Науковий вісник Ужгородського університету
Серія Біологія, випуск 55 (2023): 28–33

DOI https://doi.org/ 10.32782/1998-6475.2023.55.28-33

PGP: ПЛАТФОРМА ДЛЯ КОМПЛЕКСНОГО АНАЛІЗУ ГЕНОМНОЇ РІЗНОМАНІТНОСТІ

Валтер ВОЛФСБЕРГЕР[1], Христина ЩУБЕЛКА[1,2], Ольга Т. ОЛЕКСИК[4], Ярослава ГАСИНЕЦЬ[2], Сільвія ПАЦКУН[3], Михайло ВАКЕРИЧ[2,6], Роман КІШ[2], Віолета МІРУТЕНКО[2], Владислав МІРУТЕНКО[2], Коралія Адіна КОТОРАЧІ[5], Калін ПОП[5], Олімпія НЕАГУ[5], Корнель БАЛТЕ[5], Хільдегарда ГЕРМАН[5], Паула МАРЕ[5], Сімона ДУМІТРА[5], Гораціу ПАПІУ[5], Анка ГЕРМЕНЕАН[5], Тарас К. ОЛЕКСИК[1,2]

Популяційні геномні проєкти відіграють ключову роль у глобальних зусиллях з дослідження різноманіття геномів людських популяцій. Різноманітні бар'єри ускладнюють такі ініціативи, серед яких відсутність біоінформатичних знань та відтворюваних стандартизованих методів популяційного аналізу становлять одну з основних перешкод, що обмежують їхній потенціал. Масштабовані, автоматизовані та зручні у використанні обчислювальні конвеєри можуть допомогти дослідникам із мінімальними навичками програмування подолати ці виклики без необхідності глибокого вивчення біоінформатики. PopGenPlayground (PGP) – це оптимізований обчислювальний конвеєр, що працює за допомогою єдиної команди, розроблений для аналізу геноміки людських популяцій з використанням системи управління робочими процесами Snakemake. Створений для автоматизації вторинного аналізу даних національних геномних проєктів, він використовує загальнодоступні геномні бази даних для порівняльного аналізу та анотації варіантів. PGP є мультиплатформенним і надійним обчислювальним конвеєром для популяційного аналізу, який спрощує процес аналізу та знижує вимоги до рівня знань, необхідних для проведення початкового популяційного аналізу в рамках національного геномного проєкту. PGP забезпечує комплексний інструментарій для вторинного аналізу, який можна використовувати як на персональному комп'ютері, так і на віддалених високопродуктивних обчислювальних платформах.
Ключові слова: *обчислювальний конвеєр, біоінформатика, геноми.*

[1]*Відділ біологічних наук, Оклендський університет, Рочестер, Мічиган 48309, США;*

[2]*Біологічний факультет, Ужгородський національний університет, Ужгород, 88000, Україна; e-mail: khrystyna.shchubelka@uzhnu.edu.ua;*

[3]*Медичний факультет № 2, Ужгородський національний університет, Ужгород, 88000, Україна;*

[4]*Закарпатська обласна клінічна лікарня імені А. Новака, Ужгород, 88000, Україна;*

[5]*«Vasile Goldiş» Західний університет Арада, 94–96, Revoluţiei Bld., Арад, 310025, Румунія;*

[6]*Закарпатський науково-дослідний експертно-криміналістичний центр МВС України, Слов'янська наб., 25, Ужгород, 88000, Україна.*

PGP: A Platform for Comprehensive Analysis of Genomic Diversity. Shchubelka K.[1,2], Wolfsberger W.[1], Oleksyk O. T.[4], Hasynets Ya.[2], Patskun S.[3], Vakerych M.[2,6], Kish R.[2], Mirutenko V.[2], Mirutenko Vl.[2], Cotoraci C. A.[5], Pop C.[5], Neagu O.[5], Baltă C.[5], Herman H.,[5] Mare P.[5], Dumitra S.[5], Papiu H.[5], Hermenean A.[5], Oleksyk T.[1,2]

Population genomic projects are essential in the current drive to map the genome diversity of human populations across the globe. Various barriers persist hindering these efforts, and the lack of bioinformatic expertise and reproducible standardized population-scale analysis is one of the major challenges limiting their discovery potential. Scalable, automated, user-friendly pipelines can help researchers with minimum programming skills to tackle these issues without extensive training. PopGenPlayground (PGP), is a streamlined, single-command computation pipeline designed for human population genomics analysis based on Snakemake workflow management system. Developed to automate secondary analysis of a previously published national genome project, it leverages the publicly available genomic databases for comparative analysis and annotation of variant calls. PGP presents a multi-platform robust population analysis pipeline, that reduces the time and the expertise levels to perform the main core of population

analysis for a national genome project. PGP provides a comprehensive secondary analysis tool and can be used to perform analysis on a personal computer or using a remote high-performance computing platform.
Key words: *computational pipeline, bioinformatics, genomes.*
[1]*Department of Biological Sciences, Oakland University, Rochester, MI 48309, USA.*
[2]*Faculty of Biology, Uzhhorod National University, Uzhhorod, 88000, Ukraine; e-mail: khrystyna.shchubelka@uzhnu.edu.ua;*
[3]*Faculty of Medicine #2, Uzhhorod National University, Uzhhorod, 88000, Ukraine;*
[4]*A. Novak Transcarpathian Regional Clinical Hospital, Uzhhorod, 88000, Ukraine;*
[5]*"Vasile Goldiş" Western University of Arad, 94–96, Revoluţiei Bld., Arad 310025, Romania;*
[6]*Transcarpathian scientific research expert and forensic center of the Ministry of Internal Affairs of Ukraine, 25, Slovianska nab., Uzhhorod, 88000, Ukraine.*

Вступ

Стрімкий розвиток популяційної геноміки значною мірою можна пояснити розвитком технологій секвенування й аналізу даних та експоненціальним збільшенням обсягу даних, що генеруються в цій галузі (McGuire et al. 2020). За своєю суттю (тобто аналіз великих масивів даних із численних зразків) вона тісно пов'язана з розвитком біоінформатичних підходів та впровадженням сучасних обчислювальних методів аналізу даних (Bartlett et al. 2017). Використання новітнього програмного забезпечення та алгоритмів біоінформатики дозволяє дослідникам ефективно аналізувати потоки необроблених даних послідовностей, ідентифікувати варіанти, прогнозувати їх функціональні наслідки та досліджувати розподіл генетичних варіацій як у середині, так і між популяціями, що представляють інтерес. Проєкти з геноміки людських популяцій є важливим етапом на шляху до розроблення інформаційних ресурсів та підходів у сфері персоналізованої медицини та передових медичних технологій.

Біоінформатика спростила інтеграцію різноманітних типів даних у дослідженнях популяційної геноміки та відкрила шлях для пошуку відповідей на питання, що охоплюють кілька дисциплін одночасно. Наприклад, сучасні підходи інтегрують геномні дані з екологічними, фенотиповими та географічними даними для аналізу впливу різних факторів на генетичні варіації як у межах, так і між популяціями (Van Assche et al. 2015). Цей інтегративний підхід значно поглибив наше розуміння популяційної геноміки. Однак щораз більша складність сучасних досліджень, поєднана з відносною новизною та відсутністю стандартизованих методик аналізу, створює значні перешкоди для дослідницьких груп, які розпочинають свої проєкти в галузі популяційної геноміки. Біоінформатика відкрила доступ до цих можливостей, але це не означає, що дані можливості є доступними більшості науковців.

Типовий біоінформатичний аналіз будь-якого геномного дослідження включає перелік інструментів та проміжних етапів маніпуляції даними, що диктує вимоги, які постійно зростають, до рівня підготовки фахівців та змушує їх спеціалізуватися у вузьких нішах (Bartlett et al. 2017). Одним із вирішень цієї проблеми є інтеграція кількох кроків та повторюваних процедур аналізу в єдиний робочий процес чи конвеєр. Оптимізація, стандартизація та масштабування біоінформатичних конвеєрів у поєднанні з обміном знаннями дозволяють значно покращити доступність біоінформатики. Такі ініціативи мають потенціал демократизувати дослідження в галузі геноміки, дозволяючи більшій кількості дослідників брати участь в аналізі повногеномних даних та спрощують процес масштабування аналізу для великих даних.

Призначення

PopGenPlayground (PGP) створений для спрощення вторинного аналізу геноміки популяцій на основі файлів виклику варіантів (VCF). Пакет PGP розміщено на GitHub (Wolfsberger 2023), і він інтегрує широкий спектр методів популяційної геноміки (Табл. 1) в єдиний, зручний для користувача робочий процес. Цей конвеєр автоматизує ключові процедури в аналізі популяційної геноміки включно з обробленням даних, візуалізацією, конвертацією форматів і фазуванням, зменшуючи потребу в ручній роботі та підвищуючи ефективність. PGP використовує систему управління робочими процесами *Snakemake* для забезпечення ізольованих обчислювальних середовищ на кожному етапі аналізу, мінімізуючи вплив зовнішніх факторів на результат аналізу (Köster et al. 2021). Конвеєр підтримує масштабування для роботи у високопродуктивних обчислювальних системах, що робить його придатним для аналізу великих даних. PGP має просту конфігурацію і не вимагає від користувачів глибоких знань з біоінформатики для його запуску.

Застосування

PGP було розроблено з використанням системи управління робочими процесами *Snakemake* та мови опису правил, що базується на *Python* (Köster et al. 2021). Конвеєр інтегрує різноманітні біоінформатичні інструменти, які детально описано в Таблиці 1, для проведення аналізу даних та виконує проміжкові кроки з трансформації даних за допомогою команд Unix shell, мови програмування Python з використанням бібліотеки *Pandas* і модуля виконання точного статистичного тесту Фішера (FET). Конвеєр використовує загальнодоступні геномні бази даних. Порівняльний аналіз конвеєра використовує бази даних проєкту *International Genome Sample Resource* (IGSR) (Fairley et al. 2020), анотація – базу *Ensembl VEP* (McLaren et al. 2016), а дані *NCBI ClinVar* (Landrum et al. 2016) інтегровані для анотації клінічних медичних варіантів.

Етапи аналізу PGP визначені як правила, які система виконує для створення необхідних вихідних файлів. Залежності між правилами встановлюються автоматично на основі заданих вхідних та вихідних файлів. Інтеграція системи управління пакетами Conda (Anaconda Software Distribution 2020) дозволяє обробляти програмні залежності кожного етапу робочого процесу. PGP-конвеєр застосовує «лінивий» підхід до виконання правил, розв'язуючи їх у зворотному порядку, починаючи з кінцевого бажаного вихідного файлу та перевіряючи папку з даними на наявність результатів попередніх виконаних етапів. Це дозволяє відновлювати аналіз після переривання, без необхідності повторного виконання успішно завершених етапів. Автономні звіти забезпечують прозорість результатів та моніторинг процесу виконання етапів, параметрів, коду та програмного забезпечення.

У своїй конфігурації конвеєр PGP має мінімалістичні вимоги. Після первинної інсталяції *Snakemake* та його залежностей (Köster et al. 2021) користувач завантажує геномний файл з даними у форматі VCF. Вхідні дані включають усі генетичні варіанти досліджуваної популяції та тризначні коди популяцій з проєкту IGSR (Fairley et al. 2020). Якщо немає потреби виконувати кожен етап аналізу, в конфігураційному файлі можна

Таблиця 1. Аналізи, включені в конвеєр PopGenPlayground

Table 1. Analyzes included in the PopGenPlayground pipeline

Програмне забезпечення\ Інструмент	**Клас**	**Мета**	**Посилання**
BCFtools	Оброблення даних, VCF-аналіз	Надає можливість об'єднання та перетину виклику файлів варіантів, виводить статистику для порівняльних досліджень	(Danecek et al. 2021)
Whole Genome Sequencing Variant Call Files IGSR database	Порівняльний аналіз	Надає можливість проводити порівняльний популяційний аналіз певними інструментами	(Fairley et al. 2020)
PLINK2	Ідентифікація близькості за походженням, робота з даними, тест PCA	Оцінювання інбридингу та спорідненості, структури популяції, обрізка геномних зчеплень, конвертація даних	(Chang et al. 2015)
ADMIXTURE	Кластеризація і характеристика змішування	Групування індивідуумів у кластери з максимізацією рівноваги HW та LD між локусами	(Alexander et al. 2009)
detectRUNS (R package)	Прогони гомозиготності, Інбридинг на основі прогонів гомозиготності	Оцінювання інбридингу і спорідненості, виявлення сигналів добору	(Chang et al. 2015; Marras et al. 2015)
Shapeit4	Фазування викликів варіантів	Виявлення гаплотипів, імпутація даних	(Delaneau et al. 2008)
Ensembl VEP	Анотація варіантів	Поєднує кілька підходів до анотації для оцінювання впливу варіацій на фенотип	(McLaren et al. 2016)
Python programming language and Pandas data analysis library	Інструменти аналізу даних	Генерація звітів, візуалізація та оброблення даних	-

встановити бінарну змінну для активації або деактивації визначених аналітичних етапів. Процеси в конвеєрі виконуються відповідно до порядку команд, наведеного на рисунку 1.

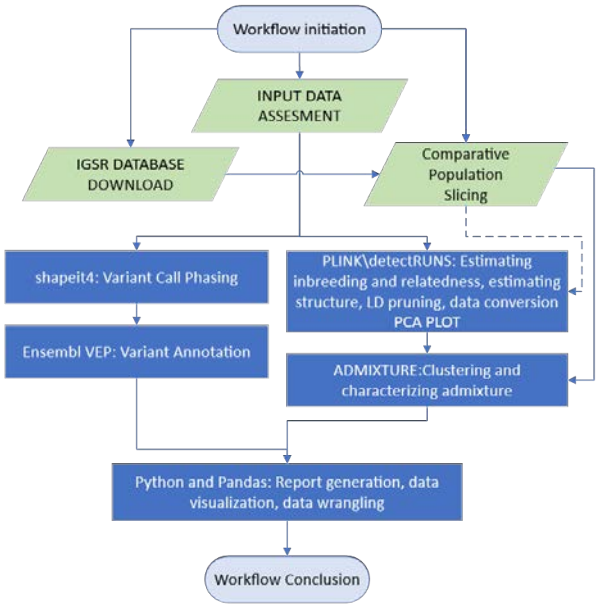

Рис. 1. Схема обчислювального конвеєра PopGenPlayground (PGP)

Fig. 1. Schematic of the PopGenPlayground (PGP) computational pipeline

Результати

Розроблений для відтворення результатів наявного аналізу в рамках національного проєкту з геноміки, обчислювальний конвеєр PGP здатен відтворювати повний аналіз популяційної геноміки, проведений у попередньому дослідженні населення України (Oleksyk et al. 2021). Для відтворення результатів за допомогою конвеєра потрібен мінімальний досвід та базове розуміння командного рядка, аналогічного до Unix shell. PGP використовує публічні наукові бази даних і генерує результати, які надають цінну інформацію для популяційних проєктів.

PGP надає детальний огляд варіацій популяційних послідовностей цілого геному, даючи уявлення про загальну кількість послідовностей, середнє покриття та варіації, такі як однонуклеотидні поліморфізми (SNP) біалельні, мультиалельні варіанти, малі індели, делеції, вставки та структурні варіанти. Додаткова таблиця детально описує зведену анотацію різних геномних елементів, включаючи кількість алелів та їх розподіл у різних геномних локаціях, як-от екзони, інтрони та міжгенні області. Нарешті, він надає дані для порівняльного попарного аналізу популяційних досліджень та всіх популяцій, виділених з IGSR (Fairley et al. 2020), виділяючи варіанти, які статистично відрізняються за частотою згідно з точним тестом Фішера.

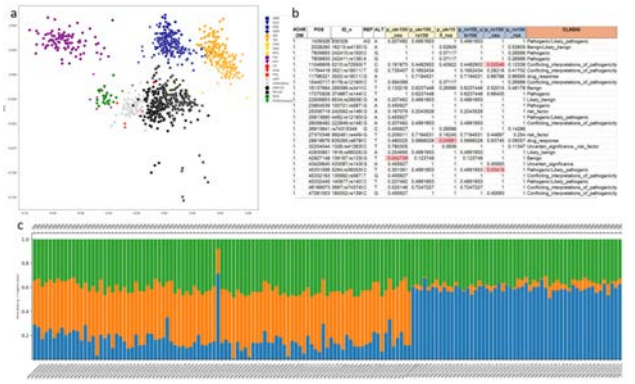

Рис. 2. Приклад ілюстрації, створеної за допомогою обчислювального конвеєра PopGenPlayground (PGP): a) графік принципових компонентів PCA геномної різноманітності досліджуваних популяцій (Oleksyk et al. 2021) у поєднанні з популяціями з бази даних IGSR (Fairley et al. 2020); b) виклики варіантів з анотацією та попарною інформацією про точний тест Фішера на частоти алелів; c) фрагмент графіка структури досліджуваної популяції (Oleksyk et al. 2021) та IGSR-популяції (Fairley et al. 2020)

Fig. 2. An example of an illustration created using the PopGenPlayground (PGP) computational pipeline: a) graph of the principal components of PCA genomic diversity of the studied populations (Oleksyk et al. 2021) in combination with populations from the IGSR database (Fairley et al. 2020); b) variant calls with annotation and pairwise information about Fisher's exact test on allele frequencies; c) a fragment of the graph of the structure of the studied population (Oleksyk et al. 2021) and the IGSR population (Fairley et al. 2020)

На додаток, PGP створює різноманітні графіки та пов'язані з ними набори даних, використовуючи мову програмування *Python* та бібліотеки для аналізу даних. Результати кластерного аналізу узагальнено на графіках PCA (Рис. 2a) та ADMIXTURE (Рис. 2c), що ілюструють взаємозв'язки між зразками всередині генеральної сукупності та їх порівняльний аналіз з популяціями, виділеними з IGSR. Варіанти з анотаціями та попарною інформацією про точний тест Фішера для частот алелів можна завантажити у форматі таблиці. Приклад виводу графіків, створених за допомогою PGP, показано на рисунку 2.

Висновки, подальші напрямки

Щоб задовольнити щораз більший попит та підвищити доступність сучасних методів

аналізу популяційної геноміки, ми розробили біоінформатичний обчислювальний конвеєр *PopGenPlayground* (PGP). PGP забезпечує зручний та ефективний доступ до аналізу повногеномних даних для популяційної геноміки. За допомогою системи управління робочими процесами *Snakemake* PGP ефективно інтегрує різноманітні типи даних та використовує кілька біоінформатичних інструментів, оптимізуючи процес аналізу. Структура конвеєра базується на попередньому досвіді аналізу геноміки популяцій в опублікованому національному проєкті з геноміки (Oleksyk et al. 2021) та інтегрує публічні геномні бази даних для порівняльного аналізу та анотації варіантів. Розміщений на GitHub (Wolfsberger 2023), PGP сприяє співпраці та подальшому розвитку конвеєра. Сфера популяційної геноміки постійно розвивається, включаючи нові інструменти та важливі набори даних. Можливість інтеграції нових інструментів в аналіз забезпечує потенціал щодо розвитку PGP, розширюючи його аналітичні можливості завдяки включенню новіших наборів даних та інструментів.

Доступність програмного забезпечення

Повний конвеєр та інструкції з використання конвеєра PopGenPlayground (PGP) доступні на GitHub (Wolfsberger 2023).

Фінансування

Фінансування проєкту було надано проєктом 2SOFT/1.2/48 «Партнерство для геномних досліджень в Україні та Румунії» Спільної операційної програми Румунія-Україна через Європейський інструмент сусідства (ENI).

Подяки

Цей конвеєр є частиною інфраструктури біоінформатики, що розвивається в Україні. Дякуємо всім учасникам BioinformaticsForUkraine.com та Консорціуму «Геномне різноманіття в Україні», які працювали з нами над розробленням інструментів для цього проєкту.

ALEXANDER, D.H., NOVEMBRE, J., LANGE, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9), 1655–1664. DOI: 10.1101/GR.094052.109

ANACONDA SOFTWARE DISTRIBUTION (2020) *In: Anaconda Documentation*. Anaconda Inc. Available at: https://docs.anaconda.com/

BARTLETT, A., PENDERS, B., LEWIS, J. (2017) Bioinformatics: Indispensable, yet hidden in plain sight? *BMC Bioinformatics*, 18(1), 1–4. DOI: 10.1186/S12859-017-1730-9/METRICS

CHANG, C.C., CHOW, C.C., TELLIER, L.C.A.M., VATTIKUTI, S., PURCELL, S.M., LEE, J.J. (2015) Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), 7. DOI: 10.1186/s13742-015-0047-8

DANECEK, P., BONFIELD, J.K., LIDDLE, J., MARSHALL, J., OHAN, V., POLLARD, M.O., WHITWHAM, A., KEANE, T., MCCARTHY, S.A., DAVIES, R.M. (2021) Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2). DOI: 10.1093/GIGASCIENCE/GIAB008

DELANEAU, O., COULONGES, C., ZAGURY, J.F. (2008) Shape-IT: New rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics*, 9, 1–14. DOI: 10.1186/1471-2105-9-540

FAIRLEY, S., LOWY-GALLEGO, E., PERRY, E., FLICEK, P. (2020) The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Research*, 48(D1), D941–D947. DOI: 10.1093/NAR/GKZ836

KÖSTER, J., MÖLDER, F., JABLONSKI, K.P., LETCHER, B., HALL, M.B., TOMKINSTINCH, C.H., SOCHAT, V., FORSTER, J., LEE, S., TWARDZIOK, S.O., KANITZ, A., WILM, A., HOLTGREWE, M., RAHMANN, S., NAHNSEN, S. (2021) Sustainable data analysis with Snakemake. *F1000Research*, 10, 33. DOI: 10.12688/f1000research.29032.2

LANDRUM, M.J., LEE, J.M., BENSON, M., BROWN, G., CHAO, C., CHITIPIRALLA, S., GU, B., HART, J., HOFFMAN, D., HOOVER, J., JANG, W., KATZ, K., OVETSKY, M., RILEY, G., SETHI, A., TULLY, R., VILLAMARIN-SALOMON, R., RUBINSTEIN, W., MAGLOTT, D. R. (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1), D862–D868. DOI: 10.1093/nar/gkv1222

MARRAS, G., GASPA, G., SORBOLINI, S., DIMAURO, C., AJMONE-MARSAN, P., VALENTINI, A., WILLIAMS, J.L., MACCIOTTA, N.P.P. (2015) Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy. *Animal Genetics*, 46(2), 110–121. DOI: 10.1111/AGE.12259

MCGUIRE, A.L., GABRIEL, S., TISHKOFF, S.A., WONKAM, A., CHAKRAVARTI, A., FURLONG, E.E.M., TREUTLEIN, B., MEISSNER, A., CHANG, H., LÓPEZ-BIGAS, N., SEGAL, E., KIM, J.-S. (2020) The road ahead in genetics and genomics. *Nature Reviews Genetics*, 21(10), 581–596. DOI: 10.1038/s41576-020-0272-6

MCLAREN, W., GIL, L., HUNT, S.E., RIAT, H.S., RITCHIE, G.R.S., THORMANN, A., FLICEK, P., CUNNINGHAM, F. (2016) The Ensemble Variant Effect Predictor. *Genome Biology,* 17(122). DOI: 10.1186/S13059-016-0974-4

OLEKSYK, T.K., WOLFSBERGER, W.W., WEBER, A.M., SHCHUBELKA, K., OLEKSYK, O.T., LEVCHUK, O., PATRUS, A., LAZAR, N., CASTRO-MARQUEZ, S.O., HASYNETS, Y., BOLDYZHAR, P., NEYMET, M., URBANOVYCH, A., STAKHOVSKA, V., MALYAR, K., CHERVYAKOVA, S., PODOROHA, O., KOVALCHUK, N., RODRIGUEZ-FLORES, J.L., ZHOU, W., MEDLEY, S., BATTISTUZZI, F., LIU, R., HOU, Y., CHEN, S., YANG, H., YEAGER, M., DEAN, M., MILLS, R.E., SMOLANKA, V. (2021) Genome diversity in Ukraine. *GigaScience*, 10(1), 1–14. DOI: 10.1093/GIGASCIENCE/GIAA159

VAN ASSCHE, R., BROECKX, V., BOONEN, K., MAES, E., DE HAES, W., SCHOOFS, L., TEMMERMAN, L. (2015) Integrating – Omics: Systems Biology as Explored Through *C. elegans* Research. *Journal of Molecular Biology*, 427(21), 3441–3451. DOI: 10.1016/J.JMB.2015.03.015

WOLFSBERGER, W.W. (2023) *PopGenPlayground* (0.1). Available at: https://github.com/wwolfsberger/OU_popgen_playground