

Boquila: NGS read simulator to eliminate read nucleotide bias in sequence analysis

Umit Akkose¹, Ogun Adebali^{1,2*}

¹ Faculty of Engineering And Natural Sciences, Sabanci University, 34956 Istanbul, Turkey

² TÜBİTAK Research Institute for Fundamental Sciences, 41470 Gebze, Turkey

* To whom correspondence should be addressed: oadebali@sabanciuniv.edu

Abstract

Sequence content is heterogeneous throughout genomes. Therefore, Genome-wide NGS reads biased towards specific nucleotide profiles are affected by the genome-wide heterogeneous nucleotide distribution. Boquila generates sequences that mimic the nucleotide profile of true reads, which can be used to correct the nucleotide-based bias of genome-wide distribution of NGS reads. Boquila can be configured to generate reads from only specified regions of the reference genome. It also allows the use of input DNA sequencing to correct the bias due to the copy number variations in the genome. Boquila uses standard file formats for input and output data, and it can be easily integrated into any workflow for high-throughput sequencing applications.

Introduction

Simulating genomic data for benchmarking bioinformatics programs has become increasingly popular, particularly for read alignment, genome assembly, and variant and RNA-seq analysis [1]. Using such an approach allows for systematic performance assessment even in the absence of gold-standard data. Most currently available simulation tools are heavily geared towards benchmarking; they concentrate on generating reads produced by a specific sequencing experiment by modeling the characteristics of reads produced by sequencing machinery. Consequently, the correction metrics are primarily associated with artificial errors commonly introduced by these specific sequencing protocols.

Although most of the tools use some profile for simulating, these profiles are used for simulating characteristics of sequencing protocols rather than the biological experiments. No simulation tool utilizes nucleotide content profile to our knowledge. SomatoSim [2], VarSim [3], SimuSCoP [4], and many other tools [5–9] were specifically designed to simulate genomic variation. ART [10] and SInC [5] generate profiles based on specific error models and quality score distribution extracted from empirical data. pIRS [11] and Mitty [12] generate quality profiles based on mapped reads and empirical data. NanoSim [13], a nanopore sequence simulator, also uses error profiles and length distributions. Gargammel, ancient DNA sequencing simulator, uses sequencing errors and quality profiles can model base compositions. It can mimic UV damage by adding deamination. However, it is specifically designed for simulation of ancient DNA sequences [14]. BEAR [15], focused on metagenomics, generates error, quality, and abundance profiles. However, the nucleotide content of the reads could be biased for

several reasons. First, if sequence library preparation involves immunoprecipitation, 24
antibodies might be biased towards pulling down specific nucleotide profiles. Moreover, 25
ligation efficiencies could be different across varying nucleotides on both 5' and 3' ends, 26
which would result in some nucleotide enrichment at the read ends. Furthermore, the 27
nature of sequencing technology might supposedly yield a particular biased nucleotide 28
profile. For instance, sequencing methods yielding the maps of UV damage naturally 29
result in dipyrimidine-enriched reads [16,17]. Additionally, the GC content of the reads 30
might vary depending on the sequencing platform [18]. Finally, the PCR step might 31
introduce another nucleotide bias due to differential efficiencies of universal primers 32
towards specific nucleotides [19]. Considering such factors that could affect the genomic 33
distribution of the reads, there is a clear need for a sequencing read simulation tool that 34
utilizes nucleotide content profile although simulated reads can also be used to eliminate 35
nucleotide content bias in experiments whose results are affected by nucleotide content. 36
Simulation tools mentioned above can account for error and quality profiles of 37
sequencing platforms and GC content biases. However, most of them were designed to 38
simulate reads based on sequencing instruments. There was no other option if aimed to 39
generate a simulated dataset that mimics the nucleotide content of input reads. Here we 40
present boquila, a nucleotide content-based NGS read simulator that can produce 41
simulated reads with a similar nucleotide content profile to input reads. They can be 42
used to normalize the nucleotide content bias in actual reads. 43

Genomic regions with higher copy numbers have a greater chance of being pulled 44
down during library preparation, whereas those with lower copy numbers are harder to 45
detect. Boquila can also use data from input sequencing as input while generating 46
simulated reads. With this approach, we can also use generated simulated reads to 47
normalize the impacts of copy number variations. 48

Features and methods

Boquila was specifically developed to generate synthetic reads with the same nucleotide 50
content as input reads. It operates with FASTA or FASTQ files as an input and 51
generates reads according to the nucleotide content of the input. The number of 52
generated reads and their length distribution will equal the input reads. The nucleotide 53
profile can be calculated based on user-defined kmer length or single nucleotides. 54
Boquila can use the entire genome or pre-defined genomic intervals while randomly 55
selecting reads from the reference genome, thus providing fine-grained control over the 56
regions where simulated reads are generated. Alternative to the reference genome, input 57
sequencing reads can also be used, if a user has raw genome sequence data as a control 58
retrieved from the same experimental setup (cell type, conditions, etc). In this case, 59
reads are randomly generated from input sequencing reads. When generating simulated 60
reads, the nucleotide profile obtained from input reads is adjusted dynamically based on 61
the nucleotide profile obtained so far from simulated reads. In this manner, simulated 62
reads can be further conformed to input reads. 63

Simulated reads are exported in FASTA or FASTQ format based on the input read 64
format. Simulated reads can also be exported in BED format. Quality scores are copied 65
over from input reads if input reads are present in FASTQ format. FASTA and FASTQ 66
are standard formats for high-throughput sequencing reads, making boquila easier to 67
integrate into existing workflows. Additionally, obtaining output in BED format can 68
help bypass alignment, which is one of the most time- and resource-intensive steps in 69
any NGS workflow. 70

Results

71
72
73
74
75
76
77
78
79
80

We used boquila to generate simulated reads for the data sets of two published studies: XR-seq data from [20] and damage-seq data from [21]. Simulated reads have the same nucleotide content as true reads Fig 1B (XR-seq Hotelling's T-square: 0.25, p-value: 0.99, damage-seq Hotelling's T-square: 5.2, p-value: 0.4), but they are randomly generated from the genome. Both XR-seq and damage-seq reads have a certain nucleotide bias due to the UV-induced damage site. The simulated reads generated by boquila can be used to normalize the true repair signal and the true damage signal obtained from XR-seq and damage-seq, respectively Fig 1C. Using this method, we can eliminate the potential bias of the nucleotide content in subsequent analyses.

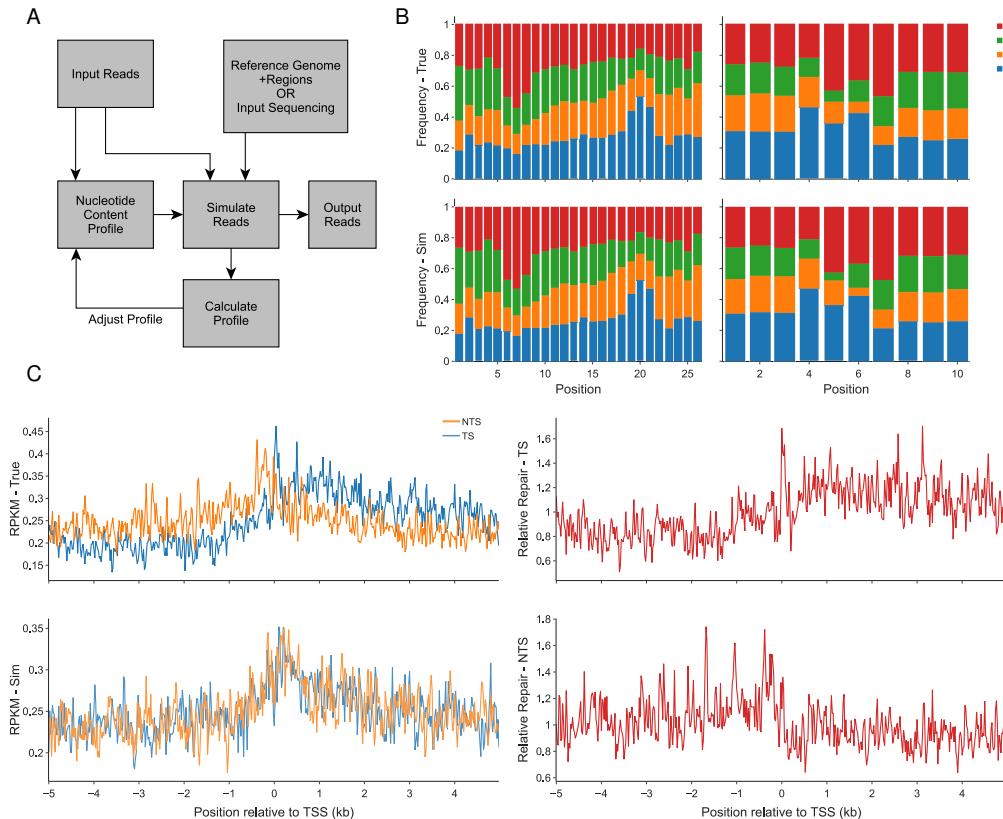


Fig 1. A: Overall workflow of boquila. B: Nucleotide frequency of simulated and true reads for XR-seq (left) and damage-seq (right). XR-seq reads are enriched with thymines at 19-21st positions of the reads, whereas centered damage-seq reads are enriched with pyrimidines at 5-6th positions. C: UV-induced repair (XR-seq) profiles around transcription start sites (TSS). Due to transcription-coupled repair (TCR), transcribed strand (TS) is expected to have higher repair signals relative to the non-transcribed strand (NTS). Whether sequence bias affect the TCR profile around TSS is investigated with the simulated reads (bottom). The profiles showing observed/expected (due to sequence context) ratios (on the right) indicate the TS (top) and NTS (bottom) TCR profiles.

Discussion

81
82
83
84
85
86
87
88
89
90
91

In this study, we propose a novel sequencing simulator utilizing nucleotide content. Some simulators can account for bias produced by sequencing machinery and GC content. Furthermore, some focused on introducing genomic variation into simulated reads, Table 1. However, no simulation tool can yield random reads reflecting the nucleotide content profile of the input reads. Alternative to the reference genome, boquila can also use input sequencing data during simulation to account for CNV bias, which most other simulation tools neglect. With the development of NGS technologies for each specific research question, specialized correction tools must be used to remove technical artifacts of the methodology. In that respect, boquila fills an essential gap in NGS analysis.

Table 1. Brief summary of existing simulation tools

Simulator	Output Layout	Output Format	GC bias	CNV bias	Nucleotide Content	Genomic Variation	Simulated Profile
MetaSim [22]	SE	FA					empirical error, genome abundance profile
Mason [23]	SE,PE	FA,FQ					empirical error and quality profiles
BEERS [24]	PE	FA					intron signal profile, error profile
ART [10]	SE,PE	FQ,SAM					empirical length, error, quality profiles
GemSIM [25]	SE,PE	FQ,SAM	✓			✓	empirical quality and error profiles
Grinder [26]	SE,PE	FQ,FA		✓			position dependent error profile
pIRS [11]	PE	FQ	✓			✓	empirical base-calling, gc%-depth profiles
Wessim [27]	SE,PE	FQ,SAM	✓				fragment generation models
NeSSM [28]	SE,PE	FQ		✓			error and coverage profiles
xs [29]	SE,PE	FQ					read-length and quality profiles
SInC [5]	SE,PE	FQ				✓	quality and error profiles
CuReSim [30]	SE	FQ		✓		✓	error profile
FASTQSim [31]	SE	FQ					read-length, quality and error profiles
BEAR [15]	SE,PE	FQ					empirical length, error, quality profiles
VarSim [3]		FQ				✓	mutation profile
SCNVSim [6]	SE,PE	FA,VCF				SV,CNV	-
IntSim [7]	SE,PE	FQ	✓			✓	read-length, quality and error profiles
tHapMix [8]		BAM				✓	error profile
gargammel [14]	SE,PE	FQ	✓				damage, fragmentation, contamination profiles
NEAT [32]	SE,PE	FQ	✓			✓	read-length, quality and error profiles
NanoSim [13]	SE	FQ					error profile, length distribution profile
Pysim-sv [9]	SE,PE	FQ	✓			✓	-
Xome-Blender [33]	PE	BAM				✓	-
Mitty [12]	SE,PE	FQ, VCF	✓			✓	empirical quality and length profiles
PASS [34]	SE,PE	FQ					read-length and error profiles
SimuSCoP [4]	SE,PE	FQ		✓		✓	quality and error profiles
PBSIM2 [35]	SE	FQ					quality and error profiles
SomatoSim [2]	PE	SAM,BAM				SNV	error profile
boquila	SE	FQ,FA,BED	✓	✓	✓		nucleotide content profile

SE: single end, PE: paired end, FA: fasta, FQ: fastq, SNV: single nucleotide variation, CNV: copy number variation, SV: structural variation.

Availability

Boquila is written in Rust and freely available at
<https://github.com/CompGenomeLab/boquila>.

92
93
94

Funding

This publication has been produced benefitting from the 2232 International Fellowship for Outstanding Researchers Program of TUBITAK (Project No: 118C320 to OA). UA and OA are supported by EMBO Installation Grant (Grant No: 4163 to OA) funded by TUBITAK.

95
96
97
98
99

References

1. Mangul S, Martin LS, Hill BL, Lam AKM, Distler MG, Zelikovsky A, et al. Systematic benchmarking of omics computational tools. *Nature Communications*. 2019;10(1):1393. doi:10.1038/s41467-019-09406-4.
2. Hawari MA, Hong CS, Biesecker LG. SomatoSim: precision simulation of somatic single nucleotide variants. *BMC Bioinformatics*. 2021;22(1):109. doi:10.1186/s12859-021-04024-8.
3. Mu JC, Mohiyuddin M, Li J, Bani Asadi N, Gerstein MB, Abyzov A, et al. VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications. *Bioinformatics*. 2015;31(9):1469–1471. doi:10.1093/bioinformatics/btu828.
4. Yu Z, Du F, Ban R, Zhang Y. SimuSCoP: reliably simulate Illumina sequencing data based on position and context dependent profiles. *BMC Bioinformatics*. 2020;21(1):331. doi:10.1186/s12859-020-03665-5.
5. Pattnaik S, Gupta S, Rao AA, Panda B. SInC: an accurate and fast error-model based simulator for SNPs, Indels and CNVs coupled with a read generator for short-read sequence data. *BMC Bioinformatics*. 2014;15(1):40. doi:10.1186/1471-2105-15-40.
6. Qin M, Liu B, Conroy JM, Morrison CD, Hu Q, Cheng Y, et al. SCNVSim: somatic copy number variation and structure variation simulator. *BMC Bioinformatics*. 2015;16(1):66. doi:10.1186/s12859-015-0502-7.
7. Yuan X, Zhang J, Yang L. IntSIM: An Integrated Simulator of Next-Generation Sequencing Data. *IEEE Transactions on Biomedical Engineering*. 2017;64(2):441–451. doi:10.1109/TBME.2016.2560939.
8. Ivakhno S, Colombo C, Tanner S, Tedder P, Berri S, Cox AJ. tHapMix: simulating tumour samples through haplotype mixtures. *Bioinformatics*. 2017;33(2):280–282. doi:10.1093/bioinformatics/btw589.
9. Xia Y, Liu Y, Deng M, Xi R. Pysim-sv: a package for simulating structural variation data with GC-biases. *BMC Bioinformatics*. 2017;18(3):53. doi:10.1186/s12859-017-1464-8.
10. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. *Bioinformatics (Oxford, England)*. 2012;28(4):593–594. doi:10.1093/bioinformatics/btr708.

11. Hu X, Yuan J, Shi Y, Lu J, Liu B, Li Z, et al. pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics*. 2012;28(11):1533–1535. doi:10.1093/bioinformatics/bts187.
12. Ghose K, Caner, Ozemsbg. sbg/Mitty: v2.29.0; 2018. Available from: <https://doi.org/10.5281/zenodo.1479963>.
13. Yang C, Chu J, Warren RL, Birol I. NanoSim: nanopore sequence read simulator based on statistical characterization. *GigaScience*. 2017;6(4):gix010. doi:10.1093/gigascience/gix010.
14. Renaud G, Hanghøj K, Willerslev E, Orlando L. gargammel: a sequence simulator for ancient DNA. *Bioinformatics*. 2017;33(4):577–579. doi:10.1093/bioinformatics/btw670.
15. Johnson S, Trost B, Long JR, Pittet V, Kusalik A. A better sequence-read simulator program for metagenomics. *BMC Bioinformatics*. 2014;15(9):S14. doi:10.1186/1471-2105-15-S9-S14.
16. Hu J, Lieb JD, Sancar A, Adar S. Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution. *Proceedings of the National Academy of Sciences*. 2016;113(41):11507 LP – 11512. doi:10.1073/pnas.1614430113.
17. Mao P, Smerdon MJ, Roberts SA, Wyrick JJ. Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution. *Proceedings of the National Academy of Sciences*. 2016;113(32):9057 LP – 9062. doi:10.1073/pnas.1606667113.
18. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. *Genome Biology*. 2013;14(5):R51. doi:10.1186/gb-2013-14-5-r51.
19. Polz MF, Cavanaugh CM. Bias in Template-to-Product Ratios in Multitemplate PCR. *Applied and Environmental Microbiology*. 1998;64(10):3724–3730. doi:10.1128/AEM.64.10.3724-3730.1998.
20. Adar S, Hu J, Lieb JD, Sancar A. Genome-wide kinetics of DNA excision repair in relation to chromatin state and mutagenesis. *Proceedings of the National Academy of Sciences*. 2016;113(15):E2124 LP – E2133. doi:10.1073/pnas.1603388113.
21. Hu J, Adebali O, Adar S, Sancar A. Dynamic maps of UV damage formation and repair for the human genome. *Proceedings of the National Academy of Sciences*. 2017;114(26):6758 LP – 6763. doi:10.1073/pnas.1706522114.
22. Richter DC, Ott F, Auch AF, Schmid R, Huson DH. MetaSim—A Sequencing Simulator for Genomics and Metagenomics. *PLOS ONE*. 2008;3(10):e3373.
23. Holtgrewe M. Mason: A Read Simulator for Second Generation Sequencing Data. Technical Report FU Berlin. 2010;.
24. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, et al. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*. 2011;27(18):2518–2528. doi:10.1093/bioinformatics/btr427.

25. McElroy KE, Luciani F, Thomas T. GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics*. 2012;13(1):74. doi:10.1186/1471-2164-13-74.
26. Angly FE, Willner D, Rohwer F, Hugenholtz P, Tyson GW. Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Research*. 2012;40(12):e94–e94. doi:10.1093/nar/gks251.
27. Kim S, Jeong K, Bafna V. Wessim: a whole-exome sequencing simulator based on in silico exome capture. *Bioinformatics*. 2013;29(8):1076–1077. doi:10.1093/bioinformatics/btt074.
28. Jia B, Xuan L, Cai K, Hu Z, Ma L, Wei C. NeSSM: A Next-Generation Sequencing Simulator for Metagenomics. *PLOS ONE*. 2013;8(10):e75448.
29. Pratas D, Pinho AJ, O S Rodrigues JM. XS: a FASTQ read simulator. *BMC Research Notes*. 2014;7(1):40. doi:10.1186/1756-0500-7-40.
30. Caboche S, Audebert C, Lemoine Y, Hot D. Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data. *BMC Genomics*. 2014;15(1):264. doi:10.1186/1471-2164-15-264.
31. Shcherbina A. FASTQSim: platform-independent data characterization and in silico read generation for NGS datasets. *BMC Research Notes*. 2014;7(1):533. doi:10.1186/1756-0500-7-533.
32. Stephens ZD, Hudson ME, Mainzer LS, Taschuk M, Weber MR, Iyer RK. Simulating Next-Generation Sequencing Datasets from Empirical Mutation and Sequencing Models. *PLOS ONE*. 2016;11(11):e0167047.
33. Semeraro R, Orlandini V, Magi A. Xome-Blender: A novel cancer genome simulator. *PLOS ONE*. 2018;13(4):e0194472.
34. Zhang W, Jia B, Wei C. PaSS: a sequencing simulator for PacBio sequencing. *BMC Bioinformatics*. 2019;20(1):352. doi:10.1186/s12859-019-2901-7.
35. Ono Y, Asai K, Hamada M. PBSIM2: a simulator for long-read sequencers with a novel generative model of quality scores. *Bioinformatics*. 2021;37(5):589–595. doi:10.1093/bioinformatics/btaa835.