

# Clustered patterns of species origins of nature-derived drugs and clues for future bioprospecting

Feng Zhu<sup>a,b,c</sup>, Chu Qin<sup>b,c,d</sup>, Lin Tao<sup>b,c,d</sup>, Xin Liu<sup>b,c</sup>, Zhe Shi<sup>b,c</sup>, Xiaohua Ma<sup>b,c</sup>, Jia Jia<sup>b,c</sup>, Ying Tan<sup>a</sup>, Cheng Cui<sup>a</sup>, Jinshun Lin<sup>a</sup>, Chunyan Tan<sup>a</sup>, Yuyang Jiang<sup>a,e,1</sup>, and Yuzong Chen<sup>a,b,c,1</sup>

<sup>a</sup>Key Laboratory of Chemical Biology, Guangdong Province Graduate School at Shenzhen, Tsinghua University, Shenzhen, Guangdong 518055, People's Republic of China; <sup>b</sup>Bioinformatics and Drug Design Group, Department of Pharmacy, and <sup>c</sup>Center for Computational Science and Engineering, National University of Singapore, Singapore 117543; <sup>d</sup>Department of Pharmacy, National University of Singapore Graduate School for Integrative Sciences and Engineering, Singapore 117456; and <sup>e</sup>School of Medicine and Department of Chemistry, Tsinghua University, Beijing 100084, P. R. China

Edited by Jerrold Meinwald, Cornell University, Ithaca, NY, and approved June 27, 2011 (received for review May 10, 2011)

**Many drugs are nature derived. Low drug productivity has renewed interest in natural products as drug-discovery sources. Nature-derived drugs are composed of dozens of molecular scaffolds generated by specific secondary-metabolite gene clusters in selected species. It can be hypothesized that drug-like structures probably are distributed in selective groups of species. We compared the species origins of 939 approved and 369 clinical-trial drugs with those of 119 preclinical drugs and 19,721 bioactive natural products. In contrast to the scattered distribution of bioactive natural products, these drugs are clustered into 144 of the 6,763 known species families in nature, with 80% of the approved drugs and 67% of the clinical-trial drugs concentrated in 17 and 30 drug-prolific families, respectively. Four lines of evidence from historical drug data, 13,548 marine natural products, 767 medicinal plants, and 19,721 bioactive natural products suggest that drugs are derived mostly from preexisting drug-productive families. Drug-productive clusters expand slowly by conventional technologies. The lack of drugs outside drug-productive families is not necessarily the result of under-exploration or late exploration by conventional technologies. New technologies that explore cryptic gene clusters, pathways, interspecies crosstalk, and high-throughput fermentation enable the discovery of novel natural products. The potential impact of these technologies on drug productivity and on the distribution patterns of drug-productive families is yet to be revealed.**

biodiversity | drug-prolific species | phylogenetic tree | herb

**M**any approved and clinical-trial drugs are derived from natural products (1, 2). During the past 2 decades, the focus of drug-discovery efforts has shifted from natural products to synthetic compounds, but this shift has not led to the anticipated increase in drug productivity (3). Despite the shifted focus, nature-derived drugs still constitute a substantial percentage of recently approved drugs. For instance, 12 (26%) of the 46 molecular entities approved by the Food and Drug Administration (FDA) in 2009–2010 are nature derived (*SI Appendix, Table S1*). There is a renewed interest in natural products as drug-discovery sources (4). The scope of biodiversity and extinction rates (5) demands bioprospecting efforts be prioritized toward the groups of species that are likely to yield new drugs.

Clues to drug-productive species can be obtained from the species-distribution profiles of nature-derived approved and clinical-trial drugs. Although particular species yield potent bioactive compounds at higher rates than others, additional drug-like properties are important for developing these compounds into marketable drugs (6). The nature-derived approved and clinical-trial drugs are composed primarily of several dozen molecular scaffolds (7–9) rather than the numerous bioactive natural-product scaffolds (10, 11). Like other bioactive natural-product scaffolds, the nature-derived privileged drug-like scaffolds are generated by enzymes partly encoded in specific secondary-metabolite gene clusters in selected groups of species (12–14). Questions arise as to whether the privileged drug-like

structures are distributed in selective species-groups rather than being scattered in the phylogenetic tree (15) and whether the distribution shows certain traceable patterns that can be explored in future bioprospecting efforts.

A large number of bioactive natural products have been identified (16, 17). Many more are likely to be discovered (17) because of their interaction with specific targets (18). A small percentage of bioactive natural products has been carried forward to derive approved (1) and clinical-trial (2, 19) drugs via direct exploration, semisynthetic modification, structural mimicking, or pharmacophore mapping. The potential for drug development of a natural product depends not only on its bioactivity but also on the drug-likeness of its structure [optimized for enhanced drug-like (6) and reduced unwanted (20–22) properties] and the susceptibility of its target (“druggability”) to drugs [324 targets are confirmed druggable in yielding approved drugs (23), and 292 targets have yielded drugs in clinical trials (24)]. The odds for finding novel drug-like natural products may be improved if one can identify new drug-productive species, particularly endangered ones, before their extinction.

Structurally diverse bioactive natural products are composed of many molecular scaffolds (16, 17). Each scaffold is generated by specific enzyme assemblies (25) encoded in the secondary-metabolite gene clusters of specific species groups (12–14, 26). Partly as a survival strategy (12), structurally diverse natural products are generated by genetic variations and repositioning (27), posttranslational modifications (28), and assembly-line regulation (28). Nonetheless, bioactive secondary metabolites of an individual scaffold typically are produced by species from a specific family (e.g., anthraquinones in Polygonaceae) (14, 29) or from a few families of a specific order (e.g., sordarins of Xylariales) (14, 30). Specifically, 14 of the 26 drug-productive scaffolds from Actinomycetales are from a unique family, six are from a few families within the order, and four are from a few families in this and a few other orders (*SI Appendix, Table S2*).

Natural products active against individual targets or classes of targets may be composed of multiple scaffolds, many of which are from only a few families. For instance, 53 nicotinic acetylcholine receptor ligands are reported from diverse species (16). Our analysis (*SI Appendix, Table S3*) showed that these ligands are clustered into 29 scaffolds; 23 of these scaffolds are from a unique family, three are from a few families within a specific

Author contributions: F.Z., Y.J., and Y.C. designed research; F.Z., Y.J., and Y.C. performed research; F.Z., C.Q., L.T., X.L., Z.S., X.M., J.J., Y.T., C.C., J.L., C.T., Y.J., and Y.C. analyzed data; and F.Z., Y.J., and Y.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

<sup>1</sup>To whom correspondence may be addressed. E-mail: jiangyy@sz.tsinghua.edu.cn or phacyz@nus.edu.sg.

This article contains supporting information online at [www.pnas.org/lookup/suppl/doi:10.1073/pnas.1107336108/-DCSupplemental](http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1107336108/-DCSupplemental).

order, and three are from a few orders of a specific class. Each of the five approved drugs in the group is from a specific family. Similarly, the 12 nature-derived FDA-approved kinase inhibitor drugs (1, 2) are grouped into three scaffold groups, with each scaffold derived from a few families distributed among only a few orders (*SI Appendix, Table S4*).

Some families such as Streptomycetaceae, Pseudonocardia- ceae, and Trichocomaceae and some genera such as *Acremonium* and *Emericellopsis* are highly drug-prolific (14, 15). Compounds synthesized by a specific metabolic pathway typically are active against only a few targets (13). It thus can be hypothesized that privileged drug-like structures targeting selective druggable targets are likely to be concentrated in specific families. This hypothesis can be evaluated, and the distribution patterns of drug-productive species can be revealed by comparing the species origins of the approved and clinical-trial drugs (1, 2) with those of preclinical drugs and bioactive natural products.

We analyzed the species origins of 939 approved drugs (*SI Appendix, Table S5*) (1), 369 clinical-trial drugs (*SI Appendix, Table S6*) (2, 19), 119 preclinical drugs (*SI Appendix, Table S7*) (31, 32), and 19,721 bioactive natural products (*SI Appendix, Table S8*) with particular focus on their distribution patterns in the phylogenetic trees of the Bacteria, Viridiplantae, Fungi, and Metazoa kingdoms or superkingdoms. Nature-derived approved and clinical-trial drugs and their species origins were obtained from published reviews (1, 2, 19) and our own literature search (*SI Appendix, Table S8*). Preclinical drugs are drug candidates that have entered preclinical studies such as safety, pharmacokinetics/absorption, distribution, metabolism, and excretion, active pharmaceutical ingredient preparation, and formulation (33). Following the seminal works on nature-derived drugs (1, 2), we included in our analysis biologics, natural products and their semisynthetic derivatives, mimics, and peptidomimetics. Biologics include peptides (34), recombinant proteins (35), and monoclonal antibodies (36) except for RNA-based drugs (37). The inclusion or exclusion of biologics and RNA-based drugs had limited effect on our analysis because they primarily are of human or viral origin. For semisynthetic derivatives, mimics, and peptidomimetics, the host species of the parent natural-product leads were analyzed (1, 2).

We also evaluated four lines of evidence from (i) historical drug data, (ii) 13,548 natural products of marine species and their nonmarine counterparts, (iii) 767 medicinal plants, and (iv) 19,721 bioactive natural products to determine whether the distribution patterns of drug-productive families are different from those of bioactive natural products and if the lack of drugs outside drug-productive families is the result of under-exploration or late exploration.

The natural-product leads of drugs and bioactive natural products in this analysis were discovered mostly by conventional technologies rather than by the new technologies that are expected to identify many previously unrecognized bioactive natural products. These technologies are based on the exploration of cryptic metabolic gene clusters (via genomic mining, epigenetic modification, and proteomics) (38–40), metabolic pathway engineering (41, 42), interspecies crosstalk (40, 43, 44), and high-throughput fermentation and screening (45). Their potential contribution to discovery of new drugs is highly anticipated (40, 42, 45). The distribution patterns of the drugs discovered by these technologies may not follow those of the existing drugs derived from conventional technologies.

## Results and Discussion

Apart from 12 drugs from widely distributed species, 933 approved and 363 clinical-trial drugs were concentrated in 144 drug-productive families of the 6,763 families known in nature. Of the 144 drug-productive families, 99 contained approved drugs. Many of these families contain endangered species. In particular, 80% of the approved drugs are concentrated in 17 drug-prolific families, and 67% of the clinical-trial drugs are

concentrated in 30 such families. Most (82.4%) of the clinical-trial drugs are distributed within 59 families that also contain approved drugs. Thus, the presence of drug-like structures, as represented by the approved and clinical-trial drugs, is highly concentrated in selective families.

Fig. 1 presents the drug-prolific families with the highest numbers of approved drugs. The well-known drug-prolific families Streptomycetaceae and Pseudonocardia- ceae of the Bacteria, the family Trichocomaceae and the genera *Acremonium* and *Emericellopsis* of the Fungi, and the family Hominidae (great apes and humans) of the Metazoa superkingdom are among the most prolific, with 59.1% of the approved drugs, followed by the Viridiplantae (green plant) families Fabaceae (legumes), Ephedraceae (Mormon tea), Papaveraceae (poppies), Asteraceae (daisy), Solanaceae (potato), Rubiaceae (coffee), and Apocynaceae (dogbane), and the Metazoan families Viperidae (venomous snakes) and Muridae (rodents). Two Bacteria families (Streptomycetaceae and Pseudonocardia- ceae), four plant families (Fabaceae, Rubiaceae, Asteraceae, and Apocynaceae), and one Metazoa family (Hominidae) also are prolific in clinical-trial drugs. The enriched number of clinical-trial drugs from these families (2, 19) arises partly from the exploration of sources such as marine actinomycete bacteria (15, 46) and plants (19). The Hominidae family is the highest ranked drug-productive family, largely because of the inclusion of biologics. It becomes the second-ranked family if biologics are excluded.

We also tentatively analyzed the ranking of drug-productive families based on the ratio of the approved drugs to the searchable bioactive natural products (including leads of the approved and clinical-trial drugs) from each family. Partly because of the limited data from the available databases and our literature search, our searched natural products are insufficient to reflect the true ratios adequately. Nonetheless, our analysis of families with >20 searchable bioactive natural products showed that 70% of the top-ranked drug-productive families in Fig. 1 are among the families with highest drug-to-natural product ratios (*SI Appendix, Fig. S1*), suggesting that the high productivity of some of these top-ranked families may result from a higher frequency of derived drugs rather than from a higher number of bioactive natural products explored. The two Fungi genera *Acremonium* and *Emericellopsis*, in the top-ranked six families in Fig. 1, were excluded from our analysis because they have fewer searchable bioactive natural products (12 and 5, respectively).

*SI Appendix, Fig. S2* shows the top-ranked families without an approved drug that nonetheless yield high numbers of clinical-

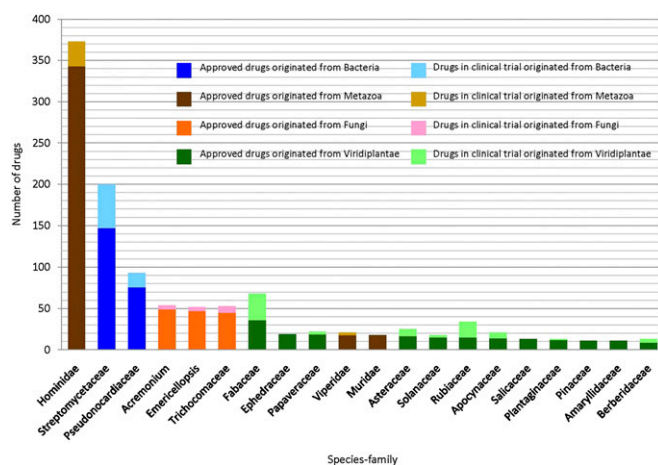
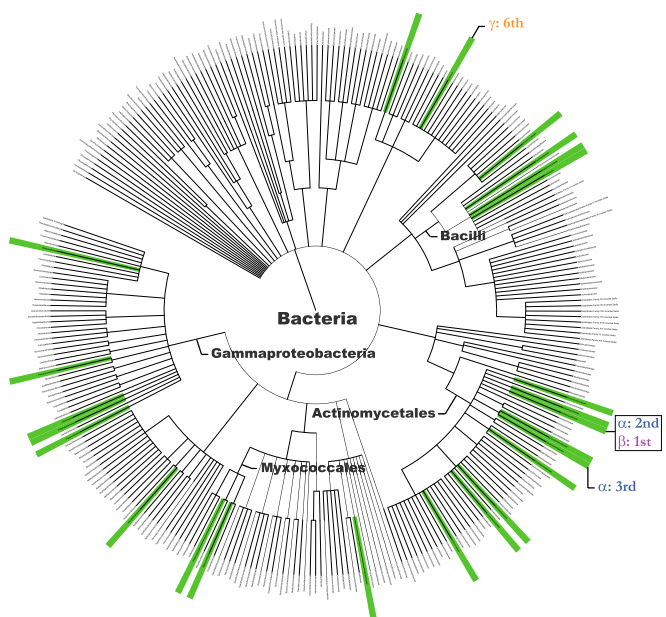


Fig. 1. Top-ranked drug-prolific families that produced high numbers of approved drugs.

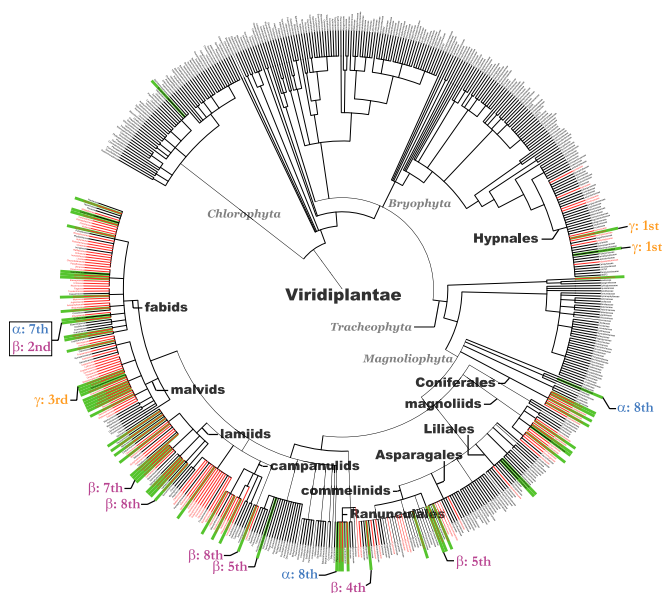
trial drugs. These families include six plant families [Leskeaceae (a moss family), Anomodontaceae (a moss family), Combretaceae (almond family), Quillajaceae, Cephalotaxaceae (plum-yew family), and Bryopsidaceae (a green algae family)], seven families of Metazoa [Aplysiidae (sea hares), Bugulidae (marine moss animals), Dendrobatidae (poison frogs), Petrosiidae (a sponge family), Axinellidae (a sponge family), Squalidae (dogfish sharks), and Hemiasterellidae (a sponge family)], and a genus of Bacteria, *Symploca*. Many of these families, particularly the marine ones, have been explored actively and reportedly have shown good drug-discovery potential (4, 19, 46–48).

Further analysis of the species origins of approved and clinical-trial drugs in the phylogenetic trees of the Bacteria (Fig. 2), Viridiplantae (Fig. 3), Fungi (SI Appendix, Fig. S3), and Metazoa (SI Appendix, Fig. S4) kingdoms or superkingdoms revealed clustered patterns. One can define a drug-productive cluster as a relatively small branch of a phylogenetic tree with two or more drug-productive families. The bacteria superkingdom (289 known families) contains 23 drug-productive families, 18 of which, including the drug-prolific ones, are concentrated in four drug-productive clusters in the Actinomycetales order, Bacillales order, Gammaproteobacteria class, and Myxococcales order (Fig. 2). Bioprospecting efforts can be extended to other subbranches in these drug-productive clusters, and such efforts have been reported. For instance, two genera of the drug-productive Actinomycetaceae family have been discovered recently and found to produce promising anticancer and antibiotic compounds (46).

The Viridiplantae kingdom (740 known families) contains 66 drug-productive families, 61 of which, including the drug-prolific ones, are concentrated in 11 drug-productive clusters. The 11 drug-productive clusters are the Fabid and Malvid groups of the Rosidae subclass; the Lamiid and Campanulid groups of the Asterid subclass; the Ranunculales order, Magnoliids clade,



**Fig. 2.** Distribution of drug-productive families (green background color) in the phylogenetic tree of the Bacteria superkingdom. The 10 families with highest number of approved drugs ( $\alpha$ ), highest number of clinical-trial drugs and at least one approved drug ( $\beta$ ), and highest number of clinical-trial drugs without an approved drug ( $\gamma$ ) are marked. Drug-productive clusters that contain two or more drug-productive families are labeled with the name of species group (class or order) in black. The family names are provided at branch ends, which can be viewed more clearly by enlarging the figure in the electronic version.



**Fig. 3.** Distribution of drug-productive families in the phylogenetic tree of the Viridiplantae kingdom. Coloring and labeling schemes are as in Fig. 2. Red lines indicate families containing endangered species.

Coniferales order, Commelinids clade, and Asparagales order of the Monocots group; the Hypnales order (mosses), and the Liliales order (Fig. 3). The Fungi kingdom (521 known families) contains 16 drug-productive families, 12 of which are concentrated in two drug-productive clusters in the Hypocerales order and Agaricales order (gilled mushrooms).

The Metazoa kingdom (4,468 known families) contains 38 drug-productive families, 30 of which, including drug-prolific ones, are concentrated in nine drug-productive clusters. The nine drug-productive clusters are the Euarchontoglires clade (rodents, lagomorphs, tree shrews, colugos, and primates including humans), Cetartiodactyla clade (whales, dolphins, and even-toed ungulates), Scleroglossa suborder in the Squamata order (scaled reptiles), Demospongiae class (demosponges), Hirudinea subclass (leeches), Gastropoda class (snails and slugs), Enterogona order (marine saclike filter feeder animals), Batrachia superorder (vertebrate amphibians without tails), and Monostilifera order (ribbon worms) (SI Appendix, Fig. S4).

The clustered patterns of drug-productive families shown in Figs. 2 and 3 and SI Appendix, Figs. S3 and S4 are in contrast to the scattered distribution patterns of the families of the 19,721 searchable bioactive natural products (SI Appendix, Figs. S5–S8). Our collected data on bioactive natural products are inadequate to represent all bioactive ingredients, but some useful indications were revealed. These natural products are distributed in 792 families (67 Bacteria, 273 Viridiplantae, 110 Fungi, and 345 Metazoa families) scattered in the phylogenetic trees, 57.1% of which (46.3% Bacteria, 37.0% Viridiplantae, 72.7% Fungi, and 69.6% Metazoa) are outside drug-productive clusters. These data suggest that the current drug-productive families form distinctive groups that tend to cluster together in phylogenetic trees. It may be hypothesized that these patterns partly reflect the distribution of the secondary-metabolite gene clusters that produce drug-like scaffolds (12–14) in selected species (7–9).

The clustered patterns of drug-productive families also were compared tentatively with the distribution patterns of the 119 nature-derived preclinical drugs (SI Appendix, Figs. S9–S12). Although the number of our collected preclinical drugs is too small to reveal their distribution patterns clearly, some useful indications may be obtained. The natural-product leads of these

drugs are from 63 families, 73.0% of which are drug productive (58.7%) or are non-drug productive but are distributed in drug-productive clusters (14.3%). This percentage is significantly higher than that (42.9%) of the 19,721 bioactive natural products. These data seem to indicate that the preclinical drug families may be more concentrated in the phylogenetic trees than those of bioactive natural products but less concentrated than the drug-productive families and clusters.

Most of the preclinical drug families outside the drug-productive clusters are from the Fungi and Metazoa kingdoms (specifically soft corals, scorpions, and tube-dwelling spiders). It is unclear whether this biased distribution is the result of limited sampling of preclinical drugs or whether it partially reflects more recent efforts investigating such sources as fungal (49), soft coral (50), and invertebrate (51) species. The drug-productive families and clusters in these two kingdoms are in significantly smaller regions of the phylogenetic trees than those in the Bacteria superkingdom and the Viridiplantae kingdom; this distribution may leave open more branches for deriving preclinical drugs.

Four lines of evidence suggest that the recognition of drug-productive clusters expands slowly using conventional technologies and that the paucity of drugs outside drug-productive families is not necessarily the result of under-exploration or late exploration by conventional technologies. The first line of evidence is from the historical data of nature-derived approved drugs. As shown in Table 1, most (80.9–97.2%) of the 39–141 nature-derived drugs approved during every 5-y period from 1961 to 2010 are from previous drug-productive families or from newly identified families in previous drug-productive clusters. In particular, most (82.1–96.3%) of the drugs approved during every 5-y period from 1961–1975 and 1981–2010 and the majority (66.3%) of the drugs approved from 1976–1980 are from previous drug-productive families. These data suggest that, regardless the varying levels of exploration, preexisting drug-productive families and clusters are the most prolific sources for yielding approved drugs.

Comparison of the newly identified drug-productive families that have emerged since 1981 with pre-1981 drug-productive clusters (*SI Appendix, Figs. S13–S16*) showed that most (80.4%) of the 28 families identified post-1981 outside the pre-1981 clusters are clustered with one or more existing or newly identified drug-productive families, suggesting that the newly identified drug-productive families tend to form clusters with existing drug-productive families or among themselves. In every 5-y period from 1961–2010, small numbers of newly identified drug-productive families (5.5 families per 5-y period from 1961–1990

and 2.75 families per 5-y period from 1991–2010) and clusters (1.5 clusters per 5-y period from 1961–2010) have emerged outside previous drug-productive clusters (Table 1).

The patterns we revealed may be influenced predominantly by older drugs. Drug-discovery focuses are shifting in terms of targets, chemotypes, diseases, and therapeutic strategies (e.g., multitarget and RNA-based drugs) (34–37, 52, 53). The patterns we discerned might change significantly as new drugs with new targets are derived from natural products extracted by conventional technologies. To study this possibility, we further analyzed 74 nature-derived approved drugs. These drugs include 12 drugs approved in 2008–2009 (*SI Appendix, Table S1*), 10 kinase inhibitors (*SI Appendix, Table S4*), and 52 drugs targeting 26 targets first explored successfully since 1990 (*SI Appendix, Table S9*) (24). To some extent, these drugs represent novel drugs for new targets, chemotypes, diseases, and therapeutic approaches. At the time of their approval, 71 (95.9%) of these drugs were from preexisting drug-productive families (91.9%) or were from non-drug-productive families within preexisting drug-productive clusters (4.0%), and 44 drugs (59%) were from families with drugs approved before 1980. These 71 drugs target 32 targets; the first drug approved for 26 (81.3%) of these targets have been derived from a preexisting drug-productive family. Thus, the evolving focus of drug discovery appears to have a limited effect on the expansion of drug-productive clusters by conventional technologies, and existing drug-productive families and clusters are the primary sources of novel drugs.

The second line of evidence, which comes from the distribution of 13,548 natural products from marine phyla (including nonmarine species in these phyla), indicates that the paucity of drugs outside drug-productive families is not necessarily the result of under-exploration by conventional technologies. Some non-drug-productive phyla produce high numbers of compounds, comparable to the number in drug-productive phyla (*SI Appendix, Table S10*). This evidence is supported further by the third line of evidence from published data from medicinal plant research. Plant extracts from >3,000 species have been assayed extensively for anticancer activity alone (54), and ~10% of the 250,000–500,000 plant species have been studied (55). From medicinal plant databases and literature searches, we identified 767 medicinal plants from 172 non-drug-productive families, 104 (60.5%) of which are outside recognized drug-productive clusters (*SI Appendix, Table S11*).

The fourth line of evidence is from the distribution profile and exploration timeline of 19,721 bioactive natural products (in-

**Table 1. Number of nature-derived approved drugs, drug-productive families, and drug-productive clusters (excluding potential drug-productive clusters) during every 5-y period, 1961–2010**

Period	No. of approved drugs in period			No. of drug-productive families			No. of drug-productive clusters	
	From previous drug-productive families	From newly identified drug-productive families in previous drug-productive clusters	From newly identified drug-productive families not in previous drug-productive clusters	No. of existing families	No. of newly identified families in previous drug-productive clusters	No. of newly identified families not in previous drug-productive clusters	No. of existing clusters	No. of newly identified clusters
1961–1965	32	0	7	21	0	5	6	1
1966–1970	26	0	1	26	0	1	7	0
1971–1975	32	3	1	27	3	1	7	0
1976–1980	59	13	17	31	8	12	7	3
1981–1985	98	1	11	51	1	8	10	3
1986–1990	128	7	6	60	6	6	13	2
1991–1995	111	9	5	72	7	3	15	1
1996–2000	129	4	4	82	3	2	16	2
2001–2005	124	1	7	87	1	4	18	1
2006–2010	44	3	2	92	4	2	19	0

cluding leads of nature-derived drugs). Although these natural products inadequately represent all bioactive natural products, some useful indications may be revealed. Consistent with the distribution profiles of the 13,548 natural products from marine phyla and 767 medicinal plants, a number of non-drug-productive families outside drug-productive clusters contain high numbers of bioactive natural products comparable to those in drug-productive families and in non-drug-productive families inside drug-productive clusters (*SI Appendix, Table S12*). In particular, non-drug-productive families outside and inside drug-productive clusters constitute 22% and 10%, respectively, of the 50 families with the highest number of searchable bioactive natural products; 60% of the 20 non-drug-productive families outside drug-productive clusters and 80% of the 10 non-drug-productive families within drug-productive clusters contain bioactive natural products that were reported before 1990. This evidence further suggests that the paucity of drugs identified outside drug-productive families and clusters is not necessarily the result of under-exploration and late exploration by conventional technologies.

The approved drugs, grouped into drug-target classes with three or more drugs against each target, are clustered in specific therapeutic regions of the phylogenetic subbranches of drug-productive families (*SI Appendix, Fig. S17*). Consistent with the report that compounds synthesized by a specific pathways typically are active against a limited number of targets (13), drugs in each target class are distributed mostly in one to three families. Exceptions are three classes of drugs used to treat infection (targeting penicillin-binding protein, ribosome, and bacterial outer membrane), two classes of drugs used to treat circulation disorders (targeting maltase-glucoamylase and adenosine receptor), three classes of drugs used to treat cancer (targeting tubulin, DNA topoisomerase, an thymidylate synthase), one drug used to treat disease of the nervous system (targeting the opioid receptor), and one drug used to treat inflammation (targeting the 5-HT receptor); these classes are distributed among five or six families.

Antibiotic drugs are primarily from one large Bacteria cluster of four families (Streptomycetaceae, Pseudonocardiaceae, Actinosynnemataceae, and Actinomycetaceae), one Fungi cluster of two families (Acremonium and Emericellopsis), and another Fungi family (Trichocomaceae). Anticancer drugs are largely from one Bacteria family (Streptomycetaceae), one Metazoa cluster of two families (Perophoridae and Hominidae), and three plant clusters consisting respectively of three families (Fabaceae, Betulaceae, and Moraceae), three families (Apocynaceae, Rubiaceae, and Icacinaceae), and two families (Poaceae and Arecaceae). Drugs targeting the reproductive system are mostly from the Hominidae family. Anti-inflammatory drugs are primarily from the Hominidae family and several plant families. Drugs targeting the nervous system are largely from a plant cluster of four families (Papaveraceae, Menispermaceae, Berberidaceae, and Fabaceae) and from several other plant (Ephedraceae, Amaranthaceae, Asteraceae, Solanaceae, and Amaryllidaceae), Metazoa, and Fungi families. Immunity drugs are primarily from three Fungi families (Clavicipitaceae, Nectriaceae, and Trichocomaceae) with a few drugs from one Bacteria and one Metazoa family. Circulation drugs are primarily from one Bacteria cluster consisting of two families (Streptomycetaceae and Pseudonocardiaceae), from one plant cluster of four families (Malvaceae, Theaceae, Plantaginaceae, and Loganiaceae), and from two Metazoa clusters of three families (Hominidae, Muridae, and Viperidae) and two families (Glossiphoniidae and Hirudinidae), respectively.

These drug-productive families typically produce drugs against multiple targets in multiple therapeutic areas. For instance, drugs from Streptomycetaceae are distributed in 12 target classes of four therapeutic areas (infection, cancer, circulation, and immunity), and drugs from Hominidae are distributed in 16 target classes of six

therapeutic areas (inflammation, reproduction, cancers, circulation, immunity, and nervous system). These multiple targets in multiple therapeutic areas suggest that usually there are multiple drug-producing secondary-metabolite gene clusters among the many gene clusters in specific families (14). Two species, *S. avermitilis* and *S. griseus*, of the *Streptomyces* genus in the drug-productive Streptomycetaceae family have been observed to produce two or three natural products, but analysis of their sequenced genome suggests that they might harbor ~25–30 predicted biosynthetic gene clusters (28). Efforts have been made to activate the silent “cryptic” gene clusters in certain drug-productive families (38, 39). These efforts may enable the discovery of additional drug leads with different structures and therapeutic applications.

From the observed tendency of newly recognized drug-productive families to cluster with preexisting drug-productive families or among themselves, one may speculate that the families with only clinical-trial drugs that are clustered with another drug-productive family may have a higher probability of yielding approved drugs. Examples of these families are the Myxococcaceae, Micrococcaceae, and Streptosporangiaceae families in the Myxococcales and Actinomycetales orders of the Bacteria superkingdom; the Melanthiaceae, Calophyllaceae, Quillajaceae, Bignoniaceae, Oleaceae, Anomodontaceae, and Leskeaceae families in the Liliales and the Fabid, Lamiid, and Hypnales orders or superorders of Viridiplantae kingdom; the Omphalotaceae family in the Agaricales order of the Fungi kingdom; and the Didemnidae, Polyclinidae, Aplysiidae, Kentrodrorididae, Placobranchidae, Ancorinidae, Aplysinellidae, Axinellidae, Dysideidae, Hemiasterellidae, Petrosiidae, Pseudoceratinidae, Amphiporidae, Bufonidae, Dendrobatidae, and Salamandridae families, and the *Paranemertes* genus in the Enterogona, Gastropoda, Demospongiae, and Monostilifera classes, suborders, or orders, of the Metazoa kingdom.

## Conclusion

Our analysis revealed that nature-derived drugs have been derived mostly from drug-productive families that tend to be clustered rather than scattered in the phylogenetic tree. Several indications from historical drug data and extracted natural products suggest that the identification of these clusters expands slowly when conventional technologies are used, and the paucity of drugs outside these clusters is not necessarily the result of under-exploration or late exploration by conventional technologies. New technologies (38–45) are expected to expand the pool of drug leads significantly (40, 42, 45). The impact of these technologies on drug productivity and the distribution of drug-productive families is yet to be determined. The distribution patterns of the future drug-productive families may differ from those of the existing drugs derived from conventional technologies. A particular question is whether the future drug-productive families are clustered, probably on a larger scale. Analysis of the species origins of 321 previously unrecognized secondary metabolites published in 2001–2011 (*SI Appendix, Table S13*) showed that 93 (29%) are distributed in 22 families outside drug-productive clusters, and 228 (71%) are distributed in 43 families within drug-productive clusters. Those outside drug-productive clusters are distributed in the relevant families at comparable densities as those within drug-productive clusters. It remains to be determined if a similar concentrated distribution is found for the secondary metabolites generated by new technologies. The clustered patterns revealed in this work provide useful information about the groups of species that are drug productive or potentially productive. This information, coupled with expanded knowledge of drug-like structures (7–9) and drug-productive species and with new technologies (4, 40–45, 56, 57), may enable more prioritized, focused, rational, and environmentally friendly bioprospecting for novel drug-like natural products.

## Materials and Methods

The species origins of 939 approved and 369 clinical-trial drugs were identified as follows. First Therapeutic Target Database (24) was checked to confirm the current approval or clinical-trial status of the literature-reported approved drugs (1) and clinical-trial drugs (2, 19) of natural origin. Then the species origin of every drug was searched in books and review and regular articles using combinations of keywords such as the drug name and alternative names, species, "natural product," and "nature." The literature searched is listed in *SI Appendix, Table S8*. The species origin of a drug was confirmed by a specific statement in the literature that the drug "originates from," "is derived from," "is isolated from," or "comes from" a species or species group (e.g., a genus or family). For drugs from semisynthetic derivatives, mimics, and peptidomimetics, the parent natural-product leads were searched first, followed by a search of host species as described above. The families of the host species of these drugs as well as all the known families in nature are from the National Center for Biotechnology Information

(NCBI) taxonomy database (58). Families with endangered species were identified by mapping the species against the International Union for the Conservation of Nature Red List of Threatened Species Version-2010.3 (<http://www.iucnredlist.org>). The phylogenetic trees were generated by using the NCBI taxonomy-based automatic tree generator in iTOL version 1.8.1 (59) against known families in the Bacteria, Viridiplantae, Fungi, and Metazoa kingdoms or superkingdoms. Family names are provided at branch ends. Red branches indicate families with endangered species. Drug-productive clusters and some of the drug-productive families are labeled or marked in the phylogenetic trees.

**ACKNOWLEDGMENTS.** This work was supported by the Chinese National Natural Science Foundation (20872077 and 90813013), by the Chinese Ministry of Science and Technology Key Special Project Grant 2009ZX09501-004, and by Singapore Academic Research Fund Grants R-148-000-141-750 and R-148-000-141-646.

- Newman DJ, Cragg GM (2007) Natural products as sources of new drugs over the last 25 years. *J Nat Prod* 70:461–477.
- Butler MS (2008) Natural products to drugs: Natural product-derived compounds in clinical trials. *Nat Prod Rep* 25:475–516.
- Paul SM, et al. (2010) How to improve R&D productivity: The pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* 9:203–214.
- Li JW, Vederas JC (2009) Drug discovery and natural products: End of an era or an endless frontier? *Science* 325:161–165.
- Butchart SH, et al. (2010) Global biodiversity: Indicators of recent declines. *Science* 328:1164–1168.
- Vistoli G, Pedretti A, Testa B (2008) Assessing drug-likeness—what are we missing? *Drug Discov Today* 13:285–294.
- Bemis GW, Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. *J Med Chem* 39:2887–2893.
- Wang J, Hou T (2010) Drug and drug candidate building block analysis. *J Chem Inf Model* 50:55–67.
- Duarte CD, Barreiro EJ, Fraga CA (2007) Privileged structures: A useful concept for the rational design of new lead drug candidates. *Mini Rev Med Chem* 7:1108–1119.
- Koch MA, et al. (2005) Charting biologically relevant chemical space: A structural classification of natural products (SCONP). *Proc Natl Acad Sci USA* 102:17272–17277.
- Kong DX, Jiang YY, Zhang HY (2010) Marine natural products as sources of novel scaffolds: Achievement and concern. *Drug Discov Today* 15:884–886.
- Fischbach MA, Clardy J (2007) One pathway, many products. *Nat Chem Biol* 3:353–355.
- Beghyn T, Deprez-Poulain R, Willand N, Folleas B, Deprez B (2008) Natural compounds: Leads or ideas? Bioinspired molecules for drug discovery. *Chem Biol Drug Des* 72:3–15.
- Nett M, Ikeda H, Moore BS (2009) Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Nat Prod Rep* 26:1362–1384.
- Fenical W, Jensen PR (2006) Developing a new resource for drug discovery: Marine actinomycete bacteria. *Nat Chem Biol* 2:666–673.
- Tulp M, Bohlin L (2002) Functional versus chemical diversity: Is biodiversity important for drug discovery? *Trends Pharmacol Sci* 23:225–231.
- Bérdy J (2005) Bioactive microbial metabolites. *J Antibiot (Tokyo)* 58:1–26.
- Tulp M, Bohlin L (2005) Rediscovery of known natural compounds: Nuisance or goldmine? *Trends Pharmacol Sci* 26:175–177.
- Saklani A, Kutty SK (2008) Plant-derived compounds in clinical trials. *Drug Discov Today* 13:161–171.
- Baell JB, Holloway GA (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J Med Chem* 53:2719–2740.
- Ekins S, Honeycutt JD, Metz JT (2010) Evolving molecules using multi-objective optimization: Applying to ADME/Tox. *Drug Discov Today* 15:451–460.
- Wang J (2009) Comprehensive assessment of ADMET risks in drug discovery. *Curr Pharm Des* 15:2195–2219.
- Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? *Nat Rev Drug Discov* 5:993–996.
- Zhu F, et al. (2010) Update of TTD: Therapeutic Target Database. *Nucleic Acids Res* 38(Database issue):D787–D791.
- Gu L, et al. (2009) Metamorphic enzyme assembly in polyketide diversification. *Nature* 459:731–735.
- Gontang EA, Gaudêncio SP, Fenical W, Jensen PR (2010) Sequence-based analysis of secondary-metabolite biosynthesis in marine actinobacteria. *Appl Environ Microbiol* 76:2487–2499.
- Perić-Concha N, Long PF (2003) Mining the microbial metabolome: A new frontier for natural product lead discovery. *Drug Discov Today* 8:1078–1084.
- Walsh CT, Fischbach MA (2010) Natural products version 2.0: Connecting genes to molecules. *J Am Chem Soc* 132:2469–2493.
- Ososki AL, Kennelly EJ (2003) Phytoestrogens: A review of the present state of research. *Phytother Res* 17:845–869.
- Vicente F, et al. (2009) Distribution of the antifungal agents sordarins across filamentous fungi. *Mycol Res* 113:754–770.
- Simmons TL, Andrianasolo E, McPhail K, Flatt P, Gerwick WH (2005) Marine natural products as anticancer drugs. *Mol Cancer Ther* 4:333–342.
- Pfisterer PH, Wolber G, Efferth T, Rollinger JM, Stuppner H (2010) Natural products in structure-assisted design of molecular cancer therapeutics. *Curr Pharm Des* 16:1718–1741.
- Steinmetz KL, Spack EG (2009) The basics of preclinical drug development for neurodegenerative disease indications. *BMC Neurol* 9(Suppl 1):S2 1–13.
- Saladin PM, Zhang BD, Reichert JM (2009) Current trends in the clinical development of peptide therapeutics. *IDrugs* 12:779–784.
- Woodcock J, et al. (2007) The FDA's assessment of follow-on protein products: A historical perspective. *Nat Rev Drug Discov* 6:437–442.
- Nelson AL, Dhimolea E, Reichert JM (2010) Development trends for human monoclonal antibody therapeutics. *Nat Rev Drug Discov* 9:767–774.
- Melnikova I (2007) RNA-based therapies. *Nat Rev Drug Discov* 6:863–864.
- Bergmann S, et al. (2007) Genomics-driven discovery of PKS-NRPS hybrid metabolites from *Aspergillus nidulans*. *Nat Chem Biol* 3:213–217.
- Lautru S, Deeth RJ, Bailey LM, Challis GL (2005) Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. *Nat Chem Biol* 1:265–269.
- Chiang YM, Chang SL, Oakley BR, Wang CC (2011) Recent advances in awakening silent biosynthetic gene clusters and linking orphan clusters to natural products in microorganisms. *Curr Opin Chem Biol* 15:137–143.
- Wenzel SC, Müller R (2005) Formation of novel secondary metabolites by bacterial multimodular assembly lines: Deviations from textbook biosynthetic logic. *Curr Opin Chem Biol* 9:447–458.
- Wilkinson B, Micklefield J (2007) Mining and engineering natural-product biosynthetic pathways. *Nat Chem Biol* 3:379–386.
- Pullen CB, et al. (2003) Occurrence and non-detectability of maytansinoids in individual plants of the genera *Maytenus* and *Putterlickia*. *Phytochemistry* 62:377–387.
- Schroeckh V, et al. (2009) Intimate bacterial-fungal interaction triggers biosynthesis of archetypal polyketides in *Aspergillus nidulans*. *Proc Natl Acad Sci USA* 106:14558–14563.
- Baltz RH (2008) Renaissance in antibacterial discovery from actinomycetes. *Curr Opin Pharmacol* 8:557–563.
- Marris E (2006) Marine natural products: Drugs from the deep. *Nature* 443:904–905.
- Mayer AM, et al. (2010) The odyssey of marine pharmaceuticals: A current pipeline perspective. *Trends Pharmacol Sci* 31:255–265.
- Singh S (2007) From exotic spice to modern drug? *Cell* 130:765–768.
- Saleem M, et al. (2007) Marine natural products of fungal origin. *Nat Prod Rep* 24:1142–1152.
- Cheng SY, et al. (2010) Antiviral and anti-inflammatory metabolites from the soft coral *Sinularia capillosa*. *J Nat Prod* 73:771–775.
- Mortari MR, Cunha AO, Ferreira LB, dos Santos WF (2007) Neurotoxins from invertebrates as anticonvulsants: From basic research to therapeutic application. *Pharmacol Ther* 114:171–183.
- Zheng CJ, et al. (2006) Therapeutic targets: Progress of their exploration and investigation of their characteristics. *Pharmacol Rev* 58:259–279.
- Bilanges B, Torbett N, Vanhaesebroeck B (2008) Killing two kinase families with one stone. *Nat Chem Biol* 4:648–649.
- Cragg GM, Newman DJ (2005) Plants as a source of anti-cancer agents. *J Ethnopharmacol* 100:72–79.
- Harvey A (2000) Strategies for discovering drugs from previously unexplored natural products. *Drug Discov Today* 5:294–300.
- Scherlach K, Hertweck C (2009) Triggering cryptic natural product biosynthesis in microorganisms. *Org Biomol Chem* 7:1753–1760.
- Chiang YM, Lee KH, Sanchez JF, Keller NP, Wang CC (2009) Unlocking fungal cryptic natural products. *Nat Prod Commun* 4:1505–1510.
- Sayers EW, et al. (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 37(Database issue):D5–D15.
- Ciccarelli FD, et al. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311:1283–1287.