marine drugs

MDPI

Review

Next Generation Sequencing of Actinobacteria for the Discovery of Novel Natural Products

Juan Pablo Gomez-Escribano *, Silke Alt and Mervyn J. Bibb

Department of Molecular Microbiology, John Innes Centre, Norwich Research Park, Norwich, NR4 7UH, UK; silke.alt@jic.ac.uk (S.A.); mervyn.bibb@jic.ac.uk (M.J.B.)

* Correspondence: juan-pablo.gomez-escribano@jic.ac.uk; Tel.: +44-(0)1603-450000

Academic Editor: Paul Long

Received: 25 February 2016; Accepted: 6 April 2016; Published: 13 April 2016

Abstract: Like many fields of the biosciences, actinomycete natural products research has been revolutionised by next-generation DNA sequencing (NGS). Hundreds of new genome sequences from actinobacteria are made public every year, many of them as a result of projects aimed at identifying new natural products and their biosynthetic pathways through genome mining. Advances in these technologies in the last five years have meant not only a reduction in the cost of whole genome sequencing, but also a substantial increase in the quality of the data, having moved from obtaining a draft genome sequence comprised of several hundred short contigs, sometimes of doubtful reliability, to the possibility of obtaining an almost complete and accurate chromosome sequence in a single contig, allowing a detailed study of gene clusters and the design of strategies for refactoring and full gene cluster synthesis. The impact that these technologies are having in the discovery and study of natural products from actinobacteria, including those from the marine environment, is only starting to be realised. In this review we provide a historical perspective of the field, analyse the strengths and limitations of the most relevant technologies, and share the insights acquired during our genome mining projects.

Keywords: actinomycetes; *Streptomyces*; genome mining; PacBio; Illumina; next generation sequencing

1. Introduction

Actinobacteria produce more than 70% of all natural product scaffolds used for the manufacture of clinically-relevant anti-infectives [1]. During the final decades of the last century, efforts to discover new anti-infectives of microbial origin were almost abandoned; however, the beginning of this century has seen a return to the search for novel bioactive natural products from actinobacteria, due to advances in cultivation and activity screening [2,3] and, in particular for this review, the realisation of the biosynthetic potential of actinobacteria, encoded in the genome but not expressed as compounds, and the development of genetic approaches to access this hidden potential [3,4].

During the sequencing of the *Streptomyces coelicolor* genome in the late 1990s [5] it became evident that actinomycetes carry the genetic potential to produce many more natural products than those detected under laboratory conditions, and during the following years many previously-unknown metabolites produced by *S. coelicolor* were identified and characterised [6]. It was realised then that access to the genome sequence of a strain could be used to unlock the biosynthetic potential of the micro-organism using different molecular genetic approaches [1,6–8] in a strategy that has been called "genome mining" [7]. However, in the early 2000s the only automated technique for DNA sequencing was the dideoxynucleotide method developed by Sanger and co-workers [9]; even with the developments reached in the 2000s, including those in computing and assembly algorithms [10], Sanger sequencing was too expensive and labour intensive to provide sufficient coverage for a routine

whole-genome shotgun approach, requiring the creation, sorting, and sequencing of genomic libraries prior to full genome assembly [5,10].

Since the late 2000s we have seen the continuous release of new DNA sequencing technologies that have pushed forward both sequencing capacity and accuracy, and lowered the cost per sequenced nucleotide. These technologies are referred to as next-generation sequencing (NGS) and are defined as "non-Sanger-based, high-throughput, and eliminating the need for fragment-cloning and amplification in *Escherichia coli* prior to sequencing" (adapted from [11]). NGS technologies make affordable the high-throughput sequencing of bacterial genomes which, when coupled with a continuous advance in computing algorithms and databases for the automated scanning and annotation of specialised metabolite gene clusters like AntiSMASH [12] and MIBiG [13], are enabling the continuous discovery and study of natural products biosynthetic pathways by genome mining.

The relatively recent realisation that actinobacteria from our oceans and seas are much more abundant than previously thought provides a new opportunity for the discovery of drugs from the marine environment. Coupled with the difficulty of growing some of these organisms at a large scale, the sequence-based approach to natural product discovery described here should be particularly pertinent and useful.

2. A Short Walk through NGS Technologies

The first NGS technologies to appear, referred to as second-generation sequencing (SGS), relied on cycles of the termination of DNA polymerisation and recording of the incorporated nucleotides in each cycle. The first SGS technology to be commercialised, in mid-2005, was 454 pyrosequencing (by 454 Life Sciences, now a subsidiary of Roche Diagnostics) [14], followed in 2006 by the reversible-terminator chemistry of Solexa/Illumina (by Solexa Ltd., now Illumina Inc.) [15] which, because of its lower cost, high throughput, and accuracy [16], has become the first choice sequencing technology across many fields of research and medical diagnostics. Other SGS technologies have been commercialised but have not attained the popularity of these two. Examples that have been used for actinobacteria genome sequencing are SOLiD (released in 2006 by Applied Biosystems Inc., now Thermo Fisher Scientific/Life Technologies) and Ion Torrent (released in 2010 by Ion Torrent Systems Inc., now Thermo Fisher Scientific/Life Technologies).

The year 2011 saw the commercialisation of the first third-generation sequencing (TGS) technology: Single Molecule Real Time (SMRT) (by Pacific Biosciences of California, Inc.) usually referred to as "PacBio" [17]. Another TGS technology currently in development, but already very promising, is Nanopore DNA sequencing (by Oxford Nanopore Technologies Ltd.); this platform is accessible through an early access program [18]; early reports indicate that the technology, despite read accuracy under 70%, is already useful for scaffolding, thanks to read lengths of over 1 kb and up to 98 kb; however, the application to actinobacterial genome sequencing is still heavily hampered by the high mol% G+C of the organisms [19,20].

TGS technologies, as opposed to SGS, rely on sequencing single molecules without amplification (which can create problems with even genome representation) and in real-time (no cycles of polymerisation/termination) and are capable of providing read lengths of several kilobases [21] facilitating the assembly of whole-genome shotgun projects.

3. Challenges of Actinobacterial Genomics

The high mol% G+C content of actinomycete genomes poses difficulties not just for the sequencing technology itself but also to the computing algorithms used for the assembly [22], although many of the errors and sequencing biases due to mol% G+C bias have been lessened by improvements in library preparation [23,24]. A more specific problem is presented by the linear chromosome and plasmids of many important actinobacteria, like the streptomycetes, with long terminal inverted repeats that can reach over one megabase [25], impossible to resolve with current sequencing technologies. In addition, extraction of high molecular weight DNA of the high quality required for NGS library

construction, and in particular of TGS, is not trivial and in many cases currently not feasible, especially from actinobacteria difficult to culture and resilient to cell wall digestion.

Many of the most relevant natural products belong to the chemical families of Type I polyketides and non-ribosomal peptides [26]. The backbone of these compounds is synthesised by large enzymes, polyketide synthases (PKS), and non-ribosomal peptide synthetases (NRPS), which consist of a highly-conserved modular enzymatic architecture that is reflected at the nucleotide sequence level by highly similar intragenic and intergenic tandem repeats, frequently spanning over 700 bp; e.g., *S. coelicolor* coelimycin PKS gene *sco6274* positions 3879–4533 and *sco6275* 11986–12639 share 99% intergenic identity (Figure 1); and the calcium-dependent antibiotic NRPS gene *sco3230*, positions 13187–14121 and 16307–17241, share 95% intragenic identity. These repeats are, in many cases, longer than the read-length of all SGS technologies, making it very difficult, if not impossible, to correctly assemble these very important gene clusters.

4. The Read-Length Problem

When we talk about "assembly" we mean the determination of the correct order of the reads, the thousands of short overlapping fragments in which the whole genome sequence would be represented (see [27] for a good review). Due to simple statistics, the longer and more accurate the reads are, the longer and more specific the overlap between two reads will be and, therefore, a more reliable assembly can be computed [28].

Of the SGS technologies, 454 has provided, consistently, the longest read length and dominated actinobacterial genome sequencing until 2012 when Illumina took over with an increasing high-accuracy read-length, together with the highest output and lowest cost per base. Currently, with maximum paired-end (*i.e.*, both ends of the same DNA molecule are sequenced) read lengths of around 2 × 150 nt for Illumina's highest throughput HiSeq sequencers, and around 2 × 300 nt for the lower throughput MiSeq (source: www.Illumina.com; accessed on October 2015), Illumina technology can typically deliver a maximum of 500 nt of contiguous reliable sequence, insufficient to resolve the highly repetitive PKS and NRPS genes, ribosomal RNA operons (which span about 5.5 kb), and terminal inverted repeats. Of the other SGS technologies, only 454-pyrosequencing provides longer read lengths, 700 nt and up to 1 kb according to the manufacturer (GS FLX+ instrument; source: http://454.com; accessed on October 2015). Ion Torrent is currently offering up to 400 nt reads and presents an alternative to Illumina for *de novo* sequencing of small genomes [29].

This situation has changed dramatically with the advent of TGS technologies. PacBio SMRT, the only TGS technology currently commercialised, is still unique in its ability to resolve highly-similar repetitive sequences thanks to an average read-length of over 10 kb [30] at an affordable 100 fold coverage required, not just for a reliable assembly, but also for an accurate base-calling of the consensus sequence.

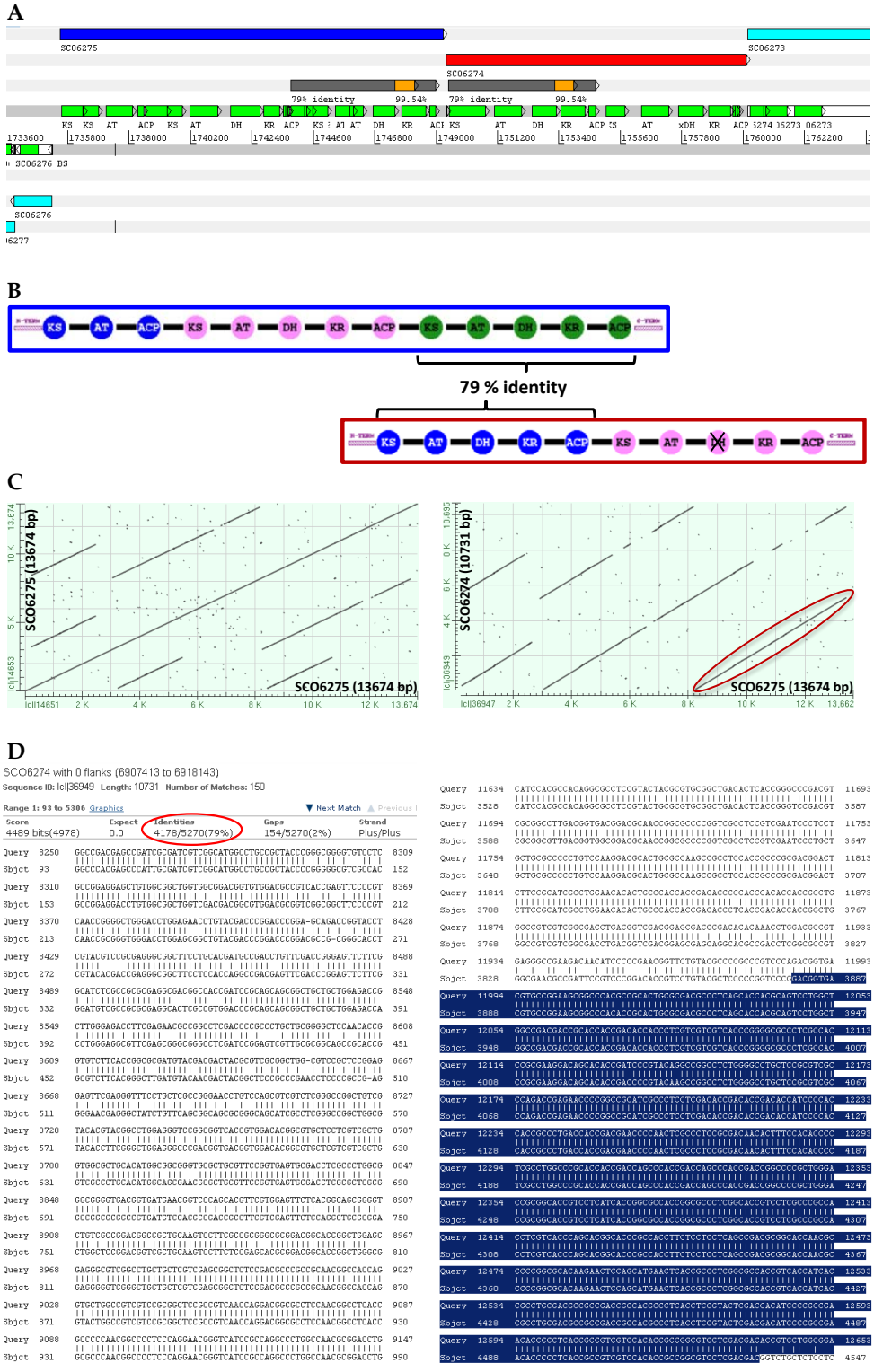


Figure 1. Example to illustrate the problem of intergenic and intragenic sequence identity in genes encoding modular polyketide synthases. (**A**) Screenshot from Artemis software [31] showing the genetic organisation of the first two PKS genes of the coelimycin gene cluster; the highlighted segments correspond to the last module of Sco6275 and first of Sco6274 (**B**) which share a sequence identity of 79% over 5 kb (grey) and more than 99.5% identity over 650 bp (orange). A closer look at the nucleotide sequence of these genes by BLAST (dot plots and alignments generated at NCBI server) shows the intragenic (**C**, left) and intergenic (**C**, right) regions of high identity, and the nucleotide alignment of the region in grey (**D**, left) and orange (**D**, right, highlighted in blue). KS, ketosynthase; AT, acyltransferase; DH, dehydratase; KR, ketoreductase; ACP, acyl-carrier protein.

5. Historical Perspective of Actinobacterial Genome Sequencing

The first actinobacterial genome fully sequenced was that of *Mycobacterium tuberculosis* [32], and the first genome of an actinobacterium relevant for natural products was that of the model streptomycete *S. coelicolor* [5], followed by the industrially important *Streptomyces avermitilis* [33], both using Sanger sequencing. As standardly practiced at the time, the *S. coelicolor* genome was assembled by sequencing an ordered cosmid and BAC genomic library. In contrast, sequencing of the genome of *S. avermitilis* was first attempted using a whole-genome shotgun approach; the published article provides enough detail for us to grasp the difficulties of a whole-genome shotgun approach only a decade ago; two genome contigs of 9,025,608 nt and 94,287 nt (the chromosome and a linear plasmid, respectively; a total of 9,119,895 nt) were assembled initially from 186,619 sequences from random ~2 kb fragments (estimating 600 nt of high-quality base-calls, (186,619 × 600) nt/9,119,895 nt = 12.3 fold coverage); this data was not sufficient for a good quality assembly and the authors had to rely on the end-sequencing of more than 10,000 cosmids from a genomic library, over a hundred full cosmid inserts, and 162 PCR-product sequences, altogether reaching 13.3 fold coverage with directed gap filling [33]. This example demonstrates that shotgun sequencing with exclusive use of Sanger-based techniques was extremely laborious and not sufficient for efficient full genome shotgun assembly.

The SGS technologies did allow shotgun sequencing of whole genomes at much higher coverage, but the short read-length combined with the intrinsic difficulties of actinobacterial genomes hampered the assembly of full replicons in single contigs. Due to the longer read-length, 454-pyrosequencing has been the technology of choice for achieving a more complete genome assembly, but full replicon assembly into a single contig has required the use of Sanger end-sequencing of large-insert genomic libraries and directed gap-filling by PCR-amplification. Examples are the sequencing of the genomes of *Streptomyces hygroscopicus* 5008, in which Illumina data was used to correct base-call errors [34], and of *Streptomyces albus* J1074 for which, even with a 377 fold coverage from combined Illumina-454 sequencing data of three libraries of different insert-size, a large BAC library was end-sequenced and chromosome-walking used to fill gaps and merge the initial 76 contigs [35].

6. The Explosion of Genome Mining

Despite the difficulty in completing an accurate full genome sequence of an actinobacterium, the number of available genome assemblies (of different completeness and quality level) in the databases has grown exponentially during the past decade (Figure 2). The main reason is that a complete and accurate genome sequence is not a prerequisite for the application of the genome mining approach to natural product discovery; multi-contig genome drafts can be equally useful when it comes to identifying the gene cluster that drives the synthesis of a known compound, or gene clusters for the biosynthesis of unknown but potentially interesting natural products. The most common automated pipeline for searching and annotating natural products biosynthetic gene clusters, AntiSMASH [12], can easily deal with multi-contig genome assemblies.

Even in the early days of Solexa technology, the information obtained could suffice for designing strategies for genome mining. For the identification and cloning of the microbisporicin biosynthetic gene cluster, almost 900 Mb of sequence data for the *Microbispora corallina* genome was obtained, but despite the high coverage (greater than 100×) the short read length at the time, only 36 nt, allowed only a very fragmented assembly (see Table 1). However, this provided enough information to identify putative homologs of lantibiotic biosynthetic genes, to design probes for Southern Blotting, and to successfully screen a cosmid library to obtain the entire gene cluster [36,37]. Genome sequencing with Solexa was also crucial in the identification and cloning of the cypemycin gene cluster after strategies based on hybridisation screening of a cosmid library were unsuccessful [38].

At the same time, 454 pyrosequencing, with its longer reads (200–400 nt at the time), provided a clear advantage for actinobacterial genome assembly; sequencing of *M. corallina* genome with one quarter of a run of 454 yielded much less data than Solexa, but led to an assembly with fewer and longer contigs that allowed the identification of more putative lantibiotic biosynthetic genes and,

crucially, the gene encoding the putative precursor peptide by using the amino acid sequence as a query for a tBLASTn search [37]. The planosporicin [39,40], tunicamycin [41], and bottromycin [42] biosynthetic gene clusters were also cloned on the basis of draft genome sequences obtained with 454 (see Table 1 for a summary of our lab's experience).

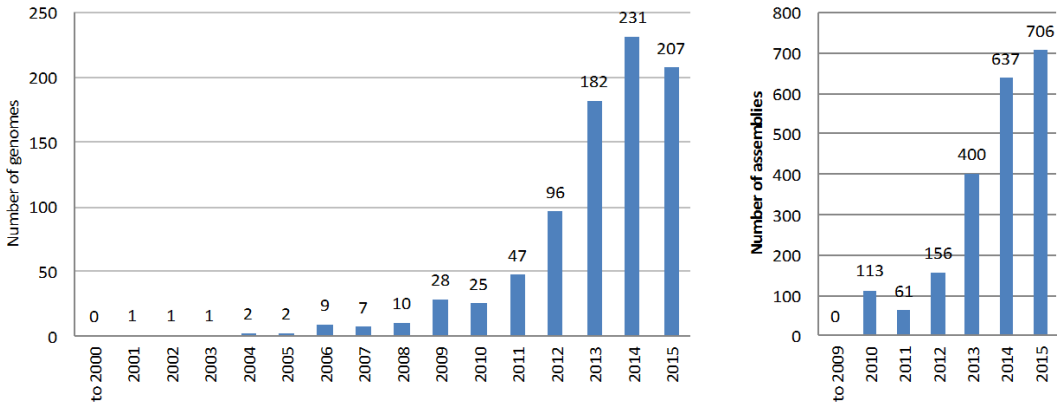


Figure 2. Number of genomes (left) and assemblies (right) of actinobacteria species relevant to natural products research deposited at NCBI databases per year.

Table 1. Summary of our group's experience using NGS for genome mining. Only the *Streptomyces leeuwenhoekii* genome was published in full; for the others, only the relevant and confidently re-sequenced segments were published.

Microorganism	Technology	Year	Number of contigs[1]	N50 contig (nt)	Longest contig (nt)	Total sum of contigs	Total data	Lanes/ runs	Ref.
Microbispora corallina NRRL 30420	Solexa/ Illumina	Mid 2007	14395	163	4436	2.93 Mb	881 Mb	7	[37]
Microbispora corallina NRRL 30420	454	Mid 2008	7580 (3027)	1219	8913	4.64 Mb	28 Mb	1/4	[37]
Streptomyces sp. OH-4156	Solexa/ Illumina	Mid 2007	15,471	378	7830	8.5 Mb			[38]
Planomonospora alba NRRL 18924	454	Mid 2009	3066 (1618)	756	14,767	2.32 Mb	13 Mb	1/8	[40]
Planomonospora alba NRRL 18924	454	Mid 2011	1017 (944)	17,314	141,100	7.3 Mb	72 Mb	1/4	[40]
Streptomyces chartreusis NRRL 3882	454	2008	3112	4582	53,916	7.95 Mb	286 Mb	1/2	[41]
Streptomyces bottropensis DSM 40262	454	End 2010	463 (427)	40,440	183,403	8.85 Mb	115 Mb	1/4	[42]
Streptomyces leeuwenhoekii DSM 42122	Illumina MiSeq	Mid 2013	387 (279) (175 scaf.)[2]	59,284	157,225	8.1 Mb	712 Mb (1.25 Gb)	Full (500 cycles)	[43]
Streptomyces leeuwenhoekii DSM 42122	PacBio	End 2013	3	7,895,833	7,895,833	8 Mb	966 Mb	2 (3) cells	[43]

[1] Total number of contigs is given first; when available, the number of contigs larger than 500 nt is given in brackets. [2] Number of scaffolds in which contigs were joined.

Partial genome sequence information has also been used to design successful strategies to obtain or identify the unknown metabolic product of a biosynthetic gene cluster. Three main strategies are pursued: the activation of the expression of the gene cluster in the natural producer, cloning and expression in a heterologous host, and mutation followed by metabolite profiling. An elegant example of the first approach is the identification and characterisation of stambomycin; Laureti and co-workers found a cryptic gene cluster (*i.e.*, without a known product) encoding the biosynthesis of an unknown polyketide in the draft genome sequence of *Streptomyces ambofaciens* and, after inferring

the transcriptional regulatory network, they induced the production of the metabolic product by constitutive expression of a cluster-situated gene encoding a transcriptional activator of the LAL family [44]. A similar example is the discovery and characterisation of ansamycin compounds with novel chemistry from *Streptomyces* sp. LZ35 [45]. An example of the second approach is the characterisation of grisemycin, a linaridin from *Streptomyces griseus* IFO 13,350, which was purified after heterologous expression of the cloned gene cluster [46]. The third approach can be used when the gene cluster is expressed under laboratory culture conditions, but its product is not known: the genome sequence can be used to design gene knock-outs that, together with comparative metabolite profiling of the producing and non-producing strains, allows the identification of the metabolic product (e.g., [47]).

Thus, despite the large number of contigs, and most likely many misassembly issues, draft genome sequences can, indeed, be sufficient to successfully identify new natural products or the biosynthetic gene clusters of known compounds. Dozens of research groups across the globe have adopted one or more of these approaches, with more than 120 papers on this topic published by the end of 2015 (number of records found in PubMed with the search string "("genome mining") AND (streptomyces OR actinobacteria OR actinomycetes OR streptomycetes)") of which 70 were in the last two years (note that since PubMed does not index many journals relevant to the field, this is likely to be an under estimate of the number of relevant publications). Therefore, it should not be surprising that the nucleotide sequence databases contain over 2000 genome assemblies of actinobacterial species.

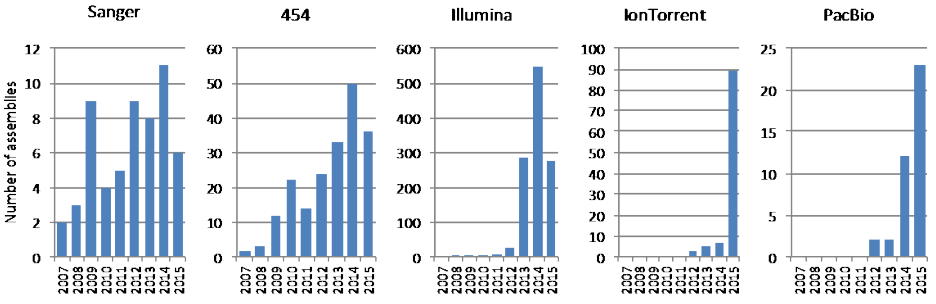
Although the situation is improving with more stringent and precise requirements during sequence submission, it is still difficult to obtain accurate characteristics about genome sequences available in public databases. There is also confusion about the terminology used by each database; e.g., NCBI hosts a database named "Assembly" that contains released sequences at different stages of completion; entries in the "Genome" database contain "assemblies" grouped at the species level, even with different strains in the same "genome" entry. It is also difficult to easily search for completed genomes, rather than drafts with hundreds of contigs; even metagenomic projects are included as a single entry in both "Genome" and "Assembly" databases. Even more difficult is to search by sequencing technology, due sometimes to the lack of information or to the use of different terms by different researchers when submitting to databases. Thus, while we have tried to ensure the accuracy of the numbers presented in the following paragraphs and Figures 2 and 3 they are not intended to be an absolutely precise description of the databases content.

At the end of 2015, a search for "actinobacteria" in the NCBI databases (search string "txid1760 [Organism:exp]") identified 1065 genomes and 7057 assemblies. Of these, 83 genomes and 4268 assemblies (over 60%) belong to *Mycobacterium* species alone, mostly to clinical isolates of *M. tuberculosis* (with 3635 assemblies). Since this review focuses on actinobacteria relevant for natural products research, genomes and assemblies from species belonging to the clinically-important genera *Mycobacterium* (Taxonomy ID: 1763), *Propionibacterium* (Taxonomy ID: 1743), *Gardnerella* (Taxonomy ID: 2701), and *Corynebacterium* (Taxonomy ID: 1716), and species of *Bifidobacterium* (Taxonomy ID: 1678), a genus that is becoming more abundant in databases because of the projects on human microflora, were filtered-out from the searches; this is not because of lack of potential relevance to natural products research but because the large number of assemblies would introduce a misleading bias in the analysis reported here. The result was that there are 849 genomes and 2073 assemblies of actinobacteria strains potentially relevant to natural products research available in the NCBI databases (Figure 2). Of the 2073 assemblies, 1865 assemblies are only completed to contig or scaffold level.

The PATRIC database [48] search tool allows a more comprehensive and precise search of bacterial genomes and, importantly, includes the completion achieved and sequencing technology used. At the time of writing, the newest entries in the PATRIC database were dated November 2015, and contained 7148 genome assemblies of actinobacteria (taxon ID 1760). Almost half corresponded to clinical isolates of *Mycobacterium tuberculosis* (3785) and a total of 5330 corresponded to the clinically-relevant strains of *Mycobacterium*, *Corynebacterium*, and the increasingly represented *Gardnerella*, *Propionibacterium*,

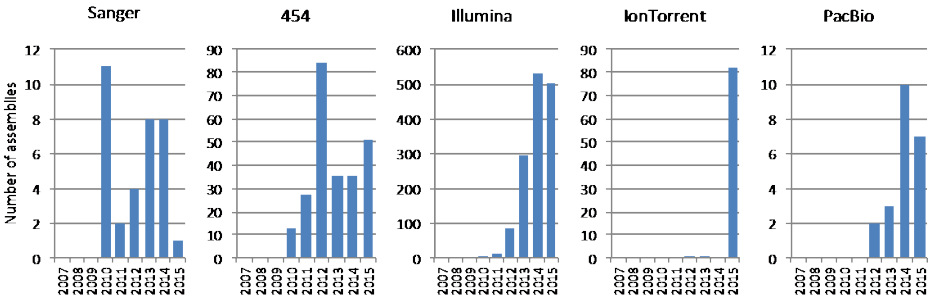
and *Bifidobacterium* genera. As explained above, in order to provide a more faithful representation of the genome sequencing projects relevant to natural products research, all assemblies of strains belonging to these genera were discarded from the statistics shown in this review; while many of these actinobacteria might be relevant to natural products research, the low number of actual species filtered out is not likely to influence the conclusions of this review.

A



Technology	Sanger	454	Illumina	IonTorrent	PacBio	SOLiD
Year 2007	2	2				
2008	3	3	1			
2009	9	12	4			
2010	4	22	2			
2011	5	14	7			
2012	9	24	27	3	2	0
2013	8	33	286	5	2	2
2014	11	50	546	7	12	0
Nov 2015	6	36	274	89	23	0

B



Technology	Sanger	454	Illumina	IonTorrent	PacBio	SOLiD
Year 2007						
2008						
2009						
2010	11	13	2			
2011	2	27	16			1
2012	4	84	85	1	2	1
2013	8	35	292	1	3	1
2014	8	35	529	0	10	0
Nov 2015	1	51	502	82	7	1

Figure 3. Distribution of assemblies included in the PATRIC (**A**) and NCBI (**B**) databases at the end of 2015, according to the technology used. Note that some assemblies have been obtained with a combination of several technologies and have been accounted for in each of them. Prior to 2007 there are 18 assemblies not included in the figure because the sequencing technology is not stated, although most certainly all were obtained by Sanger sequencing, or at least the 15 with complete status.

PATRIC allows a more comprehensive perspective on the use of technologies than NCBI (which seems not to contain searchable information prior to 2010); however, it is still not easy to analyse the

information (the names used vary, e.g., 454 can be referred to with just "GS FLX") and make any search and sorting very difficult, and there were 118 genomes sequenced with more than one technology, almost always involving 454 and Illumina, and there were over 300 entries without an associated technology) and, consequently, we chose to download the Actinobacteria genomes table and analyse it manually. The picture painted by our analysis (Figure 3) shows that Illumina is by far the most commonly used technology, due to the low cost per nucleotide, the very high throughput, and the wide availability of suppliers. As mentioned above, the longer reads of 454 pyrosequencing made this technology, despite its higher cost, the choice when a more precise assembly was desired, in particular for PKS and NRPS gene clusters, and so it was the main technology used until just two years ago, when Illumina took off. A continuous presence of Sanger sequencing reflects the use of this technology to complement the limitations of NGS when the goal is to obtain a finished genome; end-sequencing of genomic library clones and sequencing of PCR amplicons to fill gaps in the assembly are the most common applications of Sanger. SOLiD technology has been a very minor player with only 2–4 entries in PATRIC and NCBI, respectively. Interestingly, Ion Torrent seems to have gained momentum in the past year; since most of the submitted projects originated at the same institution, this may not reflect wider adoption, although improvements in library construction with high mol% G+C DNA [49] and bioinformatics [50] might help this technology to be more widely embraced in the future.

Assemblies obtained using TGS PacBio were submitted to the databases from 2012, but the first examples made use primarily of Illumina and 454 SGS, using limited coverage obtained with PacBio for scaffolding of contigs and gap closing [51,52].

7. Pacific Biosciences SMRT Platform

The first streptomycete genome that appears in the literature as sequenced with PacBio is an unfinished sequence of *Streptomyces* sp. strain Mg1 [53]; despite the relatively low coverage (20×) most of the chromosome was obtained as a single contig (accession GCA_000412265.1); however, the authors did not present a detailed analysis and only highlighted the improvement of the single contig as opposed to a previous 466-contigs assembly released by the Broad Institute (ABJF00000000) using Illumina; a more detailed comparison of both assemblies was included in a review article by Harrison and Studholme [54] but, since Illumina technology had evolved enormously during the five years that separate both assemblies, the comparison may be misleading.

Our group has recently published the sequencing of the genome of *Streptomyces leeuwenhoekii* C34, a strain isolated from the Atacama Desert in Chile [43]. The genome was sequenced and assembled independently with PacBio and Illumina MiSeq (paired-end 250 nt reads) between August and November 2013. In addition, PAC clones from a genomic library were also sequenced with 454 Junior [55], and we also have sequencing information obtained with Sanger during the study of specific gene clusters ([55] and Razmilic, manuscript in preparation). These data, obtained at almost the same time, provide a useful comparison of the relative strengths and limitations of all three technologies; some of these were also discussed in [43] but a more detailed view will be provided here.

8. PKS Modularity can be Resolved with PacBio

The genome of *S. leeuwenhoekii* C34 was first assembled by Busarakam and co-workers using Illumina 100 nt paired-end reads in to 658 contigs totalling 7.86 Mb [56]; analysis of this draft revealed numerous miss-assembly issues in the PKS genes, not surprising bearing in mind the short read length used [43,55]. We also sequenced this genome with PacBio RSII. After a quality-filter of the reads, we obtained almost 1 Gb of sequence data from 2 SMRT cells (plus a first cell that failed and generated only 77 Mb of sequence, so the full assembly was essentially derived from just 2 SMRT cells) which was assembled into just 3 contigs of 7.9 Mb, 95 kb and 10 kb. Oddly, the smallest contig matched a stretch of the largest contig with over 90% identity; these small and error-prone contigs have been observed in other PacBio sequencing projects (e.g., [57]). The largest contig contained an almost

complete chromosome, in a single uninterrupted sequence without the "Ns" typical of scaffolding with Illumina sequencing; only the ends of the terminal inverted repeats were not fully sequenced.

Almost concomitantly, we sequenced this genome with Illumina MiSeq 250 nt paired-end reads, obtained 1.25 Gb of data of which 712 Mb were assembled into 279 contigs, merged into 175 scaffolds, totalling 8.1 Mb. This assembly had a striking similarity with the PacBio assembly, and most of the studied PKS genes fitted the expected modularity ([55] and unpublished data). However, even with such high coverage, some PKS modules had been misassembled [43].

9. Identification of Circular or Linear Replicons

The assemblers produce linear molecules, but do not make decisions on the topology of the DNA molecule. Due to the long reads of PacBio, some reads will run into the terminal inverted repeats (TIRs) of linear replicons, allowing the identification of at least the beginning of the TIRs [43]; similarly, the presence of direct repeats at the ends of the contig will undoubtedly indicate that this is a circular replicon and allow the circularisation of the molecule to remove the duplicated non-existent sequence [43]. Illumina did an excellent job at assembling the circular plasmid into a single molecule, but without the direct terminal repeat of the PacBio contig it would not have been possible to identify it as a circular molecule. Another interesting finding was that PacBio could read through stretches of sequence difficult or impossible to read even by Sanger sequencing, apparently due to the formation of complex secondary structures (Figure S2).

10. Limitations of PacBio

10.1. Insertions and Deletions: Shifts in the Reading Frame

The main known limitation of PacBio in actinobacteria is the resolution of G or C homopolymers; in our experience, the final consensus sequence tends to miss a G or C, perhaps reflecting an issue with the assembly algorithm rather than a limitation of the current sequencing chemistry; if so, it might be possible to fix this in future releases of the software. We observed the opposite problem with the 454 technology, where we have observed a consistent insertion of a G or C in homopolymers (Figure S1). In both cases, the insertion or deletion causes a shift in the reading frame within a protein coding sequence (PCS), which is easily identifiable by studying the "GC Frame plot". Due to the high mol% G+C of actinobacterial DNA, there is a biased nucleotide composition at each of the three positions of a codon in a PCS: G or C are present at the third position in over 90% of codons, while they are in the second position in only around 50% of codons, and in the first position in about 70% of codons [58]. The software FRAME plots three lines corresponding to the mol% G+C of each of the three positions in the codons contained within a selected window-size [59]; a change in the distribution of the three lines (a crossing of lines) within a protein coding sequence is indicative of a shift in the reading frame caused by an insertion or deletion (Figure 4). Frame-shifts are usually easy to identify, unless they occur at the beginning or end of the PCS with a start or stop codon in proximity. If the encoded protein shows closely-related homologous proteins in the databases, a BLAST search easily identifies frame-shifts as well and can be used to confirm the FRAME plot analysis.

The insertions and deletions can be corrected by means of directed PCR amplification and Sanger sequencing (Figure 4) or at a genome-wide scale by deep-sequencing with Illumina and mapping of reads to the PacBio assembly [43,60,61]. In any case, they are easily addressable and do not pose such a problem as misassembly of PKS modules.

10.2. Sequence Missing from the Final Assembly

The most striking finding during comparison of our PacBio and Illumina assemblies was sequence missing from the PacBio assembly that was present in the pre-assembled data. During sequencing of the *S. leeuwenhoekii* C34 genome we found that over 130 kb of sequence, most likely representing a linear plasmid, was present in the Illumina but not the PacBio assembly; however, this sequence was

present in the so called PacBio corrected or pre-assembled reads, an intermediate step in the PacBio assembly pipeline [62]; the reason why this data was not assembled is still not known, but comparison of recent assemblies obtained from the same data with version 2 or 3 of the HGAP pipeline suggests that this issue has been resolved in HGAP3 (unpublished data). Another piece of sequence missing from the PacBio assembly, and also from the raw data originated from the end of linear replicons; we found that Illumina collected about 5 kb of additional unique telomere sequence located towards the end of the chromosome than PacBio; this may simply reflect the longer insert library size (over 20 kb) used for PacBio compared with the 500 bp used for Illumina, diminishing coverage of the terminal sequences. In both cases, it is important to stress that the missing sequence at the end of the chromosome, and the almost 130 kb likely linear plasmid (containing an interesting biosynthetic gene cluster), would not have been identified had we not also assembled the Illumina data, and instead used just the reads to improve the accuracy of the PacBio assembly.

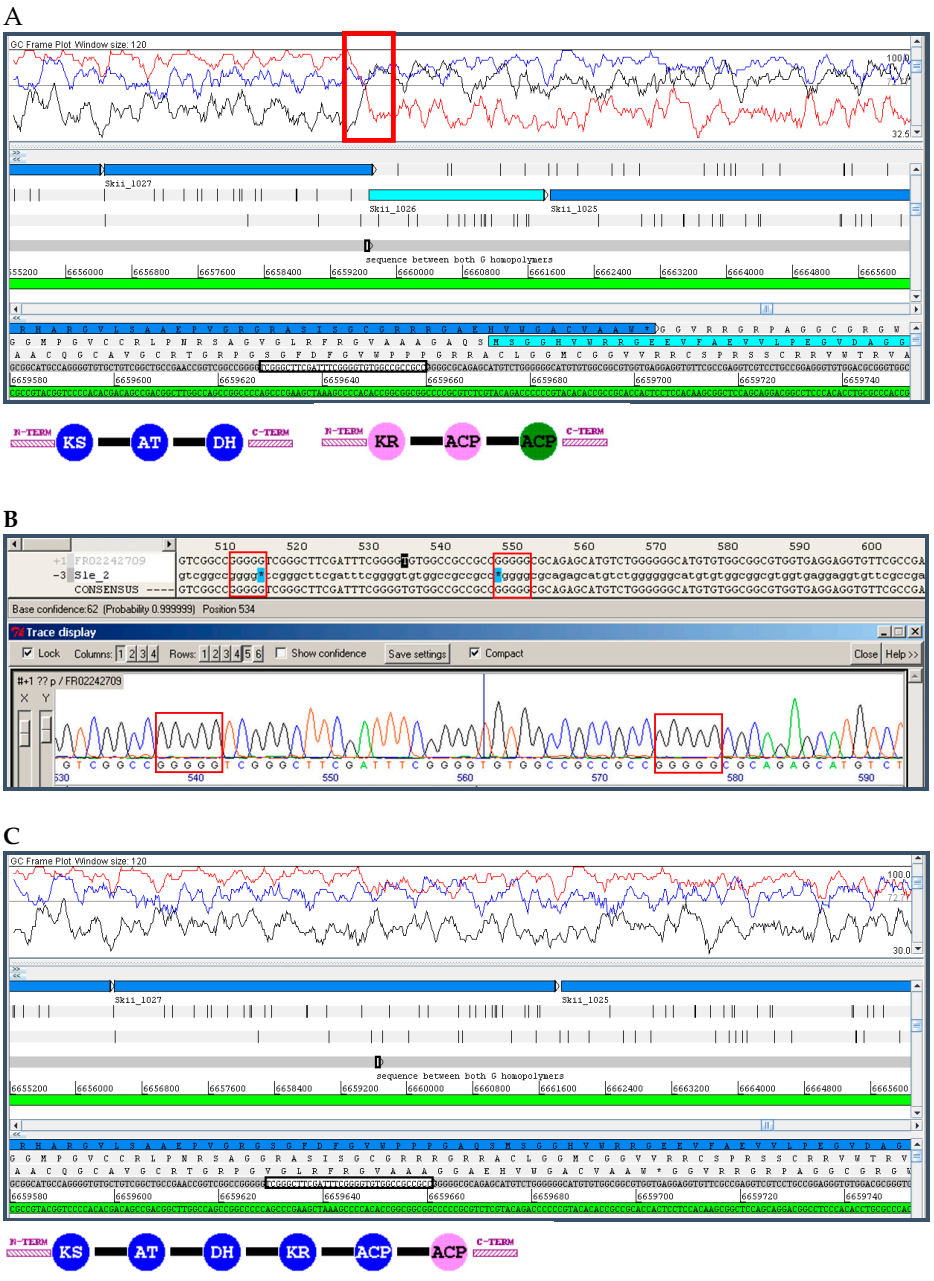


Figure 4. *Cont.*

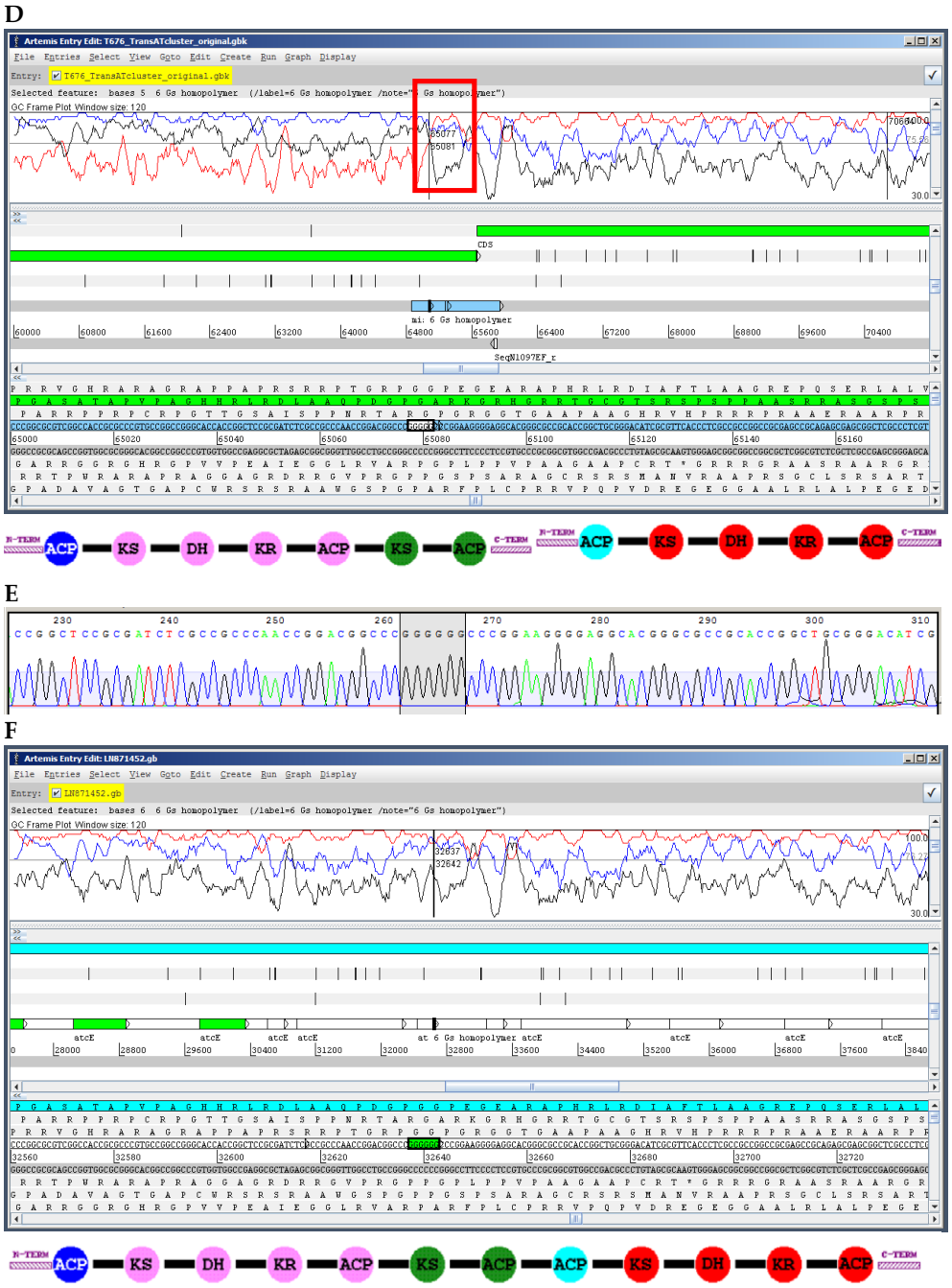


Figure 4. Illustration of shifts in the reading frame caused by omission of one nucleotide in homopolymers of G or C in PacBio assemblies, and their effect on the modularity of PKS genes of the chaxamycin [43,55] (**A**–**C**) and anthracimycin [57] (**D**–**F**) gene clusters. **A**,**D**. Original PacBio sequence showing the frame-shift (red box) and the break in the modularity of the PKS domains; in A, the frame shift is obvious since it splits the module of PKS domains into two protein coding sequences; in **D**, while each PKS protein exhibits an appropriate modular domain structure (non-canonical domain arrangements are not unusual in *trans*-AT PKSs [57]) note that the 3′ end of the first protein coding sequence continues in the wrong reading frame which would correspond to a highly non-canonical codon usage. **E**. Sanger sequencing of the affected region showing the affected homopolymer; in the chaxamycin gene cluster these errors had been also corrected with Illumina reads [43]; in the anthracimycin example, the entire region in the cyan box (**D**) was sequenced with Sanger, and the only error found was the one shown in **E**. **C**,**F**. Organisation of the PKS gene and domain modularity after error correction. PKS domains identified with the SBSPKS server [63]: KS, ketosynthase; AT, acyltransferase; DH, dehydratase; KR, ketoreductase; ACP, acyl-carrier protein; Note that there is a methyltransferase domain in the last module that has not been identified by SBSPKS.

11. Application to Actinobacteria from Marine Environments

The marine environment has proved to be a very prolific source of natural products diversity, much of which is of microbial origin [64,65]. The study of the anthracimycin gene cluster [57] (Figure 4) is just one of many examples of bioactive natural products isolated from marine actinobacteria [66,67]. Problems with reproducing natural environmental conditions in the laboratory often make the cultivation and maintenance of marine isolates difficult, and so it is not surprising that investigators are increasingly adopting the approaches outlined here to capture and exploit the biosynthetic potential of marine micro-organisms [68–73]. A genus of marine actinobacteria particularly worth highlighting for yielding potentially pharmaceutically useful natural products is *Salinispora* [74]. *S. tropica* CNB-392 is the producer of salinosporamide A [75], a promising anticancer compound currently in clinical trials [66]. While some members of the *Salinispora* genus are amenable for culturing in the laboratory [76], heterologous expression of biosynthetic gene clusters from these species has also been achieved [77] in host strains of the model actinomycete *Streptomyces coelicolor* [78]. We predict that the application of now well-established techniques for genome mining will prove particularly effective for the analysis of biosynthetic gene clusters from these and other marine micro-organisms [1,7,8].

12. Concluding Remarks

One of the main misconceptions about NGS-derived genome assemblies is that they faithfully represent the complete sequence of a genome or even of a gene cluster. Wrongly assembled segments of a genome are quite frequent in draft assemblies obtained with SGS technologies like Illumina [79]. This poses a problem not only for specific research projects, but populates nucleotide sequence databases with poorly finished drafts of actinobacterial genomes, compromising automated computerised analyses, including homology searches and annotation. The problem of misassembly has been tackled using two main approaches: sequencing two different libraries of short and long insert size (e.g., the use of mate-pair reads in Illumina [80,81]) or using optical mapping technology to generate a genome-wide restriction map [82,83].

We should also be aware that a complete genome sequence may not be represented in a shotgun assembly, even if obtained with PacBio and in a single contig per replicon. The ends of linear replicons are particularly difficult to obtain without resorting to manual curation or even directed cloning and Sanger sequencing [84] although Illumina sequencing (and presumably other technologies using short insert-size libraries) does seem better able to approach the ends than PacBio [43]. It is also worth noting that the PacBio "corrected reads" may contain unassembled sequence which can, given their high accuracy, be readily identified by querying with a protein sequence (using the tBLASTn program).

454 pyrosequencing will be discontinued by Roche during 2016 [85], and the suitability of Ion Torrent for sequencing G+C rich genomes, has yet to be firmly established. Consequently, at the moment, Pacific Biosciences SMRT (PacBio) and Illumina MiSeq are the technologies of choice for *de novo* genome sequencing of actinobacteria. The final choice will depend on many factors, mainly the financing available and the goals of the sequencing project, but also the availability of each technology (Illumina is currently more widely available than PacBio, with bench-top instruments affordable by medium-sized laboratories and the presence of suppliers offering the technology). Based on our experience, PacBio currently provides a far better assembly of similar accuracy to Illumina MiSeq but at a higher cost; Illumina cannot currently match PacBio assembly using merely a short insert library and paired-end 2 × 300 nt reads.

The long reads provided by TGS PacBio, with a consensus accuracy of over 99%, makes it a very suitable technology capable of resolving the precise organisation of modular PKS and NRPS genes, even if we need to manually correct frame-shifts that are normally easily identified. As an alternative, a combination of Illumina paired-end (short insert size library) and mate-pair (long insert size library) has been reported to lead to correctly assembled modular PKS genes, but may rely on the use of customised assembly pipelines [81].

Based on our experiences, the aim of the project and the finances available, we recommend the following options for *de novo* actinobacterial genome sequencing:

1. If the aim is to obtain the highest quality genome sequence currently possible using NGS, we recommend a combination of PacBio and Illumina sequencing, with the unassembled Illumina data used to correct missing bases, and the assembled contigs used to add possibly missing sequences in the PacBio assembly.
2. If the aim is to obtain an accurate overall assembly of the genome or to obtain reliable single-contig assemblies of biosynthetic gene clusters that encode modular PKS or NRPS enzymes with a tolerable number of missing bases, then we recommend PacBio.
3. If the goal is to identify interesting putative biosynthetic gene clusters, and to design strategies for cloning or activating gene expression, or to identify a metabolic product by gene inactivation and comparative metabolite profiling, then Illumina MiSeq will be the most cost effective choice, with potentially just paired-end sequencing of a short-insert library sufficing.

With the increasing output of PacBio at lower cost, we do not favour the use of a limited amount of PacBio long read data to scaffold Illumina contigs or scaffolds, since this approach would not help with the likely misassembly problems in repetitive regions of the genome. Instead, the PacBio assembly should form the foundation for correction with Illumina reads or extension with Illumina contigs.

Using any of the mentioned technologies to obtain a reliable sequence, we can now make confident predictions of the likely products of PKS and NRPS gene clusters and more reliably apply strategies such as heterologous expression of synthetic gene clusters to identify the encoded metabolites [86]. A recent example of the crucial importance of obtaining an accurate sequence of a modular PKS gene cluster (obtained with PacBio) is the lobosamide biosynthetic gene cluster from the marine isolate *Micromonospora* sp. RL09-050-HVF-A. The gene cluster contains seven large highly repetitive modular PKS genes. Bioinformatic analysis of conserved residues in the amino acid sequences of the ketoreductase (KR) domains contained therein allowed the absolute configurations of the resultant hydroxyl groups to be accurately predicted and, together with complementary mass-spectrometry and NMR analyses, resulted in the determination of a precise molecular structure [87]. Marine actinobacteria are arguably one of the most promising, exciting and yet relatively untapped sources of novel natural products with the potential for development into a broad range of pharmaceutically-useful drugs [66]. We believe that the sequence-based approach that we have described here will play a major role in fulfilling this promise.

Supplementary Materials: The following are available online at www.mdpi.com/link, Figure S1: Examples of insertions and deletions in PacBio and 454 assembled sequence in homopolymeric runs of G or C. Figure S2: Region posing difficulty for sequencing.

Acknowledgments: Our research is supported financially by the Biotechnology and Biological Sciences Research Council (BBSRC, United Kingdom) Institute Strategic Programme Grant "Understanding and Exploiting Plant and Microbial Secondary Metabolism" (BB/J004561/1).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NGS	Next Generation Sequencing
NRPS	Non-Ribosomal Peptide Synthetase
PKS	PolyKetide Synthase
PacBio	Pacific Biosciences SMRT technology
SGS	Second Generation Sequencing
TGS	Third Generation Sequencing
TIR	Terminal Inverted Repeat

References

1. Gomez-Escribano, J.P.; Bibb, M.J. Heterologous expression of natural product biosynthetic gene clusters in *Streptomyces coelicolor*: From genome mining to manipulation of biosynthetic pathways. *J. Ind. Microbiol. Biotechnol.* **2014**, *41*, 425–431. [CrossRef] [PubMed]
2. Xiong, Z.-Q.; Wang, J.-F.; Hao, Y.-Y.; Wang, Y. Recent advances in the discovery and development of marine microbial natural products. *Mar. Drugs* **2013**, *11*, 700–717. [CrossRef] [PubMed]
3. Harvey, A.L.; Edrada-Ebel, R.; Quinn, R.J. The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discov.* **2015**, *14*, 111–129. [CrossRef] [PubMed]
4. Baltz, R.H. Renaissance in antibacterial discovery from actinomycetes. *Curr. Opin. Pharmacol.* **2008**, *8*, 557–563. [CrossRef] [PubMed]
5. Bentley, S.D.; Chater, K.F.; Cerdeño-Tárraga, A.-M.; Challis, G.L.; Thomson, N.R.; James, K.D.; Harris, D.E.; Quail, M.A.; Kieser, H.; Harper, D.; *et al.* Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **2002**, *417*, 141–147. [CrossRef] [PubMed]
6. Challis, G.L. Exploitation of the *Streptomyces coelicolor* A3(2) genome sequence for discovery of new natural products and biosynthetic pathways. *J. Ind. Microbiol. Biotechnol.* **2014**, *41*, 219–232. [CrossRef] [PubMed]
7. Zerikly, M.; Challis, G.L. Strategies for the discovery of new natural products by genome mining. *ChemBioChem* **2009**, *10*, 625–633. [CrossRef] [PubMed]
8. Gomez-Escribano, J.P.; Bibb, M.J. *Streptomyces coelicolor* as an expression host for heterologous gene clusters. *Methods Enzymol.* **2012**, *517*, 279–300. [PubMed]
9. Sanger, F.; Nicklen, S.; Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 5463–5467. [CrossRef] [PubMed]
10. Myers, E.W.; Sutton, G.G.; Delcher, A.L.; Dew, I.M.; Fasulo, D.P.; Flanigan, M.J.; Kravitz, S.A.; Mobarry, C.M.; Reinert, K.H.; Remington, K.A.; *et al.* A whole-genome assembly of *Drosophila*. *Science* **2000**, *287*, 2196–2204. [CrossRef] [PubMed]
11. Next-generation-sequencing. Available online: http://www.nature.com/subjects/next-generation-sequencing (accessed on 1 February 2016).
12. Weber, T.; Blin, K.; Duddela, S.; Krug, D.; Kim, H.U.; Bruccoleri, R.; Lee, S.Y.; Fischbach, M.A.; Müller, R.; Wohlleben, W.; *et al.* antiSMASH 3.0—Comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* **2015**, *43*, W237–W243. [CrossRef] [PubMed]
13. Medema, M.H.; Kottmann, R.; Yilmaz, P.; Cummings, M.; Biggins, J.B.; Blin, K.; de Bruijn, I.; Chooi, Y.H.; Claesen, J.; Coates, R.C.; *et al.* Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.* **2015**, *11*, 625–631. [CrossRef] [PubMed]
14. 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.; *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **2005**, *437*, 376–380. [CrossRef] [PubMed]
15. 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.; *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **2008**, *456*, 53–59. [CrossRef] [PubMed]
16. Loman, N.J.; Misra, R.V.; Dallman, T.J.; Constantinidou, C.; Gharbia, S.E.; Wain, J.; Pallen, M.J. Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **2012**, *30*, 434–439. [CrossRef] [PubMed]
17. Eid, J.; Fehr, A.; Gray, J.; Luong, K.; Lyle, J.; Otto, G.; Peluso, P.; Rank, D.; Baybayan, P.; Bettman, B.; *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **2009**, *323*, 133–138. [CrossRef] [PubMed]
18. Minion-access-programme. Available online: https://nanoporetech.com/community/the-minion-access-programme (accessed on 1 February 2016).
19. Feng, Y.; Zhang, Y.; Ying, C.; Wang, D.; Du, C. Nanopore-based fourth-generation DNA sequencing technology. *Genom. Proteom. Bioinform.* **2015**, *13*, 4–16. [CrossRef] [PubMed]
20. Laver, T.; Harrison, J.; O'Neill, P.A.; Moore, K.; Farbos, A.; Paszkiewicz, K.; Studholme, D.J. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect Quantif.* **2015**, *3*, 1–8. [CrossRef] [PubMed]

21. Schadt, E.E.; Turner, S.; Kasarskis, A. A window into third-generation sequencing. *Hum. Mol. Genet.* **2010**, *19*, R227–R240. [CrossRef] [PubMed]
22. Nakamura, K.; Oshima, T.; Morimoto, T.; Ikeda, S.; Yoshikawa, H.; Shiwa, Y.; Ishikawa, S.; Linak, M.C.; Hirai, A.; Takahashi, H.; *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* **2011**, *39*, e90. [CrossRef] [PubMed]
23. Kozarewa, I.; Ning, Z.; Quail, M.A.; Sanders, M.J.; Berriman, M.; Turner, D.J. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods* **2009**, *6*, 291–295. [CrossRef] [PubMed]
24. Quail, M.A.; Smith, M.; Coupland, P.; Otto, T.D.; Harris, S.R.; Connor, T.R.; Bertoni, A.; Swerdlow, H.P.; Gu, Y. A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **2012**, *13*, 341. [CrossRef] [PubMed]
25. Weaver, D.; Karoonuthaisiri, N.; Tsai, H.-H.; Huang, C.-H.; Ho, M.-L.; Gai, S.; Patel, K.G.; Huang, J.; Cohen, S.N.; Hopwood, D.A.; *et al.* Genome plasticity in *Streptomyces*: Identification of 1 Mb TIRs in the *S. coelicolor* A3(2) chromosome. *Mol. Microbiol.* **2004**, *51*, 1535–1550. [CrossRef] [PubMed]
26. Doroghazi, J.R.; Metcalf, W.W. Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC Genomics* **2013**, *14*, 611. [CrossRef] [PubMed]
27. Commins, J.; Toft, C.; Fares, M.A. Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biol. Proced. Online* **2009**, *11*, 52–78. [CrossRef] [PubMed]
28. Koren, S.; Phillippy, A.M. One chromosome, one contig: Complete microbial genomes from long-read sequencing and assembly. *Curr. Opin. Microbiol.* **2015**, *23*, 110–120. [CrossRef] [PubMed]
29. Thermofisher. Available online: http://www.thermofisher.com (accessed on 1 February 2016).
30. Pacific Biosciences. Available online: http://www.pacb.com/smrt-science/smrt-sequencing/read-lengths (accessed on 1 February 2016).
31. Rutherford, K.; Parkhill, J.; Crook, J.; Horsnell, T.; Rice, P.; Rajandream, M.A.; Barrell, B. Artemis: Sequence visualization and annotation. *Bioinformatics* **2000**, *16*, 944–945. [CrossRef] [PubMed]
32. Cole, S.T.; Brosch, R.; Parkhill, J.; Garnier, T.; Churcher, C.; Harris, D.; Gordon, S.V.; Eiglmeier, K.; Gas, S.; Barry, C., 3rd; *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **1998**, *393*, 537–544. [CrossRef] [PubMed]
33. Ikeda, H.; Ishikawa, J.; Hanamoto, A.; Shinose, M.; Kikuchi, H.; Shiba, T.; Sakaki, Y.; Hattori, M.; Omura, S. Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nat. Biotechnol.* **2003**, *21*, 526–531. [CrossRef] [PubMed]
34. Wu, H.; Qu, S.; Lu, C.; Zheng, H.; Zhou, X.; Bai, L.; Deng, Z. Genomic and transcriptomic insights into the thermo-regulated biosynthesis of validamycin in *Streptomyces hygroscopicus* 5008. *BMC Genomics* **2012**, *13*, 337. [CrossRef] [PubMed]
35. Zaburannyi, N.; Rabyk, M.; Ostash, B.; Fedorenko, V.; Luzhetskyy, A. Insights into naturally minimised *Streptomyces albus* J1074 genome. *BMC Genomics* **2014**, *15*, 97. [CrossRef] [PubMed]
36. Foulston, L.C.; Bibb, M.J. Microbisporicin gene cluster reveals unusual features of lantibiotic biosynthesis in actinomycetes. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 13461–13466. [CrossRef] [PubMed]
37. Foulston, L. Cloning and analysis of the microbisporicin lantibiotic gene cluster from *Microbispora corallina*. Ph.D. Thesis, University of East Anglia, Norwich, UK, 2010.
38. Claesen, J.; Bibb, M. Genome mining and genetic analysis of cypemycin biosynthesis reveal an unusual class of posttranslationally modified peptides. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 16297–16302. [CrossRef] [PubMed]
39. Sherwood, E.J.; Hesketh, A.R.; Bibb, M.J. Cloning and analysis of the planosporicin lantibiotic biosynthetic gene cluster of *Planomonospora alba*. *J. Bacteriol.* **2013**, *195*, 2309–2321. [CrossRef] [PubMed]
40. Sherwood, E. The planosporicin gene cluster from *Planomonospora alba*. Ph.D. Thesis, University of East Anglia, Norwich, UK, 2011.
41. Wyszynski, F.J.; Hesketh, A.R.; Bibb, M.J.; Davis, B.G. Dissecting tunicamycin biosynthesis by genome mining: Cloning and heterologous expression of a minimal gene cluster. *Chem. Sci.* **2010**, *1*, 581–589. [CrossRef]

42. Gomez-Escribano, J.P.; Song, L.; Bibb, M.J.; Challis, G.L. Posttranslational [small beta]-methylation and macrolactamidination in the biosynthesis of the bottromycin complex of ribosomal peptide antibiotics. *Chem. Sci.* **2012**, *3*, 3522–3525. [CrossRef]
43. Gomez-Escribano, J.P.; Castro, J.F.; Razmilic, V.; Chandra, G.; Andrews, B.; Asenjo, J.A.; Bibb, M.J. The *Streptomyces leeuwenhoekii* genome: De novo sequencing and assembly in single contigs of the chromosome, circular plasmid pSLE1 and linear plasmid pSLE2. *BMC Genomics* **2015**, *16*, 485. [CrossRef] [PubMed]
44. Laureti, L.; Song, L.; Huang, S.; Corre, C.; Leblond, P.; Challis, G.L.; Aigle, B. Identification of a bioactive 51-membered macrolide complex by activation of a silent polyketide synthase in Streptomyces ambofaciens. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 6258–6263. [CrossRef] [PubMed]
45. Li, S.; Li, Y.; Lu, C.; Zhang, J.; Zhu, J.; Wang, H.; Shen, Y. Activating a cryptic ansamycin biosynthetic gene cluster to produce three new naphthalenic octaketide ansamycins with *n*-pentyl and *n*-butyl side chains. *Org. Lett.* **2015**, *17*, 3706–3709. [CrossRef] [PubMed]
46. Claesen, J.; Bibb, M.J. Biosynthesis and regulation of grisemycin, a new member of the linaridin family of ribosomally synthesized peptides produced by *Streptomyces griseus* IFO 13350. *J. Bacteriol.* **2011**, *193*, 2510–2516. [CrossRef] [PubMed]
47. Jiang, Y.; Wang, H.; Lu, C.; Ding, Y.; Li, Y.; Shen, Y. Identification and characterization of the cuevaene A biosynthetic gene cluster in *Streptomyces* sp. LZ35. *ChemBioChem* **2013**, *14*, 1468–1475. [CrossRef] [PubMed]
48. Wattam, A.R.; Abraham, D.; Dalay, O.; Disz, T.L.; Driscoll, T.; Gabbard, J.L.; Gillespie, J.J.; Gough, R.; Hix, D.; Kenyon, R.; *et al.* PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* **2014**, *42*, D581–D591. [CrossRef] [PubMed]
49. Schorn, M.; Zettler, J.; Noel, J.P.; Dorrestein, P.C.; Moore, B.S.; Kaysser, L. Genetic basis for the biosynthesis of the pharmaceutically important class of epoxyketone proteasome inhibitors. *ACS Chem. Biol.* **2014**, *9*, 301–309. [CrossRef] [PubMed]
50. Bragg, L.M.; Stone, G.; Butler, M.K.; Hugenholtz, P.; Tyson, G.W. Shining a light on dark sequencing: Characterising errors in Ion Torrent PGM data. *PLoS Comput. Biol.* **2013**, *9*, e1003031. [CrossRef] [PubMed]
51. Cruz-Morales, P.; Vijgenboom, E.; Iruegas-Bocardo, F.; Girard, G.; Yácez-Guerra, L.A.; Ramos-Aboites, H.E.; Pernodet, J.-L.; Anné, J.; van Wezel, G.P.; Barona-Gómez, F. The genome sequence of *Streptomyces lividans* 66 reveals a novel tRNA-dependent peptide biosynthetic system within a metal-related genomic island. *Genome Biol. Evol.* **2013**, *5*, 1165–1175. [CrossRef] [PubMed]
52. Girard, G.; Willemse, J.; Zhu, H.; Claessen, D.; Bukarasam, K.; Goodfellow, M.; van Wezel, G.P. Analysis of novel *Kitasatosporae* reveals significant evolutionary changes in conserved developmental genes between *Kitasatospora* and *Streptomyces*. *Antonie Van Leeuwenhoek* **2014**, *106*, 365–380. [CrossRef] [PubMed]
53. Hoefler, B.C.; Konganti, K.; Straight, P.D. De Novo Assembly of the *Streptomyces* sp. Strain Mg1 Genome Using PacBio Single-Molecule Sequencing. *Genome Announc.* **2013**, *1*. [CrossRef] [PubMed]
54. Harrison, J.; Studholme, D.J. Recently published *Streptomyces* genome sequences. *Microb. Biotechnol.* **2014**, *7*, 373–380. [CrossRef] [PubMed]
55. Castro, J.F.; Razmilic, V.; Gomez-Escribano, J.P.; Andrews, B.; Asenjo, J.A.; Bibb, M.J. Identification and heterologous expression of the chaxamycin biosynthesis gene cluster from *Streptomyces leeuwenhoekii*. *Appl. Environ. Microbiol.* **2015**, *81*, 5820–5831. [CrossRef] [PubMed]
56. Busarakam, K.; Bull, A.T.; Girard, G.; Labeda, D.P.; van Wezel, G.P.; Goodfellow, M. *Streptomyces leeuwenhoekii* sp. nov., the producer of chaxalactins and chaxamycins, forms a distinct branch in *Streptomyces* gene trees. *Antonie Van Leeuwenhoek* **2014**, *105*, 849–861. [CrossRef] [PubMed]
57. Alt, S.; Wilkinson, B. Biosynthesis of the novel macrolide antibiotic anthracimycin. *ACS Chem. Biol.* **2015**, *10*, 2468–2479. [CrossRef] [PubMed]
58. Wright, F.; Bibb, M.J. Codon usage in the G+C-rich *Streptomyces* genome. *Gene* **1992**, *113*, 55–65. [CrossRef]
59. Bibb, M.J.; Findlay, P.R.; Johnson, M.W. The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. *Gene* **1984**, *30*, 157–166. [CrossRef]
60. Otto, T.D.; Sanders, M.; Berriman, M.; Newbold, C. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* **2010**, *26*, 1704–1707. [CrossRef] [PubMed]
61. Koren, S.; Schatz, M.C.; Walenz, B.P.; Martin, J.; Howard, J.T.; Ganapathy, G.; Wang, Z.; Rasko, D.A.; McCombie, W.R.; Jarvis, E.D.; *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **2012**, *30*, 693–700. [CrossRef] [PubMed]

62. Chin, C.-S.; Alexander, D.H.; Marks, P.; Klammer, A.A.; Drake, J.; Heiner, C.; Clum, A.; Copeland, A.; Huddleston, J.; Eichler, E.E.; *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **2013**, *10*, 563–569. [CrossRef] [PubMed]
63. Anand, S.; Prasad, M.V.R.; Yadav, G.; Kumar, N.; Shehara, J.; Ansari, M.Z.; Mohanty, D. SBSPKS: Structure based sequence analysis of polyketide synthases. *Nucleic Acids Res.* **2010**, *38*, W487–W496. [CrossRef] [PubMed]
64. Bhatnagar, I.; Kim, S.-K. Immense essence of excellence: Marine microbial bioactive compounds. *Mar. Drugs* **2010**, *8*, 2673–2701. [CrossRef] [PubMed]
65. Imhoff, J.F.; Labes, A.; Wiese, J. Bio-mining the microbial treasures of the ocean: New natural products. *Biotechnol. Adv.* **2011**, *29*, 468–482. [CrossRef] [PubMed]
66. Manivasagan, P.; Venkatesan, J.; Sivakumar, K.; Kim, S.-K. Pharmaceutically active secondary metabolites of marine actinobacteria. *Microbiol. Res.* **2014**, *169*, 262–278. [CrossRef] [PubMed]
67. Xu, D.-B.; Ye, W.-W.; Han, Y.; Deng, Z.-X.; Hong, K. Natural products from mangrove actinomycetes. *Mar. Drugs* **2014**, *12*, 2590–2613. [CrossRef] [PubMed]
68. Kennedy, J.; Flemer, B.; Jackson, S.A.; Lejon, D.P.H.; Morrissey, J.P.; O'Gara, F.; Dobson, A.D.W. Marine metagenomics: New tools for the study and exploitation of marine microbial metabolism. *Mar. Drugs* **2010**, *8*, 608–628. [CrossRef] [PubMed]
69. Kennedy, J.; Marchesi, J.R.; Dobson, A.D.W. Metagenomic approaches to exploit the biotechnological potential of the microbial consortia of marine sponges. *Appl. Microbiol. Biotechnol.* **2007**, *75*, 11–20. [CrossRef] [PubMed]
70. Kleigrewe, K.; Almaliti, J.; Tian, I.Y.; Kinnel, R.B.; Korobeynikov, A.; Monroe, E.A.; Duggan, B.M.; Di Marzo, V.; Sherman, D.H.; Dorrestein, P.C.; *et al.* Combining mass spectrometric metabolic profiling with genomic analysis: A powerful approach for discovering natural products from cyanobacteria. *J. Nat. Prod.* **2015**, *78*, 1671–1682. [CrossRef] [PubMed]
71. Piel, J. Approaches to capturing and designing biologically active small molecules produced by uncultured microbes. *Annu. Rev. Microbiol.* **2011**, *65*, 431–453. [CrossRef] [PubMed]
72. Reen, F.J.; Romano, S.; Dobson, A.D.W.; O'Gara, F. The sound of silence: Activating silent biosynthetic gene clusters in marine microorganisms. *Mar. Drugs* **2015**, *13*, 4754–4783. [CrossRef] [PubMed]
73. Trindade, M.; van Zyl, L.J.; Navarro-Fernández, J.; Abd Elrazak, A. Targeted metagenomics as a tool to tap into marine natural product diversity for the discovery and production of drug candidates. *Front. Microbiol.* **2015**, *6*, 890. [CrossRef] [PubMed]
74. Jensen, P.R.; Moore, B.S.; Fenical, W. The marine actinomycete genus *Salinispora*: A model organism for secondary metabolite discovery. *Nat. Prod. Rep.* **2015**, *32*, 738–751. [CrossRef] [PubMed]
75. Feling, R.H.; Buchanan, G.O.; Mincer, T.J.; Kauffman, C.A.; Jensen, P.R.; Fenical, W. Salinosporamide A: A highly cytotoxic proteasome inhibitor from a novel microbial source, a marine bacterium of the new genus *Salinospora*. *Angew. Chem. Int. Ed. Engl.* **2003**, *42*, 355–357. [CrossRef] [PubMed]
76. Mincer, T.J.; Jensen, P.R.; Kauffman, C.A.; Fenical, W. Widespread and persistent populations of a major new marine actinomycete taxon in ocean sediments. *Appl. Environ. Microbiol.* **2002**, *68*, 5005–5011. [CrossRef] [PubMed]
77. Bonet, B.; Teufel, R.; Crüsemann, M.; Ziemert, N.; Moore, B.S. Direct capture and heterologous expression of *Salinispora* natural product genes for the biosynthesis of enterocin. *J. Nat. Prod.* **2015**, *78*, 539–542. [CrossRef] [PubMed]
78. Gomez-Escribano, J.P.; Bibb, M.J. Engineering *Streptomyces coelicolor* for heterologous expression of secondary metabolite gene clusters. *Microb. Biotechnol.* **2011**, *4*, 207–215. [CrossRef] [PubMed]
79. Salzberg, S.L.; Yorke, J.A. Beware of mis-assembled genomes. *Bioinformatics* **2005**, *21*, 4320–4321. [CrossRef] [PubMed]
80. Murphy, R.R.; O'Connell, J.; Cox, A.J.; Schulz-Trieglaff, O. NxRepair: Error correction in de novo sequence assembly using Nextera mate pairs. *PeerJ* **2015**, *3*, e996. [CrossRef] [PubMed]
81. Tao, W.; Yurkovich, M.E.; Wen, S.; Lebe, K.E.; Samborskyy, M.; Liu, Y.; Yang, A.; Liu, Y.; Ju, Y.; Deng, Z.; *et al.* A genomics-led approach to deciphering the mechanism of thiotetronate antibiotic biosynthesis. *Chem. Sci.* **2016**, *7*, 376–385. [CrossRef]

82. Latreille, P.; Norton, S.; Goldman, B.S.; Henkhaus, J.; Miller, N.; Barbazuk, B.; Bode, H.B.; Darby, C.; Du, Z.; Forst, S.; *et al.* Optical mapping as a routine tool for bacterial genome sequence finishing. *BMC Genomics* **2007**, *8*, 321. [CrossRef] [PubMed]
83. Muggli, M.D.; Puglisi, S.J.; Ronen, R.; Boucher, C. Misassembly detection using paired-end sequence reads and optical mapping data. *Bioinformatics* **2015**, *31*, i80–i88. [CrossRef] [PubMed]
84. Ohnishi, Y.; Ishikawa, J.; Hara, H.; Suzuki, H.; Ikenoya, M.; Ikeda, H.; Yamashita, A.; Hattori, M.; Horinouchi, S. Genome sequence of the streptomycin-producing microorganism *Streptomyces griseus* IFO 13350. *J. Bacteriol.* **2008**, *190*, 4050–4060. [CrossRef] [PubMed]
85. Seven days. The news in brief. Business. End sequence. Available online: http://www.nature.com/news/seven-days-18-24-october-2013-1.13994 (accessed on 8 April 2016).
86. Frasch, H.-J.; Medema, M.H.; Takano, E.; Breitling, R. Design-based re-engineering of biosynthetic gene clusters: Plug-and-play in practice. *Curr. Opin. Biotechnol.* **2013**, *24*, 1144–1150. [CrossRef] [PubMed]
87. Schulze, C.J.; Donia, M.S.; Siqueira-Neto, J.L.; Ray, D.; Raskatov, J.A.; Green, R.E.; McKerrow, J.H.; Fischbach, M.A.; Linington, R.G. Genome-directed lead discovery: Biosynthesis, structure elucidation, and biological evaluation of two families of polyene macrolactams against *Trypanosoma brucei*. *ACS Chem. Biol.* **2015**, *10*, 2373–2381. [CrossRef] [PubMed]