et al. 2005, Caboche et al. 2008, Bachmann and Ravel 2009, Medema et al. 2011). The logic of this genome mining approach is illustrated in **Fig. 5**, by the line and frame in red.

Table 2. List of complete cyanobacterial genomes.

Organism	SubGroup	Size (Mb)	GC %	# Gene	# Protein	Release Date	Reference
Acaryochloris marina MBIC11017	Oscillatoriophycideae	8.36	46.99	8571	8383	16.10.2007	Swingley et al. 2008b
Anabaena sp. 90	Nostocales	5.31	38.09	4797	4511	13.11.2012	**I**
Anabaena cylindrica PCC 7122	Nostocales	7.06	38.8	6258	5838	7.12.2012	Shih et al. 2013
Anabaena variabilis ATCC 29413	Nostocales	7.11	41.39	5813	5710	15.9.2005	
Arthrospira platensis NIES-39	Oscillatoriophycideae	6.79	44.3	6676	6630	26.3.2010	Fujisawa et al. 2010
Calothrix sp. PCC 6303	Nostocales	6.96	39.8	5841	5535	6.12.2012	Shih et al. 2013
Calothrix sp. PCC 7507	Nostocales	7.02	42.2	6250	5950	4.12.2012	Shih et al. 2013
Chamaesiphon minutus PCC 6605	Oscillatoriophycideae	6.76	45.65	6426	5945	5.12.2012	Shih et al. 2013
Chroococcidiopsis thermalis PCC 7203	Pleurocapsales	6.69	44.44	6033	5752	5.12.2012	Shih et al. 2013
Crinalium epipsammum PCC 9333	Oscillatoriophycideae	5.62	40.2	5059	4752	6.12.2012	Shih et al. 2013
Cyanobacterium aponinum PCC 10605	Oscillatoriophycideae	4.18	34.97	3614	3431	7.12.2012	Shih et al. 2013
Cyanobacterium stanieri PCC 7202	Oscillatoriophycideae	3.16	38.7	2941	2837	7.12.2012	Shih et al. 2013
Cyanobium gracile PCC 6307	Oscillatoriophycideae	3.34	68.7	3437	3280	4.12.2012	Shih et al. 2013
Cyanothece sp. ATCC 51142	Oscillatoriophycideae	5.46	37.97	5364	5303	1.4.2008	Welsh et al. 2008
Cyanothece sp. PCC 7424	Oscillatoriophycideae	6.55	38.5	5942	5710	17.12.2008	Bandyopadhyay et al. 2011
Cyanothece sp. PCC 7425	Oscillatoriophycideae	5.79	50.66	5507	5327	9.1.2009	Bandyopadhyay et al. 2011
Cyanothece sp. PCC 7822	Oscillatoriophycideae	7.84	39.87	7042	6642	15.9.2010	Bandyopadhyay et al. 2011
Cyanothece sp. PCC 8801	Oscillatoriophycideae	4.79	39.8	4619	4367	17.12.2008	Bandyopadhyay et al. 2011
Cyanothece sp. PCC 8802	Oscillatoriophycideae	4.8	39.8	4700	4444	26.8.2009	Bandyopadhyay et al. 2011
Cylindrospermum stagnale PCC 7417	Nostocales	7.61	42.2	6738	6229	6.12.2012	Shih et al. 2013
Dactylococcopsis salina PCC 8305	Oscillatoriophycideae	3.78	42.4	3684	3337	7.12.2012	Shih et al. 2013
Geitlerinema sp. PCC 7407	Oscillatoriophycideae	4.68	58.5	3912	3815	5.12.2012	Shih et al. 2013
Gloeobacter kilaueensis JS1	Gloeobacteria	4.72	60.5	4562	4507	28.10.2013	Saw et al. 2013

Gloeobacter violaceus PCC 7421	Gloeobacteria	4.66	62	4482	4430	25.9.2003	Nakamura et al. 2003
Gloeocapsa sp. PCC 7428	Oscillatoriophycideae	5.88	43.39	5304	5011	6.12.2012	Shih et al. 2013
Halothece sp. PCC 7418	Oscillatoriophycideae	4.18	42.9	3920	3708	7.12.2012	Shih et al. 2013
Leptolyngbya sp. PCC 7376	Oscillatoriophycideae	5.13	43.9	4654	4228	4.12.2012	Shih et al. 2013
Microcoleus sp. PCC 7113	Oscillatoriophycideae	7.97	46.2	6821	6441	6.12.2012	Shih et al. 2013
Microcystis aeruginosa NIES-843	Oscillatoriophycideae	5.84	42.3	6364	6312	25.1.2008	Kaneko et al. 2007
Nostoc sp. PCC 7107	Nostocales	6.33	40.4	5538	5237	4.12.2012	Shih et al. 2013
Nostoc sp. PCC 7120	Nostocales	7.21	41.22	6213	6129	28.11.2001	Kaneko et al. 2001
Nostoc sp. PCC 7524	Nostocales	6.72	41.5	5687	5449	4.12.2012	Shih et al. 2013
Nostoc punctiforme PCC 73102	Nostocales	9.06	41.34	7164	6689	24.4.2008	Meeks et al. 2001
Oscillatoria acuminata PCC 6304	Oscillatoriophycideae	7.8	47.61	6100	5796	5.12.2012	Shih et al. 2013
Oscillatoria nigro-viridis PCC 7112	Oscillatoriophycideae	8.27	45.81	7006	6360	6.12.2012	Shih et al. 2013
Pleurocapsa sp. PCC 7327	Pleurocapsales	4.99	45.2	4665	4268	5.12.2012	Shih et al. 2013
Prochlorococcus marinus str. AS9601	Prochlorales	1.67	31.3	1965	1920	19.1.2007	Kettler et al. 2007
Prochlorococcus marinus str. MIT 9211	Prochlorales	1.69	38	1900	1854	13.11.2007	Kettler et al. 2007
Prochlorococcus marinus str. MIT 9215	Prochlorales	1.74	31.1	2054	1982	20.9.2007	Kettler et al. 2007
Prochlorococcus marinus str. MIT 9301	Prochlorales	1.64	31.3	1962	1906	5.3.2007	Kettler et al. 2007
Prochlorococcus marinus str. MIT 9303	Prochlorales	2.68	50	3136	2997	22.1.2007	Kettler et al. 2007
Prochlorococcus marinus str. MIT 9312	Prochlorales	1.71	31.2	1856	1810	2.11.2005	Coleman et al. 2006
Prochlorococcus marinus str. MIT 9313	Prochlorales	2.41	50.7	2330	2269	14.8.2003	Rocap et al. 2003
Prochlorococcus marinus str. MIT 9515	Prochlorales	1.7	30.8	1964	1905	19.1.2007	Kettler et al. 2007
Prochlorococcus marinus str. NATL1A	Prochlorales	1.86	35	2250	2193	22.1.2007	Kettler et al. 2007
Prochlorococcus marinus str. *NATL2A*	Prochlorales	1.84	35.1	2228	2162	9.8.2005	Kettler et al. 2007
Prochlorococcus marinus str. CCMP1375	Prochlorales	1.75	36.4	1913	1863	25.7.2003	Dufresne et al. 2003
Prochlorococcus pastoris str. CCMP1986	Prochlorales	1.66	30.8	1762	1717	14.8.2003	Rocap et al. 2003
Pseudanabaena sp. PCC 7367	Oscillatoriophycideae	4.89	46.21	4014	3854	5.12.2012	Shih et al. 2013

Rivularia sp. PCC 7116	Nostocales	8.73	37.5	6946	6644	4.12.2012	Shih et al. 2013
Stanieria cyanosphaera PCC 7437	Pleurocapsales	5.54	36.27	5041	4781	6.12.2012	Shih et al. 2013
Synechococcus sp. CC9311	Oscillatoriophycideae	2.61	52.4	2944	2892	1.9.2006	Palenik et al. 2006
Synechococcus sp. CC9605	Oscillatoriophycideae	2.51	59.2	2756	2645	27.10.2005	Dufresne et al. 2008
Synechococcus sp. CC9902	Oscillatoriophycideae	2.23	54.2	2357	2306	26.10.2005	Dufresne et al. 2008
Synechococcus sp. JA-2-3B'a(2-13)	Oscillatoriophycideae	3.05	58.5	2942	2862	6.2.2006	Bhaya et al. 2007
Synechococcus sp. JA-3-3Ab	Oscillatoriophycideae	2.93	60.2	2897	2760	6.2.2006	Bhaya et al. 2007
Synechococcus sp. PCC 6312	Oscillatoriophycideae	3.72	48.49	3794	3545	4.12.2012	Shih et al. 2013
Synechococcus sp. PCC 7002	Oscillatoriophycideae	3.41	49.16	3238	3187	14.3.2008	
Synechococcus sp. PCC 7502	Oscillatoriophycideae	3.58	40.6	3666	3318	5.12.2012	Shih et al. 2013
Synechococcus sp. RCC307	Oscillatoriophycideae	2.22	60.8	2581	2533	19.5.2007	Dufresne et al. 2008
Synechococcus sp. WH 7803	Oscillatoriophycideae	2.37	60.2	2586	2533	19.5.2007	Dufresne et al. 2008
Synechococcus sp. WH 8102	Oscillatoriophycideae	2.43	59.4	2581	2519	15.8.2003	Palenik et al. 2003
Synechococcus elongatus PCC 6301	Oscillatoriophycideae	2.7	55.5	2581	2522	14.12.2004	Sugita et al. 2007
Synechococcus elongatus PCC 7942	Oscillatoriophycideae	2.74	55.46	2715	2662	8.11.2005	
Synechocystis sp. PCC 6803	Oscillatoriophycideae	3.95	47.35	3625	3575	31.8.1995	Kaneko et al. 1996
Synechocystis sp. PCC 6803 GT-S	Oscillatoriophycideae	3.57	47.7	3219	3170	2.7.2011	Tajima et al. 2011
Synechocystis sp. PCC 6803 PCC-M	Oscillatoriophycideae	3.95	47.35	3610	3561	19.2.2013	Trautmann et al. 2012
Synechocystis sp. PCC 6803 substr. GT-I	Oscillatoriophycideae	3.57	47.7	3217	3168	2.12.2011	Kanesaki et al. 2012
Synechocystis sp. PCC 6803 substr. PCC-N	Oscillatoriophycideae	3.57	47.7	3217	3168	2.12.2011	Kanesaki et al. 2012
Synechocystis sp. PCC 6803 substr. PCC-P	Oscillatoriophycideae	3.57	47.7	3218	3169	2.12.2011	Kanesaki et al. 2012
Thermosynechococcus elongatus BP-1	Oscillatoriophycideae	2.59	53.9	2525	2476	17.8.2002	Nakamura et al. 2002
Trichodesmium erythraeum IMS101	Oscillatoriophycideae	7.75	34.1	5126	4451	6.7.2006	
'Nostoc azollae' 0708	Nostocales	5.49	38.33	5380	3651	14.6.2010	Ran et al. 2010

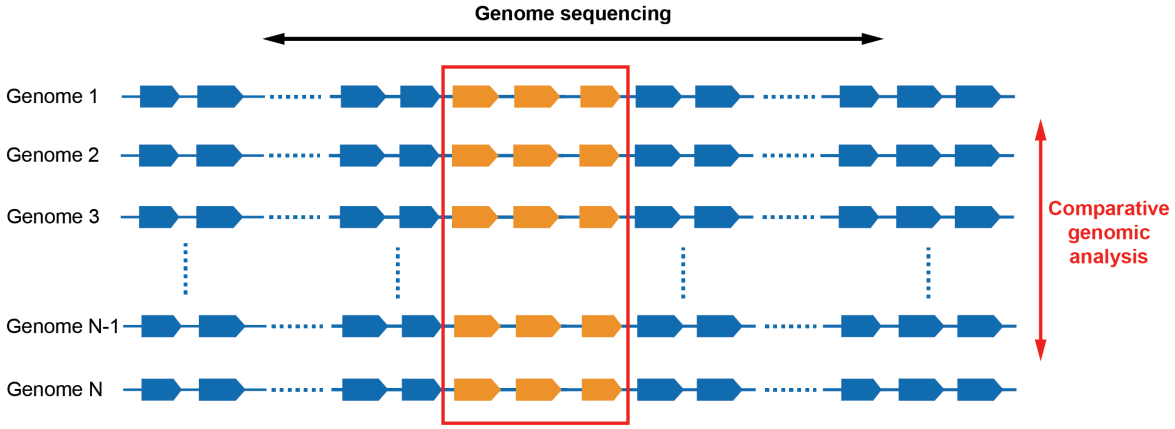

Fig. 5. Illustration of comparative genomic analysis for natural product biosynthetic pathways. Genome sequencing provides genetic information of genes and their locations for a specific organism, which are indicated by the horizontal lines and solid boxes in blue. The continuously increasing genomes are becoming valuable resources for natural product research. Their biosynthetic genes, which are indicated by orange solid boxes inside the red frame, can be located using software tools and databases. Comparative analysis of these gene clusters will yield biological insights into their biosynthetic mechanism, regulation and modifications.

2. Aim of the study

In the study, I aim to explore the genetic machineries of natural products synthesized both from ribosomally and nonribosomally pathways, by using genome sequencing and comparative analysis approaches.

The specific aims of this thesis were:

1. To complete and annotate the genome of hepatotoxin-producing and filamentous cyanobacterium *Anabaena* sp. 90 in order to characterize of the biosynthetic machinery used to assemble its full array of bioactive peptides. (**I**)

2. To study the distribution and evolution of ribosomally synthetized bacteriocin pathways in cyanobacteria, and to discover bacteriocin precursor genes encoded by the discovered gene clusters. (**II**)

3. To conduct systematically comparative analysis of biosynthetic pathways of natural products synthesized by nonribosomal peptide synthetases and polyketide synthases in the three domains of life. (**III**)

3. Materials and methods

The materials and methods used in this study are listed. Detailed descriptions are in the original papers, **I** – **III**.

3.1 Strains and organisms

The cyanobacterium *Anabaena* sp. 90 was selected for whole genome sequencing (**I**). A total of 58 cyanobacterial genomes were downloaded from the Genbank database for bacteriocin gene cluster analysis (**II**). Genome mining of NRPS/PKS biosynthetic pathways was based on a total of 2,699 complete archaean, bacterial, fungal and protozoan genomes and partial genomes of other eukaryotic organisms (**III**).

3.2 Methods used in this study

In this study, whole genome shotgun approach was applied in sequencing the *Anabaena* sp. 90 genome (**I**), and comparative genomic analysis for studying the natural product biosynthetic pathways (**II**, **III**).

The databases and software tools for specific bioinformatic analysis are listed in **Table 3**.

Table 3. Bioinformatic analysis tools and databases used in the original publications (**I** – **III**).

Bioinformatic Analysis	Tools/packages/databases	Articles
Base calling	Phred	I
Genome assembly	Phrap	I
Genome finishing	Consed	I
Gene finding	Glimmer	I
tRNA locating	tRNAscan-SE	I
rRNA locating	blastn search	I
Role category	Comprehensive Microbial Resoures	I
RM system	REBASE	I
IS identification	ISFinder	I
Metabolic pathway prediction	Pathway Tools	I
Genome view and editing	Artemis	I, II, III
Protein search	Blastp	I, II, III
Domain recongnition	Interpro	I
Protein classification	COG	I
Motif search	HMMER	III
NRPS, PKS prediction	2metDB	III
2nd metabolic pathway identification	antiSMASH	III
Multiple alignment	ClustalW, MUSCLE	I, II, III
Phylogenetic analysis	MEGA, PhyML	II, III

4. Results and discussion

4.1 Complete genome of *Anabaena* sp. 90

A rare multichromosome compositon of *Anabaena* sp. 90 was discovered in the first time for the order *Nostocales* (**I**). So far, there is only one other cyanobacterial strain (*Cyanothece* sp. ATCC 51142) in cyanobacteria that has two chromosomes, one circular and one linear (Welsh et al. 2008). Highly abundant mobile genetic elements with diverse sizes were found in the genome of *Anabaena* sp. 90 (**I**) and other bloom-forming cyanobacteria (Kaneko et al. 2007, Frangeul et al. 2008, Rounge et al. 2009, Stucken et al. 2010, Voss et al. 2013). These elements are present as repeats in the genomes and cause serious problems in genome assembly, so that many genomes of bloom-forming cyanobacteria are incomplete (Frangeul et al. 2008, Rounge et al. 2009, Stucken et al. 2010, Voss et al. 2013). The *Anabaena* sp. 90 genome was successfully completed with high sequence quality by using large insert paired-end data and elaborate finishing (**I**). Moreover, detailed annotation revealed that nearly five percent of the genes were found with mutations or disrupted open reading frames, and thus identified as pseudogenes. Their occurrences can be associated with the activities of various mobile genetic elements, such as MITEs (**I**). Therefore, pseudogenes caused by mobile genetic elements might be common among cyanobacteria, given their richness of transposable elements (Lin et al. 2011), especially in bloom-forming *Microcystis aeruginosa* strains (Kaneko et al. 2007, Frangeul et al. 2008). However, the reports of pseudogenes have remained undocumented in cyanobacterial genomes previously, except for an *Azolla* symbiont (Ran et al. 2010). Furthermore, based on metabolic pathway analysis, it was revealed that these mobile genetic elements have been trimming nonessential genes and pathways of *Anabaena* sp. 90 during cultivation under laboratory conditions, in which optimal conditions and nutrients are available. The truncated *gvpG* gene in the *gvp* operon appeared as an example, which led to the phenotype with loss of buoyancy after introducing to a purified (axenic) culture (Rouhiainen et al. 1995). Because the mobile elements are usually considered as foreign origins and RM systems function as microbial defence systems against foreign DNAs (Ershova et al. 2012). Thus it was surprising to find a pronounced set of restriction-modification (RM) systems in *Anabaena* sp. 90 (**I**). This might be the outcome of inefficiency of RM systems.

Anabaena sp. 90 was previously known to produce nonribosomal peptides anabaenopeptilides (Rouhiainen et al. 2000), anabaenopeptins (Rouhiainen et al. 2010) and microcystins (Rouhiainen et al. 2004). The genome project revealed another large gene cluster responsible for production of glycolipopeptides (hassallidins), which were remained undetected because the gene cluster was inactivated by a deletion event (**I**). Unlike previous works that usually just consider nonribosomal biosynthetic pathways of bioactive peptides, gene clusters of ribosomally synthesized natural products were also taken into account in this study (**I**). An anacyclamide-encoding cyanobactin gene cluster (Leikoski et al. 2010) and seven putative bacteriocin gene clusters were located. The mutants with an inactive hassallidin biosynthetic pathway were prevailed over the cells with functional gene cluster in the culture. Hence it appeared that the cells with inactivated hassallidin gene cluster obtained certain growth advantages, perhaps due to a lower metabolic burden after stopping hassallidin production. And the parallel culture of mutant *Anabaena* sp. 90, which has an inactivated anabaenopeptilide synthetase gene cluster (Rouhiainen et al. 2000), still has the intact hassallidin gene cluster.

Anabaena is a genus of filamentous diazotrophic cyanobacteria that are capable of both photosynthesis and nitrogen fixation. These two processes are spatially separated into vegetative cells and dispersed heterocysts in the filaments (Flores and Herrero 2010). The heterocyst differentiation process involves regulated cleavage of excision elements and re-joining of chromosomes (Golden et al. 1991). Previously, three excision elements (*nifD*, *fdxN* and *hupL*) were known in *Anabaena/Nostoc* PCC 7120 (Carrasco et al. 1995, Henson et al. 2011). These excision elements may function in insulating nitrogenase and hydrogenase from oxygen generated in vegetative cells (Fay 1992). Here, the simultaneous presence of four excision elements, including two novel *nifH* elements, was found in *Anabaena* sp. 90 (**I**). One of the *nifH* elements has the largest known size (80 kb). These elements extend the *nif* operon of *Anabaena* sp. 90 to 122 kb, on which precise genomic splicing occurs during the heterocyst differentiation (**I**). Inactivation of *patS* and *hetN* genes lead to overproduction of HetR and heterocysts, which is lethal to cells of *Anabaena*/*Nostoc* sp. PCC 7120 (Borthakur et al. 2005). Thus, the absence of the *patS* and *hetN* genes in *Anabaena* sp. 90 (**I**) and *Cylindrospermopsis raciborskii* CS-505 (Stucken et al. 2010) suggest a new pattern of heterocyst spatial development.

4.2 Widespread occurrences of cyanobacterial bacteriocin gene clusters and the classification

The analysis of *Anabaena* sp. 90 genome (**I**) promoted the discovery of widespread existence of bacteriocins biosynthetic pathways in cyanobacteria (**II**). A total of 145 bacteriocin gene clusters were discovered from genomes of nearly all examined cyanobacterial species. These gene clusters were classified into seven groups based on their gene organization and domain composition. This classification is supported by the phylogenetic analysis.

Lantibiotics have specific intramolecular ring structures (Bierbaum and Sahl 2009). The counterparts of lanthionine modification enzyme, LanM, have been found in many cyanobacterial genomes (Begley et al. 2009, Goto et al. 2010, O'Sullivan et al. 2011). In this study, the LanM-containing gene clusters were assigned to group IV. There are nine *lanM* genes found without associated bacteriocin biosynthetic machinery, such as the one in *Prochlorococcus marinus* MIT9313 that is the first known bacteriocin-producing strain in cyanobacteria (Li et al. 2010b). Since the phylogenetic analysis suggests that LanM-encoding gene clusters may work together with gene clusters in groups III, V, and VII (**II**). Therefore, the lantipeptides produced in *Prochlorococcus marinus* MIT9313 may be synthesized by the stand-alone LanM together with the other two group V gene clusters.

4.3 Cyanobacterial bacteriocin precursor genes and their discovery

In cyanobacteria, two new families of double-glycine-type precursors, NHLP and N11P, were recently discovered (Haft et al. 2010). Metagenomic analysis suggested that huge amount of lantibiotics might be produced by strains of *Prochlorococcus* and *Synechococcus* in oceans (Li et al. 2010b). In this study, hundreds of new putative precursors were found through genome mining (**II**). As a result, more attention will be

attracted to cyanobacteria for ribosomal natural products research (Velasquez and van der Donk 2011).

Unlike previous bacteriocin finding tools that directly look for precursor genes (de Jong et al. 2006, 2010), which are difficult to be located due to their compact sizes and often absent from conventional genome annotation, a new approach was developed in this study by firstly locating large biosynthetic genes and then looking for precursors in the surrounding regions. The effectiveness of this method was validated by the large amount of novel precursors predicted.

4.4 Distribution of NRPS and PKS biosynthetic pathways

The genome mining presented an atlas of NRPS/PKS biosynthetic pathways (**III**). A total of 3,339 NRPS, PKS and hybrid NRPS/PKS gene clusters were discovered from all three domains of life, by scanning a total of 15.72 Gb genomic sequences (**Table 4**). The majority of the gene clusters (2976, 89.1%) were from bacteria. In contrast, only three clusters (0.1%) were detected from archaean genomes. The remaining gene clusters are mostly from fungi (307, 9.2%), in addition to 53 (1.6%) clusters from other eukaryotic organisms.

Table 4. Summary of genomes and gene clusters analyzed in this study.

Domain	Gene cluster				Genome	
	#	# protein	# domain	Cumulative szie (Mb)	#	Cumulative size (Gb)
Bacteria	2,976	15,889	56,269	95.88	2,478	8,69
Archaea	3	6	27	0.03	160	0.38
Eukarya	360	699	3,304	6.44	61	6.65
Total	**3,339**	**16,594**	**59,600**	**102.35**	**2,699**	**15.72**

The close relationship between NPRS and PKS was demonstrated by the discovery that one third (1145, 34.5%) of the gene clusters are hybrid NRPS-PKS type (**III**), from which a total of 462 hybrid NRPS-PKS megasynthetases were found. These hybrid gene clusters tend to possess more domains than the stand-alone NRPS/PKS clusters. For example, nearly all trans-AT gene clusters, in which the discrete AT protein is iteratively used for the substrate loading to AT-less PKSs (Cheng et al. 2002, 2003), were found as hybrids.

Highly diverse structural organizations were observed in these gene clusters. The protein and domain number are mostly less than 10 and 20, respectively. Data mining also revealed the extremes of highly modular NRPS and PKS biosynthetic pathways. The largest gene cluster is a hybrid one, which contains 189 domains spread over 23 proteins, including 6 large PKSs, 4 NRPSs and talorizing enzymes, gathering in a range of ~250 kb in the chromosome of *Actinoplanes* sp. N902-109. While the largest NRPS and PKS gene clusters have 99 domains (7 enzymes) and 92 domains (3 enzymes), and from *Pseudomonas syringae* B728a and *Mycobacterium ulcerans* Agy99, respectively. At protein level, the largest PKS (MULP_065, 17,019 aa) from *Mycobacterium liflandii* 128FXT has nine modules constituting 47 domains, while the largest NRPS (plu2670, 16,367 aa) from *Photorhabdus luminescens* TTO1 was found with 15 modules consisting of 46 domains, and the longest hybrid enzyme (MXAN_3779, 14,274 aa) from *Myxococcus xanthus* DK 1622 possesses 39 domains fused into 1 PKS and 11 NRPS

modules.

Sequence homologies of condensation domain can be evidently observed within the major bacterial lineages that produce nonribosomal peptides, in addition to traces of horizontal gene transfer events (**III**). This suggested that the evolutionary process of nonribosomal machineries involved both common descent and horizontal gene transfer.

4.5 Common occurrence of nonmodular NRPS and PKS biosynthetic machineries

Through examining the domain composition of NRPSs and PKSs discovered from these pathways, a great number (8,906) of monodomain enzymes that just carry a single functional domain were found in bacteria (**III**). Their genetic organization differs from the current classification, by which NRPSs and type I PKSs are defined as multidomain enzymes. Many of these monodomain enzymes were further discovered in close genomic vicinity and form around one tenth of bacterial gene clusters, which are similar to type II PKSs in genetic organization. The finding of these type II-like gene clusters revealed common occurrence of nonmodular NRPS and PKS biosynthetic machineries, which may represent the primitive form of modular NRPS/PKS biosynthetic machineries. And the hybrid between NRPS and PKS already occurred in these gene clusters. Although the biosynthetic mechanism of these novel pathways is unclear, they had been known for the biosynthesis of nonribosomal peptides, such as acinetobactin (Mihara et al. 2004).

5. Conclusions and future perspectives

The sequencing project of *Anabaena* sp. 90 generated its complete genome with high sequence quality and detailed annotation. This genome contains a high number of pseudogenes that could be attributed to the activities of diverse mobile genetic elements. The study indicated that the selective pressure imposed by these mobile genetic elements have been trimming the nonessential genes and pathways, including the secondary metabolites pathways, under the laboratory conditions. In addition, the array of biosynthetic gene clusters that represent the full capacity of bioactive compounds production revealed by this study will serve as a basis to study the regulation of natural product biosynthesis by functional genomic analysis.

Comparative genomic analysis in cyanobacteria revealed the widespread occurrence of bacteriocin gene clusters. This indicates that cyanobacteria are a prolific source of the posttranslationally modified peptides, in addition to end products from nonribosomal pathways. The classification and phylogeny of these newly discovered bacteriocin gene clusters in cyanobacteria represent a major expansion of bacteriocin research from the previous focus mainly in lactic acid bacteria. The study also showed that bioinformatic analysis is becoming an effective approach in locating candidates of novel natural products. A hint from this study is that there might be many gene clusters for ribosomal natural product biosynthesis in other bacterial lineages, which are awaiting for identification by more extensive genome sequencing and mining.

Consequently, a genome-mining study for NRPS and PKS biosynthetic gene clusters was carried out. This study presented a comprehensive atlas of nonribosomal peptide and modular polyketide biosynthetic pathways in nature, and demonstrated their widespread distribution. The data generated laid a solid basis for future studies in refining biosynthetic mechanism, finding new tailoring domains/enzymes and novel natural products, as well as the chronological phylogeny of the biosynthetic pathways. The surprising finding of common presence of nonmodular peptide synthetase and polyketide synthase further corroborates the capacity of genome-mining in creation of new knowledge.

In the future, there is no doubt that genome-based natural product research will expand its scope both in breadth (from bacteriocins, NRPs and PKs to others, type II and III PKs, terpenoids, lantipeptides etc.) and depth (from gene clusters to structure prediction of products and drug lead screening). This study represents a trend in the post-genomic era, when high amount of genomic data has been accumulated and data mining is urgently required. The comparative genomic analysis is playing a more pronounced role in sorting and mining the massive information for natural product research.

6. Acknowledgements

This work was financially supported by Viikki Doctoral Programme in Molecular Biosciences, the Academy of Finland (Research Center of Excellence grant 118637 and 1258827) and the University of Helsinki (Microbial Resources 53305).

At first, I want to express my most sincere gratitude to Professor Kaarina Sivonen for her guidance and continuous support to my PhD study. I would like to attribute the achievements of this study to your excellent supervision.

I really want to give my great thanks to Dr. David Fewer and Dr. Leo Rouhiainen for your valuable intellectual input to this study. David, your insightful ideas and linguistic help guarantee the smooth proceeding of the work. Leo, I still remember the first lesson about the biosynthetic gene clusters you lectured to me seven years ago. I have to say that this study is based on many of your previous works. Here, I want to thank both of you again for the discussions and a lot of useful comments to this study. It was nice to work with you in these years.

I want to express my gratitude to all other co-authors: Christina Lyra, Anne Rantala-Ylinen, Johanna Vestola, Jouni Jokela, Kaisa Rantasärkkä, Zhijie Li and Bin Liu. I thank Lyudmila Saari for maintaining and growing the cyanobacterial strains used in this study. My grateful thanks go to all the previous and current members of the Cyanobacteria group for your kind help to me, and I am happy to work with you!

Finally, I would like to present my special thanks to my dear wife, Yanhua. You did more contribution to this study than me by taking care of the family and allowing me to work freely. I am also indebted to my two lovely and beautiful princesses; it is my fortune to be with you! Here, I would like to render this thesis to the new member of my family and celebrate your coming!

7. References

1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E. & McVean, G.A. (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061-1073.

Alkan, C., Sajjadian, S. & Eichler, E.E. (2011) Limitations of next-generation genome sequence assembly. *Nature Methods* 8:61-65.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25:3389-3402.

Arnison, P.G., Bibb, M.J., Bierbaum, G., Bowers, A.A., Bugni, T.S., Bulaj, G., Camarero, J.A., Campopiano, D.J., Challis, G.L., Clardy, J., Cotter, P.D., Craik, D.J., Dawson, M., Dittmann, E., Donadio, S., Dorrestein, P.C., Entian, K.D., Fischbach, M.A., Garavelli, J.S., Goransson, U., Gruber, C.W., Haft, D.H., Hemscheidt, T.K., Hertweck, C., Hill, C., Horswill, A.R., Jaspars, M., Kelly, W.L., Klinman, J.P., Kuipers, O.P., Link, A.J., Liu, W., Marahiel, M.A., Mitchell, D.A., Moll, G.N., Moore, B.S., Muller, R., Nair, S.K., Nes, I.F., Norris, G.E., Olivera, B.M., Onaka, H., Patchett, M.L., Piel, J., Reaney, M.J., Rebuffat, S., Ross, R.P., Sahl, H.G., Schmidt, E.W., Selsted, M.E., Severinov, K., Shen, B., Sivonen, K., Smith, L., Stein, T., Sussmuth, R.D., Tagg, J.R., Tang, G.L., Truman, A.W., Vederas, J.C., Walsh, C.T., Walton, J.D., Wenzel, S.C., Willey, J.M. & van der Donk, W.A. (2013) Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Natural Product Reports* 30:108-160.

Atsumi, S., Higashide, W. & Liao, J.C. (2009) Direct photosynthetic recycling of carbon dioxide to isobutyraldehyde. *Nature Biotechnology* 27:1177-1180.

Bachmann, B.O. & Ravel, J. (2009) Chapter 8. Methods for *in silico* prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods in Enzymology* 458:181-217.

Bandyopadhyay, A., Elvitigala, T., Welsh, E., Stockel, J., Liberton, M., Min, H., Sherman, L.A. & Pakrasi, H.B. (2011) Novel metabolic attributes of the genus *cyanothece*, comprising a group of unicellular nitrogen-fixing cyanobacteria. *mBio* 2: e00214-11.

Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P. & Lander, E.S. (2002) ARACHNE: A Whole-Genome Shotgun Assembler. *Genome Research* 12:177-189.

Begley, M., Cotter, P.D., Hill, C. & Ross, R.P. (2009) Identification of a novel two-peptide lantibiotic, lichenicidin, following rational genome mining for LanM proteins. *Applied and Environmental Microbiology* 75:5451-5460.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fuentes Fajardo, K.V., Scott Furey, W., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., Hoschler, K., Hurwitz, S., Ivanov, D.V., Johnson, M.Q., James, T., Huw Jones, T.A., Kang, G.D., Kerelska, T.H., Kersey, A.D., Khrebtukova, I., Kindwall, A.P., Kingsbury, Z., Kokko-Gonzales, P.I., Kumar, A., Laurent, M.A., Lawley, C.T., Lee, S.E., Lee, X., Liao, A.K., Loch, J.A., Lok, M., Luo, S., Mammen, R.M., Martin, J.W., McCauley, P.G., McNitt, P., Mehta, P., Moon, K.W., Mullens, J.W., Newington, T., Ning, Z., Ling Ng, B., Novo, S.M., O'Neill, M.J., Osborne, M.A., Osnowski, A., Ostadan, O., Paraschos, L.L., Pickering, L., Pike, A.C., Pike, A.C., Chris Pinkard, D., Pliskin, D.P., Podhasky, J., Quijano, V.J., Raczy, C., Rae, V.H., Rawlings, S.R., Chiva Rodriguez, A., Roe, P.M., Rogers, J., Rogert Bacigalupo, M.C., Romanov, N., Romieu, A., Roth, R.K., Rourke, N.J., Ruediger, S.T., Rusman, E., Sanches-Kuiper, R.M., Schenker, M.R., Seoane, J.M., Shaw, R.J., Shiver, M.K., Short, S.W., Sizto, N.L., Sluis, J.P., Smith, M.A., Ernest Sohna Sohna, J., Spence, E.J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C.L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S.M., Wakelin, S., Walcott, G.C., Wang, J., Worsley, G.J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J.C., Hurles, M.E., McCooke, N.J., West, J.S., Oaks, F.L., Lundberg, P.L., Klenerman, D., Durbin, R. & Smith, A.J. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53-59.

Bhaya, D., Grossman, A.R., Steunou, A.S., Khuri, N., Cohan, F.M., Hamamura, N., Melendrez, M.C., Bateson, M.M., Ward, D.M. & Heidelberg, J.F. (2007) Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. *The ISME Journal* 1:703-713.

Bierbaum, G. & Sahl, H.G. (2009) Lantibiotics: mode of action, biosynthesis and bioengineering. *Current Pharmaceutical Biotechnology* 10:2-18.

Borthakur, P.B., Orozco, C.C., Young-Robbins, S.S., Haselkorn, R. & Callahan, S.M. (2005) Inactivation of *patS* and *hetN* causes lethal levels of heterocyst differentiation in the filamentous cyanobacterium *Anabaena* sp. PCC 7120. *Molecular Microbiology* 57:111-123.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S.R., Moon, K., Burcham, T., Pallas, M., DuBridge, R.B., Kirchner, J., Fearon, K., Mao, J. & Corcoran, K. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology* 18:630-634.

Bryant, D.A. (1994) The molecular biology of cyanobacteria. Kluwer Academic Publishers, Boston.

Burge, C. & Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* 268:78-94.

Caboche, S., Pupin, M., Leclere, V., Fontaine, A., Jacques, P. & Kucherov, G. (2008) NORINE: a database of nonribosomal peptides. *Nucleic Acids Research* 36:D326-31.

Carrasco, C.D., Buettner, J.A. & Golden, J.W. (1995) Programmed DNA rearrangement of a cyanobacterial *hupL* gene in heterocysts. *Proceedings of the National Academy of Sciences of the United States of America* 92:791-795.

Cheng, Y.Q., Tang, G.L. & Shen, B. (2002) Identification and localization of the gene cluster encoding biosynthesis of the antitumor macrolactam leinamycin in *Streptomyces atroolivaceus* S-140. *Journal of Bacteriology* 184:7013-7024.

Cheng, Y.Q., Tang, G.L. & Shen, B. (2003) Type I polyketide synthase requiring a discrete acyltransferase for polyketide biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America* 100:3149-3154.

Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Karra, K., Krieger, C.J., Miyasato, S.R., Nash, R.S., Park, J., Skrzypek, M.S., Simison, M., Weng, S. & Wong, E.D. (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Research* 40:D700-5.

Clarke, J., Wu, H.C., Jayasinghe, L., Patel, A., Reid, S. & Bayley, H. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology* 4:265-270.

Coleman, M.L., Sullivan, M.B., Martiny, A.C., Steglich, C., Barry, K., Delong, E.F. & Chisholm, S.W. (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* 311:1768-1770.

Cotter, P.D., Hill, C. & Ross, R.P. (2005) Bacteriocins: developing innate immunity for food. *Nature Reviews Microbiology* 3:777-788.

Cotter, P.D., Ross, R.P. & Hill, C. (2013) Bacteriocins - a viable alternative to antibiotics? *Nature Reviews Microbiology* 11:95-105.

de Jong, A., van Heel, A.J., Kok, J. & Kuipers, O.P. (2010) BAGEL2: mining for bacteriocins in genomic data. *Nucleic Acids Research* 38:W647-51.

de Jong, A., van Hijum, S.A., Bijlsma, J.J., Kok, J. & Kuipers, O.P. (2006) BAGEL: a web-based bacteriocin genome mining tool. *Nucleic Acids Research* 34:W273-9.

Delcher, A.L., Bratke, K.A., Powers, E.C. & Salzberg, S.L. (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23:673-679.

Di Ventra, M. (2013) Fast DNA sequencing by electrical means inches closer. *Nanotechnology* 24:342501.

Dirix, G., Monsieurs, P., Dombrecht, B., Daniels, R., Marchal, K., Vanderleyden, J. & Michiels, J. (2004) Peptide signal molecules and bacteriocins in Gram-negative bacteria: a genome-wide *in silico* screening for peptides containing a double-glycine leader sequence and their cognate transporters. *Peptides* 25:1425-1440.

Dittmann, E., Fewer, D.P. & Neilan, B.A. (2013) Cyanobacterial toxins: biosynthetic routes and evolutionary roots. *FEMS Microbiology Reviews* 37:23-43.

Dolled-Filhart, M.P., Lee, M.,Jr, Ou-Yang, C.W., Haraksingh, R.R. & Lin, J.C. (2013) Computational and bioinformatics frameworks for next-generation whole exome and genome sequencing. *The Scientific World Journal* 2013:730210.

Donadio, S., Busti, E., Monciardini, P., Bamonte, R., Mazza, P., Sosio, M. & Cavaletti, L. (2005) Sources of polyketides and non-ribosomal peptides. In: Wohlleben, W., Spellig, T., Muller-Tiemann, B. (eds.) *Biocombinatorial approaches for drug finding.* Heidelberg, Germany: Springer. p. 19-41.

Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., Dahl, F., Fernandez, A., Staker, B., Pant, K.P., Baccash, J., Borcherding, A.P., Brownley, A., Cedeno, R., Chen, L., Chernikoff, D., Cheung, A., Chirita, R., Curson, B., Ebert, J.C., Hacker, C.R., Hartlage, R., Hauser, B., Huang, S., Jiang, Y., Karpinchyk, V., Koenig, M., Kong, C., Landers, T., Le, C., Liu, J., McBride, C.E., Morenzoni, M., Morey, R.E., Mutch, K., Perazich, H., Perry, K., Peters, B.A., Peterson, J., Pethiyagoda, C.L., Pothuraju, K., Richter, C., Rosenbaum, A.M., Roy, S., Shafto, J., Sharanhovich, U., Shannon, K.W., Sheppy, C.G., Sun, M., Thakuria, J.V., Tran, A., Vu, D., Zaranek, A.W., Wu, X., Drmanac, S., Oliphant, A.R., Banyai, W.C., Martin, B., Ballinger, D.G., Church, G.M. & Reid, C.A. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327:78-81.

Du, L., Sanchez, C. & Shen, B. (2001) Hybrid peptide-polyketide natural products: biosynthesis and prospects toward engineering novel molecules. *Metabolic Engineering* 3:78-95.

Dufresne, A., Ostrowski, M., Scanlan, D.J., Garczarek, L., Mazard, S., Palenik, B.P., Paulsen, I.T., de Marsac, N.T., Wincker, P., Dossat, C., Ferriera, S., Johnson, J.,

Post, A.F., Hess, W.R. & Partensky, F. (2008) Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biology* 9:R90.

Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I.M., Barbe, V., Duprat, S., Galperin, M.Y., Koonin, E.V., Le Gall, F., Makarova, K.S., Ostrowski, M., Oztas, S., Robert, C., Rogozin, I.B., Scanlan, D.J., Tandeau de Marsac, N., Weissenbach, J., Wincker, P., Wolf, Y.I. & Hess, W.R. (2003) Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proceedings of the National Academy of Sciences of the United States of America* 100:10020-10025.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J. & Turner, S. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323:133-138.

English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C. & Gibbs, R.A. (2012) Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PloS One* 7:e47768.

Ershova, A.S., Karyagina, A.S., Vasiliev, M.O., Lyashchuk, A.M., Lunin, V.G., Spirin, S.A. & Alexeevski, A.V. (2012) Solitary restriction endonucleases in prokaryotic genomes. *Nucleic Acids Research* 40:10107-10115.

Fay, P. (1992) Oxygen relations of nitrogen fixation in cyanobacteria. *Microbiological Reviews* 56:340-373.

Finking, R. & Marahiel, M.A. (2004) Biosynthesis of nonribosomal peptides. *Annual Review of Microbiology* 58:453-488.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L., Tate, J. & Punta, M. (2014) Pfam: the protein families database. *Nucleic Acids Research* 42:D222-30.

Fischbach, M.A. & Walsh, C.T. (2006) Assembly-line enzymology for polyketide and nonribosomal Peptide antibiotics: logic, machinery, and mechanisms. *Chemical Reviews* 106:3468-3496.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A. & Merrick, J.M. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269:496-512.

Flores, E. & Herrero, A. (2010) Compartmentalized function through cell differentiation in filamentous cyanobacteria. *Nature Reviews Microbiology* 8:39-50.

Frangeul, L., Quillardet, P., Castets, A.M., Humbert, J.F., Matthijs, H.C., Cortez, D., Tolonen, A., Zhang, C.C., Gribaldo, S., Kehr, J.C., Zilliges, Y., Ziemert, N., Becker, S., Talla, E., Latifi, A., Billault, A., Lepelletier, A., Dittmann, E., Bouchier, C. & de Marsac, N.T. (2008) Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. *BMC Genomics* 9:274.

Fujisawa, T., Narikawa, R., Okamoto, S., Ehira, S., Yoshimura, H., Suzuki, I., Masuda, T., Mochimaru, M., Takaichi, S., Awai, K., Sekine, M., Horikawa, H., Yashiro, I., Omata, S., Takarada, H., Katano, Y., Kosugi, H., Tanikawa, S., Ohmori, K., Sato, N., Ikeuchi, M., Fujita, N. & Ohmori, M. (2010) Genomic structure of an economically important cyanobacterium, *Arthrospira* (*Spirulina*) *platensis* NIES-39. *DNA Research* 17:85-103.

Galvez, A., Lopez, R.L., Abriouel, H., Valdivia, E. & Omar, N.B. (2008) Application of bacteriocins in the control of foodborne pathogenic and spoilage bacteria. *Critical Reviews in Biotechnology* 28:125-152.

Genome 10K Community of Scientists (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *The Journal of Heredity* 100:659-674.

Golden, J.W., Whorff, L.L. & Wiest, D.R. (1991) Independent regulation of *nifHDK* operon transcription and DNA rearrangement during heterocyst differentiation in the cyanobacterium *Anabaena* sp. strain PCC 7120. *Journal of Bacteriology* 173:7098-7105.

Goto, Y., Li, B., Claesen, J., Shi, Y., Bibb, M.J. & van der Donk, W.A. (2010) Discovery of unique lanthionine synthetases reveals new mechanistic and evolutionary insights. *PLoS Biology* 8:e1000339.

Green, P. (1994) "PHRAP documentation: ALGORITHMS". (URL: http://www.phrap.org/phredphrap/phrap.html).

Grigoriev, I.V., Nordberg, H., Shabalov, I., Aerts, A., Cantor, M., Goodstein, D., Kuo, A., Minovitsky, S., Nikitin, R., Ohm, R.A., Otillar, R., Poliakov, A., Ratnere, I., Riley, R., Smirnova, T., Rokhsar, D. & Dubchak, I. (2012) The genome portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Research* 40:D26-32.

Haft, D.H., Basu, M.K. & Mitchell, D.A. (2010) Expansion of ribosomally produced natural products: a nitrile hydratase- and Nif11-related precursor family. *BMC Biology* 8:70.

Haft, D.H., Selengut, J.D. & White, O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Research* 31:371-373.

Hagen, J.B. (2000) The origins of bioinformatics. *Nature Reviews Genetics* 1:231-236.

Halinen, K., Fewer, D.P., Sihvonen, L.M., Lyra, C., Eronen, E. & Sivonen, K. (2008) Genetic diversity in strains of the genus *Anabaena* isolated from planktonic and benthic habitats of the Gulf of Finland (Baltic Sea). *FEMS Microbiology Ecology* 64:199-208.

Hammami, R. (2007) BACTIBASE: a new web-accessible database for bacteriocin characterization. *BMC Microbiology* 7:89.

Hammami, R., Zouhir, A., Le Lay, C., Ben Hamida, J. & Fliss, I. (2010) BACTIBASE second release: a database and tool platform for bacteriocin characterization. *BMC Microbiology* 10:22.

Harris, T.D., Buzby, P.R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J.W., Giladi, E., Gill, J., Healy, J., Jarosz, M., Lapen, D., Moulton, K., Quake, S.R., Steinmann, K., Thayer, E., Tyurina, A., Ward, R., Weiss, H. & Xie, Z. (2008) Single-molecule DNA sequencing of a viral genome. *Science* 320:106-109.

Håvarstein, L.S., Diep, D.B. & Nes, I.F. (1995) A family of bacteriocin ABC transporters carry out proteolytic processing of their substrates concomitant with export. *Molecular Microbiology* 16:229-240.

Henson, B.J., Hartman, L., Watson, L.E. & Barnum, S.R. (2011) Evolution and variation of the *nifD* and *hupL* elements in the heterocystous cyanobacteria. *International Journal of Systematic and Evolutionary Microbiology* 61:2938-2949.

Hertweck, C., Luzhetskyy, A., Rebets, Y. & Bechthold, A. (2007) Type II polyketide synthases: gaining a deeper insight into enzymatic teamwork. *Natural Product Reports* 24:162-190.

Hess, W.R. (2008) Comparative genomics of marine cyanobacteria and their phages. In: Herrero, A. & Flores, E. (eds.) *The cyanobacteria: molecular biology, genomics, and evolution.* Norfolk, UK: Caister Academic Press. p. 89-116.

Hess, W.R. (2011) Cyanobacterial genomics for ecology and biotechnology. *Current Opinion in Microbiology* 14:608-614.

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R.D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A.F., Selengut, J.D., Sigrist, C.J., Thimma, M., Thomas, P.D., Valentin, F., Wilson, D., Wu, C.H. & Yeats, C. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Research* 37:D211-5.

Jenke-Kodama, H., Sandmann, A., Muller, R. & Dittmann, E. (2005) Evolutionary implications of bacterial polyketide synthases. *Molecular Biology and Evolution* 22:2027-2039.

Kaneko, T., Nakajima, N., Okamoto, S., Suzuki, I., Tanabe, Y., Tamaoki, M., Nakamura, Y., Kasai, F., Watanabe, A., Kawashima, K., Kishida, Y., Ono, A., Shimizu, Y., Takahashi, C., Minami, C., Fujishiro, T., Kohara, M., Katoh, M., Nakazaki, N., Nakayama, S., Yamada, M., Tabata, S. & Watanabe, M.M. (2007) Complete genomic structure of the bloom-forming toxic cyanobacterium *Microcystis aeruginosa* NIES-843. *DNA Research* 14:247-256.

Kaneko, T., Nakamura, Y., Wolk, C.P., Kuritz, T., Sasamoto, S., Watanabe, A., Iriguchi, M., Ishikawa, A., Kawashima, K., Kimura, T., Kishida, Y., Kohara, M., Matsumoto, M., Matsuno, A., Muraki, A., Nakazaki, N., Shimpo, S., Sugimoto, M., Takazawa, M., Yamada, M., Yasuda, M. & Tabata, S. (2001) Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Research* 8:205-13; 227-253.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M. & Tabata, S. (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Research* 3:109-136.

Kanesaki, Y., Shiwa, Y., Tajima, N., Suzuki, M., Watanabe, S., Sato, N., Ikeuchi, M. & Yoshikawa, H. (2012) Identification of substrain-specific mutations by massively parallel whole-genome resequencing of *Synechocystis* sp. PCC 6803. *DNA Research* 19:67-79.

Keller, O., Kollmar, M., Stanke, M. & Waack, S. (2011) A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* 27:757-763.

Kettler, G.C., Martiny, A.C., Huang, K., Zucker, J., Coleman, M.L., Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., Johnson, J., Steglich, C., Church, G.M., Richardson, P. & Chisholm, S.W. (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genetics* 3:e231.

Kwon, H.J., Smith, W.C., Xiang, L. & Shen, B. (2001) Cloning and heterologous expression of the macrotetrolide biosynthetic gene cluster revealed a novel polyketide synthase that lacks an acyl carrier protein. *Journal of the American Chemical Society* 123:3385-3386.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M. & FitzHugh, W. (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860-921.

Lander, E.S. & Waterman, M.S. (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2:231-239.

Leikoski, N., Fewer, D.P., Jokela, J., Wahlsten, M., Rouhiainen, L. & Sivonen, K. (2010) Highly diverse cyanobactins in strains of the genus *Anabaena*. *Applied and Environmental Microbiology* 76:701-709.

Li, B., Sher, D., Kelly, L., Shi, Y., Huang, K., Knerr, P.J., Joewono, I., Rusch, D., Chisholm, S.W. & van der Donk, W.A. (2010b) Catalytic promiscuity in the biosynthesis of cyclic peptide secondary metabolites in planktonic marine cyanobacteria. *Proceedings of the National Academy of Sciences of the United States of America* 107:10430-10435.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J. & Wang, J. (2010a) *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Research* 20:265-272.

Li, Y.M., Milne, J.C., Madison, L.L., Kolter, R. & Walsh, C.T. (1996) From peptide precursors to oxazole and thiazole-containing peptide antibiotics: microcin B17 synthase. *Science* 274:1188-1193.

Lim, J.S., Choi, B.S., Lee, J.S., Shin, C., Yang, T.J., Rhee, J.S., Lee, J.S. & Choi, I.Y. (2012) Survey of the applications of NGS to whole-genome sequencing and expression profiling. *Genomics & Informatics* 10:1-8.

Lin, S., Haas, S., Zemojtel, T., Xiao, P., Vingron, M. & Li, R. (2011) Genome-wide comparison of cyanobacterial transposable elements, potential genetic diversity indicators. *Gene* 473:139-149.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. & Law, M. (2012) Comparison of next-generation sequencing systems. *Journal of Biomedicine & Biotechnology* 2012:251364.

Lowe, T.M. & Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* 25:955-964.

Majoros, W.H., Pertea, M. & Salzberg, S.L. (2004) TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics* 20:2878-2879.

Marahiel, M.A., Stachelhaus, T. & Mootz, H.D. (1997) Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chemical Reviews* 97:2651-2674.

Mardis, E.R. (2008) Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics* 9:387-402.

Mardis, E.R. (2013) Next-generation sequencing platforms. *Annual Review of Analytical Chemistry* 6:287-303.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J. & Chen, Z. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-380.

McPherson, J.D., Marra, M., Hillier, L., Waterston, R.H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E.R., Wilson, R.K., Fulton, R., Kucaba, T.A., Wagner-McPherson, C., Barbazuk, W.B., Gregory, S.G., Humphray, S.J., French, L., Evans, R.S., Bethel, G., Whittaker, A., Holden, J.L., McCann, O.T., Dunham, A., Soderlund, C., Scott, C.E., Bentley, D.R., Schuler, G., Chen, H.C., Jang, W., Green, E.D., Idol, J.R., Maduro, V.V., Montgomery, K.T., Lee, E., Miller, A., Emerling, S., Kucherlapati, Gibbs, R., Scherer, S., Gorrell, J.H., Sodergren, E., Clerc-Blankenburg, K., Tabor, P., Naylor, S., Garcia, D., de Jong, P.J., Catanese, J.J., Nowak, N., Osoegawa, K., Qin, S., Rowen, L., Madan, A., Dors, M., Hood, L., Trask, B., Friedman, C., Massa, H., Cheung, V.G., Kirsch, I.R., Reid, T., Yonescu, R., Weissenbach, J., Bruls, T., Heilig, R., Branscomb, E., Olsen, A., Doggett, N., Cheng, J.F., Hawkins, T., Myers, R.M., Shang, J., Ramirez, L., Schmutz, J., Velasquez, O., Dixon, K., Stone, N.E., Cox, D.R., Haussler, D., Kent, W.J., Furey, T., Rogic, S., Kennedy, S., Jones, S., Rosenthal, A., Wen, G., Schilhabel, M., Gloeckner, G., Nyakatura, G., Siebert, R., Schlegelberger, B., Korenberg, J., Chen, X.N., Fujiyama, A., Hattori, M., Toyoda, A., Yada, T., Park, H.S., Sakaki, Y., Shimizu, N., Asakawa, S., Kawasaki, K., Sasaki, T., Shintani, A., Shimizu, A., Shibuya, K., Kudoh, J., Minoshima, S., Ramser, J., Seranski, P., Hoff, C., Poustka, A., Reinhardt, R., Lehrach, H. & International Human Genome Mapping Consortium (2001) A physical map of the human genome. *Nature* 409:934-941.

Medema, M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E. & Breitling, R. (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research* 39:W339-46.

Meeks, J.C., Elhai, J., Thiel, T., Potts, M., Larimer, F., Lamerdin, J., Predki, P. & Atlas, R. (2001) An overview of the genome of *Nostoc punctiforme*, a multicellular, symbiotic cyanobacterium. *Photosynthesis Research* 70:85-106.

Meier, J.L. & Burkart, M.D. (2009) The chemical biology of modular biosynthetic enzymes. *Chemical Society Reviews* 38:2012-2045.

Mihara, K., Tanabe, T., Yamakawa, Y., Funahashi, T., Nakao, H., Narimatsu, S. & Yamamoto, S. (2004) Identification and transcriptional organization of a gene cluster involved in biosynthesis and transport of acinetobactin, a siderophore produced by *Acinetobacter baumannii* ATCC 19606T. *Microbiology* 150:2587-2597.

Milne, J.C., Roy, R.S., Eliot, A.C., Kelleher, N.L., Wokhlu, A., Nickels, B. & Walsh, C.T. (1999) Cofactor requirements and reconstitution of microcin B17 synthetase: a multienzyme complex that catalyzes the formation of oxazoles and thiazoles in the antibiotic microcin B17. *Biochemistry* 38:4768-4781.

Mootz, H.D., Schwarzer, D. & Marahiel, M.A. (2002) Ways of assembling complex natural products on modular nonribosomal peptide synthetases. *Chembiochem* 3:490-504.

Moss, S.J., Martin, C.J. & Wilkinson, B. (2004) Loss of co-linearity by modular polyketide synthases: a mechanism for the evolution of chemical diversity. *Natural Product Reports* 21:575-593.

Mulder, N. & Apweiler, R. (2007) InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods in Molecular Biology* 396:59-70.

Muller, R. (2004) Don't classify polyketide synthases. *Chemistry & Biology* 11:4-6.

Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H., Remington, K.A., Anson, E.L., Bolanos, R.A., Chou, H.H., Jordan, C.M., Halpern, A.L., Lonardi, S., Beasley, E.M., Brandon, R.C., Chen, L., Dunn, P.J., Lai, Z., Liang, Y., Nusskern, D.R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G.M., Adams, M.D. & Venter, J.C. (2000) A whole-genome assembly of *Drosophila*. *Science* 287:2196-2204.

Nakamura, Y., Kaneko, T., Sato, S., Ikeuchi, M., Katoh, H., Sasamoto, S., Watanabe, A., Iriguchi, M., Kawashima, K., Kimura, T., Kishida, Y., Kiyokawa, C., Kohara, M., Matsumoto, M., Matsuno, A., Nakazaki, N., Shimpo, S., Sugimoto, M., Takeuchi, C., Yamada, M. & Tabata, S. (2002) Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1. *DNA Research* 9:123-130.

Nakamura, Y., Kaneko, T., Sato, S., Mimuro, M., Miyashita, H., Tsuchiya, T., Sasamoto, S., Watanabe, A., Kawashima, K., Kishida, Y., Kiyokawa, C., Kohara, M., Matsumoto, M., Matsuno, A., Nakazaki, N., Shimpo, S., Takeuchi, C., Yamada, M. & Tabata, S. (2003) Complete genome structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids. *DNA Research* 10:137-145.

Nes, I.F., Diep, D.B. & Holo, H. (2007) Bacteriocin diversity in *Streptococcus* and *Enterococcus*. *Journal of Bacteriology* 189:1189-1198.

Newman, D.J. & Cragg, G.M. (2007) Natural products as sources of new drugs over the last 25 years. *Journal of Natural Products* 70:461-477.

Nunnery, J.K., Mevers, E. & Gerwick, W.H. (2010) Biologically active secondary metabolites from marine cyanobacteria. *Current Opinion in Biotechnology* 21:787-793.

O'Sullivan, O., Begley, M., Ross, R.P., Cotter, P.D. & Hill, C. (2011) Further identification of novel lantibiotic operons using LanM-based genome mining. 3:27-40.

Olson, J.M. (2006) Photosynthesis in the Archean era. *Photosynthesis Research* 88:109-117.

Oman, T.J. & van der Donk, W.A. (2010) Follow the leader: the use of leader peptides to guide natural product biosynthesis. *Nature Chemical Biology* 6:9-18.

Oman, T.J. (2011) Sublancin is not a lantibiotic but an S-linked glycopeptide. *Nature Chemical Biology* 7:78-80.

Onaka, H., Nakaho, M., Hayashi, K., Igarashi, Y. & Furumai, T. (2005) Cloning and characterization of the goadsporin biosynthetic gene cluster from *Streptomyces* sp. TP-A0584. *Microbiology* 151:3923-3933.

Palenik, B., Brahamsha, B., Larimer, F.W., Land, M., Hauser, L., Chain, P., Lamerdin, J., Regala, W., Allen, E.E., McCarren, J., Paulsen, I., Dufresne, A., Partensky, F., Webb, E.A. & Waterbury, J. (2003) The genome of a motile marine *Synechococcus*. *Nature* 424:1037-1042.

Palenik, B., Ren, Q., Dupont, C.L., Myers, G.S., Heidelberg, J.F., Badger, J.H., Madupu, R., Nelson, W.C., Brinkac, L.M., Dodson, R.J., Durkin, A.S., Daugherty, S.C., Sullivan, S.A., Khouri, H., Mohamoud, Y., Halpin, R. & Paulsen, I.T. (2006) Genome sequence of *Synechococcus* CC9311: Insights into adaptation to a coastal environment. *Proceedings of the National Academy of Sciences of the United States of America* 103:13555-13559.

Peterson, J.D., Umayam, L.A., Dickinson, T., Hickey, E.K. & White, O. (2001) The Comprehensive Microbial Resource. *Nucleic Acids Research* 29:123-125.

Pop, M., Phillippy, A., Delcher, A.L. & Salzberg, S.L. (2004) Comparative genome assembly. *Briefings in Bioinformatics* 5:237-248.

Pruitt, K.D., Tatusova, T., Klimke, W. & Maglott, D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Research* 37:D32-6.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P. & Gu, Y. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13:341.

Ran, L., Larsson, J., Vigil-Stenman, T., Nylander, J.A., Ininbergs, K., Zheng, W.W., Lapidus, A., Lowry, S., Haselkorn, R. & Bergman, B. (2010) Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. *PloS One* 5:e11486.

Rantala-Ylinen, A., Kana, S., Wang, H., Rouhiainen, L., Wahlsten, M., Rizzi, E., Berg, K., Gugger, M. & Sivonen, K. (2011) Anatoxin-a synthetase gene cluster of the cyanobacterium *Anabaena* sp. strain 37 and molecular methods to detect potential producers. *Applied and Environmental Microbiology* 77:7271-7278.

Rastogi, R.P. & Sinha, R.P. (2009) Biotechnological and industrial significance of cyanobacterial secondary metabolites. *Biotechnology Advances* 27:521-539.

Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W. & Huson, D.H. (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Research* 33:5799-5808.

Ridley, M. (2000) Genome: the autobiography of a species in 23 chapters. Harper Collins Publishers, New York

Riley, M.A. & Wertz, J.E. (2002) Bacteriocins: evolution, ecology, and application. *Annual Review of Microbiology* 56:117-137.

Rippka, R., Castenholz, R.W., Iteman, I. & Herdman, M. (2001) Form-genus I. *Anabaena*. In: Boone, D.R. et al. (eds.). *Bergey's manual of systematic bacteriology. Vol. 1*. New York: Springer. p. 566-568.

Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A., Arellano, A., Coleman, M., Hauser, L., Hess, W.R., Johnson, Z.I., Land, M., Lindell, D., Post, A.F., Regala, W., Shah, M., Shaw, S.L., Steglich, C., Sullivan, M.B., Ting, C.S., Tolonen, A., Webb, E.A., Zinser, E.R. & Chisholm, S.W. (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* 424:1042-1047.

Rouhiainen, L., Vakkilainen, T., Siemer, B.L., Buikema, W., Haselkorn, R. & Sivonen, K. (2004) Genes coding for hepatotoxic heptapeptides (microcystins) in the cyanobacterium *Anabaena* strain 90. *Appl.Environ.Microbiol* 70:686-692.

Rouhiainen, L., Jokela, J., Fewer, D.P., Urmann, M. & Sivonen, K. (2010) Two alternative starter modules for the non-ribosomal biosynthesis of specific anabaenopeptin variants in *Anabaena* (Cyanobacteria). *Chemistry & Biology* 17:265-273.

Rouhiainen, L., Paulin, L., Suomalainen, S., Hyytiainen, H., Buikema, W., Haselkorn, R. & Sivonen, K. (2000) Genes encoding synthetases of cyclic depsipeptides, anabaenopeptilides, in *Anabaena* strain 90. *Molecular Microbiology* 37:156-167.

Rouhiainen, L., Sivonen, K., Buikema, W.J. & Haselkorn, R. (1995) Characterization of toxin-producing cyanobacteria by using an oligonucleotide probe containing a tandemly repeated heptamer. *Journal of Bacteriology* 177:6021-6026.

Rounge, T.B., Rohrlack, T., Nederbragt, A.J., Kristensen, T. & Jakobsen, K.S. (2009) A genome-wide analysis of nonribosomal peptide synthetase gene clusters and their peptides in a *Planktothrix rubescens* strain. *BMC Genomics* 10:396.

Rusk, N. (2011) Torrents of sequence. *Nature Method* 8:44-44.

Salzberg, S.L. & Yorke, J.A. (2005) Beware of mis-assembled genomes. *Bioinformatics* 21:4320-4321.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M. & Smith, M. (1977a) Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265:687-695.

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

Shen, B. (2003) Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Current Opinion in Chemical Biology* 7:285-295.

Shendure, J. & Ji, H. (2008) Next-generation DNA sequencing. *Nature Biotechnology* 26:1135-1145.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D. & Church, G.M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309:1728-1732.

Shih, P.M., Wu, D., Latifi, A., Axen, S.D., Fewer, D.P., Talla, E., Calteau, A., Cai, F., Tandeau de Marsac, N., Rippka, R., Herdman, M., Sivonen, K., Coursin, T., Laurent, T., Goodwin, L., Nolan, M., Davenport, K.W., Han, C.S., Rubin, E.M., Eisen, J.A., Woyke, T., Gugger, M. & Kerfeld, C.A. (2013) Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proceedings of the National Academy of Sciences of the United States of America* 110:1053-1058.

Shizuya, H., Birren, B., Kim, U.J., Mancino, V., Slepak, T., Tachiiri, Y. & Simon, M. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proceedings of the National Academy of Sciences of the United States of America* 89:8794-8797.

Sivonen, K. & Börner, T. (2008) Bioactive compounds produced by cyanobacteria. In: Herrero, A. & Flores, E. (eds.) *The cyanobacteria: molecular biology, genomics, and evolution.* Norfolk, UK: Caister Academic Press. p. 159-197.

Sivonen, K. (2009) Cyanobacterial toxins. In: Schaechter M. (ed.). *Encyclopedia of Microbiology.* Oxford, UK: Academic Press. p. 290-307.

Sivonen, K., Leikoski, N., Fewer, D.P. & Jokela, J. (2010) Cyanobactins-ribosomal cyclic peptides produced by cyanobacteria. *Applied Microbiology and Biotechnology* 86:1213-1225.

Sivonen, K., Niemelä, S.I., Niemi, R.M., Lepistö, L., Luoma, T.H. & Räsänen, L.A. (1990) Toxic cyanobacteria (blue-green algae) in Finnish fresh and coastal waters. *Hydrobiologia* 190:267-275.

Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B. & Hood, L.E. (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* 321:674-679.

Smith, S. & Tsai, S.C. (2007) The type I fatty acid and polyketide synthases: a tale of two megasynthases. *Natural Product Reports* 24:1041-1072.

Staden, R. (1979) A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research* 6:2601-2610.

Staunton, J. & Weissman, K.J. (2001) Polyketide biosynthesis: a millennium review. *Natural Product Reports* 18:380-416.

Stein, L. (2001) Genome annotation: from sequence to biology. *Nature Reviews Genetics* 2:493-503.

Stepper, J. (2011) Cysteine S-glycosylation, a new post-translational modification found in glycopeptide bacteriocins. *FEBS Letters* 585:645-650.

Stucken, K., John, U., Cembella, A., Murillo, A.A., Soto-Liebe, K., Fuentes-Valdes, J.J., Friedel, M., Plominsky, A.M., Vasquez, M. & Glockner, G. (2010) The smallest known genomes of multicellular and toxic cyanobacteria: comparison, minimal gene sets for linked traits and the evolutionary implications. *PloS One* 5:e9235.

Sugita, C., Ogata, K., Shikata, M., Jikuya, H., Takano, J., Furumichi, M., Kanehisa, M., Omata, T., Sugiura, M. & Sugita, M. (2007) Complete nucleotide sequence of the freshwater unicellular cyanobacterium *Synechococcus elongatus* PCC 6301 chromosome: gene content and organization. *Photosynthesis Research* 93:55-67.

Swingley, W.D., Blankenship, R.E. & Raymond, J. (2008a) Insights into cyanobacterial evolution from comparative genomics. In: Herrero, A. & Flores, E. (eds.) *The cyanobacteria: molecular biology, genomics, and evolution.* Norfolk, UK: Caister Academic Press. p. 21-43.

Swingley, W.D., Chen, M., Cheung, P.C., Conrad, A.L., Dejesa, L.C., Hao, J., Honchak, B.M., Karbach, L.E., Kurdoglu, A., Lahiri, S., Mastrian, S.D., Miyashita, H., Page, L., Ramakrishna, P., Satoh, S., Sattley, W.M., Shimada, Y., Taylor, H.L., Tomo, T., Tsuchiya, T., Wang, Z.T., Raymond, J., Mimuro, M., Blankenship, R.E. & Touchman, J.W. (2008b) Niche adaptation and genome expansion in the chlorophyll d-producing cyanobacterium *Acaryochloris marina*. *Proceedings of the National Academy of Sciences of the United States of America* 105:2005-2010.

Tajima, N., Sato, S., Maruyama, F., Kaneko, T., Sasaki, N.V., Kurokawa, K., Ohta, H., Kanesaki, Y., Yoshikawa, H., Tabata, S., Ikeuchi, M. & Sato, N. (2011) Genomic structure of the cyanobacterium *Synechocystis* sp. PCC 6803 strain GT-S. *DNA Research* 18:393-399.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J. & Natale, D.A. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4:41.

Trautmann, D., Voss, B., Wilde, A., Al-Babili, S. & Hess, W.R. (2012) Microevolution in cyanobacteria: re-sequencing a motile substrain of *Synechocystis* sp. PCC 6803. *DNA Research* 19:435-448.

Treangen, T.J. & Salzberg, S.L. (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews.Genetics* 13:36-46.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K., Sidow, A., Fire, A. & Johnson, S.M. (2008) A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Research* 18:1051-1063.

Velasquez, J.E. & van der Donk, W.A. (2011) Genome mining for ribosomally synthesized natural products. *Current Opinion in Chemical Biology* 15:11-21.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. & Zhu, X. (2001) The sequence of the human genome. *Science* 291:1304-1351.

Venter, J.C., Adams, M.D., Sutton, G.G., Kerlavage, A.R., Smith, H.O. & Hunkapiller, M. (1998) Shotgun sequencing of the human genome. *Science* 280:1540-1542.

Voss, B., Bolhuis, H., Fewer, D.P., Kopf, M., Moke, F., Haas, F., El-Shehawy, R., Hayes, P., Bergman, B., Sivonen, K., Dittmann, E., Scanlan, D.J., Hagemann, M., Stal, L.J. & Hess, W.R. (2013) Insights into the physiology and ecology of the brackish-water-adapted cyanobacterium *Nodularia spumigena* CCY9414 based on a genome-transcriptome analysis. *PloS One* 8:e60224.

Walsh, C.T. & Fischbach, M.A. (2010) Natural products version 2.0: connecting genes to molecules. *Journal of the American Chemical Society* 132:2469-2493.

Waterston, R.H., Lander, E.S. & Sulston, J.E. (2002) On the sequencing of the human genome. *Proceedings of the National Academy of Sciences* 99:3712-3716.

Welker, M. & von Döhren, H. (2006) Cyanobacterial peptides - nature's own combinatorial biosynthesis. *FEMS Microbiology Reviews* 30:530-563.

Welsh, E.A., Liberton, M., Stockel, J., Loh, T., Elvitigala, T., Wang, C., Wollam, A., Fulton, R.S., Clifton, S.W., Jacobs, J.M., Aurora, R., Ghosh, B.K., Sherman, L.A., Smith, R.D., Wilson, R.K. & Pakrasi, H.B. (2008) The genome of *Cyanothece* 51142, a unicellular diazotrophic cyanobacterium important in the marine nitrogen cycle. *Proceedings of the National Academy of Sciences of the United States of America* 105:15094-15099.

Wenzel, S.C. & Muller, R. (2005) Formation of novel secondary metabolites by bacterial multimodular assembly lines: deviations from textbook biosynthetic logic. *Current Opinion in Chemical Biology* 9:447-458.

Wetzel, J., Kingsford, C. & Pop, M. (2011) Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies. *BMC Bioinformatics* 12:95.

Willey, J.M. & van der Donk, W.A. (2007) Lantibiotics: peptides of diverse structure and function. *Annual Review of Microbiology* 61:477-501.

Yu, D., Xu, F., Zeng, J. & Zhan, J. (2012) Type III polyketide synthases in natural product biosynthesis. *IUBMB Life* 64:285-295.

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