

PCNE: A Tool for Plasmid Copy Number Estimation

Bioinformatics and Biology Insights
Volume 20: 1–7
© The Author(s) 2026
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11779322251410037
journals.sagepub.com/home/bbi



Riccardo Bollini^{1,2}  and Valeria Cento^{2,3} 

Abstract

The identification of plasmids from assembled genomes is well supported by numerous different tools, yet very few incorporate a plasmid copy number estimation step. This limits a comprehensive plasmid analysis, often leaving researchers to perform copy number estimation independently, leading to a lack of standardization. Plasmid Copy Number Estimator (PCNE) addresses this by providing an accessible and versatile command-line tool for estimating plasmid copy numbers directly from short-read sequencing data. Starting from standard input data like raw reads and a genome assembly, PCNE allows to apply a flexible normalization strategy, including an optional GC-bias correction, and is designed to complement existing plasmid detection pipelines. By simplifying and standardizing copy number estimation, PCNE, through the integration of state-of-art methodologies, aims to empower researchers to gain deeper insights into plasmid biology, particularly in studies of antimicrobial resistance and horizontal gene transfer.

Keywords

Plasmid copy number, software, microbiology, bioinformatics

Received: 26 August 2025; accepted: 7 December 2025

Introduction

Plasmids are key drivers of bacterial evolution, mediating horizontal gene transfer¹ and playing crucial roles in virulence and antibiotic resistance,² posing significant clinical challenges. In this context, an important role is mediated by their copy number, representing an important evolutionary parameter, influencing gene dosage, expression levels, fitness costs, and stability.³ Indeed, high-copy-number plasmids can amplify resistance or virulence phenotypes,⁴ while low-copy-number plasmids may exhibit tighter host control and long-term persistence with minimal burden.^{5,6} The widespread adoption of sequence-based methods, in both research and clinical settings, has led to the development of several tools^{7–9} to identify plasmids, from both short and long reads, but the estimation of plasmid copy number remains underexplored, with no standardized tool existing to perform plasmid copy number (PCN) estimation. Consequently, few genomic studies report PCN alongside plasmid identification, unless performed via the standard quantitative polymerase chain reaction (qPCR) quantification, despite its importance for understanding plasmid stability, host burden, and gene dosage effects.³ To address this gap, we developed Plasmid Copy Number Estimator (PCNE), a tool designed to estimate PCNs directly from short-read whole-genome sequencing (WGS) data. The PCNE automates PCN estimation by integrating established bioinformatic tools into a cohesive, multi-step pipeline, facilitating its use for a wide range of users. It

requires standard inputs, such as a draft or complete genome assembly and short reads in paired-end format. The PCNE aims to complement well-established plasmid identification tools by providing a standardized yet versatile approach to PCN estimation.

Methods

Plasmid copy number estimator overview

The PCNE accepts 2 primary input configurations, shortly defined as mode 1 and mode 2. Mode 1 requires separate chromosome and plasmid FASTA files, while mode 2 requires the whole-assembled genome in FASTA format, and a file listing plasmid-contigs, with the option to also provide the list of chromosomal contigs. These 2 inputs ensure a straightforward integration with the common

¹Department of Public Health, Experimental and Forensic Medicine, University of Pavia, Pavia, Italy

²Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Italy

³Microbiology and Virology, IRCCS Humanitas Research Hospital, Rozzano, Italy

Corresponding author:

Riccardo Bollini, Department of Public Health, Experimental and Forensic Medicine, University of Pavia, Pavia 27100, Italy.
Email: riccardo.bollini@hunimed.eu



outputs of plasmid identification tools. In both cases, short reads in paired-end format are required.

Normalization and bias correction

The PCNE provides 1 normalization method and an optional correction feature to ensure precise PCN estimates. The default normalization is Whole Chromosome based, which calculates the ratio between the median plasmid coverage depth (D_p) and the median coverage depth of chromosomal contigs (D_{chr}), as described by the following equations:

$$\text{Plasmid Copy Number} = \frac{D_p}{D_{chr}}$$

The way D_p is calculated depends on the parameter `–single-plasmid`, which if selected treats all contigs determined as plasmid-borne as part of the same plasmid, and performs a length-weighted mean depth:

$$D_p = \frac{\sum_{p=1}^n L_p \cdot D_p}{\sum_{p=1}^n L_p}$$

where L_p is the length of plasmid contig p , and D_p is the median depth of plasmid contig p . If `–single-plasmid` is not selected, each contig represents a unique plasmid, and the median window-based depth is calculated:

$$D_p = \text{median} \{d_i \mid i \in W_p\}$$

where W_p is the selected window for contig p , and d_i is the window's depth. Chromosomal depth (D_{chr}) is always computed as:

$$D_{chr} = \text{median} \{d_i \mid i \in W_{chr}\}$$

This approach was selected as the default due to its simplicity and robustness, as it utilizes all available chromosomal data to establish a comprehensive coverage baseline.

The GC content variations are known to introduce systematic biases in sequencing coverage when using short-read sequencer,¹⁰ especially when GC content exceeds the 45% to 65% values.¹¹ This systematic bias can distort coverage depth and, consequently, affect the accuracy of PCN estimates. To mitigate this bias, PCNE implements an optional GC-bias correction module. When applied, it performs a Locally Estimated Scatterplot Smoothing (LOESS) regression to model the relationship between GC content and observed mean depth of chromosomal windows (Supplemental Figure S1). If not otherwise specified (`–gc-frac`), the LOESS span parameter α (also called the smoothing fraction) is selected automatically by k-fold cross-validation with $k=5$. The best α is defined as the one with the lowest Mean Squared Error (MSE). The LOESS model is then used to predict the expected depth based on

GC content and used to correct the observed depth of all windowed data, using the following equation:

$$d_i^* = d_i \cdot \frac{M}{d_i^{\wedge}}$$

with d_i^* being the GC-corrected depth, M the overall median depth, d_i the window's depth, and $d_i^{\wedge}$ the expected depth.

Plasmid copy number estimator workflow

The initial step involves the processing of short-read sequencing data. BWA¹² is used for indexing the reference FASTA file and aligning the reads. The resulting SAM file is then converted to BAM format, sorted, and indexed using samtools.¹³ User can apply optional filtering based on reads-mapping quality (`–mapping-quality`) and SAM flags (`–filter`). The reference FASTA file is then divided into windows (default: 1k bp), and for each window, both GC content and coverage depth are calculated. If no GC-correction is applied, then the plasmid and chromosomal depths are calculated, and the copy number is estimated by a ratio between the 2 depths. When GC-correction (`–gc-correction`) is applied, then the LOESS regression is applied on chromosomal windows, calculating the expected depth based on GC content. The expected depth is then used on all windowed data to adjust observed depth. The final step involves copy number estimation by calculating the ratio of the plasmid's median depth to the chromosomal median depth. The results are formatted and written to a tab-separated values (.tsv) file.

The complete workflow is summarized in Figure 1.

Plasmid copy number estimator validation

To evaluate the accuracy and versatility of PCNE, we conducted analyses using simulated sequencing data with known PCNs and on real data with PCN ground truth estimated via qPCR.

Simulated data. Reference sequences for *Klebsiella pneumoniae* subsp. *pneumoniae* HS11286 (GCF_000240185.1), specifically the chromosome (CP003200.1) and one of its native plasmids (CP003223.1), were downloaded from the National Center for Biotechnology Information (NCBI). Paired-end short reads (150 bp) were simulated using ART-illumina (v.2.5.8)¹⁴ with the MSv3 error profile, a mean fragment size of 500 bp (`–m 500`), and a standard deviation of 50 bp (`–s 50`). To introduce gc-bias, InSilicoSeq (v2.0.1)¹⁵ was used to simulate reads from the same 2 sequences, using the `–gc_bias` and `–model nextseq` parameters. A total of 6 distinct sets of reads were created with $1\times$, $2\times$, and $5\times$ of PCN, and a chromosome coverage depth of $50\times$. For each data set, PCNE was tested using 2 distinct parameters: (1) whole-chromosome normalization and (2) whole-chromosome normalization with GC-correction enabled. The accuracy and reliability of PCNE results were quantified using the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). To assess PCNE integration and

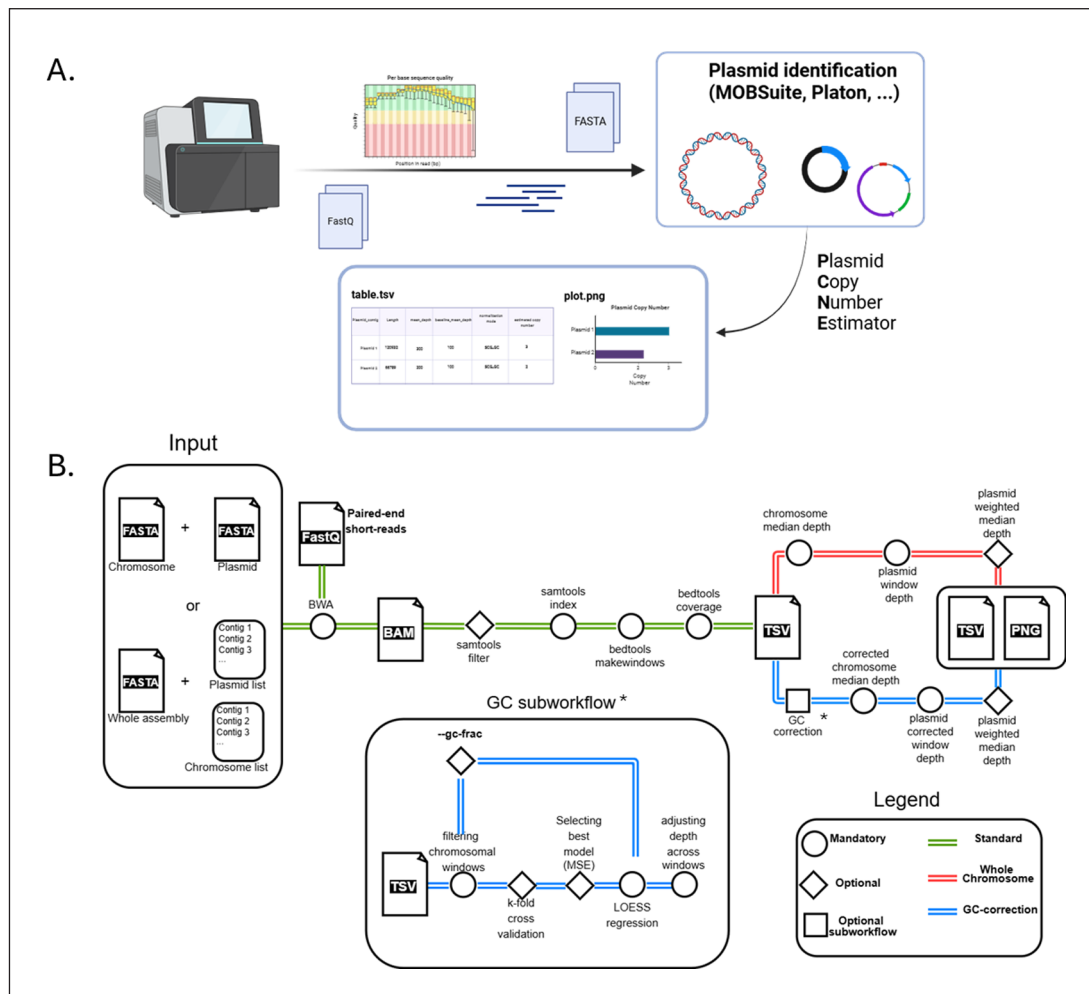


Figure 1. Plasmid copy number estimator workflow. (A) General workflow illustrating how PCNE integrates into a standard sequencing analysis. This panel was created with BioRender.com. (B) In detail, all steps performed by PCNE, including optional filters and optional sub workflow. This panel was created with Draw.io.

performance with already established tools, all the generated reads were assembled *de novo* using Shovill (v1.1.0) (<https://github.com/tseemann/shovill>) (–assembler spades), and the assemblies were processed independently by Platon (default parameters) (v1.7)⁹ and MOB-suite (mob_recon was used) (v3.1.9).⁸ The outputs from these tools were used directly as input for PCNE (–single-plasmid) to demonstrate workflow compatibility and assess performance consistency.

Real data. To validate PCNE on a real data set, raw reads from the Xiang et al, Emerg Microbes Infect. 2024, were retrieved from SRA (SRP473989), with data deposited under BioProject numbers PRJNA1044738. Raw reads were trimmed using fastP (v1.0.1)¹⁶ (–q 30) and checked for quality using fastQC (v0.12.1) (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Filtered reads were then assembled using Shovill (v1.1.0) (<https://github.com/tseemann/shovill>) (–assembler spades), and plasmids were identified using Platon (v1.7)⁹ and MOB-suite (mob_recon was used) (v3.1.9).⁸ Since MOB-suite failed to identify any plasmid, it was excluded from further analysis. Then, PCNE was run (–single-plasmid) on the generated output,

and the reliability of the outcome was evaluated using Pearson correlation, with qPCR results as reference.

Computational performance. The PCNE has been evaluated in terms of runtime (minutes) and maximum memory required (gigabytes) per run. To test the performance, 2 distinct groups were compared: the 36 samples from the simulated data set and the 14 samples from the real-life data set. Within each group, we compared the standard workflow with the GC-bias correction workflow. All analyses have been performed on HPC, running 1 job per isolate and giving 5 CPUs and 32 GB of memory per job.

Case study

We also tested PCNE in a real-life case study, where no qPCR data were available, to show its practical utility in a common scenario. To do this, we selected an outbreak study, involving hypervirulent and multidrug-resistant strains of *K. pneumoniae*, presented by Jian et al, Scientific Reports 2024. Raw reads were retrieved from SRA (SRP473988), with data deposited under BioProject

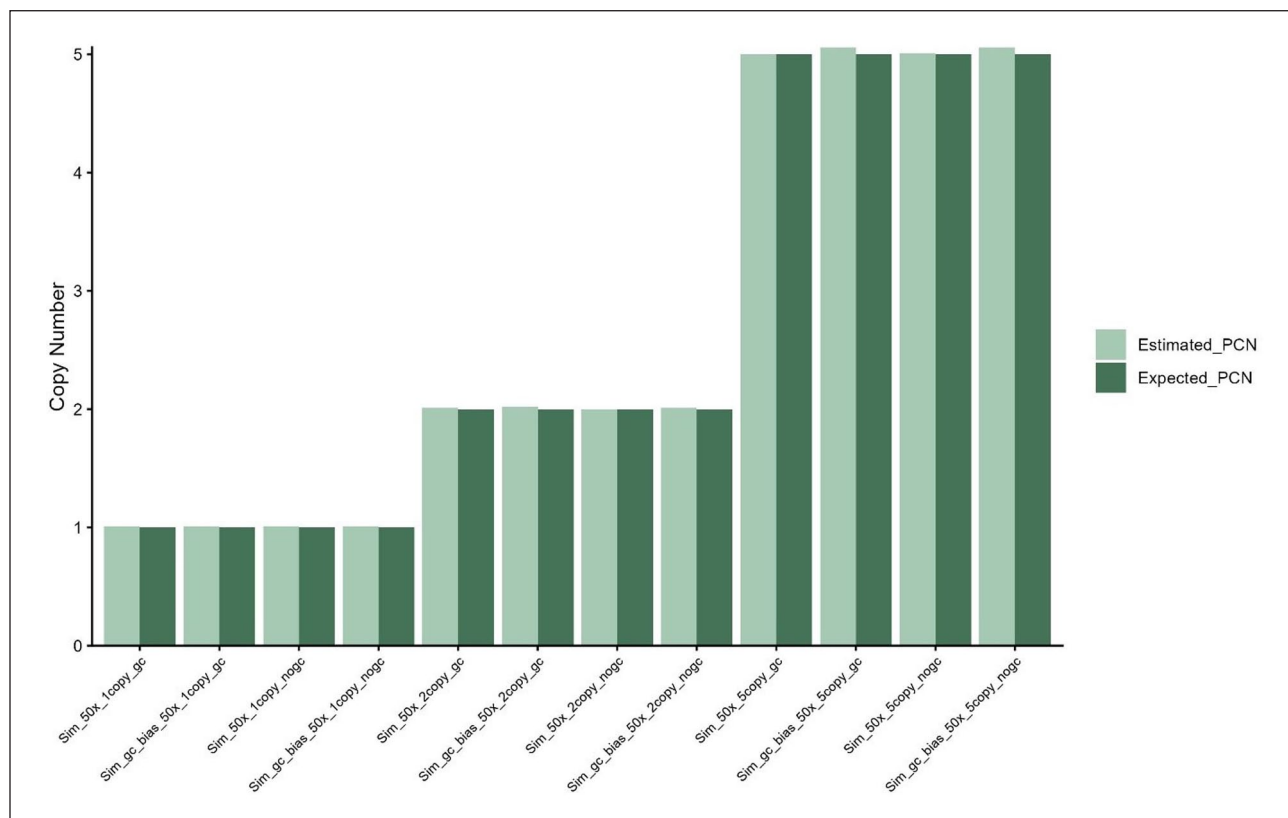


Figure 2. PCNE results on simulated data set.

Bar plot showing the expected and estimated plasmid copy numbers for simulated samples. Samples are ordered by the expected copy number (1, 2, and 5). The *_gc* suffix indicates samples processed with GC-correction, while *_nogc* indicates no GC-correction.

number PRJNA1043816. Reads were trimmed using fastP (v1.0.1)¹⁶ (-q 30). Filtered reads were then assembled using Shovill (v1.1.0) (-assembler spades), and plasmids were identified using MOB-suite (v3.1.9)⁸ (mob_recon was used). The PCNE was run (-single-plasmid) on the generated output. Resistance and virulence genes were detected with Abricate (v1.0.1) (<https://github.com/tseemann/abricate>), using the card¹⁷ and the vfdb¹⁸ databases, respectively.

Results and Discussion

Simulated data set

The PCNE's performance was first validated using simulated short-read data where the true PCN was known. We tested the PCNE across datasets with chromosomal coverage of 50× and PCNs of 1×, 2×, and 5×. Quantitative analysis revealed that PCNE consistently produced estimates close to the true simulated values (Figure 2), with an overall MAE of 0.017 ± 0.02 . There are no major differences between the application of GC-correction or not, with GC-correction yielding an overall MAE of 0.018 and no GC-correction yielding an overall MAE of 0.016. Data estimated from reads with GC bias were the least accurate (MAE=0.028 and RMSE=0.03), while those without GC bias yield an MAE of 0.006 and an RMSE of 0.008, with no differences between the application or not of GC-correction. A detailed table is available as Supplemental Table S1.

To simulate a standard workflow, the simulated reads were assembled and classified using Platon and MobSuite, separately. The PCNE was able to directly use the outputs from both tools to provide copy number estimates, confirming its easy integration into different analysis pipelines. Results were concordant for both tools, with an overall MAE of 0.08 ± 0.06 and an RMSE of 0.1 ± 0.01 and are summarized in Supplemental Table S2.

Real data set

All assemblies generated were of high quality. Platon correctly identified 1 AMR-plasmid per isolate, as expected. The PCNE was run with and without GC-correction, correctly identifying all isolates for which an increase in copy numbers was expected, with a Pearson's correlation of 0.88 (Figure 3). Regarding the accuracy of estimates, the overall MAE was 0.798 ± 1.06 , 0.15 for 1 copy isolates and 1.28 for the multiple copies isolates. The reason for the reduced accuracy is due to 2 isolates, the samples EM4N3 and EM4N4, for which the estimation was lower than expected. Given the high accuracy in all other simulated and real cases, a possible explanation relies on the qPCR results taken as reference and the WGS sequences used. Indeed, as the authors reported, the estimation was taken over several days, during which copy numbers varied. A hypothesis is that for the 2 discordant isolates, the sequence record does not correspond to the time point from which the qPCR data were reported in the main text of the reference manuscript but corresponds

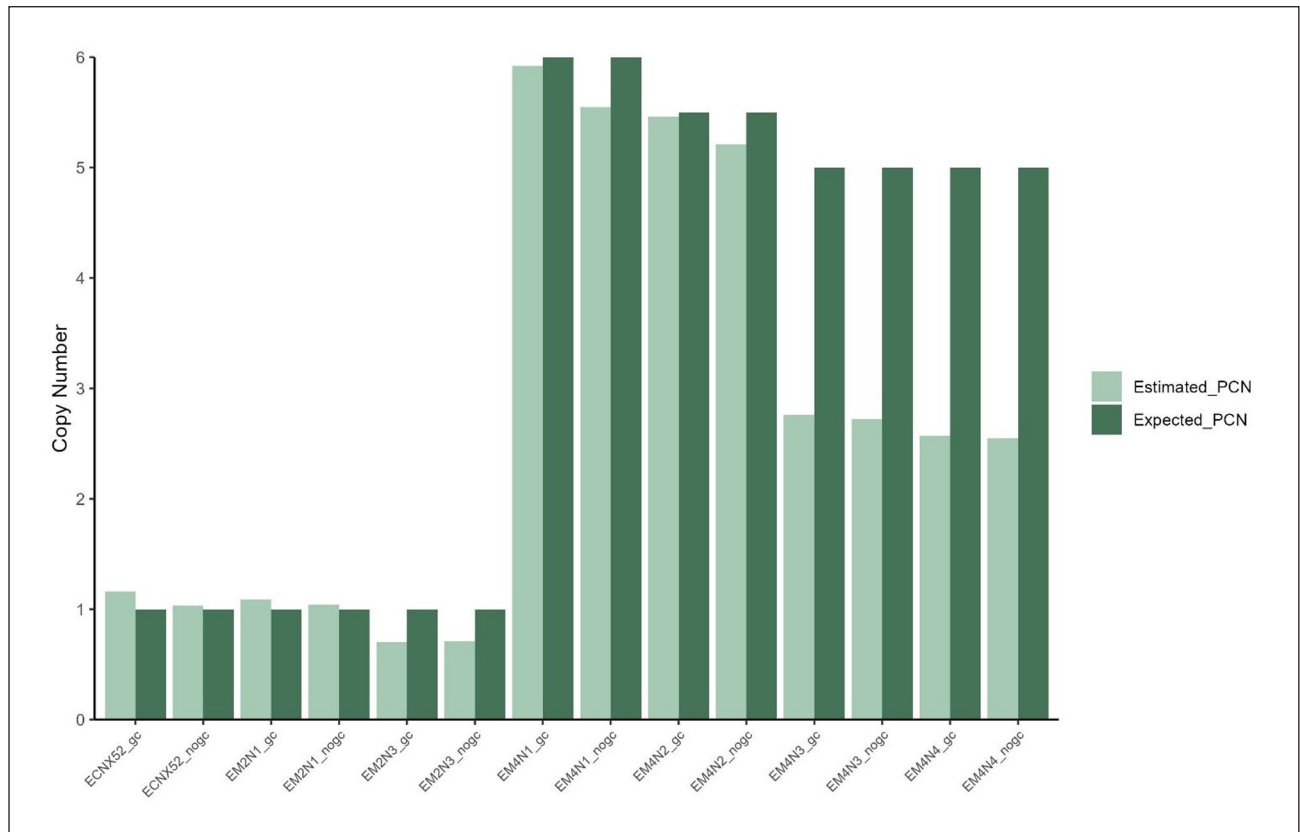


Figure 3. PCNE results on a real data set.

Bar plot showing the expected and estimated plasmid copy numbers for the real data set. The *_gc* suffix indicates samples processed with GC-correction, while *_nogc* indicates no GC-correction.

to one of the time points reported in their Supplementary Material. Despite this minor inconvenience, which we acknowledge as a limitation of this test, PCNE correctly identified the significant trend of increased copy number in isolates EM4N1-EM4N4, as well as the stability of isolates EM2N1 and EM2N3, and the control ECNX52. A detailed table is available as Supplemental Table S3.

Computational performance

The overall runtime was low for all datasets, with a mean of $01:36 \pm 01:04$ minutes. The simulated data set was processed faster and with lower memory usage than real-world data set, with a mean runtime (minutes) of $01:00 \pm 00:15$, compared to $02:59 \pm 00:36$ (Figure 4A). In terms of required memory, more variability has been found, depending mainly on chromosomal coverage depth and GC-correction. The simulated data set required a less sensitive amount than the real-world data set, with a mean memory (Gb) usage of 0.90 ± 0.49 , compared to 10.1 ± 5.8 (Figure 4B). As expected, GC-correction runs required the longest runtime and the highest amount of memory, for both data sets. All data are summarized in Supplemental Table S4.

Case study

All assemblies generated were of high quality. The MOB-suite identified the same 5 plasmids in all isolates,

confirming the isolate's closeness as reported in the original study. The application of PCNE, however, expanded this genomic characterization by revealing heterogeneity in the PCNs. Our analysis identified 2 distinct quantitative profiles: isolates KP_01, KP_02, and KP_05 maintained the *bla*_{KPC-2}-carrying plasmids at a low copy number (1.1-1.3 copies), while isolates KP_03 and KP_04 exhibited a 2- to 3-fold amplification of these same plasmids (2.3-3.1 copies). The presence of the primary virulence plasmid carrying the aerobactin siderophore genes (*iucABCD*, *iutA*) and the capsule regulator gene (*rmpA2*) was confirmed across all 5 isolates, with a slight increase in PCN for isolates KP_03 and KP_04 (Supplemental Table S5). Critically, the copy number variability did not correlate with the carbapenem and colistin resistant or hypervirulent phenotypes. This finding decouples plasmid-dosage-drive effect from other mechanisms, like the chromosomal mutations found in the study, responsible for the colistin resistance. Moreover, the high, consistent virulence phenotype, despite varying virulence PCNs, suggests that for these hypervirulent *K. pneumoniae* strains, the mere presence of the virulence plasmid and its associated genes is the primary determinant of pathogenicity, and that an increase in copy number beyond a certain threshold offers no measurable fitness advantage in the in vivo model used in the study. This demonstrates PCNE's utility in providing the quantitative data necessary to delineate the true drivers of a specific phenotype in an outbreak setting.

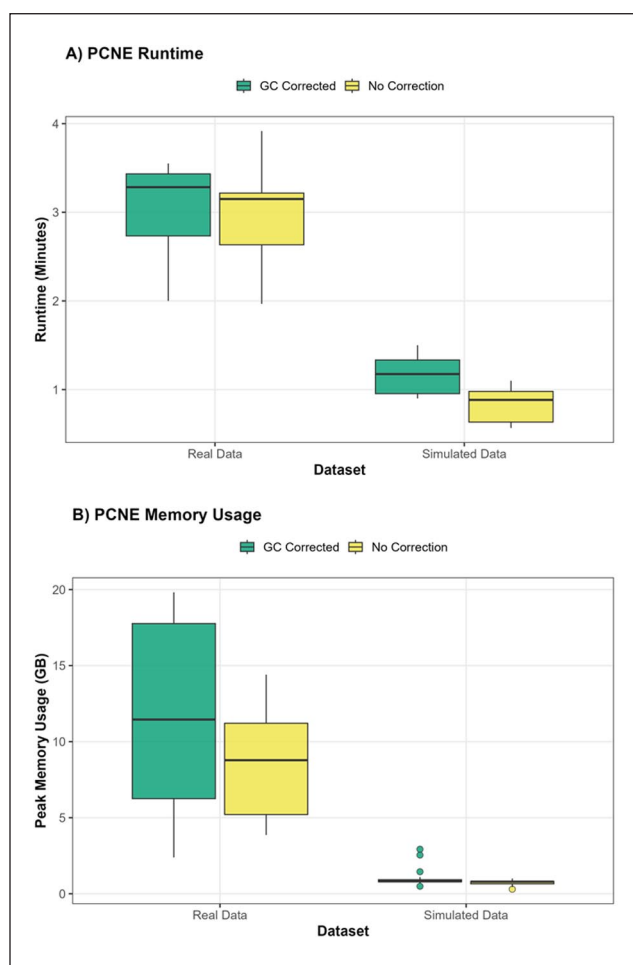


Figure 4. PCNE computational performance. (A) PCNE runtime, real data vs simulated data. (B) PCNE peak memory usage, real data vs simulated data.

Limitations

A limitation of all coverage-based copy number estimation methods, including PCNE and qPCR-based approaches, is the chromosomal replication dynamic. During active replication, bacteria possess multiple replication forks, leading to a gradient of DNA copy number across the chromosome: regions near the origin of replication (*oriC*) are present in higher copies than regions near the terminus (*ter*).¹⁹ This phenomenon introduces a systematic bias, as the chromosomal coverage used for normalization is not uniform. The PCNE attempts to mitigate the effects of this gradient by using the median coverage of all chromosomal windows as the baseline for normalization, rather than the mean. The median is inherently more robust to the high-coverage outliers near the *oriC* and low-coverage outliers near the *ter*. However, this statistical approach does not fully correct for the underlying biological bias, which remains a caveat for any sequencing data derived from a replicating cell population. Another limitation of the current PCNE version is its dependency on the limited accuracy of upstream assembly and classification tools for short-reads data. This limitation is partially overcome by the possibility to provide hybrid

assemblies. However, we acknowledge that the possibility and the need to perform hybrid sequencing are very limited and not cost-effective; for this reason, future work will focus on expanding its functionality to long reads only sequencing data, for which we aim to overcome the current major limitations.

Conclusion

Plasmid Copy Number Estimator was developed to enhance and standardize current short-read-based plasmid analysis workflow, via an easy-to-use, versatile, and efficient pipeline. The validation on simulated data demonstrates that PCNE provides reliable results across all its different configurations. Its flexible input modes allow for seamless integration with the outputs of widely used plasmid identification tools like Platon and MOB-suite. This was confirmed also by the test on real datasets, for which PCNE results were concordant. Furthermore, in our case study application, PCNE successfully uncovered significant PCN heterogeneity within a clonal outbreak, providing the quantitative data necessary to decouple plasmid dosage from the observed phenotypic resistance and virulence.

ORCID iDs

Riccardo Bollini  <https://orcid.org/0009-0001-9839-4380>
Valeria Cento  <https://orcid.org/0000-0002-3272-4252>

Author Contributions

Riccardo Bollini: Conceptualization; Writing – original draft; Methodology; Validation; Visualization; Writing – review & editing; Software; Data curation.

Valeria Cento: Funding acquisition; Writing – review & editing; Supervision.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project was supported by EU funding within the NextGenerationEU-MUR PNRR Extended Partnership initiative on Emerging Infectious Diseases (project no. PE00000007, INF-ACT).

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statement

The source and documentation for PCNE are available at <https://github.com/riccabolla/PCNE>. All the scripts to reproduce data in this paper are available at https://github.com/riccabolla/PCNE_paper_script/. The chromosome and the plasmid sequence used for the validation are deposited on the NCBI GenBank under accession numbers CP003200 and CP003223, respectively. All the data used in this manuscript are available at <https://doi.org/10.5281/zenodo.17550873>.

Supplemental Material

Supplemental material for this article is available online.

References

1. San Millan A. Evolution of plasmid-mediated antibiotic resistance in the clinical context. *Trends Microbiol.* 2018;26:978-985.
2. Tao S, Chen H, Li N, Wang T, Liang W. The spread of antibiotic resistance genes in vivo model. *Can J Infect Dis Med Microbiol.* 2022;2022:3348695.
3. Dimitriu T, Matthews AC, Buckling A. Increased copy number couples the evolution of plasmid horizontal transmission and plasmid-encoded antibiotic resistance. *Proc Natl Acad Sci U S A.* 2021;118:e2107818118. doi:10.1073/pnas.2107818118
4. Xiang G, Zhao Z, Zhang S, et al. Porin deficiency or plasmid copy number increase mediated carbapenem-resistant *Escherichia coli* resistance evolution. *Emerg Microbes Infect.* 2024;13:2352432.
5. Wein T, Hülter NF, Mizrahi I, Dagan T. Emergence of plasmid stability under non-selective conditions maintains antibiotic resistance. *Nat Commun.* 2019;10:2595.
6. Liu YT, Sau S, Ma CH, et al. The partitioning and copy number control systems of the selfish yeast plasmid: an optimized molecular design for stable persistence in host cells. *Microbiol Spectr.* 2014;2:10.1128/microbiolspec.PLAS-0003-2013.
7. Carattoli A, Zankari E, García-Fernández A, et al. In silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother.* 2014;58:3895-3903.
8. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom.* 2018;4:e000206.
9. Schwengers O, Barth P, Falgenhauer L, Hain T, Chakraborty T, Goesmann A. Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microb Genom.* 2020;6:mgen000398.
10. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* 2012;40:e72.
11. Browne PD, Nielsen TK, Kot W, et al. GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. *Gigascience.* 2020;9:giaa008.
12. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013 Mar 16. Accessed December 16, 2025. <http://arxiv.org/abs/1303.3997>
13. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *Gigascience.* 2021;10:giab008.
14. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. *Bioinformatics.* 2012;2:593-594. doi:10.1093/bioinformatics/btr708
15. Gourel H, Karlsson-Lindsjö O, Hayer J, Bongcam-Rudloff E. Simulating illumina metagenomic data with InSilicoSeq. *Bioinformatics.* 2019;35:521-522.
16. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics.* 2018;34:i884-i890. doi:10.1093/bioinformatics/bty560
17. Alcock BP, Huynh W, Chalil R, et al. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 2023;51:D690-D699.
18. Zhou S, Liu B, Zheng D, Chen L, Yang J. VFDB 2025: an integrated resource for exploring anti-virulence compounds. *Nucleic Acids Res.* 2025;53:D871-D877.
19. Slager J, Kjos M, Attaiech L, Veening JW. Antibiotic-induced replication stress triggers bacterial competence by increasing gene dosage near the origin. *Cell.* 2014;157:395-406.