# Supplementary material

## Table of Contents



## Readme

This document contains supplementary information for the publication "**An integrative genomic and chemical similarity approach linking fungal specialized metabolites and biosynthetic gene clusters**" by Steffen *et al*.

Under "**Supplementary information**", we compiled figures and tables mentioned in the main text and their legends. Exceptions are Figures S3, S4, S5 which are provided separately due to their size.

We further include information on raw and intermediate data files provided for reproduction and additional information.

Under "**Custom analysis code and in-house scripts**", we provide key analysis scripts for reuse and verification.

# Supplementary Information

## Supplementary figures

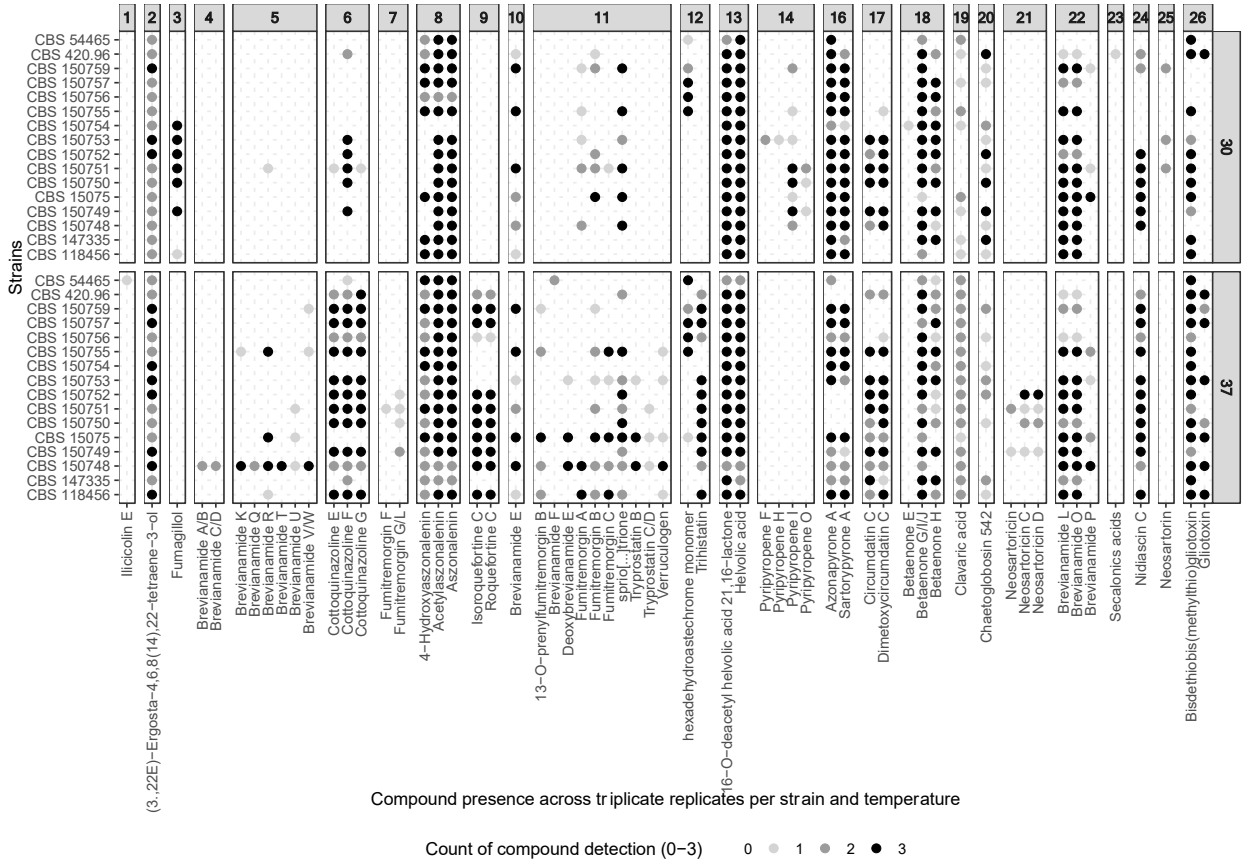


**Figure S1**: Per strain and temperature indication of SM expression/detection across the triplicate replicates. The SMs are grouped by assigned compound clusters. The compound 'spiro[5H,10H-dipyrrolo-[1,2-a:1′,2′-d]pyrazine-2-(3H),2′-[2H]-indole]-3′,5,10(1′H)trione' was abbreviated 'sprio[...]trione'.

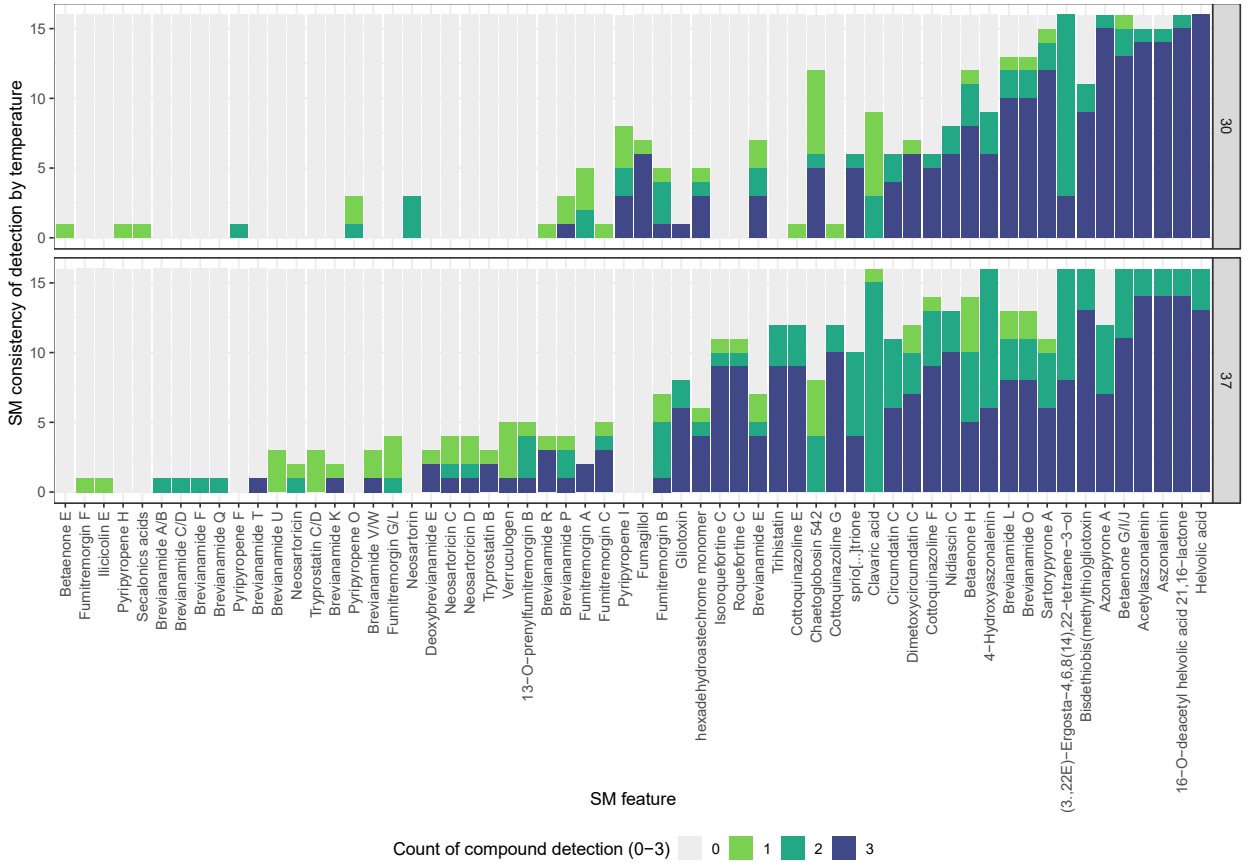


**Figure S2**: Consistency of expression/detection of SMs at 30°C and 37°C. Out of 16 strains, in how many strains was each SM detected consistently, i.e. in all replicates (3) or in none of the replicates (0), and in how many strains was the SM detected inconsistently (i.e. in one or two out of three replicates).

Integrating chemical evidence affords more nuanced insights into the phenotypes, and can be facilitated by encoding structural information via SMILES. In this study/In Rinker *et al.*, (2024), the three biological replicates for metabolomics were treated as uniformly as possible. Despite best efforts, we find that the detection of many SMs is fairly inconsistent across the replicates and strains (**Figure S1**). This adds nuance to the notion of the 'silent BGC', as it seems that in *A. fischeri*, only a handful of BGCs are active consistently and detected across all three replicates (**Figure S2**). To detect a maximum diversity of SMs, it seems that biological replicates should be considered valuable and informative together with other established methods such as co-

culturing (Knowles et al. 2022; Wang et al. 2022) or cultivation on different substrates (OSMAC 'one strain, many compounds' (Bode et al. 2002)) to elicit a diverse chemical response by the fungus.

**Figure S3**: (see Figure_S3_dendrogram_SMs.pdf) Dendrogram representing the hierarchical clustering of experimental SMs with all structures deposited in MIBiGv 3.1.

**Figure S4**: (see Figure_S4_Compound_clusters.pdf) Structures of all experimental SM features shown in the assigned compound clusters together with similar structures from known BGCs.

We refrained from closely linking transcriptomic and metabolomic data as these data were collected in two independent experiments. However, the overwhelming majority of BGC genes was expressed/transcriptionally accessed, which stands in contrast to the lack of detection in metabolomics (**Figure S1, S2**). This might be due to post-transcriptional regulation, presence of SMs at levels below the limit of detection or true variations between the two experiments. Regardless, transcriptomics highlights that the notion of BGCs being a set of co-regulated genes is highly nuanced and we note both variations in expression across the genes within a BGC and between different strains suggesting that a direct link of expression levels to product presence is unlikely (**Figure S5**).

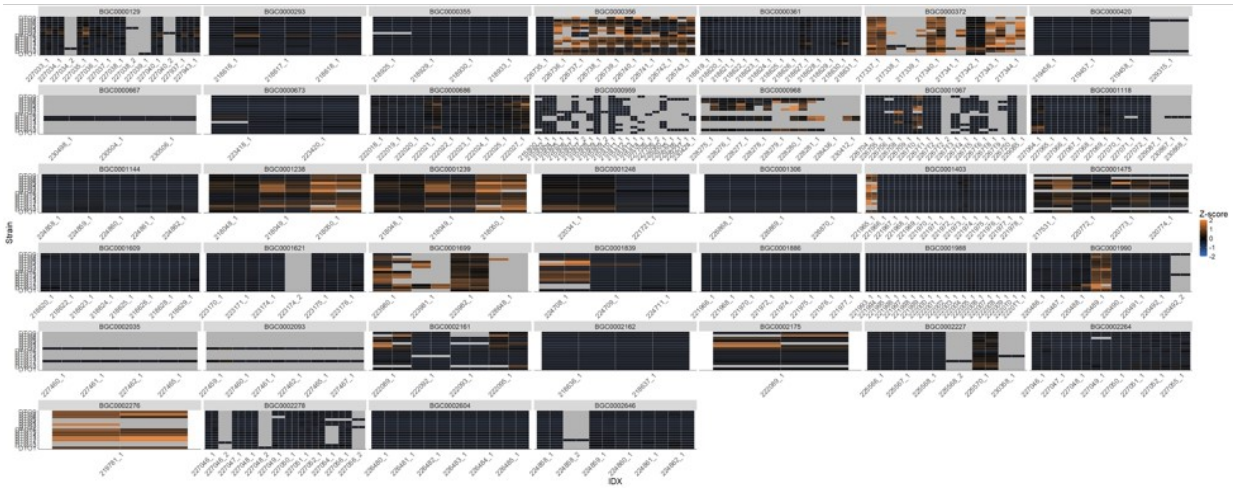

**Figure S5** (see z-score_all_BGCs.png): Normalized differential expression values across all strains for the BGCs found present or putative.

## Supplementary tables

**Table S1: Accordance of compound clusters (CC) with BGC presence.** Checking whether the hypothesized BGC is present and the linked compounds known to be produced by it or whether the link is a new hypothesis. CC number is arbitrary.

| CC | LINKED BGC(S) | BGC KNOWN/UNKNOWN | COMPOUNDS KNOWN OR ADDITIONAL ATTRIBUTIONS |
|---|---|---|---|
| **14** | BGC0000129 | known | present, predicted |
| **8** | BGC0000293 | known | present, predicted & reported |
| **6** | BGC0000355 | known | putatively present, predicted |
| **11** | BGC0000356 | known | present, predicted & reported |
| **13** | BGC0000361 | known | present, reported |
| **12** | BGC0000372 | known | present, predicted & reported |
| **9** | BGC0000420 | known | present, reported |
| **13** | BGC0000686 | known | present, predicted & reported |
| **20** [A] | BGC0000968 | known | present, predicted |
| **3** | BGC0001067 | known | present, reported |
| **7** [B] | BGC0001142, (BGC0000355) | new BGC (1) | not detected, predicted |
| **21** | BGC0001144 | known | present, predicted & reported |
| **19** | BGC0001248 | known | present, reported |
| **18** [C] | BGC0001264, BGC0002165 | new BGC (2) | not detected, predicted |
| **17** [D] | BGC0001652, BGC0000448 (BGC0000409, BGC0000303) | new BGC (3) | not detected, predicted |
| **23** | BGC0001886 | known | present, reported |
| **1** [E] | BGC0001923, BGC0001924 | new BGC (4) | not detected, predicted |
| **25** | BGC0001988 | known | present, reported |

| | | | |
|---|---|---|---|
| **22** [F] | BGC0002208, (BGC0002242) | new BGC (5) | not detected, predicted |
| **4** [G] | BGC0002254, BGC0002253, BGC0000819, BGC0000818, BGC0001084 | new BGC (6) DKP Trp-Pro | not detected, predicted |
| **24** [H] | BGC0002275, BGC0002171 | new BGC (7) | not detected, predicted |
| **16** | BGC0002604 | known | present, predicted & reported |
| **10** [I] | BGC0002617 | new BGC (8) | not detected, predicted |
| **5** | new BGC | new BGC (9) DKP Trp-Pro | not detected, predicted |
| **2** | – (sterol) | – | primary metabolism |

**A** BGC present but the presence/absence pattern of BGC and SM do not match.

**B** BGC0001142 (22 genes) not present but runs of several genes matching several BCG genes exist (eg DTO10_FUN_006044 to DTO10_FUN_006051, region not previously classified as BGC). BGC0000355 present putatively.

**C** BGC0001264 (8 genes) not present, no tentative regions. BGC0002165 (7 genes) not present but runs of several genes matching several BCG genes exist (SNOG_07866~DTO10_FUN_009035, SNOG_07864~
DTO10_FUN_009036, SNOG_07861~DTO10_FUN_009039, region not previously classified as BGC).

**D** BGC0001652 (18 genes) not present, no tentative regions. BGC0000448 (17 genes) not present but region of 5 genes matching 2 BGC core synthesis genes exist (DTO10_FUN_006047, DTO10_FUN_006049, DTO10_FUN_006050, DTO10_FUN_006051 ~ ACN39014.1 *tomA* & ACN39015.1 *tomB*, region not previously classified as BGC), albeit with low identify. BGC0000409 (34 genes) not present, no tentative regions. BGC0000303 (25 genes) not present, no tentative regions.

**E** BGC0001923 (8 genes) not present, no tentative regions. BGC0001924 (3 genes) not present, no tentative regions.

**F** BGC0002208 (6 genes) not present but several regions of interest exist (DTO10_FUN_003576, DTO10_FUN_003577, DTO10_FUN_003580, DTO10_FUN_003582, DTO10_FUN_003583 ~ HK57_00060 & HK57_00063 & HK57_00065, some genes putatively classified as BGC000355); (DTO10_FUN_006047-DTO10_FUN_006051 ~ HK57_00065, region not previously classified as BGC); (DTO10_FUN_008550 ~ HK57_00063, DTO10_FUN_008551 ~ HK57_00062, DTO10_FUN_008554 ~ HK57_00060, region not previously classified as BGC).
BGC0002242 (4 genes) not present but several regions of interest exist (DTO10_FUN_003576 ~HK57_00503, DTO10_FUN_003580~HK57_00503, DTO10_FUN_003582~HK57_00501, DTO10_FUN_003583~HK57_00503, some genes putatively classified as BGC000355); (DTO10_FUN_006047, DTO10_FUN_006048, DTO10_FUN_006050, DTO10_FUN_006051 ~HK57_00503, region not previously classified as BGC). Interestingly, there is a large overlap in the set of genes producing hits, with the two main putative BGC regions being 1 (DTO10_FUN_003576–DTO10_FUN_003583, putatively BGC0000355) and 2 (DTO10_FUN_006047-DTO10_FUN_006051).

**G** BGC0002254 (18 genes) not present but regions of interest exits (the same as for previous BGCs, matching a putative NRPS: DTO10_FUN_003576~QKG86306.1, DTO10_FUN_003580~QKG86306.1, DTO10_FUN_003582~QKG86310.1, DTO10_FUN_003583~QKG86306.1), but also (DTO10_FUN_006746~WaiD|QKG86316.1, DTO10_FUN_006747~QKG86315.1, DTO10_FUN_006748~ribosomal_protein_L31|QKG86314.1, DTO10_FUN_006749~

lysophospholipid_acyltransferase|QKG86312.1 or hypothetical_protein|QKG86313.1,
DTO10_FUN_006750~lysophospholipid_acyltransferase|QKG86312.1,
DTO10_FUN_006766~ubiquitin_carbon_terminal_hydrolase|QKG86299.1).
BGC0002253 (9 genes) not present but regions of interest exist
(DTO10_FUN_007145~QKG86291.1|laccase, DTO10_FUN_007148~QKG86292.1|enoyl_reductase,
DTO10_FUN_007150~QKG86295.1|non-red), (DTO10_FUN_008547~QKG86295.1|non-reducing_polyketide_synthase, DTO10_FUN_008551~QKG86294.1|MFS_gen,
DTO10_FUN_008553~QKG86298.1|glycosyl_transferase, DTO10_FUN_008554~QKG86290.1|O-methyltransferase). BGC0000819 (15 genes) not present but regions of interest exist (the same region as previously, but longer: 37269.1
DTO10_FUN_003576~AGA37269.1|NRPS, DTO10_FUN_003580~AGA37269.1|NRPS,
DTO10_FUN_003582~AGA37278.1|FAD_monooxygenase,
DTO10_FUN_003583~AGA37269.1|NRPS,
DTO10_FUN_003885~AGA37282.1|P450_monooxygenase,
DTO10_FUN_003887~AGA37273.1|efflux_pump, DTO10_FUN_003898~AGA37269.1|NRPS),
(another region previously observed all matching the same gene DTO10_FUN_006047–
DTO10_FUN_006051~AGA37269.1|NRPS), and
(DTO10_FUN_011150~AGA37280.1|P450_monooxygenase,
DTO10_FUN_011152~AGA37278.1|FAD_monooxygenase,
DTO10_FUN_011154~AGA37274.1|negative_regulator).
BGC0000818 (18 genes) not present but regions of interest exist
(DTO14_FUN_007678~AGC83584.1|NotM,DTO14_FUN_007679~AGC83583.1|transcription_factor,
DTO14_FUN_007680~AGC83582.1|nucleoside_transport),
(DTO14_FUN_010421~AGC83576.1|NRPS,
DTO14_FUN_010424~AGC83580.1|FAD_monooxygenase or AGC83573.1|FAD_monooxygenase,
DTO14_FUN_010425~AGC83576.1|NRPS ), and (DTO14_FUN_011439~AGC83575.1|oxidoreductase,
DTO14_FUN_011544~ prenyltransferases AGC83574.1 or AGC83577.1,
DTO14_FUN_011545~AGC83579.1|P450_monooxygenase,
DTO14_FUN_011548~AGC83577.1|prenyltransferase,
DTO14_FUN_011550~AGC83578.1|P450_monooxygenase,
DTO14_FUN_011551~AGC83576.1|NRPS,
DTO14_FUN_011656~AGC83587.1|short_chain_dehydrogenase; this region was classified as BGC0000356).
BGC0001084 (18 genes) not present but hits and regions of interest exist
(DTO15_FUN_002089~ADM34144.1|efflux_pump; high identity,
DTO15_FUN_002090~ADM34145.1|trancriptional_activator),
(DTO15_FUN_003872~ADM34137.1|oxidoreductase, DTO15_FUN_003875~ADM34138.1|non-ribosomal_peptide_synthetase, DTO15_FUN_003978~ADM34147.1|dehydrogenase),
(DTO15_FUN_005990~ADM34138.1|non-ribosomal_peptide_synthetase,
DTO15_FUN_005994~ADM34150.1, DTO15_FUN_005996~ADM34146.1,
DTO15_FUN_006001~ADM34151.1|hypothetical_protein,
DTO15_FUN_006003~ADM34149.1|metallo-beta-lactamase_domain_protein,
DTO15_FUN_006017~ADM34145.1|trancriptional_activator,
DTO15_FUN_006019~ADM34149.1|metallo-beta-lactamase_domain_protein,
DTO15_FUN_006021~ADM34150.1, DTO15_FUN_006022~ADM34146.1,
DTO15_FUN_006024~ADM34151.1,
DTO15_FUN_006027~ADM34135.1|FAD_binding_domain_protein,
DTO15_FUN_006030~ADM34144.1|efflux_pump), and (DTO15_FUN_011502~ADM34138.1|non-ribosomal_peptide_synthetase, DTO15_FUN_011503~ADM34140.1|P450,
DTO15_FUN_011508~ADM34141.1|P450, DTO15_FUN_011509~ADM34136.1|aromatic_prenyl-

transferase or ADM34139.1|aromatic_prenyl-transferase; this region was classified as BGC0000356).

**H** BGC0002275 (2 genes) not present, one gene produces hits but none of reasonable identity and coverage. BGC0002171 (2 genes) not present, one gene produces hits but none of reasonable identity and coverage.

**I** BGC0002617 (3 genes) not present. Two of the genes (AMQ36132.1|PsyA and AMQ36134.1|PsyC frequently produce hits against the same gene), there are regions of interest (DTO13_FUN_008191, DTO13_FUN_008192, DTO13_FUN_008193, DTO13_FUN_008195 ~ AMQ36132.1|PsyA or AMQ36134.1|PsyC, region not previously classified as BGC).

## Supplementary_tables.xlsx

Collection of all tables mentioned in the main publication:

- **Table 1** (as in the main publication)
- **Table S1**: extended Table 1
- **Table S2**: BGCs in *A. fischeri*
- **Table S3**: BGC_information (extensive information on all BGC genes across all strains including genomic location, blast identities to MIBiG queries, classification etc.)

## Raw and intermediate data provided separately

- BGC_grouping_aS7_PROCESSED.csv: Manually corrected results of clustering antiSMASH BGC regions by gene content.
- BGC_grouping_aS7_unprocessed.csv: Initial results of clustering antiSMASH BGC regions by gene content.
- annot.anotations_for_BGCs.extended.txt
- anti.antiSMASH7_unscaffolded.prosmash_output.tsv: results after combining antiSMASH results using in-house script prosmash.py.

- mibig_to_DTO.AA.tsv: Combined results of diamond blast searches of MIBiG v.3.1 amino acid queries to the *A. fischeri* genomes.
- pan.Fischeri_pangenome_info.csv: Pangenome information, i.e. correspondence of genes/transcripts across the 16 strains. Orthologous genes are tracked by the same IDX.
- strain_names.csv: Throughout the project and before receiving official strain designations, the different collaborating labs used internal IDs which are provided here.
- mibig_BGCs_genes.tsv: List of all genes per BGC in MIBiG v3.1
- compounds_for_clustering.csv: SMILES and names of all compounds in MIBiG v3.1 and at the bottom, the experimental SMs found in *A. fischeri*.
- compounds_strain_temperature_count.csv: count of the number of replicates (0–3) in which a given SM was detected in a given strain at a given temperature.

## Custom analysis code and in-house scripts

- Based on antiSMASH results, we ran in-house bash and python scripts to collect antiSMASH results (run_prosmash.sh, prosmash.py).
- To identifying the same antiSMASH predicted BGCs across the 16 genomes, we compared the BGC regions by gene content using antiSMASH7_BGC_clustering.R.
- For the manual BGC validations, we compiled a markdown detailing all results and decisions (Supplement_BGCs.Rmd)
- We clustered compounds using rdkit as documented in compound_clustering_A_fischeri.ipnb

# References

Bode, Helge Björn, Barbara Bethe, Regina Höfs, and Axel Zeeck. 2002. “Big Effects from Small Changes: Possible Ways to Explore Nature’s Chemical Diversity.” *ChemBioChem* 3 (7): 619. https://doi.org/10.1002/1439-7633(20020703)3:7<619::AID-CBIC619>3.0.CO;2-9.

Knowles, Sonja L., Huzefa A. Raja, Christopher D. Roberts, and Nicholas H. Oberlies. 2022. “Fungal–Fungal Co-Culture: A Primer for Generating Chemical Diversity.” *Natural Product Reports* 39 (8): 1557–73. https://doi.org/10.1039/D1NP00070E.

Rinker, David C., Thomas J. C. Sauters, Karin Steffen, Adiyantara Gumilang, Huzefa A. Raja, Manuel Rangel-Grimaldo, Camila Figueiredo Pinzan, et al. 2024. “Strain Heterogeneity in a Non-Pathogenic Aspergillus Fungus Highlights Factors Associated with Virulence.” *Communications Biology* 7 (1): 1082. https://doi.org/10.1038/s42003-024-06756-8.

Wang, Gang, Huomiao Ran, Jie Fan, Nancy P. Keller, Zhiguo Liu, Fan Wu, and Wen-Bing Yin. 2022. “Fungal-Fungal Cocultivation Leads to Widespread Secondary Metabolite Alteration Requiring the Partial Loss-of-Function VeA1 Protein.” *Science Advances* 8 (17): eabo6094. https://doi.org/10.1126/sciadv.abo6094.