DR. PHILINE G. D. FEULNER (Orcid ID : 0000-0002-8078-1788)

Article type : Opinion

Sequencing platform shifts provide opportunities but pose challenges for combining genomic datasets

Running title

Challenges for combining genomic datasets

Authors

De-Kayne, Rishi[1,2†], Frei, David[1,2,†], Greenway, Ryan[1], Mendes, Sofia L.[3], Retel Cas[1], Feulner, Philine G. D.[1,2,*]

[1]Department of Fish Ecology and Evolution, Centre of Ecology, Evolution and Biogeochemistry, EAWAG Swiss Federal Institute of Aquatic Science and Technology, Seestrasse 79, 6047 Kastanienbaum, Switzerland

[2]Division of Aquatic Ecology and Evolution, Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland

[3]Centre for Ecology, Evolution and Environmental Changes, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal

† These authors contributed equally

* Correspondence: philine.feulner@eawag.ch

Abstract

Technological advances in DNA sequencing over the last decade now permit the production and curation of large genomic datasets in an increasing number of non-model species. Additionally, this new data provides the opportunity for combining datasets, resulting in larger studies with a broader taxonomic range. Whilst the development of new sequencing platforms has been beneficial, resulting in a higher throughput of data at a lower per-base cost, shifts in sequencing technology can also pose challenges for those wishing to combine new sequencing data with data sequenced on older platforms. Here, we outline the types of studies where the use of curated data might be beneficial, and highlight potential biases that might be introduced by combining data from different sequencing platforms. As an example of the challenges associated with combining data across sequencing platforms, we focus on the impact of the shift in Illumina's base calling technology from a four-channel to a two-channel system. We caution that when data is combined from these two systems, erroneous guanine base calls that result from the two-channel chemistry can make their way through a bioinformatic pipeline, eventually leading to inaccurate and potentially misleading conclusions. We also suggest solutions for dealing with such potential artifacts, which make samples sequenced on different sequencing platforms appear more differentiated from one another than they really are. Finally, we stress the importance of archiving tissue samples and the associated sequences for the continued reproducibility and reusability of sequencing data in the face of ever-changing sequencing platform technology.

Keywords

NGS, reproducibility, reusability, poly-G, NovaSeq, HiSeq

Opportunities: Combining and extending datasets across time and space

DNA sequencing data reflecting the diversity of life is accumulating, as technological developments continue to increase the basepair yield of sequencing runs, whilst lowering the per-basepair prices. This data continues to facilitate comparative studies of genome structure for more and more organisms, spanning the tree of life (Baker et al., 2020; Cheng et al., 2018; Leebens-Mack et al., 2019; Morris et al., 2018; Peter et al., 2018; Shen et al., 2018; Shi et al., 2018; Zhang et al., 2014). Further, the field of molecular ecology is flourishing, with more and more studies investigating the genetic variation within and among closely related groups of organisms (Brawand et al., 2014; Lamichhaney et al., 2015; Tollis et al., 2018). However, for molecular ecologists working on non-model species, budgets still limit the amount of sequence data that can be produced. As a result, exhaustive experimental designs, which include the sampling of many individuals from many different populations, are rare (but are emerging; (Feulner et al., 2015; Greenway et al., 2020; Martin et al., 2016; Soria-Carrasco et al., 2014; Stankowski et al., 2019; Vijay et al., 2016)). The effort to publicly archive sequence data that has already contributed to publications helps to maintain the reproducibility of sequencing studies, whilst prolonging the value of such sequence data in perpetuity. Additionally, this practice of sequence data storage provides the opportunity to expand datasets beyond those that one laboratory is capable of producing (in terms of time, labour, and finances) to increase the impact of studies despite a potentially limited budget. Repositories like the Short Read Archive (SRA) -- part of the International Nucleotide Sequence Database Collaboration (INSDC) that includes the NCBI Sequence Read Archive (SRA), the European Bioinformatics Institute (EBI), and the DNA Database of Japan (DDBJ) -- are essential for both the reproducibility of genetic and genomic studies, and the reusability of sequencing data. Although combining datasets is challenging for many sequencing approaches, particularly those that sequenced anonymous reduced representations of the genome (i.e. microsatellites, amplified fragment length polymorphisms, and maybe even restriction site associated DNA sequencing and genotyping by sequencing; but see Leigh, Lischer, Grossen, & Keller (2018) for an example), the increasingly common approach of re-sequencing whole-genomes (even for a broader range of non-model organisms) makes the possibility of combining datasets more inviting.

Between the continued growth of sequencing data repositories and the continued ability to sequence more DNA quicker and cheaper the following types of studies are increasingly carried out:

(1) Broad macroevolutionary studies. Typically, such macroevolutionary studies benefit from a wide taxon sampling and few individuals suffice, making the combination of samples from different published datasets particularly useful. Often these analyses are restricted to more conserved regions of the genome. For example, Zhang et al. (2020) compiled a comprehensive dataset of 365 species of asterids representing all 17 orders containing published and newly sequenced whole genomes and transcriptomes to resolve the deep asterid phylogeny. In another example, Greenway et al. (2020) focus on the Poeciliidae family of fish, to demonstrate that adaptation to extreme, here sulfide-rich, environments has evolved convergently in ten independent lineages, by combining already published and newly sequenced transcriptome sequences.

(2) Microevolutionary studies investigating spatial variation across populations or closely related taxa. Such studies typically focus on one study system but rely on a larger sampling to reflect the variation within species or populations. These studies may benefit from combining newly sequenced material with archived sequence data from previous projects to produce larger within-system datasets. By taking advantage of existing sequence data, these combined datasets facilitate analyses of genomic differentiation across a much broader geographic sampling or among more individuals than would be otherwise possible. Here, the curated data is used to evaluate patterns in comparable populations to widen the perspective, i.e. to show whether a pattern is general or specific to the population under investigation. For example, Ravinet, Kume, Ishikawa, & Kitano (2020) evaluated if patterns of divergence and introgression between Japan Sea and Pacific Ocean stickleback resemble patterns at other locations where these species co-occur. In a comprehensive study conducted by Samuk et al. (2017), the authors compiled multiple genotyping by sequencing and whole genome sequencing datasets to a global evaluation of 1300 stickleback individuals across 51 populations, to show that putative

adaptive alleles tend to occur more often in regions of low recombination. Bergland, Behrman, O'Brien, Schmidt, & Petrov (2014) used curated data to check haplotypes under seasonal selection in *Drosophila melanogaster* for between-species divergence with a sister species (*D. simulans*). Most recently, Jones, Mills, Jensen, & Good (2020) combined new and published whole-genome and exome sequences with targeted genotyping of *Agouti*, a pigmentation gene introgressed from black-tailed jackrabbits, to investigate the evolutionary history of local seasonal camouflage adaptation in Snowshoe hares from the Pacific Northwest.

(3) Studies investigating temporal variation within and between population and species. Such studies involve combining datasets across time scales and often contain sequencing data that originated from a variety of sample types including museum collections, long-term preserved fossils or hard tissues, and contemporary fresh samples. For example, the use of museum specimens facilitated the investigation of independent temporal genomic contrasts spanning a century of climate change for two co-distributed chipmunk species (Bi et al., 2019) and a paleogenomics approach investigated the temporal component of adaptation to freshwater in sticklebacks by sequencing the genomes of 11-13,000-year-old bones and comparing them with 30 modern stickleback genomes (Kirch, Romundset, Gilbert, Jones, & Foote, 2020). Experimental approaches combining previous sequencing efforts with new samples are also commonly used to increase our understanding of temporal variation. Tenaillon et al. (2016) compiled sequence data from several other publications in addition to new sequences to strengthen their conclusions on the tempo and mode of *E. coli* genome evolution. Bottery, Wood, & Brockhurst (2019), after having shown that tetracycline resistance requires multiple mutations, used curated data to investigate if the mutation establishment order was repeatable. This by no means exhaustive selection of examples highlights that the growing amount of sequence data provides the opportunity for endless combinations of datasets to be analysed to address a multitude of questions.

Challenges: Biases change with technological developments

One technological advance which sped up the Illumina workflow and made it more cost-effective was a change from four-channel chemistry, where each of the four DNA bases is detected by a different fluorescent dye, to a two-channel chemistry, that uses only two different fluorescent dyes (Illumina). In these two-channel workflows, as implemented in the NextSeq and NovaSeq platforms, a guanine base (G) is called in the absence of fluorescence (Figure 1). Hence, it is difficult to differentiate between no signal and a G, resulting in an overrepresentation of poly-G strings in sequence data from both NextSeq and NovaSeq (Chen, Zhou, Chen, & Gu, 2018).

To most accurately capture biological variation in a given sample or population, it is important to differentiate between potentially erroneous and correct base calls, which is often done using base quality scores. However, erroneous poly-G base calls produced on the NextSeq and NovaSeq platforms can be difficult to detect, because, as a result of the two-colour chemistry, they are not always associated with reduced base qualities. Unfortunately, read trimming software packages that were written for the older four-colour systems do not flag or trim poly-G tails. Although one might think that mapping should remove the effect of these overrepresented Gs without the need for read trimming, it has been shown that some may still trickle through a bioinformatics pipeline and influence variant calling steps. A comprehensive empirical study making use of cancer cell lines to benchmark systematic differences between technologies revealed that NovaSeq instruments produced more stretches of Gs than HiSeqX in both paired-end reads (Arora et al., 2019). Arora et al. (2019) further confirmed that the bias remained detectable in the mapped reads and resulted in a relatively large number of T > G mutations among the variants unique to the NovaSeq instrument. To reduce the potential down-stream impact of these poly-G strings, newer trimming software packages such as fastp (Chen et al., 2018) check the source of the data and implement poly-G trimming by default for the two-colour systems. This not only improves the computational efficiency of sequence alignment, but should also reduce the impact of erroneous variant calling on these bases.

The impact of these changes in base calling and the subsequent erroneous G calls on the biological interpretation may vary with the chosen experimental design and other sources of variation such as for example DNA quality. Although the biases resulting from not trimming off or filtering out poly-G strings might be mild or irrelevant when analysing data produced from high quality input DNA from a single system, this may not be true when data from different technologies are combined across various biological units (e.g.

across populations, species, treatments, or time points). On top of variation in the quality of input DNA, a range of variation in sequencing approaches exists, along with differences in library preparation, including variation in read length or whether reads are single-end or paired-end. Where different individuals within a single dataset have been sequenced with variation in these methodological factors biases may also be exacerbated, potentially producing misleading results. Variation in length of sequences reads across a dataset for example has been shown to lead to pronounced allele frequency differences between populations and subsequently suggested false biological trends (Leight et al. 2018). Metagenomic work suggested that both library preparation and sequencing platform had systematic effects on the microbial community description (Poulsen, Pamp, Ekstrøm, & Aarestrup, 2019; Sato et al., 2019). In summary, attention should be paid to DNA quality, library preparation protocols, and the sequencing platform used when analysing and interpreting publicly available genomic data.

Although the prospect of combining datasets to improve our power to detect patterns is alluring, it is important to consider the ways in which these data may result in misleading conclusions. Combining datasets often means combining data from different sequencing platforms, as DNA sequencing technology continues to develop through time. Unfortunately, some of the developments (e.g. the change from four-channel to two-channel chemistry in Illumina sequencing machines) have changed the way in which uncertainties in base calling are presented in the sequencer's output files. If managed incorrectly, these changes hamper our ability to combine datasets obtained with different sequencing technologies, and the subsequent genotyping and analysis of these combined datasets may be biased (in the worst cases leading to erroneous conclusions). The most straightforward way to prevent this is a well-thought out experimental design, a step which can often be overlooked in a time where sequencing data is being produced so rapidly (see Mason (2017) for sound advice on experimental design). As has been shown for sequencing reduced-representation libraries, it is crucial for any type of sequencing experiment to carefully consider types of errors that may be introduced during laboratory work and data processing, and how to minimize, detect and remove these errors (O'Leary, Puritz, Willis, Hollenbeck, & Portnoy 2018). However, it may be difficult to achieve the ideal or optimal study design when an investigation integrates new information with already existing data (e.g. with individuals and treatments randomised across sequencing batches). Despite this limitation there are a number of approaches that can help to rectify some of these imbalances and allow the combination of multiple genomic datasets whilst minimising the impact of cross-platform biases.

Ways forward: Suggestions on how to minimise technological bias when integrating datasets

Despite the ease with which new datasets can be produced it is critical that researchers do not forgo project planning and experimental design steps and aim to understand and reduce the potential impact of intrinsic data biases. These planning steps should be similar to those carried out for the sequencing of new samples and could include an assessment of the dataset (1) and the pipeline for analysis (2):

(1) When compiling a combined dataset, it is important to consider the key question that is being addressed and to evaluate how many samples of each population, species, treatment, or time unit are needed to have the power to draw meaningful conclusions. It is also worth evaluating the trade-offs between sequencing new samples or using existing data (e.g. if only a handful of samples are missing could it be worthwhile to sequence more samples so that all individuals are sequenced the same way, reducing the likelihood that biases or batch effects will cause problems downstream in the analysis). If datasets will be combined to address a specific question then it is important to asses which specific sequenced samples are available and how many different datasets these samples come from. It is important to be conscious of, and carefully document, the different technologies used for library preparation and sequencing across samples and datasets, and if possible, to glean an understanding of the origin and quality of the input DNA. Ideally, the dataset would be compiled in a way that minimizes the number of differences between samples from different sources. Further, it is critical to strive to randomise samples from different biological units across different sequencing batches (Meirmans 2015). It can be particularly beneficial to repeat sequencing of one or a few representatives from a curated dataset to evaluate and correct potential biases. If feasible, repeated sequencing of the same individual allows to identify problematic loci that are not genotyped identically or consistently across technologies despite originating

from the same individual. We therefore urge researchers wherever possible to archive tissue and/or DNA. These collections can be of tremendous value, as they facilitate the repeated sequencing of past samples into newly compiled datasets to determine whether any variants or alleles may have been erroneously missed because of technological biases. Using archived tissue or DNA in this way is one of the only possibilities to verify new sequence variants found using future technologies.

(2) Once it is decided that integrating dataset from various sources provides the best power to answer a particular question, it is important to determine which checks should be implemented in the analysis pipeline to avoid misleading biological interpretation of the data. The ways in which biological and technological differences are distributed across the compiled dataset should be reported and critical steps that would identify potentially problematic sequence artifacts and biases should be implemented in the bioinformatic pipeline. It is also crucial to determine how potential artifacts and biases amongst datasets will be handled. Figure 2 provides a suggestion for a pipeline evaluating known differences between sequencing data produced with four-channel chemistry (e.g. HiSeqX) and two-channel chemistry (e.g. NovaSeq). We suggest comparing the FastQC report (Andrews, 2010) between samples sequenced with the two technologies to each other. Any systematic difference across FastQC reports might be relevant, however, when samples sequenced with different sequence chemistry that affects the base calling are combined reports on per base sequence and k-mers content are particularly worth paying attention to (see Figure 1 for an example, illustrating differences in k-mer counts). To see whether mapping reduces sequencing artefacts, FastQC can be re-run on only the reads that mapped well and will be used for genotyping. If biases persist, read trimming should be considered. Here fastp (Chen et al., 2018) could be used to trim poly-G tails efficiently. Once reads have been mapped, variants have been called, and genotypes have been determined, genotypes should be evaluated for potential batch effects. Here, we recommend identifying individuals sampled using different datasets and/or technologies with specific symbols or colours allowing the possible differences between these artificial groups to be highlighted (see section above). For example, in a Principal Component Analysis (PCA) which represents the various technological and sample differences by different symbols and biological

differences (i.e. populations or species) by colour, any PC axis separating symbols instead of colours suggests there might be some technological bias causing batch effects (Figure 1). However, biases might not always show up as batch effects and are especially problematic when one population or other biological unit is the only one sequenced with a different technology. In this scenario, artifacts and biological differences would be confounded and as a result artifacts and biases would be hard to detect (not visible as a batch effect in a PCA) and correct for. For this reason, we suggest that researchers aim to sequence biological units (species, populations, treatments, or time points) across each batch to avoid confounding biological differences with library or other technical effects. Alternatively, a bias might (although not necessarily) show up as a mutational bias relative to the reference, which can be evaluated and compared to published biases resulting from sequencing platform shifts (see Arora et al. (2019)). To reduce biases and undesired batch effects, the filtering parameters for variant calls and genotypes will need to be adjusted. One way to find the optimal filtering settings could be to determine which filtering thresholds allow you to minimize the differences between the detected batches. Specifically, it may be useful to compare distributions of quality scores between reference and alternate allele, which should look very similar in the absence of batch effects. However, we do not recommend solely relying on this to remove biases in the reads (such as poly-Gs in NovaSeq data) but mention this as one option that might help to reduce other sources of undesired batch effects. If none of these approaches suffice to identify and remove biases, one potential solution could be to define variable sites in a subset of the data, which only represents one technology, and then call genotypes on the whole dataset for only those regions. This comes with a potential ascertainment bias depending on how broadly biological units are represented in such a subset, but should reduce spurious variation caused by technological differences. Such an approach is similar to defining a SNP panel and then using SNPchips or other technologies to genotype a larger sampling (Kim et al., 2018). As all datasets are different, different approaches might be needed to reduce any effects of technological differences in compiled datasets. Critically, in each of these scenarios the identification and removal of biases associated with technological shifts serves to reduce the possibility of incorrectly or erroneously inferring biological patterns or processes.

Finally, we want to emphasise the huge value of community efforts to archive sequencing data that makes science reproducible and reusable. We hope that we have demonstrated not only how technological shifts may pose challenges for the meaningful reusability of data, but also that the removal of biases associated with such shifts allows us to address new and exciting biological questions. We highlight the importance and value of accurate documentation, archiving of tissue and DNA samples, and sequence data, and urge researchers to assess the experimental design of their research projects to ensure scientifically sound and robust results.

Acknowledgements

We thank David Marques and the Fish Ecology and Evolution group at Eawag for their helpful comments and fruitful discussions on the topic. We are grateful to three reviewers for their constructive comments on the manuscript.

RD is supported by an SNSF grant (31003A_163446) awarded to PF. DF is supported by the grant “SeeWandel: Life in Lake Constance – the past, present and future” within the framework of the Interreg V programme “Alpenrhein-Bodensee-Hochrhein (Germany/Austria/Switzerland/Liechtenstein)” which funds are provided by the European Regional Development Fund as well as the Swiss Confederation and cantons. D.F. received additional financial support from the Swiss Federal Office for the Environment and Eawag (Eawag Discretionary Funds 2018-2022). SLM is supported by an FCT scholarship (SFRH/BD/145153/2019) granted by the Portuguese National Science Foundation (Fundação para a Ciência e a Tecnologia – FCT).

The funders had no role in study design, decision to publish, or preparation of the manuscript.

References

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Arora, K., Shah, M., Johnson, M., Sanghvi, R., Shelton, J., Nagulapalli, K., . . . Robine, N. (2019). Deep whole-genome sequencing of 3 cancer cell lines on 2 sequencing platforms. *Scientific Reports, 9*, 19123. doi:10.1038/s41598-019-55636-3

Baker, B. J., De Anda, V., Seitz, K. W., Dombrowski, N., Santoro, A. E., & Lloyd, K. G. (2020). Diversity, ecology and evolution of Archaea. *Nature Microbiology, 5*(7), 887-900. doi:10.1038/s41564-020-0715-z

Bergland, A. O., Behrman, E. L., O'Brien, K. R., Schmidt, P. S., & Petrov, D. A. (2014). Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in Drosophila. *Plos Genetics*, 10(11), e1004775. doi:10.1371/journal.pgen.1004775

Bi, K., Linderoth, T., Singhal, S., Vanderpool, D., Patton, J. L., Nielsen, R., . . . Good, J. M. (2019). Temporal genomic contrasts reveal rapid evolutionary responses in an alpine mammal during recent climate change. *Plos Genetics*, 15(5), e1008119. doi:10.1371/journal.pgen.1008119

Bottery, M. J., Wood, A. J., & Brockhurst, M. A. (2019). Temporal dynamics of bacteria-plasmid coevolution under antibiotic selection. *Isme Journal, 13*(2), 559-562. doi:10.1038/s41396-018-0276-9

Brawand, D., Wagner, C. E., Li, Y. I., Malinsky, M., Keller, I., Fan, S., . . . Di Palma, F. (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nature, 513*(7518), 375-381. doi:10.1038/nature13726

Chen, S. F., Zhou, Y. Q., Chen, Y. R., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics, 34*(17), 884-890. doi:10.1093/bioinformatics/bty560

Cheng, S. F., Melkonian, M., Smith, S. A., Brockington, S., Archibald, J. M., Delaux, P. M., . . . Wong, G. K. S. (2018). 10KP: A phylodiverse genome sequencing plan. *Gigascience, 7*(3). doi:10.1093/gigascience/giy013

Feulner, P. G. D., Chain, F. J. J., Panchal, M., Huang, Y., Eizaguirre, C., Kalbe, M., . . . Milinski, M. (2015). Genomics of divergence along a continuum of parapatric population differentiation. *Plos Genetics*, 11(2), e1005414. doi:10.1371/journal.pgen.1004966

Greenway, R., Barts, N., Henpita, C., Brown, A. P., Rodriguez, L. A., Pena, C. M. R., . . . Shaw, J. H. (2020). Convergent evolution of conserved mitochondrial pathways underlies repeated adaptation to extreme environments. *Proceedings of the National Academy of Sciences of the United States of America, 117*(28), 16424-16430. doi:10.1073/pnas.2004223117

Jones, M. R., Mills, L. S., Jensen, J. D., & Good, J. M. (2020). The origin and spread of locally adaptive seasonal camouflage in snowshoe hares. *American Naturalist, 196*(3), 316-332. doi:10.1086/710022

Kim, J. M., Santure, A. W., Barton, H. J., Quinn, J. L., Cole, E. F., Visser, M. E., . . . Great Tit HapMap, C. (2018). A high-density SNP chip for genotyping great tit (*Parus major*) populations and its application to studying the genetic architecture of exploration behaviour. *Molecular Ecology Resources, 18*(4), 877-891. doi:10.1111/1755-0998.12778

Kirch, M., Romundset, A., Gilbert, M. T. P., Jones, F. C., & Foote, A. D. (2020). Pleistocene stickleback genomes reveal the constraints on parallel evolution. *bioRxiv*, 2020.2008.2012.248427. doi:10.1101/2020.08.12.248427

Lamichhaney, S., Berglund, J., Almen, M. S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., . . . Andersson, L. (2015). Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature, 518*(7539), 371-375. doi:10.1038/nature14181

Leebens-Mack, J. H., Barker, M. S., Carpenter, E. J., Deyholos, M. K., Gitzendanner, M. A., Graham, S. W., . . . One Thousand Plant, T. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature, 574*(7780), 679-685. doi:10.1038/s41586-019-1693-2

Leigh, D. M., Lischer, H. E. L., Grossen, C., & Keller, L. F. (2018). Batch effects in a multiyear sequencing study: False biological trends due to changes in read lengths. *Molecular Ecology Resources* **18**: 778-788. doi: 10.1111/1755-0998.12779

Marques, D. A., Lucek, K., Sousa, V. C., Excoffier, L., & Seehausen, O. (2019). Admixture between old lineages facilitated contemporary ecological speciation in

Lake Constance stickleback. *Nature Communications, 10*, 4240. doi:10.1038/s41467-019-12182-w

Martin, S. H., Most, M., Palmer, W. J., Salazar, C., McMillan, W. O., Jiggins, F. M., & Jiggins, C. D. (2016). Natural selection and genetic diversity in the butterfly *Heliconius melpomene*. *Genetics, 203*(1), 525-541. doi:10.1534/genetics.115.183285

Mason, C. C. (2017). Four study design principles for genetic investigations using next generation sequencing. *Bmj-British Medical Journal*, 359, j4069. doi:10.1136/bmj.j4069

Meirmans, P. G. (2015). Seven common mistakes in population genetics and how to avoid them. *Molecular Ecology* 24: 3223-3231. doi: 10.1111/mec.13243

Morris, J. L., Puttick, M. N., Clark, J. W., Edwards, D., Kenrick, P., Pressel, S., . . . Donoghue, P. C. J. (2018). The timescale of early land plant evolution. *Proceedings of the National Academy of Sciences of the United States of America, 115*(10), E2274-E2283. doi:10.1073/pnas.1719588115

O'Leary Shannon, J., Puritz Jonathan, B., Willis Stuart, C., Hollenbeck Christopher, M., & Portnoy David, S. (2018). These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology* 27:3193–3206. doi: 10.1111/mec.14792

Peter, J., De Chiara, M., Friedrich, A., Yue, J. X., Pflieger, D., Bergstrom, A., . . . Schacherer, J. (2018). Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature, 556*(7701), 339-344. doi:10.1038/s41586-018-0030-5

Poulsen, C. S., Pamp, S. J., Ekstrøm, C. T., & Aarestrup, F. M. (2019). Library preparation and sequencing platform introduce bias in metagenomics characterisation of microbial communities. *bioRxiv*, 592154. doi:10.1101/592154

Ravinet, M., Kume, M., Ishikawa, A., & Kitano, J. Patterns of genomic divergence and introgression between Japanese stickleback species with overlapping breeding habitats. *Journal of Evolutionary Biology, 00,* 1-14. doi:10.1111/jeb.13664

Samuk, K., Owens, G. L., Delmore, K. E., Miller, S. E., Rennison, D. J., & Schluter, D. (2017). Gene flow and selection interact to promote adaptive divergence in regions of low recombination. *Molecular Ecology, 26*(17), 4378-4390. doi:10.1111/mec.14226

Sato, M. P., Ogura, Y., Nakamura, K., Nishida, R., Gotoh, Y., Hayashi, M., . . . Hayashi, T. (2019). Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. *DNA Research, 26*(5), 391-398. doi:10.1093/dnares/dsz017

Shen, X. X., Opulente, D. A., Kominek, J., Zhou, X., Steenwyk, J. L., Buh, K. V., . . . Rokas, A. (2018). Tempo and mode of genome evolution in the budding yeast subphylum. *Cell, 175*(6), 1533-1545. doi:10.1016/j.cell.2018.10.023

Shi, M., Lin, X. D., Chen, X., Tian, J. H., Chen, L. J., Li, K., . . . Zhang, Y. Z. (2018). The evolutionary history of vertebrate RNA viruses. *Nature, 561*(7722), E6. doi:10.1038/s41586-018-0310-0

Soria-Carrasco, V., Gompert, Z., Comeault, A. A., Farkas, T. E., Parchman, T. L., Johnston, J. S., . . . Nosil, P. (2014). Stick insect genomes reveal natural selection's role in parallel speciation. *Science, 344*(6185), 738-742. doi:10.1126/science.1252136

Stankowski, S., Chase, M. A., Fuiten, A. M., Rodrigues, M. F., Ralph, P. L., & Streisfeld, M. A. (2019). Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers. *Plos Biology*, 17(7), e3000391. doi:10.1371/journal.pbio.3000391

Tenaillon, O., Barrick, J. E., Ribeck, N., Deatherage, D. E., Blanchard, J. L., Dasgupta, A., . . . Lenski, R. E. (2016). Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature, 536*(7615), 165-170. doi:10.1038/nature18959

Tollis, M., Hutchins, E. D., Stapley, J., Rupp, S. M., Eckalbar, W. L., Maayan, I., . . . Kusumi, K. (2018). Comparative genomics reveals accelerated evolution in conserved pathways during the diversification of anole lizards. *Genome Biology and Evolution, 10*(2), 489-506. doi:10.1093/gbe/evy013

Vijay, N., Bossu, C. M., Poelstra, J. W., Weissensteiner, M. H., Suh, A., Kryukov, A. P., & Wolf, J. B. W. (2016). Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. *Nature Communications, 7*, 10. doi:10.1038/ncomms13195

Zhang, G. J., Li, C., Li, Q. Y., Li, B., Larkin, D. M., Lee, C., . . . Avian Genome, C. (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. *Science, 346*(6215), 1311-1320. doi:10.1126/science.1251385

Zhang, C., Zhang, T., Luebert, F., Xiang, Y., Huang, C.-H., Hu, Y., . . . Ma, H. (2020). Asterid phylogenomics/phylotranscriptomics uncover morphological evolutionary histories and support phylogenetic placement for numerous whole-genome duplications. *Molecular Biology and Evolution*. doi:10.1093/molbev/msaa160

Author's contributions

RD, DF, and PF conceived of the presented ideas based on the experience and insights of DF. RD and PF drafted the manuscript. PF drafted the figures. All authors contributed to the discussion and critical revision of the final manuscript.

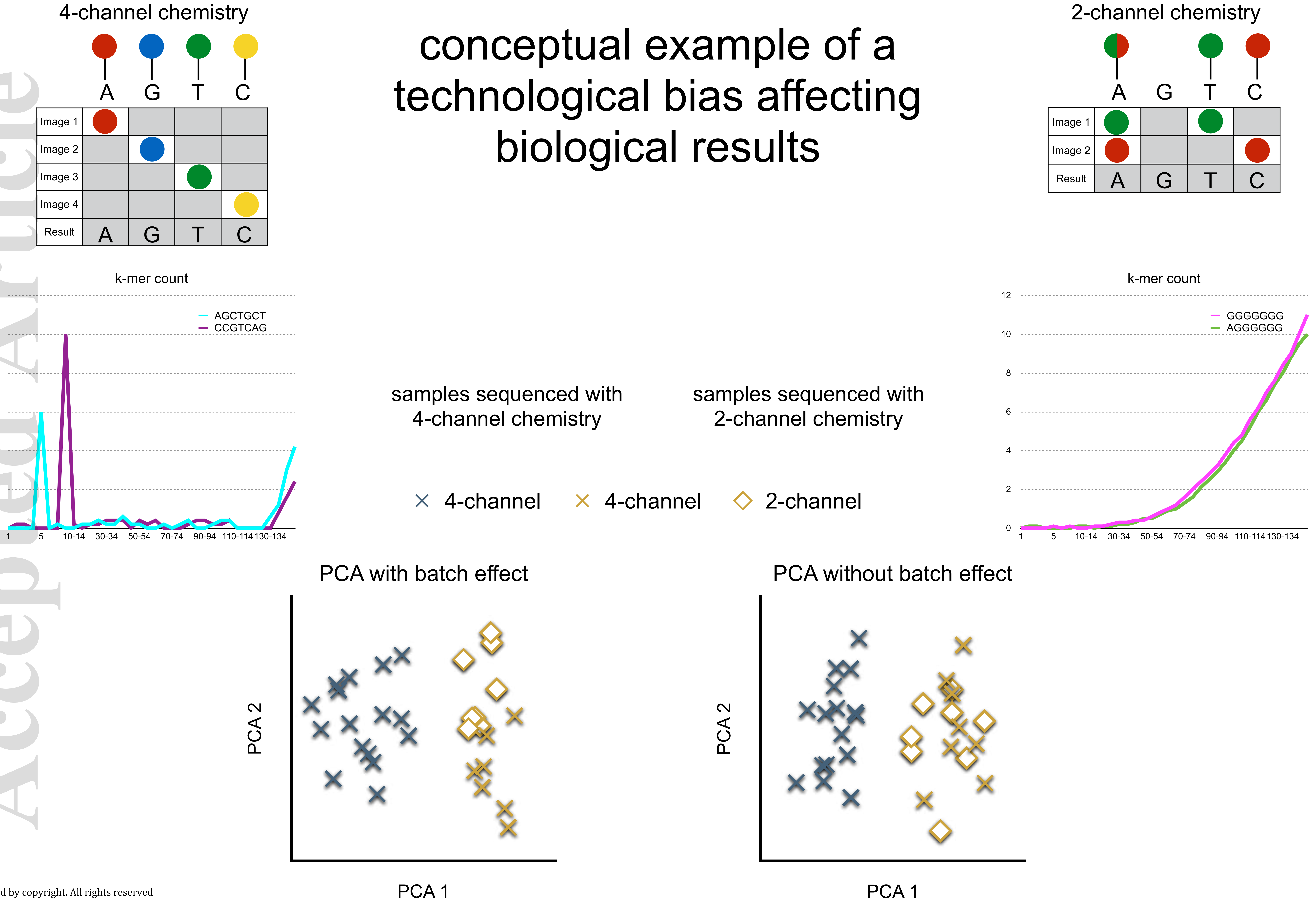
Figure 1: Example of a technological difference between sequencing chemistries, which introduces a bias (overrepresentation of G k-mers) in the sequenced reads and result in a batch effect visible when genotypes are evaluated in a principal component analysis (PCA).

Top: Schematic redrawn from Illumina representing the differences between 4-channel chemistry evaluating each of the four bases by a distinct fluorescence label, and 2-channel chemistry representing the four bases with two dyes only.

Middle: Redrawn examples of the one aspect of a typical FastQC (Andrews, 2010) report, which evaluates the count of each short nucleotide of length k (default = 7) starting at each position along the read. Any given k-mer should be evenly represented across the length of the read. The y axis reports the relative enrichment (log2 observed over expected counts) of the 7-mers over the read length (x axis). The graph presents those k-mers which appear at specific positions with greater than expected frequency. In the left panel reads sequenced with 4-channel chemistry are represented which show a slight overrepresentation of two random 7-mers represented by different colours (typically the report would plot the first six hits). The overrepresentation is small and most pronounced at the beginning of the read (to the left of the x axis), a pattern often found in high quality sequencing libraries due to slight, sequence dependent efficiency of DNA shearing or a result of random priming. In the right panel, an overrepresentation of poly-G-mers toward the end of the reads is exemplified as typical for raw reads sequenced with 2-channel chemistry. Note the difference in the logarithmic scale between left and right panel.

Bottom: Conceptual representation of a batch effect resulting from technological differences. Each sample's genotype, compiled of a large number of loci distributed across the whole genome, is represented as a coloured symbol in multivariate space, where PC axis one and two reflect two primary axes of variation in the dataset. The left panel would reflect a dataset with a batch effect. The fact that samples are separated by sequencing technology on PC axis 2 indicates the presence of a technological bias. In the right panel, batch effects have been reduced, e.g. by trimming off poly-G tails. Symbols in the PCA differentiate samples sequenced with either 2-channel (diamond) or 4-channel (cross) chemistry, colours differentiate different populations or species (biological differences). The left panel is imagined to be based on a data set of untrimmed reads, PC axis 2 separates samples due to technological differences. That effect is gone in the right panel, after read trimming was applied.

Figure 2: Flow diagram of an exemplified pipeline evaluating and accounting for biases caused by different sequencing technologies in a compiled data set. For more details see text.



conceptual example of a technological bias affecting biological results

4-channel chemistry

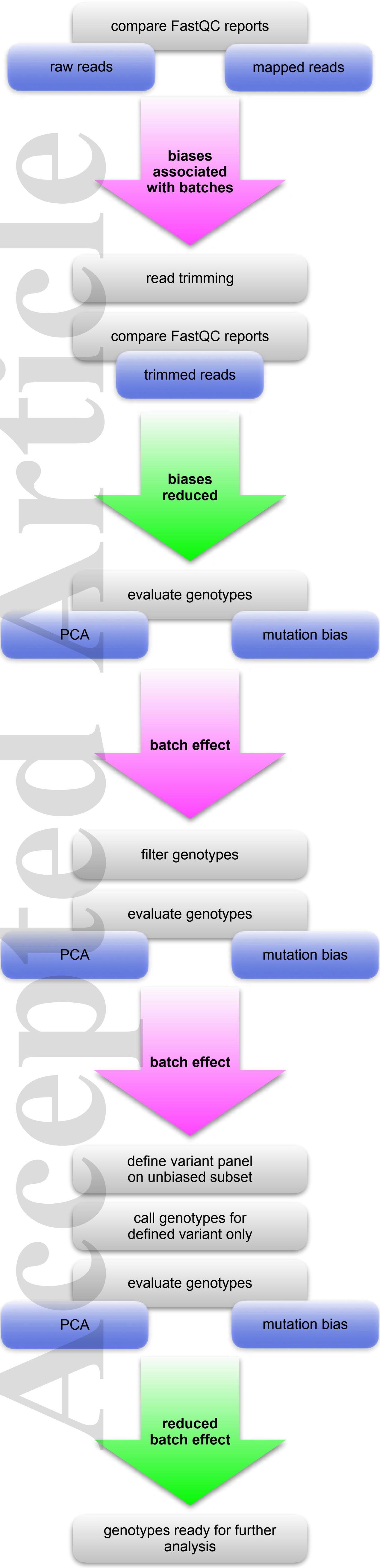
	A	G	T	C
Image 1	🔴			
Image 2		🔵		
Image 3			🟢	
Image 4				🟡
Result	A	G	T	C

2-channel chemistry

	A	G	T	C
Image 1	🟢		🟢	
Image 2	🔴			🔴
Result	A	G	T	C

samples sequenced with 4-channel chemistry

samples sequenced with 2-channel chemistry


compare FastQC reports
raw reads
mapped reads
biases associated with batches
read trimming
compare FastQC reports
trimmed reads
biases reduced
evaluate genotypes
PCA
mutation bias
batch effect
filter genotypes
evaluate genotypes
PCA
mutation bias
batch effect
define variant panel on unbiased subset
call genotypes for defined variant only
evaluate genotypes
PCA
mutation bias
reduced batch effect
genotypes ready for further analysis