Do Public Databases Need Higher Standards for Next-Generation Data Submissions?

Joseph L. Gallucci[1] and Dr. Rodolfo Aramayo[2]

[1] Department of Biochemistry and Biophysics, Texas A&M University
[2] Department of Biology; Texas A&M University

Introduction

- Genomics, an extension of Genetics, is a powerful tool to study the function and evolution of genes and genomes. When applied to the Human genome, it can play a key role in understanding the origin of many human diseases like *Cancer*.
- However, obtaining meaningful insights into any medical condition and/or pathological state requires the input of High-Quality data. Observations and/or conclusions based on incomplete and/or low quality data are not only hard to replicate and reproduce, but they are also highly questionable.
- The vast majority of the Human *Next-Generation Sequencing* (NGS) datasets have been deposited in the *National Center for Biotechnology Information* (NCBI) - *Small Read Archive* (SRA) database.
- This project started with the aim of re-analyzing a selected set of *Cancer*-related NCBI-SRA datasets in order to evaluate our ability to both reproduce and replicate previously published results, using a set of, *in-house,* newly developed algorithms.
- To our surprise, we found that the overall quality, and specially the genome coverage of these selected datasets was not only highly variable, but especially low in coverage, and non-uncommonly, contained contaminating sequences.
- In our view, these observations put into question the reproducibility and replicability potential of work based on these datasets.
- We conclude that in order to guarantee the replicability and reproducibility in Science, public databases, like the NCBI-SRA, need to set higher standards for data submission.

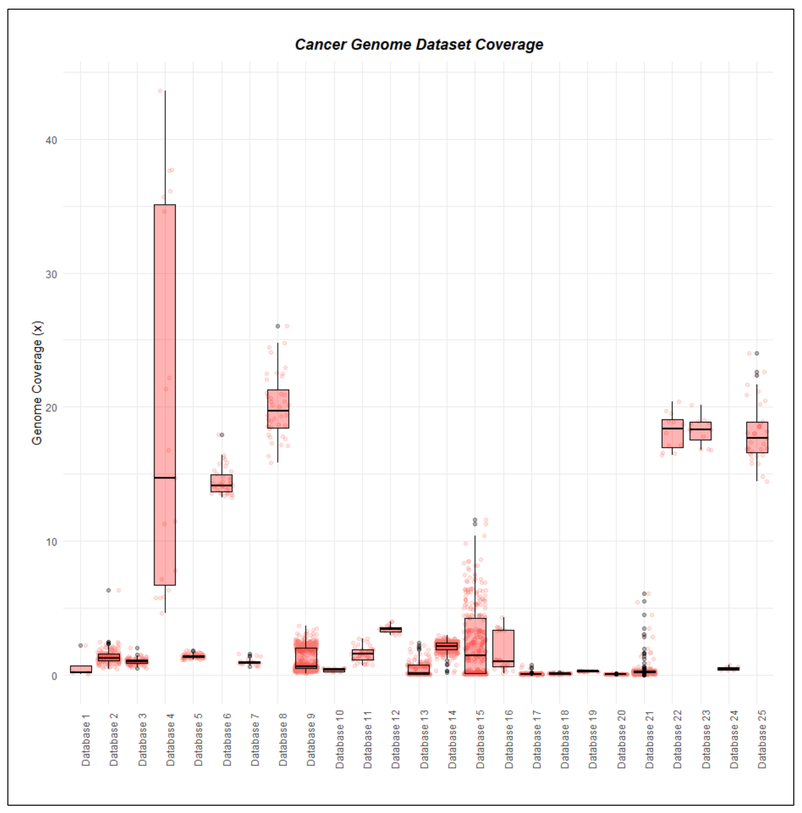
Tools and Methodology

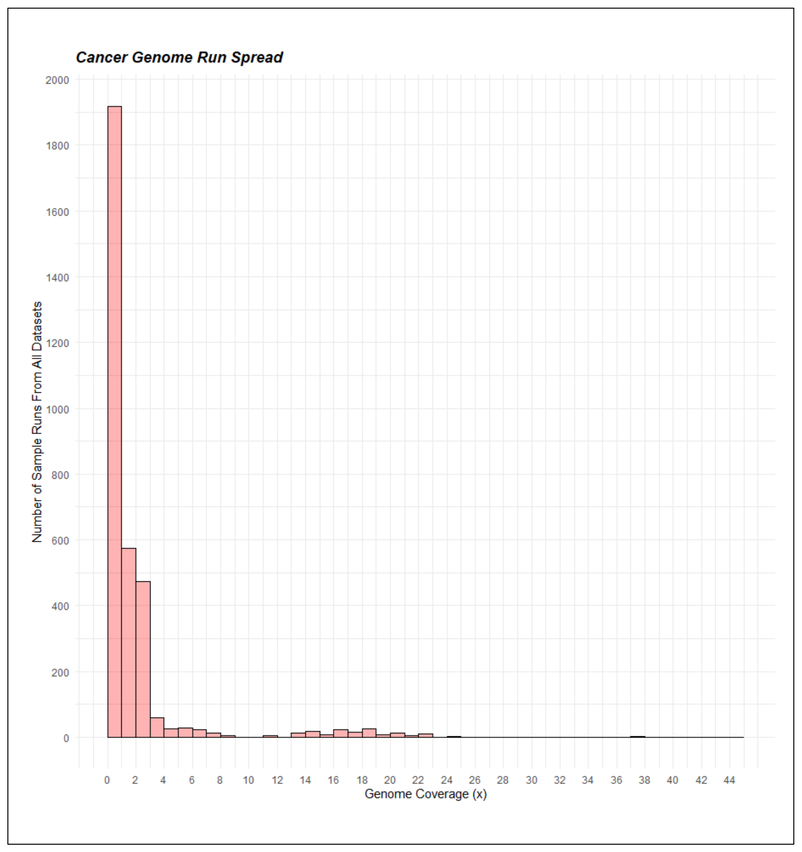
- We began by obtaining 25 random *Human* (i.e., Tax. ID 9606), *Cancer*-related datasets from NCBI-SRA database using multiple parameters:
  - To generate unbiased results, we randomly selected a set of NCBI-SRA datasets.
  - We removed datasets from *Human* tumors grafted onto Murine subjects, in order to avoid any possible *Murine* DNA contamination.
  - We removed datasets composed solely of exome and/or transcriptome data, to focus on genome data.
  - We selected datasets that contained *Paired-FastQ* data.
  - We used two different web-scrapers to parse for the datasets' metadata: *Parsehub* and *Octoparse*.
    - *Parsehub* was used for preliminary findings, as it scrapes much faster than Octoparse. During an initial exploratory scrape, ParseHub obtained 100 sample runs in a 6-hour period.
    - *Octoparse* was used to scrape the 25 randomly selected datasets for their respective metadata. As Parsehub only allows scraping for less than 20 web pages at a time, Octoparse was used because it allows for unlimited web-parsing, but at a slower rate. In a 24-hour time frame, 3,561 lines of data were extracted from NCBI SRA datasets. The data was later cleaned to only include genome data, resulting in 3,278 lines of genome data.
- Finally, we calculated the coverage for each sample run by multiplying the average fragment length by the number of reads present.

Tool References:

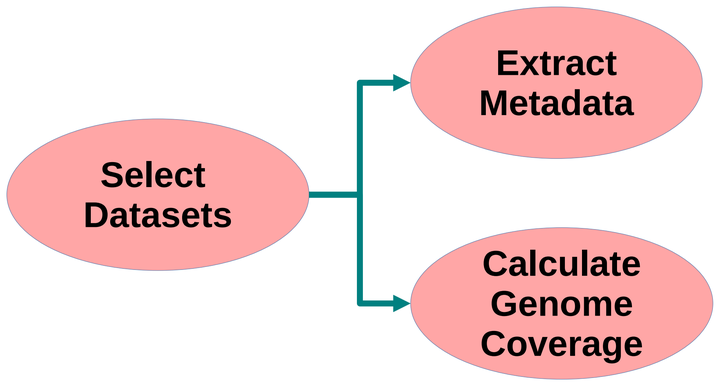
- Parsehub: https://www.parsehub.com
- Octoparse: https://www.octoparse.com

Figures





Workflow



Results

- We analyzed a total of 25 datasets.
  - The highest average coverage belonged to Database 8 with 19.96X coverage.
  - The lowest average coverage belonged to Database 20 at 0.037X coverage.
  - The sample run with the highest gene coverage by itself belonged to Database 4 at 43.59X coverage.
  - The lowest gene coverage by itself belonged to Database 17 and to Database 21, both with a coverage of less than 0.001X coverage.
  - Only 11 (i.e., 44%) datasets had greater coverage than 1X out of 25 total datasets.
  - Only 6 (i.e., 24%) datasets had greater coverage than 4X out of the 25 total datasets.
  - The majority of the individual datasets, have an average coverage of less than 1X.
  - The 6 datasets that had a value greater than 4X coverage typically had values centered around 20X coverage, with the exception of Database 04, which had a sparse and wide-range of coverage.
  - The majority of the sample runs (more than 1900), have very low genome coverage.
  - These results show that not only the average of the genome coverages are important, but also the distribution pattern of their coverages. Good datasets have tight coverage distribution among different sample runs.
  - In summary, less than 24% of the data selected had sufficient genome coverage for further genomic analysis.

Conclusions

- High-quality genomic data is a must for genomic analysis, and is especially important for meaningful medical research.
- Given that NCBI main mission is to provide access to high-quality databases to the public, it is imperative that all deposited datasets be of high quality.
- Among the databases we studied, we observed that they can contain either variable quality, and/or coverage.
- The presence of low coverage datasets places into question not only the quality of the data deposited in the NCBI-SRA database, but also the conclusions of the accompanying published studies.
- In our view, it is hard to believe that meaningful conclusions can be made from low coverage samples as this low coverage disqualify them for further meaningful analysis.
- An alternative explanation that might explain the presence of low coverage datasets is if the authors only made public a fraction of their data. If this were to be the case, it would still put into question anyone else's ability to replicate and to reproduce already published results.

Acknowledgments

- Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing.

Datasets SRA-IDs

1) SRP322573	6) SRP015139	11) ERP001476	16) ERP004006	21) SRP074289
2) SRP367159	7) SRP336096	12) SRP027366	17) SRP055057	22) SRP127747
3) ERP134554	8) SRP372487	13) SRP029757	18) SRP062373	23) SRP126796
4) SRP352436	9) SRP017787	14) SRP034680	19) SRP071834	24) SRP128158
5) ERP140186	10) ERP002259	15) SRP013572	20) SRP119412	25) SRP132286

References

1) 10.1186/s12920-021-01032-8	9) 10.1371/journal.pone.0067464	16) 10.1016/j.ccell.2014.09.010
3) 10.1016/j.isci.2022.105392	10) 10.1371/journal.pone.0021639	17) 10.1101%2Fgr.188060.114
4) 10.1016/j.trsl.2022.09.004	11) 10.1016%2Fj.cell.2013.02.032	18) 10.3390/cells9061370
7) 10.1002/2211-5463.13491	11) 10.1038%2Fnature10531	20) 10.1158/0008-5472.CAN-17-0529
8) 10.1101/2022.04.26.489523	15) 10.1038/nature13600	25) 10.1016/j.ccell.2018.06.008