Statistical reanalysis of natural products reveals increasing chemical diversity

Michael A. Skinnider[a,b] and Nathan A. Magarvey[a,b,1]

In their retrospective analysis of natural product (NP) discovery since the 1940s, Pye et al. (1) observe a gradual decline in the proportion of NPs discovered each year with low similarity to previously known compounds [defined by maximum Tanimoto coefficient (Tc) < 0.4]. Additionally, the authors report that the median maximum Tc for all NPs discovered in a given year has plateaued since the mid-1990s. Their analysis suggests that the pace of structurally unique NP discovery is decreasing.

However, previous work by Shoichet and coworkers (2) suggests that the trends Pye et al. (1) observed might be expected to hold true for any randomly growing chemical database. Intuitively, as the number of structures in a database grows, it becomes increasingly likely that any comparison with the entire database will result in at least one pair with structural similarity. It is therefore unclear to what extent the observed trends can be attributed to declining rates of structurally unique NP discovery, as opposed to the simple increase in the number of known NPs over time.

We reproduced Pye et al.'s (1) results with our own in-house database of 32,380 NPs (3): the rate of NP discovery plateaued beginning in the mid-1990s (Fig. 1*A*), whereas the proportion of molecules with low similarity to known compounds has decreased gradually over time (Fig. 1*B*), and median maximum Tcs have plateaued since the 1980s (Fig. 1*C*). However, we also observed these same trends after random permutation of the year of compound discovery (Fig. 1 *D* and *E*). Additionally, we observed the same trends when NP structures were substituted with a random sample of compounds from the ZINC database (4), despite lower structural similarity overall (Fig. 1 *F* and *G*).

These observations suggest that the observed trends may be a feature of any growing database of chemical structures, rather than reflecting trends specific to NP discovery. A more appropriate statistical null model would compare chemical similarity between novel and known NPs to random expectation. We compared the proportion of structurally unique NPs in our in-house database to the proportion defined by randomly permuting years of compound discovery and found that, since 1990, the rate of structurally novel compound discovery has dramatically outpaced random expectation (Fig. 1*H*) (Kolmogorov–Smirnov test, $P = 6.2 \times 10^{-14}$). Over the same period, the median maximum Tc has declined relative to random expectation (Fig. 1*I*) ($P = 7.6 \times 10^{-11}$). In other words, relative to a randomly growing library of NP structures, NPs discovered within the last three decades have been characterized by unprecedented chemical diversity.

Multiple factors may underlie the increase in chemical diversity relative to random expectation since the 1980s, among them the development of methods to dereplicate previously discovered compounds, a shift toward more taxonomically diverse producing organisms, or incentives to publish novel structures rather than analogs of known compounds. New genome-guided tools for NP discovery may further expand the range of known chemotypes (5, 6). Our reanalysis suggests that the future is bright for structurally novel NP discovery.

Acknowledgments

We thank Chad Johnston for helpful discussions and Nishanth Merwin for assistance preparing the dataset.

[a]Department of Biochemistry and Biomedical Sciences, Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, ON, Canada L8S 4K1; and [b]Department of Chemistry and Chemical Biology, Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, ON, Canada L8S 4K1

Author contributions: M.A.S. and N.A.M. designed research; M.A.S. performed research; M.A.S. analyzed data; and M.A.S. and N.A.M. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. Email: magarv@mcmaster.ca.

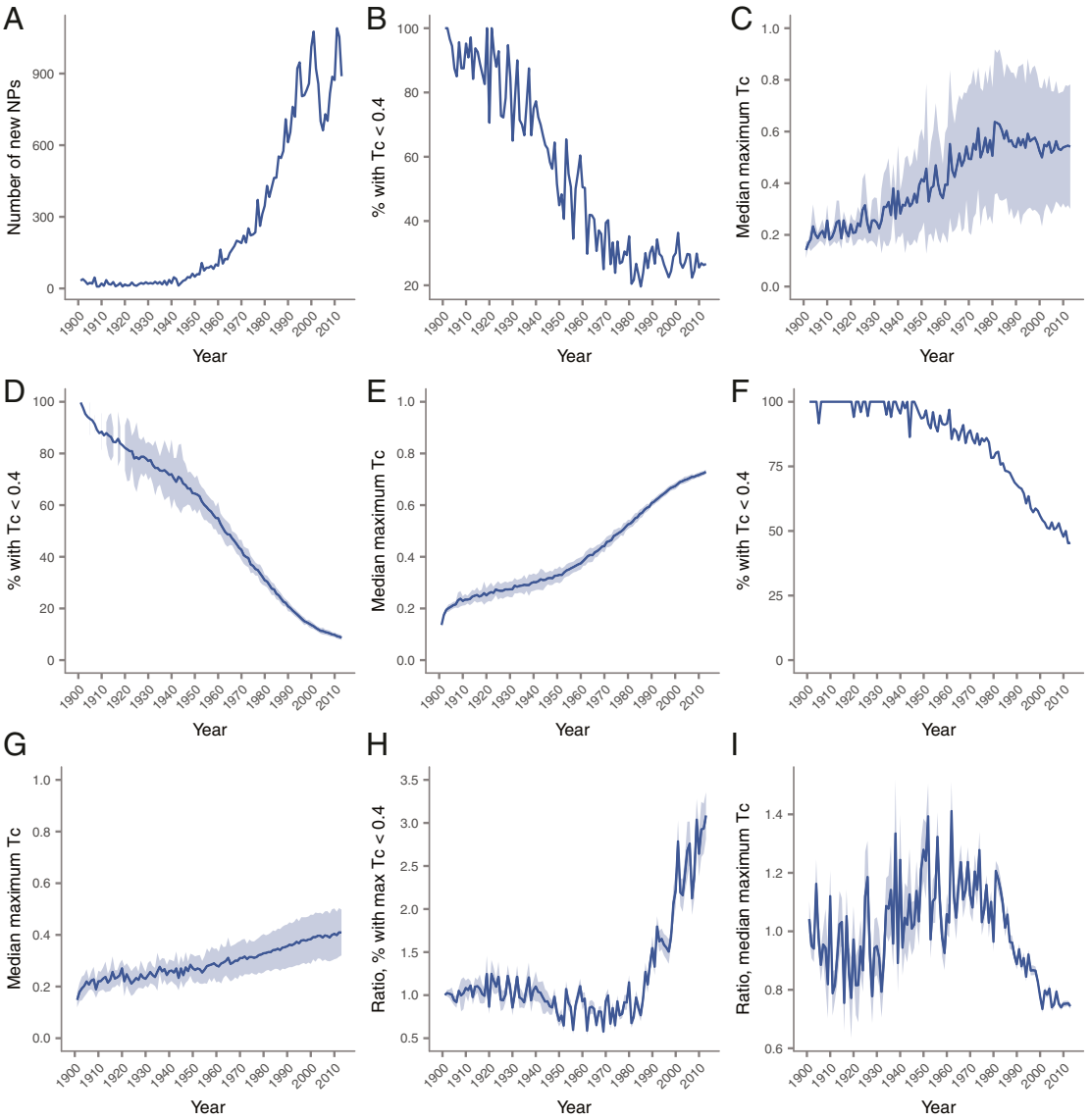

Fig. 1. Statistical reanalysis of natural product structural diversity, 1900–2013. (*A*) Number of NPs published per year in our in-house database of NP structures. (*B*) Fraction of structurally novel NPs published per year (Tc < 0.4). (*C*) Median maximum Tc between newly discovered NPs and all previously known NPs as a function of time. All shaded regions show median absolute deviation. (*D*) Same as *B*, with year of NP discovery randomly permuted. Results of 100 bootstraps are shown. (*E*) Same as *C*, with year of NP discovery randomly permuted. (*F* and *G*) Same as *B* and *C*, with NP structures replaced by a random sample of commercially available compounds from ZINC. (*H*) Ratio of structurally novel NPs published per year to random expectation. (*I*) Ratio of median maximum Tc between newly discovered NPs and all previously known NPs to random expectation as a function of time.

1 Pye CR, Bertin MJ, Lokey RS, Gerwick WH, Linington RG (2017) Retrospective analysis of natural products provides insights for future discovery trends. *Proc Natl Acad Sci USA* 114:5601–5606.

2 Keiser MJ, et al. (2007) Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* 25:197–206.

3 Dejong CA, et al. (2016) Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. *Nat Chem Biol* 12:1007–1014.

4 Irwin JJ, Shoichet BK (2005) ZINC—A free database of commercially available compounds for virtual screening. *J Chem Inf Model* 45:177–182.

5 Skinnider MA, et al. (2016) Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. *Proc Natl Acad Sci USA* 113:E6343–E6351.

6 Skinnider MA, Merwin NJ, Johnston CW, Magarvey NA (2017) PRISM 3: Expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res*, 10.1093/nar/gkx320.