UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **accepted version** of the book part:

Ramos Onsins, Sebastián E. [et al.]. «npstat : An Efficient Tool to Explore the Population Genome Variability and Divergence Using Pool Sequencing Data». *Insect Genomics. Methods and Protocols*, 2025, p. 51-66 DOI 10.1007/978-1-0716-4583-3_3

This version is available at https://ddd.uab.cat/record/324884

under the terms of the IN COPYRIGHT license.

Corresponding Authors:

Sebastian E. Ramos-Onsins[1], Sara Guirao-Rico[2,3]

[1] Plant and Animal Genomics Program, Centre for Research in Agricultural Genomics (CRAG), Consejo Superior de Investigaciones Científicas-Institute of Agrifood Research and Technology-Autonomous University of Barcelona-University of Barcelona,. sebastian.ramos@cragenomica.es

[2] Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona, Barcelona, Spain. sguirao@ub.edu

[3] Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain

npstat: An Efficient Tool to Explore the Population Genome Variability and Divergence using Pool Sequencing Data

Sebastian E. Ramos-Onsins[1], Sara Guirao-Rico[2,3], Ahmed Hafez[4], Luca Ferretti[5]

[1] Plant and Animal Genomics Program, Centre for Research in Agricultural Genomics (CRAG), Consejo Superior de Investigaciones Científicas-Institute of Agrifood Research and Technology-Autonomous University of Barcelona-University of Barcelona. sebastian.ramos@cragenomica.es

[2] Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona, Barcelona, Spain. sguirao@ub.edu

[3] Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain

[4] Biotechvana. Technology. Park of Paterna,Paterna (Valencia) Spain. ahmed.hafez@biotechvana.com

[5] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department for Medicine, University of Oxford, Oxford, UK. luca.ferretti@bdi.ox.ac.uk

Abstract

Pool sequencing has emerged as a valuable approach in ecological studies, particularly when dealing with very small organisms (with limited amount of DNA available), when distinguish individual organisms is a challenge (e.g., in colonies, microbiome), when there is a trade-off between the sequencing cost and the number of individuals to sequence, when the main goal is to estimate nucleotide variability and variant frequency patterns at the population level (that is, when individual information is not required). Estimates of variability can be efficiently explored by analyzing sequences of pooled individuals sampled from the population. When using this approach, the number of pooled individuals and the mean read depth are key choices in the experimental design. The software *npstat* calculates different estimates of nucleotide variability and neutrality tests, including the number of synonymous and nonsynonymous variants and the proportion of beneficial substitutions (alpha) using the MKT approach, these last when a GTF annotation file and an outgroup is provided.

Key Words

Polymorphism, Divergence, Population Genomics, Proportion of Beneficial substitutions, Sliding Windows Analyses.

Running Head

npstat for Pool Sequencing Variability Analysis

1. Introduction

Pool sequencing is widely employed in molecular ecology studies [e.g., 1,2,3,4], particularly when working with non-model organisms, as they typically have fewer genetic tools and resources and for which obtaining individual sequencing data remains expensive. Still, it has been used in large-scale population studies of model organisms like *Drosophila melanogaster* [4,5]. However, this approach has some limitations, such as uncertainty in the number of sequenced individuals in a given sample and the lack of individual information. Despite these constraints, pool sequencing has proven to be an effective method for assessing population variability [6]. Among its advantages is that including many individuals in the pool significantly reduces the variance in variability estimates while reducing the sequencing costs compared to individual sequencing [6]. The variance in the estimates of allele frequency arises from a two-step sampling process: first, by sampling individuals from the population, and second, by sampling sequencing reads from the pooled sample. The total variance is approximately the sum of the variances from these two sampling events [6,7]. Therefore, for experiments with the same read depth, increasing the number of pooled individuals is an effective strategy to reduce the overall variance in the allele frequency estimates.

In analysis of variation using pool sequencing, the primary goal is typically to infer the whole-genome site frequency spectrum of the population to subsequently obtain the population-level estimates and statistics derived from this data [e.g., 8,7]. In contrast, *npstat* is designed for the direct estimation of summary statistics of genetic variation commonly used in population genetics from the data, such as Watterson's, Tajima's, and Fay and Wu's estimators [9,10,11,12]. It also performs neutrality tests like Tajima's *D*, Fay and Wu's *H*, and the McDonald-Kreitman test (MKT) [11,12,13]. Some of these analyses require a GFF (Generic Feature Format, https://gmod.org) annotation file to calculate synonymous and nonsynonymous positions and an outgroup sequence.

2. Materials

Data requirements to run *npstat*

npstat requires pool sequencing DNA data from a population to perform the analysis of variability. This data must be provided in mpileup format [14]. Pool sequencing reads (*e.g.*, files in fastq format) must first be trimmed and filtered (i.e., to remove duplicated reads, adapter sequences, low-quality reads), and then aligned

against a reference sequence. The reference sequence (in FASTA format) is usually a consensus genome sequence from the same species, although a closely related species or an outgroup may also be used as a reference. Once the filtered reads are aligned against the reference (using for example bwa [15]), and optionally optimized and re filtered (e.g., removing duplicates and recalibrating base quality scores using GenomeAnalysisToolKit [16]), the resulted aligned pool sequencing data (in BAM format [14]) can be converted to mpileup format using bcftools [14,17].

Additional files can be used to perform optional analysis:
(i) A FASTA-formatted outgroup sequence aligned to the reference sequence used in the pool-sequencing dataset. This is necessary to perform analysis that needs to estimate divergence (e.g., when comparing patterns of polymorphism versus divergence). In case of using an outgroup as a reference, the same sequence can be used to perform this analysis.
(ii) A GFF (version 3) file containing the annotation of functional regions analyzed (e.g., the coordinates of genes or cds regions, the strand and the frame of exons) of the species used as a reference. This information is needed to estimate the number of nonsynonymous and synonymous variants from the pool sequencing data.
(iii) A text file containing the coordinates of specific SNP (Single Nucleotide Polymorphisms) in the analysed region. Each row should correspond to a specific position for each SNP. If this option is used, the program only analyzes the included positions in the file.

Install *npstat*

npstat repository is located in https://github.com/lucaferretti/npstat. *npstat* can be installed locally using a docker container. Using Docker to build and run *npstat* provides a consistent and reproducible environment, minimizing issues related to dependencies and system configurations. To do it, install first Docker following the instructions below:

- Installing Docker

For Linux systems, you can install Docker using the package manager of your distribution. For MacOS and Windows, you can download the Docker Desktop application from the official Docker website. For example, to install Docker on Ubuntu/Debian or CentOS, you can use the following commands:

In Ubuntu/Debian

sudo apt-get update

sudo apt-get install docker-ce

in CentOS

sudo yum install docker

For more detailed installation instructions, refer to the official Docker documentation.

In MacOS

(i) Download Docker Desktop: Visit the Docker Desktop for Mac page.

(ii) Run the Installer: Open the downloaded .dmg file and drag Docker to the Applications folder.

(iii) Start Docker Desktop: Open Docker from the Applications folder.

(iv) Verify Installation: Open a Terminal and run:

docker –version

in Windows

(i) Download Docker Desktop: Visit the Docker Desktop for Windows page.

(ii) Run the Installer: Open the downloaded .exe file and follow the installation instructions.

- Using *npstat* container

Before using the container, you need to pull the image from Docker Hub. The image is available at the following location: biotechvana/npstat.

Pull the image from docker hub

docker pull biotechvana/npstat

tag the image, if you want to use a different name or to avoid using the long image name

docker tag biotechvana/npstat npstat

Verify that the image has been successfully pulled by running the following command:

docker run --rm npstat

Using the Docker Container

To run the *npstat* application, use the following command:

docker run --rm npstat

This command will display the usage information for npstat.

- Running *npstat* with Data:

To analyze specific data, you can mount a local directory containing your data files to the container and specify the input files. For example:

docker run --rm -v /path/to/data:/data npstat -n sample_size -l window_length [options] FILE.pileup

(i) Replace *sample_size* with the haploid sample size.

(ii) Replace *window_length* with the window length in base pairs.

(iii) Replace */path/to/data* with the path to your local directory containing the data files and FILE.pileup with the name of your input file.

Note: that ":/data" specifies the mount point inside the container where the data files will be available and mounted as the current working directory. And the mount point "/data" is fixed inside the container and should not be changed. "/data" is also the default working directory inside the container.

For example, if you have a data file named *FILE.pileup* in a directory */home/user/npstat_data*, you can run *npstat* as follows:

docker run --rm -v /home/user/npstat_data:/data npstat -i /data/FILE.pileup ## as /data is the working directory inside the container, you can also run it as follows

docker run --rm -v /home/user/npstat_data:/data npstat -i FILE.pileup

Alternatively, if you have a data file named *FILE.pileup* in your current directory, you can run *npstat* as follows:

docker run --rm -v $(pwd):/data npstat -i FILE.pileup

or

docker run --rm -v ./:/data npstat -i FILE.pileup

The output file will be generated in the same directory where the input file is located.

- Building the Docker Image [Optional]

If you want to build the Docker image locally, you can follow these steps:

1. Clone the *npstat* repository from GitHub:

git clone https://github.com/ahmedihafez/npstat.git

2. Change to the *npstat* directory:

cd npstat

3. Build the Docker image using the provided Dockerfile:

docker build -t npstat .

4. Verify that the image has been successfully built by running the following command:

docker run --rm npstat

3. Methods

Statistics calculated in *npstat*

npstat implements algorithms to estimate nucleotide variation at the population level and to perform neutrality tests from pool sequencing data. The statistics included in this software were developed in [6]. *npstat* software calculates estimates of variation and neutrality tests given a specified window size (in base pairs) and considering the options defined by the user. All these statistics are computed after filtering for minimum read depth, sequence quality and a minimum allele count. Specifically, *npstat* contains the following statistics:

(i) S, the number of Segregating sites per window and its variance,

(ii) the variability Watterson estimator of θ [9] per nucleotide and window and its variance (see [6] equation 1 and 2 and its Suppl. Information section SI.3, and SI.5),

(iii) the Tajima's π estimator of heterozygosity [10] per nucleotide and window (see [6] equation 3 and its Suppl. Information section SI.2),

(iv) the neutrality tests Tajima's D [11] per window (see [6] equation 5 and its Suppl. Information section SI.5),

(v) the unnormalized Fay and Wu's *H* test [12] per nucleotide and window, in case an outgroup sequence is provided (it is also possible to obtain the Fay and Wu's *H* variability estimator just summing π, See equation 8 in [6] and its Suppl. Information section SI.3),

(vi) the normalized (unnormalized value divided by its variance) Fay and Wu's *H* test [12], see also [6] equation 7 and its Suppl. Information section SI.5),

(vii) the synonymous and non-synonymous segregating sites per window in case the annotation file with the information of the coding sequence regions is provided. Note that we assume all aminoacids to be 4-fold degenerate (i.e. non-synonymous and synonymous sites correspond to the 1st/2nd base and 3rd base in the codon, respectively),

(viii) the synonymous and non-synonymous divergence (fixed variants) in case an outgroup sequence and the annotation file are provided,

(ix) the proportion of beneficial substitutions (α, [18]) per window. This is estimated using the number of segregating sites and the divergence for synonymous and non-synonymous positions using the equation $\alpha = 1 - Sn/Ss * Ks/Kn$, where Sn is the number of non-synonymous Segregating sites, Ss is the number of synonymous Segregating sites, Ks is the synonymous divergence and Kn is the non-synonymous divergence.

Run npstat

From the working directory, use the command

./npstat -n sample_size -l window_length [options] FILE.pileup

All the available options to run *npstat* are shown in **Table 1.**

An important option is *-nolowfreq m*. This specifies how many reads with the alternative variant there must be to discard specific positions (i. e., remove positions with low-frequency alleles). The default option is m=1, which means that if the alternative variant is supported only by 1 read this position will be discarded. High depth or error rate sequencing data would need higher values, e.g. m=2 above read depth 50, m=3 above read depth 100, etc. Use m=0 only if the SNPs have already been called by an external SNP caller and input into the program through the option *-snpfile*.

Output

The output file retains the same name as the input file, with the extension *.stats*. The output file contains 18 columns, each corresponding to information regarding the analyzed window, including the above mentioned statistics, while each row refers to an individual genomic window. (**Tables 2 and 3**):

Additionally, the HKA test [19] can be obtained from the estimated S (columns 5), Var(S) (column 11) and divergence (column 13). Also, the significance for McDonald-Kreitman test [13] can be obtained from the estimated synonymous/non-synonymous polymorphism/divergence data from columns 14-17 in a 2 × 2 contingency table.

We include a R script with which the user can plot the results included in the output file. The script is called ***npstat_plot_windows.R*** and can be run with the command:

R –vanilla < ./npstat_plot_windows.R file_out file_out.pdf

where *file.out* is the output file obtained after running *npstat*, Two PDF output files are generated after running this script, one containing plots for each statistic, showing it value across windows and a the other containing the distributions of the different statistics.

Example

We include three examples to show three different types of analyses that can be performed (uploaded at https://zenodo.org/records/13364757). Here, we collected sequencing data obtained from the Drosophila Genetic Reference Panel (DGRP) [20] that were pooled (specifically, we pooled the inbreed lines with the reference RAL numbers indicated in the Table S1 from [20]), aligned to the *D. melanogaster* reference genome, version 6.12, and processed and filtered as indicated in [21]. A pileup file containing the chromosome 2L were used as the input file for running *npstat*. In the example 1, we used this pileup without an outgroup and without an annotation file, In the second example, we added an outgroup (the line of *Drosophila yakuba* SRR26246471 [22], which was aligned to the reference sequence using standard methods indicated in the Methods section). In the example 3, we additionally included a GFF annotation file for *D. melanogaster*, in order to identify non-synonymous and synonymous variants. The commands used are the following:

echo "Example 1: "

./npstat -n 16 -l 100000 -nolowfreq 2 -minqual 18 Pool_seq2.mel_mpileup_2L.txt

mv Pool_seq2.mel_mpileup_2L.txt.stats Pool_seq2.mel_mpileup_2L.txt.stats.ex1.txt

echo "Example 2: "

./npstat -n 16 -l 100000 -nolowfreq 2 -minqual 18 -outgroup Dyakuba-mel_final_cns_2L.fa Pool_seq2.mel_mpileup_2L.txt

mv Pool_seq2.mel_mpileup_2L.txt.stats Pool_seq2.mel_mpileup_2L.txt.stats.ex2.txt

echo "Example 3 "

./npstat -n 16 -l 100000 -nolowfreq 2 -minqual 18 -outgroup Dyakuba-mel_final_cns_2L.fa -annot dmel-all-r6.12_sorted_2L.gtf Pool_seq2.mel_mpileup_2L.txt

mv Pool_seq2.mel_mpileup_2L.txt.stats Pool_seq2.mel_mpileup_2L.txt.stats.ex3.txt

This generates an output file in tabular format **(**see example in **Table 1)** with 18 columns, and as many rows as windows specified in the analysis (plus a header with the name of the columns). This table can be used to compute summary statistics and plot some figures from the obtained results. Note that in this example, we change the name of the output files to avoid overwriting the output file since the name for the output file is obtained from the name of the input file and in all three examples we use the same input file.

To visualize the results, run the R script npstat_plot_window.R. For these exemples, run the following:

R --vanilla < npstat_plot_windows.R --args Pool_seq2.mel_mpileup_2L.txt.stats Pool_seq2.mel_mpileup_2L.txt.stats_ex1.pdf

R --vanilla < npstat_plot_windows.R --args Pool_seq2.mel_mpileup_2L.txt.stats.ex1.txt Pool_seq2.mel_mpileup_2L.txt.stats.ex2.pdf

R --vanilla < npstat_plot_windows.R --args Pool_seq2.mel_mpileup_2L.txt.stats.ex1.txt Pool_seq2.mel_mpileup_2L.txt.stats.ex3.pdf

The R script is designed to generate two kinds of figures: (i) sliding window plots with the values of each statistic in the y-axis and the window in the x-axis **(Figure 1),** and (ii) density (violin) plots with the density distribution of the statistics **(Figure 2)**. Other interesting plots not included in the script can be, for example, those corresponding to the analysis of autocorrelation or to comparative studies between scaffolds or to other populations.

4. Notes

Pool sequencing offers a cost-effective alternative with the advantage of higher accuracy in estimating nucleotide variability compared to individual sequencing. This is because more individuals can be included in the pool at the same cost compared to an individual sequencing sample, which reduces the variance in the estimates. This advantage makes the careful design of pool sequencing experiments essential. One critical parameter is the number of haploid samples included in the analysis, specified using the -n option in the software. If the exact number of samples in the pool is unknown, this can pose a problem depending on the sample size. For small pools (around 10 individuals), inaccurately doubling or halving the sample size can significantly alter the estimates of the statistics. However, for larger pools (e.g., $n > 100$), an error of this magnitude is less likely to substantially impact variability estimates.

Additionally, sequencing quality (controlled by the -minqual option) can influence the detection of variants, particularly if the quality threshold is set too high. While high thresholds may reduce the number of detected variants, they are unlikely to affect overall variability estimates, as sequence quality impacts both monomorphic and polymorphic positions equally. Conversely, setting a low-quality threshold can be offset by excluding low-frequency variants using the -nolowfreq option. The software *npstat* addresses these potential sequencing errors by efficiently calculating variability estimates while considering this ascertainment bias. The statistical power of the analyses using pool sequencing data has been previously discussed in detail [6].

It is important to highlight that *npstat* uses a relatively simple method to differentiate between synonymous and nonsynonymous positions. It considers nonsynonymous positions strictly as those located at the second base of codons, while positions at the first and third bases are treated as synonymous. This approach ensures accurate variability estimates for nonsynonymous positions. However, this method may miss some nonsynonymous

variants at the first and third codon positions (counted as synonymous), which could slightly bias the Syn/NSyn ratio for both polymorphism and divergence. Nonetheless, this bias is not expected to have a significant impact.

A second version of *npstat* is currently in development, which will include the analysis of multi-scaffold pool sequencing data, additional variability statistics (such as theta E from Zeng and Zeng's E test [23]), more accurate estimates of variability for synonymous and nonsynonymous sites, and improved methods for the obtaining these estimates. The new version will also include the calculation of population differentiation metrics (e.g., Fst, as per equations 11-13 in [6]), using two pool sequencing pileup files.

5. Acknowledgement

This work is based on our previous publication [6]. S.E.R.-O. was supported by grants AGL2016-78709-R funded by MICIU/AEI/10.13039/501100011033 and by "ERDF A way of making Europe" and grant PID2020-119255GB-I00 funded by MICIU/AEI/10.13039/501100011033. This work was also supported by grants SEV-2015-0533 and CEX2019-000902-S funded by MICIU/AEI/10.13039/501100011033, and by the CERCA Programme / Generalitat de Catalunya.

6. References

[1] Chen C, Parejo M, Momeni J, Langa J, Nielsen RO, Shi W, Smartbees Wp Diversity Contributors, Vingborg R, Kryger P, Bouga M, Estonba A, Meixner M. Population Structure and Diversity in European Honey Bees (Apismellifera L.)-An Empirical Comparison of Pool and Individual Whole-Genome Sequencing. Genes (Basel). 2022 Jan 21;13(2):182. doi: 10.3390/genes13020182. PMID: 35205227; PMCID: PMC8872436.

[2] Eynard SE, Vignal A, Basso B, Canale-Tabet K, Le Conte Y, Decourtye A, Genestout L, Labarthe E, Mondet F, Servin B. Reconstructing queen genotypes by pool sequencing colonies in eusocial insects: Statistical Methods and their application to honeybee. Mol Ecol Resour. 2022 Nov;22(8):3035-3048. doi: 10.1111/1755-0998.13685. Epub 2022 Aug 1. PMID: 35816386; PMCID: PMC9796407.

[3] Carvalho J, Morales HE, Faria R, Butlin RK, Sousa VC. Integrating Pool-seq uncertainties into demographic inference. Mol Ecol Resour. 2023 Oct;23(7):1737-1755. doi: 10.1111/1755-0998.13834. Epub 2023 Jul 20. PMID: 37475177.

[4] Bogaerts-Márquez M, Guirao-Rico S, Gautier M, González J. Temperature, rainfall and wind variables underlie environmental adaptation in natural populations of Drosophila melanogaster. Mol Ecol. 2021 Feb;30(4):938-954. doi: 10.1111/mec.15783. Epub 2021 Jan 26. PMID: 33350518; PMCID: PMC7986194.

[5] Kapun M, Nunez JCB, Bogaerts-Márquez M, Murga-Moreno J, Paris M, Outten J, Coronado-Zamora M, Tern C, Rota-Stabelli O, Guerreiro MPG, Casillas S, Orengo DJ, Puerma E, Kankare M, Ometto L, Loeschcke V, Onder BS, Abbott JK, Schaeffer SW, Rajpurohit S, Behrman EL, Schou MF, Merritt TJS, Lazzaro BP, Glaser-Schmitt A, Argyridou E, Staubach F, Wang Y, Tauber E, Serga SV, Fabian DK, Dyer KA, Wheat CW, Parsch J, Grath S, Veselinovic MS, Stamenkovic-Radak M, Jelic M, Buendía-Ruíz AJ, Gómez-Julián MJ, Espinosa-Jimenez ML, Gallardo-Jiménez FD, Patenkovic A, Eric K, Tanaskovic M, Ullastres A, Guio L, Merenciano M, Guirao-Rico S, Horváth V, Obbard DJ, Pasyukova E, Alatortsev VE, Vieira CP, Vieira J, Torres JR, Kozeretska I, Maistrenko OM, Montchamp-Moreau C, Mukha DV, Machado HE, Lamb K, Paulo T, Yusuf L, Barbadilla A, Petrov D, Schmidt P, Gonzalez J, Flatt T, Bergland AO. Drosophila Evolution over Space and Time (DEST): A New Population Genomics Resource. Mol Biol Evol. 2021 Dec 9;38(12):5782-5805. doi: 10.1093/molbev/msab259. Erratum in: Mol Biol Evol. 2022 Jan 7;39(1):msab337. doi: 10.1093/molbev/msab337. PMID: 34469576; PMCID: PMC8662648.

[6] Ferretti L, Ramos-Onsins SE, Pérez-Enciso M. Population genomics from pool sequencing. Mol Ecol. 2013 Nov;22(22):5561-76. doi: 10.1111/mec.12522. Epub 2013 Oct 28. PMID: 24102736.

[7] Futschik A, Schlötterer C. The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics. 2010 Sep;186(1):207-18. doi: 10.1534/genetics.110.114397. Epub 2010 May 10. PMID: 20457880; PMCID: PMC2940288.

[8] Hellmann I, Mang Y, Gu Z, Li P, de la Vega FM, Clark AG, Nielsen R. Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. Genome Res. 2008 Jul;18(7):1020-9. doi: 10.1101/gr.074187.107. Epub 2008 Apr 14. PMID: 18411405; PMCID: PMC2493391.

[9] Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 1975 Apr;7(2):256-76. doi: 10.1016/0040-5809(75)90020-9. PMID: 1145509.

[10] Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics. 1983 Oct;105(2):437-60. doi: 10.1093/genetics/105.2.437. PMID: 6628982; PMCID: PMC1202167.

[11] Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989 Nov;123(3):585-95. doi: 10.1093/genetics/123.3.585. PMID: 2513255; PMCID: PMC1203831.

[12] Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics. 2000 Jul;155(3):1405-13. doi: 10.1093/genetics/155.3.1405. PMID: 10880498; PMCID: PMC1461156.

[13] McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 1991 Jun 20;351(6328):652-4. doi: 10.1038/351652a0. PMID: 1904993.

[14] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. Epub 2009 Jun 8. PMID: 19505943; PMCID: PMC2723002.

[15] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754-60. doi: 10.1093/bioinformatics/btp324. Epub 2009 May 18. PMID: 19451168; PMCID: PMC2705234.

[16] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297-303. doi: 10.1101/gr.107524.110. Epub 2010 Jul 19. PMID: 20644199; PMCID: PMC2928508.

[17] Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. Gigascience. 2021 Feb 16;10(2):giab008. doi: 10.1093/gigascience/giab008. PMID: 33590861; PMCID: PMC7931819.

[18] Smith NG, Eyre-Walker A. Adaptive protein evolution in Drosophila. Nature. 2002 Feb 8;415(6875):1022-4. doi: 10.1038/4151022a. PMID: 11875568.

[19] Hudson RR, Kreitman M, Aguadé M. A test of neutral molecular evolution based on nucleotide data. Genetics. 1987 May;116(1):153-9. doi: 10.1093/genetics/116.1.153. PMID: 3110004; PMCID: PMC1203113.

[20] Mackay TF, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, Richardson MF, Anholt RR, Barrón M, Bess C, Blankenburg KP, Carbone MA, Castellano D, Chaboub L, Duncan L, Harris Z, Javaid M, Jayaseelan JC, Jhangiani SN, Jordan KW, Lara F, Lawrence F, Lee SL, Librado P, Linheiro RS, Lyman RF, Mackey AJ, Munidasa M, Muzny DM, Nazareth L, Newsham I, Perales L, Pu LL, Qu C, Ràmia M, Reid JG, Rollmann SM, Rozas J, Saada N, Turlapati L, Worley KC, Wu YQ, Yamamoto A, Zhu Y, Bergman CM, Thornton KR, Mittelman D, Gibbs RA. The Drosophila melanogaster Genetic Reference Panel. Nature. 2012 Feb 8;482(7384):173-8. doi: 10.1038/nature10811. PMID: 22318601; PMCID: PMC3683990.

[21] Guirao-Rico S, González J. Benchmarking the performance of Pool-seq SNP callers using simulated and real sequencing data. Mol Ecol Resour. 2021 May;21(4):1216-1229. doi: 10.1111/1755-0998.13343. Epub 2021 Mar 5. PMID: 33534960; PMCID: PMC8251607.

[22] Kim BY, Gellert HR, Church SH, Suvorov A, Anderson SS, Barmina O, Beskid SG, Comeault AA, Crown KN, Diamond SE, Dorus S, Fujichika T, Hemker JA, Hrcek J, Kankare M, Katoh T, Magnacca KN, Martin RA, Matsunaga T, Medeiros MJ, Miller DE, Pitnick S, Simoni S, Steenwinkel TE, Schiffer M, Syed ZA, Takahashi A, Wei KH, Yokoyama T, Eisen MB, Kopp A, Matute D, Obbard DJ, O'Grady PM, Price DK, Toda MJ, Werner T, Petrov DA. Single-fly assemblies fill major phylogenomic gaps across the Drosophilidae Tree of Life. bioRxiv [Preprint]. 2023 Oct 2:2023.10.02.560517. doi: 10.1101/2023.10.02.560517. Update in: PLoS

Biol. 2024 Jul 18;22(7):e3002697. doi: 10.1371/journal.pbio.3002697. PMID: 37873137; PMCID: PMC10592941.

[23] Zeng K, Fu YX, Shi S, Wu CI. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics. 2006 Nov;174(3):1431-9. doi: 10.1534/genetics.106.061432. Epub 2006 Sep 1. PMID: 16951063; PMCID: PMC1667063.

Figure Captions

Figure 1. Scatter plots obtained with the R script ***npstat_plot_windows.R***. The plost shows different estimates of nucleotide variability using non-overlapped windows of physical size 1e5 bp across the chromosome 2L in a pool of 15 *D. melanogaster* inbreed lines. (A) Levels of nucleotide variability of Pi (black) and Watterson theta (read) at windows for example 1. (B) Levels of nucleotide variability of Pi (black), Watterson theta (red) and theta H (blue) at windows for example 2.

Figure 2. Violin and box plots showing the distribution of the nucleotide variability, and using the same data and the same output than in Figure 1 (see Example section). (A) This plot shows the distributions of nucleotide variability for Pi (red) and Theta Watterson (cyan) analysed in the example 1. (B) Plot showing the distributions of nucleotide variability of Pi (red), Theta Watterson (green) and theta H (blue) analysed in the example 2. (C) Distribution of the neutrality tests Tajima's D analysed in the example 1. (D) Distribution of the neutrality tests Tajima's D (cyan) and Fay and Wu's H (red) analysed in the example 2. Each plot includes data points, box plots and density curves.

Table Captions

Table 1. List of options available to run *npstat*

Table 2. Information and statistics obtained per window (rows) in the *npstat* output file.

Table 3. *npstat* output file for example 3. The first line shows the name of each column, and each of the next rows show the values for each window analyzed by *npstat*.

Table 1.

Option	**Value**	**Definition**
-n	sample_size	haploid sample size. (only allowed even sample sizes)
-l	window_length	window length in base pairs.
-mincov	minimum coverage	Minimum coverage filter (default 4).
-maxcov	maximum coverage	Maximum coverage filter (default 100).
-minqual	minimum base quality	Base quality filter (default 10)
-nolowfreq	m	Minimum allele count filter > m (default 1).
-outgroup	file.fa	outgroup file in FASTA.
-snpfile	file.snp	consider SNPs only if present in file.snp.
-annot	file.gff3	annotation file in GFF3 format.

Table 2.

Column number	**Definition**
1	Window number
2	Number of base pairs covered by sequences
3	Number of bases covered and with known outgroup allele
4	Average read depth
5	Number of segregating sites S
6	Watterson estimator of θ
7	Tajima's Π estimator of heterozygosity
8	Tajima's D
9	Unnormalized Fay and Wu's H
10	Normalized Fay and Wu's H
11	Variance of the number of segregating sites
12	Variance of the Watterson estimator,
13	Divergence per base (from outgroup),
14	Nonsynonimous polymorphisms,
15	Synonimous polymorphisms,
16	Nonsynonimous divergence,
17	Synonimous divergence,
18	α (fraction of substitutions fixed by positive selection)

Table 3.

window	length	length_outgroup	read_depth	S	Watterson	Pi	Tajima_D	var_S	var_Watterson	unnorm_FayWu_H	FayWu_H	div	nonsyn_pol	syn_pol	nonsyn_div	syn_div	alpha
1	90183	52920	43.977446	362	0.001466	0.001423	-0.155877	23936.069934	0.000000	0.000640	0.515096	0.073375	21	28	624	1393	-0.674279
2	99953	72054	43.760267	359	0.001310	0.001207	-0.421497	23577.520614	0.000000	0.000418	0.490813	0.071613	37	31	792	2100	-2.164712
3	100000	77130	45.281160	529	0.001908	0.001889	-0.052169	50169.640868	0.000001	0.000653	0.570728	0.062868	37	89	654	2199	-0.397846
4	90786	63662	42.475800	722	0.002967	0.003052	0.154076	95863.629616	0.000002	0.000972	0.505745	0.056517	21	55	319	794	0.049644
5	99771	79387	44.565575	944	0.003445	0.003414	-0.046439	160219.351405	0.000002	0.000957	0.489224	0.052792	38	112	382	1420	-0.261219

6	99995	80622	43.591280	1768	0.006463	0.006447	-0.013118	566276.275348	0.000008	0.000828	0.230309	0.048113	99	317	398	1315	-0.031855
7	91088	72905	41.242085	1445	0.005928	0.005854	-0.068949	388685.956111	0.000007	0.000633	0.193967	0.040560	50	114	167	519	-0.363063
8	98257	78848	41.738563	1818	0.006907	0.006914	0.006187	611554.241101	0.000009	0.001211	0.320836	0.043907	39	163	160	648	0.030982
9	99949	78544	44.269908	1899	0.006920	0.006935	0.011452	648435.602548	0.000009	0.001141	0.285613	0.052467	93	334	483	1477	0.148529
10	89248	64929	42.292892	2435	0.010134	0.009982	-0.082024	1089674.657138	0.000019	0.004013	0.633532	0.037302	23				

Figure 1

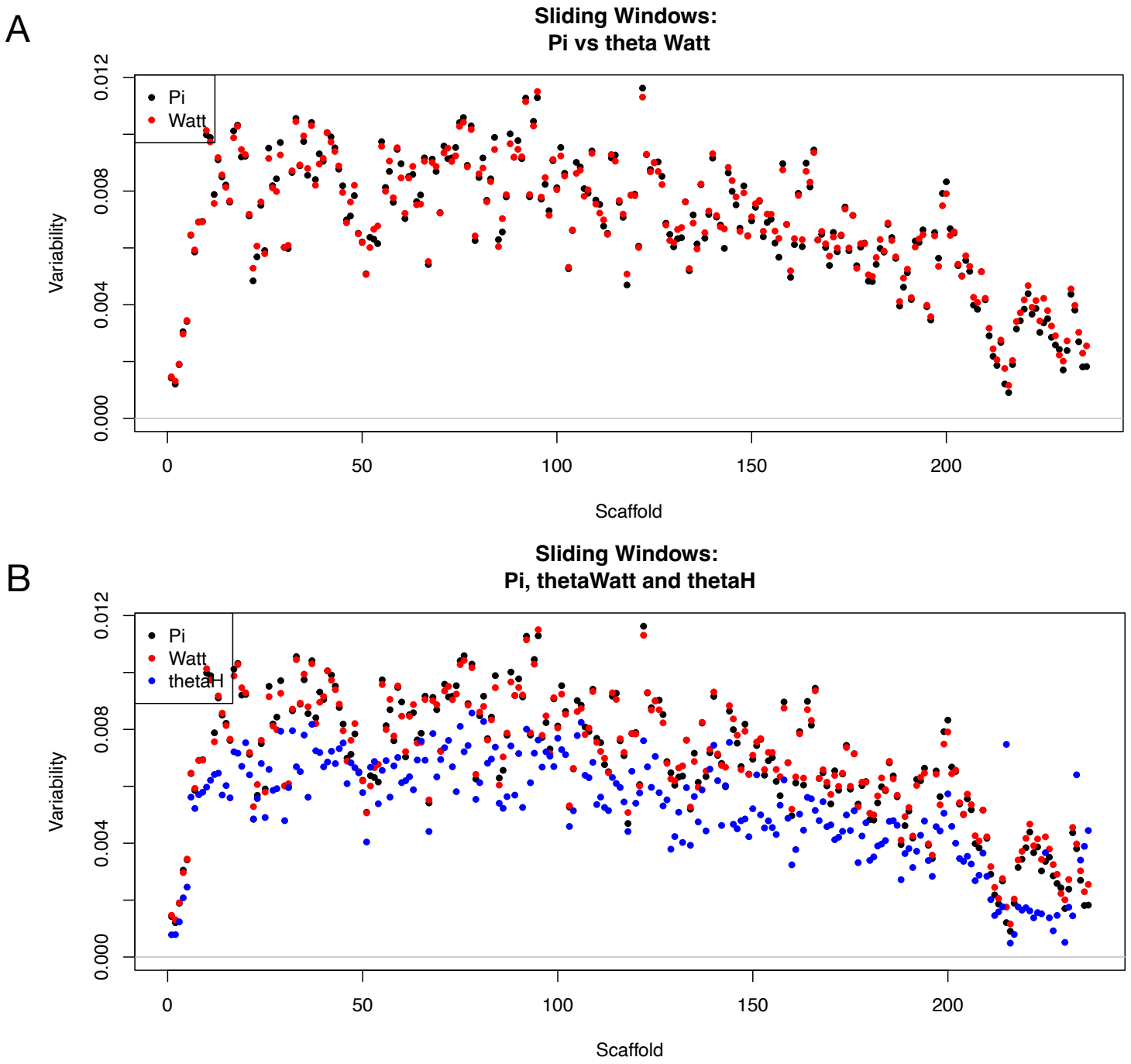
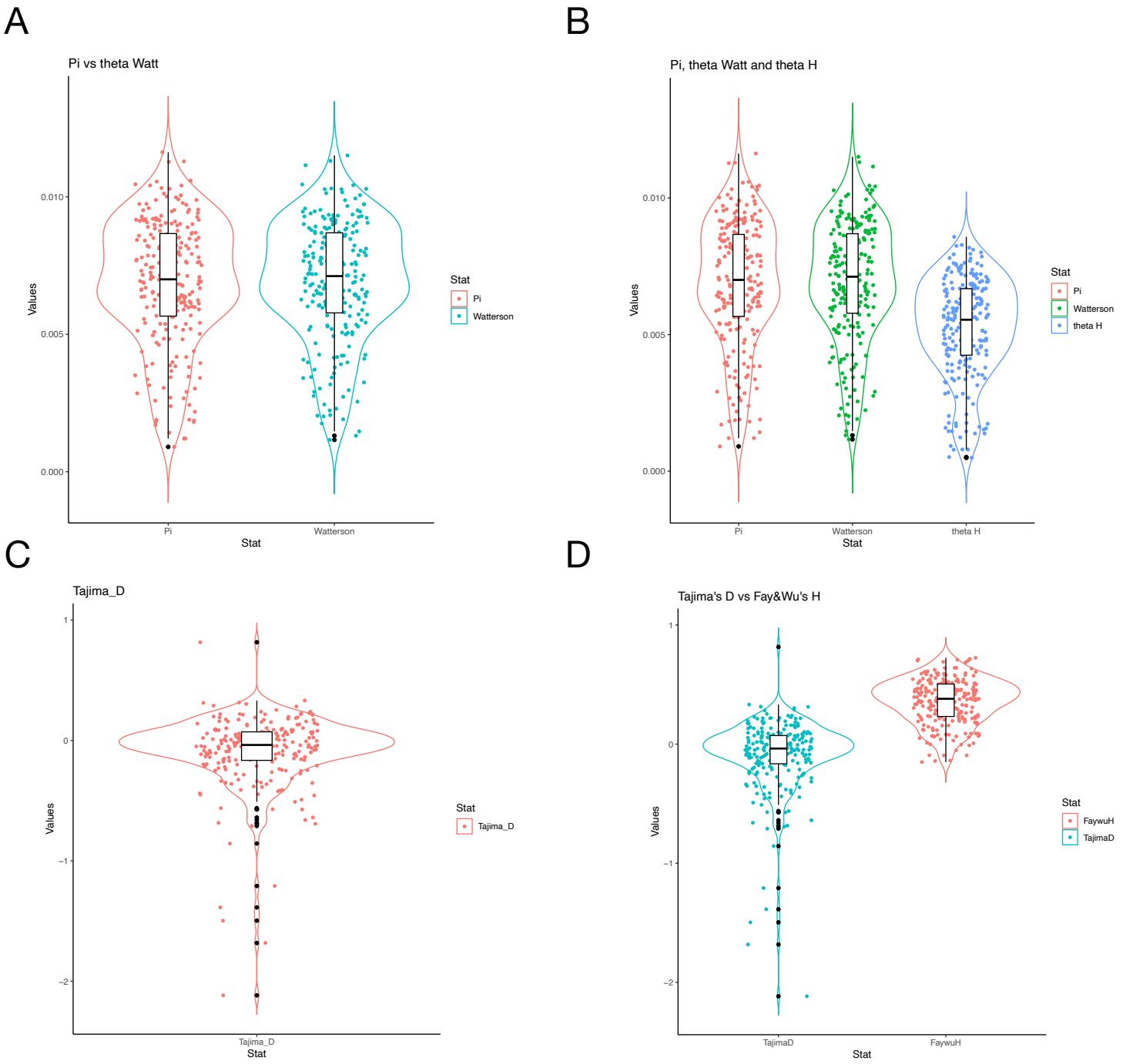

A
Sliding Windows:
Pi vs theta Watt
Pi
Watt
Variability
0.000
0.004
0.008
0.012
0
50
100
150
200
Scaffold
B
Sliding Windows:
Pi, thetaWatt and thetaH
Pi
Watt
thetaH
Variability
0.000
0.004
0.008
0.012
0
50
100
150
200
Scaffold

Figure 2


A
Pi vs theta Watt
Values
0.010
0.005
0.000
Pi
Watterson
Stat
Stat
Pi
Watterson
B
Pi, theta Watt and theta H
Values
0.010
0.005
0.000
Pi
Watterson
theta H
Stat
Stat
Pi
Watterson
theta H
C
Tajima_D
Values
1
0
−1
−2
Tajima_D
Stat
Stat
Tajima_D
D
Tajima's D vs Fay&Wu's H
Values
1
0
−1
−2
TajimaD
FaywuH
Stat
Stat
FaywuH
TajimaD