www.genominfo.org

eISSN 2234-0742

GENOMICS & INFORMATICS

Vol. 19 - No. 3, September 30 2021

Special Issue: The 7th Biomedical Linked Annotation Hackathon (BLAH7)

Genomics & Informatics is indexed/tracked/covered by PubMed, PubMed Central, Scopus, KoreaMed, KoMCI, ScienceCentral, CrossRef, BIOSIS Previews, DOAJ, and Google Scholar.

www.genominfo.org
eISSN 2234-0742

G&I
Genomics & Informatics

Volume 19 number 3, September 30, 2021

Aims and scope

Genomics & Informatics is the official journal of the Korea Genome Organization (http://kogo.or.kr). Its abbreviated title is Genomics Inform. It was launched in 2003 by the Korea Genome Organization. It aims at making a substantial contribution to the understanding of any areas of genomics or informatics. Its scope includes novel data on the topics of gene discovery, comparative genome analyses, molecular and human evolution, informatics, genome structure and function, technological innovations and applications, statistical and mathematical methods, cutting-edge genetic and physical mapping, next generation sequencing and de novo assembly, and other topics that present data where sequence information is used to address biological concerns. Especially, Clinical genomics section is for a short report of all kinds of genome analysis data from clinical field such as cancer, diverse complex diseases and genetic diseases. It encourages submission of the cancer panel analysis data for a single cancer patient or a group of patients. It also encourages deposition of the genome data into designated database. Genome archives section is for a short manuscript announcing the genetic information of recently sequenced prokaryotic and eukaryotic genomes. These genome archives data can make the rationale for sequencing a specific organism.

It is published and distributed quarterly at the last dates of March, June, September, and December. All submitted manuscripts will be reviewed and selected for publication after single blind review process. All manuscripts must be submitted online through the e-submission system available from:
http://submit.genominfo.org. It is an online-only peer reviewed open access journal. A free full text both in the XML and PDF formats is available from the journal homepage (https://genominfo.org). It has been indexed by or searchable from PubMed, PubMed Central, Scopus, BIOSIS Previews, KoreaMed, KoMCI, Korea Citation Index, CrossRef metadata, DOAJ, and Google Scholar. This journal was supported by the Korean Federation of Science and Technology Societies Grant funded by the Korean Government.

- Manuscript Editing by InfoLumi Co., Seongnam, Korea.
- E-submission system by Inforang, Seoul, Korea
- PDF layout, XML production, and homepage management by M2Community Co., Seoul, Korea

Published by the Korea Genome Organization
Contact information
Park, Taesung, Editor-in-Chief

Editorial office of Genomics & Informatics
Room No. 806, 193 Mallijae-ro, Jung-gu, Seoul 04501, Korea
Tel: +82-2-558-9394, Fax: +82-2-558-9434, email: kogo3@kogo.or.kr, URL address: https://genominfo.org

G&I
Genomics & Informatics

Editorial Board

G&I
Genomics & Informatics

Volume 19 number 3, September 30, 2021

G&I
Genomics & Informatics

Volume 19 number 3, September 30, 2021

G&I
Genomics & Informatics

Editorial

eISSN 2234-0742
Genomics Inform 2021;19(3):e20
https://doi.org/10.5808/gi.19.3.e1

Received: September 27, 2021
Accepted: September 27, 2021

*Corresponding author:
E-mail: jdkim@dbcls.rois.ac.jp

Editor's introduction to the special section on the 7th Biomedical Linked Annotation Hackathon (BLAH7)

Jin-Dong Kim[1*], Kevin Bretonnel Cohen[2], Fabio Rinaldi[3], Zhiyong Lu[4], Hyun-Seok Park[5]

[1]Database Center for Life Science (DBCLS), Research Organization of Information and Systems (ROIS), Kashiwa, Chiba 277-0871, Japan
[2]School of Medicine, University of Colorado, Aurora, CO 80045, USA
[3]Dalle Molle Institute for Artificial Intelligence Research (IDSIA), 6928 Manno, Switzerland
[4]National Center for Biotechnology Information (NCBI), National Institutes of Health (NIH), Bethesda, MD 20894, USA
[5]Center for Convergence Research of Advanced Technologies, Ewha Womans University, Seoul 03760, Korea

The special section is dedicated to reporting achievements of the *7th Biomedical Linked Annotation Hackathon (BLAH7)*. *BLAH* is an annual hackathon event which is organized to join forces of biomedical text mining for the goal to promote interoperability among text mining resources. This year, the 7th edition was held in January, 2021. Due to the pandemic, it was organized as an online event, with the special theme "*coronavirus disease 2019 (COVID-19)*". The goal was to develop text mining resources to help address the pandemic situation. During the hackathon, 47 participants from 11 countries worked on voluntarily organized projects, and the results are reported in this special collection.

This section includes seven application notes and one opinion article. The first application note by Hernandez et al. [1] presents a Twitter dataset which includes more than 120 million "potentially clinically-relevant" tweets. The tweets are automatically annotated for clinically important named entities like drugs and symptoms. The dataset is released publicly to facilitate research on mining social media data for biomedical and clinical applications. Lithgow-Serrano et al. [2] presents named entity annotation of the LitCovid [3] dataset using OntoGene's Biomedical Entity Recogniser (OGER) [4] and shows its effectiveness for document classification. Ouyang et al. [5] presents the AGAC annotation [6] added on top of the PubTator [7] and OGER annotations and shows that the addition is potentially useful to mine regulatory or causal relationships between biomedical entities. The following three papers represent efforts for multilingualism of text mining. Barros et al. [8] presents a multilingual parallel corpus of PubMed articles for the language pairs English-Portuguese and English-Spanish. Their corpus was annotated for biomedical entities and also relationships between them, which was then used to develop a multilingual recommendation dataset for recommending biomedical entities to the authors of the articles. Yamaguchi et al. [9] and Soares et al. [10] are written by the same set of authors. They developed two versions of Japanese translation of MeSH terms, one through merging of existing resources and manual curation, and another through an automatic translation method, of which the results are reported in the two separate application notes. Larmande et al. [11] reports a revision to OryzaGP [12], a corpus of PubMed articles relevant to rice species, which are automatically annotated for proteins and genes. The last one by Dohi et al. [13] presents the authors' opinion after their case study with Alexander disease towards visualizing the phenotype diversity.

Based on the spirit of sharing, most of the resulting datasets, including corpora, annotations, and dictionaries, are released through open repositories like GitHub, PubAnnotation/PubDictionaries [14], and so on. We hope that this special collection will be an opportunity for the readers of the journal *Genomics* & *Informatics* to get informed about recent biomedical text mining activities aimed at providing support in the current COVID-19 pandemic situation.

ORCID

Jin-Dong Kim: https://orcid.org/0000-0002-8877-3248
Kevin Bretonnel Cohen: https://orcid.org/0000-0003-1749-8290
Fabio Rinaldi: https://orcid.org/0000-0001-5718-5462
Zhiyong Lu: https://orcid.org/0000-0002-8301-9553
Hyun-Seok Park: https://orcid.org/0000-0002-1237-8831

References

1. Hernandez LA, Callahan TJ, Banda JM. A biomedically oriented automatically annotated Twitter COVID-19 dataset. Genomics Inform 2021;19:e21.
2. Lithgow-Serrano O, Cornelius J, Kanjirangat V, Méndez-Cruz CF, Rinaldi F. Improving classification of low-resource COVID-19 literature by using Named Entity Recognition. Genomics Inform 2021;19:e22.
3. Chen Q, Allot A, Lu Z. LitCovid: an open database of COVID-19 literature. Nucleic Acids Res 2021;49:D1534-D1540.
4. Basaldella M, Furrer L, Tasso C, Rinaldi F. Entity recognition in the biomedical domain using a hybrid approach. J Biomed Semantics 2017;8:51.
5. Ouyang S, Wang Y, Zhou K, Xia J. LitCovid-AGAC: cellular and molecular level annotation data set based on COVID-19. Genomics Inform 2021;19:e23.
6. Wang Y, Zhou K, Kim JD, Cohen KB, Gachloo M, Ren Y, et al. An active gene annotation corpus and its application on anti-epilepsy drug discovery. In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019 Nov 18-21, San Diego, CA, USA. New York: Institute of Electrical and Electronics Engineers, 2019.
7. Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res 2013;41:W518-W522.
8. Barros M, Ruas P, Sousa D, Bangash AH, Couto FM. COVID-19 recommender system based on an annotated multilingual corpus. Genomics Inform 2021;19:e24.
9. Yamaguchi A, Takatsuki T, Tateisi Y, Soares F. Constructing Japanese MeSH term dictionaries related to the COVID-19 literature. Genomics Inform 2021;19:e25.
10. Soares F, Tateisi Y, Takatsuki T, Yamaguchi A. O-JMeSH: creating a bilingual English-Japanese controlled vocabulary of MeSH UIDs through machine translation and mutual information. Genomics Inform 2021;19:e26.
11. Larmande P, Liu Y, Yao X, Xia J. OryzaGP 2021 update: a rice gene and protein dataset for named-entity recognition. Genomics Inform 2021;19:e27.
12. Larmande P, Do H, Wang Y. OryzaGP: rice gene and protein dataset for named-entity recognition. Genomics Inform 2019;17:e17.
13. Dohi E, Bangash AH. Visualizing the phenotype diversity: a case study of Alexander disease. Genomics Inform 2021;19:e28.
14. Kim JD, Wang Y, Fujiwara T, Okuda S, Callahan TJ, Cohen KB. Open Agile text mining for bioinformatics: the PubAnnotation ecosystem. Bioinformatics 2019;35:4372-4380.

https://doi.org/10.5808/gi.19.3.e1

G&I
Genomics & Informatics

BLAH7
Application note

eISSN 2234-0742
Genomics Inform 2021;19(3):e21
https://doi.org/10.5808/gi.21011

Received: March 12, 2021
Accepted: July 26, 2021

*Corresponding author:
E-mail: jbanda@gsu.edu

A biomedically oriented automatically annotated Twitter COVID-19 dataset

Luis Alberto Robles Hernandez[1], Tiffany J. Callahan[2], Juan M. Banda[1*]

[1]Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA
[2]Computational Bioscience Program, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA

The use of social media data, like Twitter, for biomedical research has been gradually increasing over the years. With the coronavirus disease 2019 (COVID-19) pandemic, researchers have turned to more non-traditional sources of clinical data to characterize the disease in near-real time, study the societal implications of interventions, as well as the sequelae that recovered COVID-19 cases present. However, manually curated social media datasets are difficult to come by due to the expensive costs of manual annotation and the efforts needed to identify the correct texts. When datasets are available, they are usually very small and their annotations don't generalize well over time or to larger sets of documents. As part of the 2021 Biomedical Linked Annotation Hackathon, we release our dataset of over 120 million automatically annotated tweets for biomedical research purposes. Incorporating best-practices, we identify tweets with potentially high clinical relevance. We evaluated our work by comparing several SpaCy-based annotation frameworks against a manually annotated gold-standard dataset. Selecting the best method to use for automatic annotation, we then annotated 120 million tweets and released them publicly for future downstream usage within the biomedical domain.

Keywords: biomedical annotations, COVID-19, datasets, social media data
Availability: All code and documentation related to this project are publicly available on GitHub (https://github.com/thepanacealab/ annotated_twitter_covid19_dataset).

Introduction

Social media platforms like Twitter, Instagram, and Facebook provide researchers with unprecedented insight into personal behavior on a global scale. Twitter is currently one of the leading social networking services with over 353 million users and reaching ~6% of the world's population over the age of 13 [1]. It is also quickly becoming one of the most popular platforms for conducting health-related research because of its use for public health surveillance, pharmacovigilance, event detection/forecasting, and disease tracking [2,3]. During the last decade, Twitter has provided substantial aid in the surveillance of pandemics, including the Zika virus [4], H1N1 (or Swine Flu) [5], H7N9 (or avian/bird flu) [6], and Ebola [7]. Twitter has been used extensively during the 2020 coronavirus disease 2019 (COVID-19) outbreak [8], providing insight into everything from monitoring communication between public health officials and world leaders [9], tracking emerging symptoms [10] and access to testing facilities [11], to understanding the public's top fears and concerns about infection rates and vaccination [12]. While it is clear that Twitter contains invaluable content that can be used for a myriad of benevolent endeavors, there are many challenges to accessing and leveraging these data for clinical research and/or applications.

Researchers face a myriad of challenges when trying to utilize Twitter data. Aside from

the potential ethical challenges, which will not be discussed in this work (see Webb et al. [13] for a review of this area), it can be difficult to obtain access to these data and hard to keep up with real-time content collection [14,15]. Once the data have been obtained, researchers must then perform several preprocessing steps to ensure the data are sufficient for analysis. Concerning COVID-19, there are several existing social media repositories [16-20]. Unfortunately, most of these repositories are infrequently updated, do not provide any preprocessing or data cleaning, and either do not provide the raw data or lack appropriate metadata or provenance. The COVID-19 Twitter Chatter dataset [20] is a robust large-scale repository of tweets that is well-maintained and frequently updated (over 50 versions released at the time of publication). Recent work utilizing this resource has shown great promise for tracking long-term patient-reported symptoms [21] as well as highlighted mentions of drugs relevant to the treatment of COVID-19 [22]. While these are compelling clinical use cases, additional work is needed to fully understand what additional biomedical and clinical utility can be obtained from these data.

This paper presents preliminary work achieved during the 2021 Biomedical Linked Annotation Hackathon (BLAH 7) [23], which aimed to enhance and extend the COVID-19 Twitter Chatter dataset [20] to include biomedical entities. By annotating symptoms and other relevant biomedical entities from COVID-19 tweets, we hope to improve the downstream clinical utility of these data and provide researchers with a means to clinically characterize personally-reported COVID-19 phenomena. We envision this work as the first step towards our larger goal of deriving mechanistic insights from specific types of entities within COVID-19 tweets by integrating these data with larger and more complex sources of biomedical knowledge, like PheKnowLator [24] and the KG-COVID-19 [25] knowledge graphs. The remainder of this paper is organized as follows: an overview of the methods and technologies utilized in this work, an overview of our findings, and a brief discussion of conclusions and future work.

Methods

To prepare the dataset released in this work, we looked for named entity recognition (NER) pipelines to identify biomedical entities in text. We opted to evaluate: MedSpaCy [26], MedaCy [27], and ScispaCy [28], alongside a traditional text annotation pipeline from Social Media Mining Toolkit (SMMT), a product of a BLAH 6 hackathon [29]. The main reason for selecting these text processing pipelines is the fact that they are all based on SpaCy [30], a widely adopted open-source library for Natural Language Processing (NLP) in Python, allowing our codebases to be streamlined, and the annotation output to be easily compared in our evaluation as well as ingested by other work utilizing similar pipelines. Several preprocessing steps like URL and emoji removal were performed on all tweets.

Please note that the selected NER pipelines are usually tuned and developed to annotate specific types of clinical/scientific text, from either electronic health records, clinical notes, or scientific literature. The only general-purpose tagger is the SMMT, which does not perform any specialized tasks other than tagging or annotating text. This fact impacted their performance in Twitter social media data, and the following comparison should not be used to evaluate the systems' performance on clinical data/scientific literature, but rather the need for appropriately tuned systems for social media data.

Datasets

As the source for this work, we used one of the largest COVID-19 Twitter Chatter datasets available [20]. We used version 44 of the dataset [20], which contains 903,223,501 unique tweets. To improve the quality and relevance of the annotations, we used the clean version of this dataset, which has all retweets removed. Leaving us with a total of 226,582,903 unique tweets to annotate. From this subset, we selected only English tweets, as all the systems evaluated were created to extract/annotate biomedical concepts in this language.

For the evaluation of the annotations from each NER system and the SMMT tagger, we will use as a gold standard, a manually annotated dataset created for symptoms, conditions, prescriptions, and measurement procedures identification in patients with long Covid phenotypes [21]. This dataset consists of 10,315 manually annotated tweets, by multiple clinicians. Currently, the dataset is not publicly available but will be released at a later date.

ScispaCy

Developed by the Allen AI institute, the pipelines and models in this package have been tuned for use on scientific documents [28]. In our evaluation, we used the following model: *en_core_sci_lg*, which consists of ~785k vocabulary and 600k word vectors. Additionally, we used the EntityLinker component to annotate the Unified Medical Language System (UMLS) concepts. Since this pipeline provides more than one match per annotation, we only selected the first match to avoid duplicates. The code used can be found in [31].

MedaCy

Developed by researchers at Virginia Commonwealth University, MedaCy is a text processing framework wrapper for spaCy. It supports extremely fast prototyping of highly predictive medical NLP models. For our evaluation, we used their provided *medacy_model_*

https://doi.org/10.5808/gi.21011

clinical_notes model, with all other default settings. The code used can be found [31].

MedSpaCy

Currently, in beta release, MedSpaCy was created as a toolkit to enable user-specific clinical NLP pipelines. In our evaluation, we wanted to use some of the out-of-the-box components instead of fine-tuning them for our Twitter annotation task. We used the *en_info_3700_i2b2_2012* model - trained on i2b2 data, and the Sectionizer [32]. We initially tried to use the demo QuickUMLS entity linker, but ultimately opted not to do this as their demo only includes 100 concepts, and building it from scratch was outside of the scope of our task. The code used can be found in [31].

SMMT tagger

As part of SMMT, the SpaCy-based tagger relies on a user-specified dictionary to annotate concepts on the provided text. This tagger does not perform any NER or section detection, but only simple string matching. Designed with simplicity and flexibility in mind, when using social media data, it is preferred to provide a concise dictionary with the desired terms for annotation, rather than using pre-trained models that may not generalize well to domain-specific tasks, or are computationally expensive. The dictionary used in this evaluation consists of a mix of SNOMED-CT [33], ICD 9/10 [34], MeSH [35], and RxNorm [36] extracted from the Observational Health Data Sciences and Informatics (OHDSI) vocabulary. This dictionary is available as part of the paper's code repository.

Results

Extraction performance

In Table 1 we show the processing time and count of annotations produced by the evaluated systems on the gold standard dataset. Note that as expected, simple text annotation from the SMMT tagger is the fastest, with MedaCy coming in second as its annotation model is small. The SMMT tagger dictionary produces plenty of annotations as it considers some of the common misspellings for COVID-19 (e.g., "fatigue" vs "fatige") as well as related symptoms and drugs that have been curated in our previous work when extracting drug mentions in Twitter data [22].

Due to the larger model utilized by ScispaCy, the processing time is nearly five-fold that of simple text annotation. However, this comes with the added benefit that abbreviations are nicely normalized to UMLS concepts, hence creating some annotations that any of the other systems will be unable to find.

Overlap between systems on gold standard dataset

To determine which system to use for the large-scale annotation of the Twitter COVID-19 chatter dataset, we evaluated all systems against the manually annotated gold-standard. Here, while we grouped the annotations into three categories: drugs, conditions/symptoms, and measurements. We did not use the systems' annotation categories, but rather their annotated terms and spans. This was done to accommodate the custom entity categories that systems like MedSpaCy and MedaCy have in their default settings and the fact that we are using only the first UMLS concepts identified by ScispaCy. Table 2 shows the annotation overlap analysis.

We would like to stress again that MedSpaCy and MedaCy are at a disadvantage as their models are trained on considerably different data that does not work well with Twitter data. ScispaCy, however, performs fairly decently (in comparison) as the larger models provide capture relevant annotations when the tweet's text is clean and well-formed. It is out of the scope of this paper to properly tune these systems to ensure that they perform well with Twitter data, but it is certainly an interesting avenue for future research.

Extraction evaluation on a limited set

While it is clear that regular text annotation performed the best in replicating the annotations that our clinicians made, we still annotated all 226,582,903 dataset tweets and evaluated the overlap of annotations made by the different systems. Table 3 shows the comparison between counts of produced annotations, processing time, and overlaps in annotations between the systems.

Conclusion

In this work we release a biomedically oriented automatically anno-

Table 1. Extraction evaluation of proposed systems

	Tweets	Annotations produced	Processing time (s)
SMMT Tagger	10,315	92,835	10,815.24
MedSpaCy	10,315	51,575	33,746.40
MedaCy	10,315	61,890	21,896.63
ScispaCy	10,315	72,205	49,168.85

SMMT, Social Media Mining Toolkit.

Table 2. Annotation overlap analysis between gold standard dataset and evaluated systems

	Drugs (%)	Conditions/ Symptoms (%)	Measurements (%)	Average (%)
SMMT Tagger	69.31	71.91	39.83	60.35
MedSpaCy	19.98	13.49	7.45	13.64
MedaCy	47.04	27.14	12.56	28.91
ScispaCy	59.71	44.65	26.98	43.78

SMMT, Social Media Mining Toolkit.

Table 3. Annotation overlap evaluation for complete dataset

	Annotations produced	Processing time (min)	Overlaps with SMMT (%)	Overlap with MedSpaCy (%)	Overlap with MedaCy (%)	Overlap with ScispaCy (%)
SMMT Tagger	751,245,366	24,120	100	20.12	33.91	72.28
MedSpaCy	582,768,145	159,267	53.48	100	42.23	55.39
MedaCy	656,311,799	26,147	51.14	44.92	100	49.73
ScispaCy	775,615,621	325,620	89.17	34.77	44.17	100

SMMT, Social Media Mining Toolkit.

tated dataset of COVID-19 chatter tweets. We demonstrate that while there are existing SpaCy-based systems for NER on clinical and scientific documents, they do not generalize well when used on non-clinical sources of data like tweets. However, we use this evaluation to justify the usage of a simple text tagger (SMMT) to produce annotations on a large set of tweets, based on its robustness when evaluated on a gold-standard manually curated dataset. The resulting dataset and biomedical annotations is the first and largest of its kind making it a substantial contribution with respect to using large-scale Twitter data for biomedical research. We have also added components for these types of tasks to SMMT, improving the usability of the resource.

As for future work, the release of this dataset will facilitate continued development of fine-tuned resources for mining social media data for biomedical and clinical applications. Recent research has shown social media data to be a valuable source of patient-reported information that is not available in similar granularity in other more traditional data sources.

ORCID

Luis Alberto Robles Hernandez: https://orcid.org/0000-0002-5396-7800
Tiffany J. Callahan: https://orcid.org/0000-0002-8169-9049
Juan M. Banda: https://orcid.org/0000-0001-8499-824X

Authors' Contribution

Conceptualization: JMB, TJC. Data curation: JMB, LARH. Formal analysis: JMB, TJC. Methodology: JMB, LARH, TJC. Writing - original draft: JMB, TJC, LARH. Writing - review & editing: JMB, TJC, LARH.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

We would like to thank Jin-Dong Kim and the organizers of the virtual Biomedical Linked Annotation Hackathon 7 for providing us a space to work on this project and their valuable feedback during the online sessions.

References

1. Newberry C. 36 Twitter statistics all marketers should know in 2021. Vancouver: Hootsuite Inc., 2021. Accessed 2021 Mar 9. Available from: https://blog.hootsuite.com/twitter-statistics/.
2. Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, Merchant RM. Twitter as a tool for health research: a systematic review. Am J Public Health 2017;107:e1-e8.
3. Edo-Osagie O, De La Iglesia B, Lake I, Edeghere O. A scoping review of the use of Twitter for public health research. Comput Biol Med 2020;122:103770.
4. Masri S, Jia J, Li C, Zhou G, Lee MC, Yan G, et al. Use of Twitter data to improve Zika virus surveillance in the United States during the 2016 epidemic. BMC Public Health 2019;19:761.
5. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One 2010;5:e14118.
6. Vos SC, Buckner MM. Social media messages in an emerging health crisis: Tweeting bird flu. J Health Commun 2016;21:301-308.
7. Tang L, Bie B, Park SE, Zhi D. Social media and outbreaks of emerging infectious diseases: a systematic review of literature. Am J Infect Control 2018;46:962-972.
8. Coronavirus: staying safe and informed on Twitter. San Francisco: Twitter Inc., 2021. Accessed 2021 Mar 9. Available from: https://blog.twitter.com/en_us/topics/company/2020/covid-19.html.
9. Rufai SR, Bunce C. World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis. J Public Health (Oxf) 2020;42:510-516.
10. Guo JW, Radloff CL, Wawrzynski SE, Cloyes KG. Mining twitter to explore the emergence of COVID-19 symptoms. Public

https://doi.org/10.5808/gi.21011

Health Nurs 2020;37:934-940.
11. Mackey T, Purushothaman V, Li J, Shah N, Nali M, Bardier C, et al. Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with COVID-19 on Twitter: retrospective big data infoveillance study. JMIR Public Health Surveill 2020;6:e19509.
12. Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z. Top concerns of Tweeters during the COVID-19 pandemic: infoveillance study. J Med Internet Res 2020;22:e19016.
13. Webb H, Jirotka M, Stahl BC, Housley W, Edwards A, Williams M, et al. The ethical challenges of publishing Twitter data for research dissemination. In: Proceedings of the 2017 ACM on Web Science Conference, 2017 Jun 25-28, Troy, NY, USA. New York: Association for Computing Machinery, 2017. pp. 339-348.
14. Hino A, Fahey RA. Representing the Twittersphere: archiving a representative sample of Twitter data under resource constraints. Int J Inf Manage 2019;48:175-184.
15. Kim Y, Nordgren R, Emery S. The story of goldilocks and three Twitter's APIs: a pilot study on Twitter data sources and disclosure. Int J Environ Res Public Health 2020;17:864.
16. Kabir MY, Madria S. CoronaVis: a real-time COVID-19 Tweets data analyzer and data repository. Preprint at: https://arxiv.org/abs/2004.13932 (2020).
17. Chen E, Lerman K, Ferrara E. Tracking social media discourse about the COVID-19 pandemic: development of a public coronavirus Twitter data set. JMIR Public Health Surveill 2020;6:e19273.
18. Gupta RK, Vishwanath A, Yang Y. Global reactions to COVID-19 on Twitter: a labelled dataset with latent topic, sentiment and emotion attributes. Preprint at: http://arxiv.org/abs/2007.06954 (2021).
19. Alqurashi S, Alhindi A, Alanazi E. Large arabic Twiter dataset on COVID-19. Preprint at: https://arxiv.org/abs/2004.04315 (2020).
20. Banda JM, Tekumalla R, Wang G, Yu J, Liu T, Ding Y, et al. A large-scale COVID-19 Twitter chatter dataset for open scientific research: an international collaboration. Epidemiologia 2021;2:315-324.
21. Banda JM, Singh SR, Alser OH, Prieto-Alhambra D. Long-term patient-reported symptoms of COVID-19: an analysis of social media data. Preprint at: https://doi.org/10.1101/2020.07.29.20164418 (2020).
22. Tekumalla R, Banda JM. Characterizing drug mentions in COVID-19 Twitter Chatter. New York: Association for Computational Linguistics, 2020. Accessed 2021 Mar 9. Available from: https://www.aclweb.org/anthology/2020.nlpcovid19-2.25/.
23. Biomedical Linked Annotation Hackathon 7. Kashiwa: Database Center for Life Science, 2021. Accessed 2021 Mar 9. Available from: https://blah7.linkedannotation.org/.
24. Callahan TJ, Tripodi IJ, Hunter LE, Baumgartner WA Jr. KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response. Preprint at: https://doi.org/10.1101/2020.04.30.071407 (2020).
25. Reese JT, Unni D, Callahan TJ, Cappelletti L, Ravanmehr V, Carbon S, et al. KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response. Patterns (N Y) 2021;2:100155.
26. medspacy. San Francisco: GitHub, 2021. Accessed 2021 Mar 9. Available from: https://github.com/medspacy/medspacy.
27. Mulyar A, Mahendran D, Maffey L, Olex A, Matteo G, Dill N, et al. TAC SRIE 2018: extracting systematic review information with MedaCy. Gaithersburg: National Institute of Standards and Technology, 2018. Accessed 2021 Mar 9. Available: https://www.researchgate.net/profile/Darshini_Mahendran/publication/340870892_TAC_SRIE_2018_Extracting_Systematic_Review_Information_with_MedaCy/links/5ea1add5a6fdcc88fc381e4c/TAC-SRIE-2018-Extracting-Systematic-Review-Information-with-MedaCy.pdf.
28. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: fast and robust models for biomedical natural language processing. New York: Association for Computational Linguistics, 2019. Accessed 2021 Mar 9. https://doi.org/10.18653/v1/W19-5034.
29. Tekumalla R, Banda JM. Social Media Mining Toolkit (SMMT). Genomics Inform 2020;18:e16.
30. Explosion AI. spaCy-Industrial-strength Natural Language Processing in Python. Explosion AI, 2017. Accessed 2021 Mar 9. Available from: https://spacy.io/.
31. Annotated_twitter_covid19_dataset. San Francisco: Github, 2021. Accessed 2021 Mar 9. Available from: https://github.com/thepanacealab/annotated_twitter_covid19_dataset.
32. medspacy. San Francisco: Github, 2021. Accessed 2021 Mar 9. Available from: https://github.com/medspacy/medspacy.
33. Donnelly K. SNOMED-CT: the advanced terminology and coding system for eHealth. Stud Health Technol Inform 2006;121:279-290.
34. International Statistical Classification of Diseases and Related Health Problems (ICD). Geneva: World Health Organization, 2020. Accessed 2021 Mar 10. Available from: https://www.who.int/standards/classifications/classification-of-diseases.
35. Medical subject headings. Bethesda: National Library of Medicine, 2020. Accessed 2021 Mar 10. Available from: https://www.nlm.nih.gov/mesh/meshhome.html.
36. RxNorm. Bethesda: National Library of Medicine, 2004. Accessed 2021 Mar 10. Available from: https://www.nlm.nih.gov/research/umls/rxnorm/index.html.

G&I
Genomics & Informatics

BLAH7
Application note

eISSN 2234-0742
Genomics Inform 2021;19(3):e22
https://doi.org/10.5808/gi.21018

Received: March 17, 2021
Revised: August 6, 2021
Accepted: August 12, 2021

*Corresponding author:
E-mail: oscarwilliam.lithgow@idsia.ch

Improving classification of low-resource COVID-19 literature by using Named Entity Recognition

Oscar Lithgow-Serrano[1*], Joseph Cornelius[1], Vani Kanjirangat[1], Carlos-Francisco Méndez-Cruz[2], Fabio Rinaldi[1]

[1]Dalle Molle Institute for Artificial Intelligence Research, IDSIA USI-SUPSI, Polo universitario Lugano-Campus Est, Via la Santa 1, CH-6962 Lugano, Switzerland
[2]Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Avenida Universidad s/n Col. Chamilpa, 62210 Cuernavaca, Mor., Mexico

Automatic document classification for highly interrelated classes is a demanding task that becomes more challenging when there is little labeled data for training. Such is the case of the coronavirus disease 2019 (COVID-19) clinical repository—a repository of classified and translated academic articles related to COVID-19 and relevant to the clinical practice—where a 3-way classification scheme is being applied to COVID-19 literature. During the 7th Biomedical Linked Annotation Hackathon (BLAH7) hackathon, we performed experiments to explore the use of named-entity-recognition (NER) to improve the classification. We processed the literature with OntoGene's Biomedical Entity Recogniser (OGER) and used the resulting identified Named Entities (NE) and their links to major biological databases as extra input features for the classifier. We compared the results with a baseline model without the OGER extracted features. In these proof-of-concept experiments, we observed a clear gain on COVID-19 literature classification. In particular, NE's origin was useful to classify document types and NE's type for clinical specialties. Due to the limitations of the small dataset, we can only conclude that our results suggests that NER would benefit this classification task. In order to accurately estimate this benefit, further experiments with a larger dataset would be needed.

Keywords: classification, COVID-19, Named Entity Recognition, NLP
Availability: The code and data are available at https://github.com/IDSIA-NLP/blah7/tree/task3/task3.

Introduction

The current pandemic took the world by surprise in all aspects, among which is the efficient broadcasting of relevant findings to reach interested researchers, clinicians and other stakeholders. The medical literature relevant to coronavirus disease 2019 (COVID-19) is growing exponentially and, doctors, clinicians, and health workers in general need tools for monitoring and prioritizing the literature to make the most of their time by allowing them to quickly identify the most relevant information.

We have witnessed an enormous effort and solidarity that experts around the world have put in contributing and sharing a broad scope of experiences and findings. In this paper we consider in particular the case of an interinstitutional COVID-19 cooperation group (https://covid19.ccg.unam.mx/) which has been working to provide tools and resources for COVID-19 literature exploration.

The main contribution of the group is a clinical repository consisting of continuously

updated academic literature related to COVID-19 (https://covid19.ccg.unam.mx/repoinfo.html). Considering the PRECEPT [1] scheme, their clinical experience and, other common classifications, the group has adopted a 3-way classification strategy to organize the repository to make it easier for potential users to find relevant literature. Despite a big effort to read and manually classify these documents, the pace of production of this literature hamper the utility of this work. In response the group has been working on automatic document classification machine learning models (https://covid19.ccg.unam.mx/machinelearning.html), however the number of classes, the interconnected nature of the clinical fields and the very little labeled examples imposes considerable challenges [2,3].

On the other hand, another important contribution has been the inclusion of a specialized COVID-19 named-entity annotation service, OntoGene's Biomedical Entity Recogniser (OGER) [4]. OGER is an annotation tool that performs Named Entity Recognition and Disambiguation (NERD) using shallow and deep dependency parsing combined with state-of-the-art machine learning techniques. It interacts with *The Bio Term Hub* (https://covid19.nlp.idsia.ch/bth.html) which enables OGER to process and link named entities from major life science databases: cellosaurus, cell ontology, ChEBI, CTD, EntrezGene, Gene Ontology, MeSH, Molecular Process Ontology, NCBI Taxonomy, Protein Ontology, RxNorm, Sequence Ontology, Swiss-Prot, Uberon.

Aiming to leverage the convergence of both contributions we hypothesized that the named entities recognized by OGER could be used as a free-lunch feature augmentation strategy to boost the classification performance [5,6].

During the Biomedical Linked Annotation Hackathon 7 (https://blah7.linkedannotation.org/) within the task "Analyzing COVID-19 literature with OGER" we proposed a subtask oriented to investigate this hypothesis (https://coree.github.io/ blah7/task3.html).

The task's aim was to classify COVID-19 literature according to three independent dimensions: clinical specialties, types, topics-and-subtopics, with special emphasis in exploring the use of Named Entities to better leverage the title, abstract and text during classification.

Here, we present a proof-of-concept experiment performed during the hackathon. We processed the literature with OGER and used the resulting identified named-entity-recognition (NER) and their links to major biological databases (NED) as extra input features for a basic classifier model. We then compared the results with the same classification model but without the OGER extracted features. Although very preliminary, the results are promising and show a clear gain on COVID-19 literature classification by using named-entity-recognition as an auxiliary feature augmentation step, suggesting the benefit of NER for this task.

Methods

In the preprocessing phase the original PDF documents were converted to text using the Linux utility *pdftotext* and no other cleaning or normalization steps were performed.

Next, the documents' title and full text were converted to features applying Term Frequency-Inverse Document Frequency (tf-idf), Latent Semantic Analisis (LSA), and using the CLS embedding (first embedding) of the pretrained BERT-base as a sentence-embedding [7].

Besides, the full-text was processed with the OGER's API specialized for COVID-19 applications (https://pub.cl.uzh.ch/ projects/ontogene/oger/upload/txt/tsv?dict=509f822aaf527390). This process resulted in an average of 1,622 annotations per article (The same token can be annotated as multiple entities because OGER uses multiple sources) including, among others, the entity type (e.g., organism, disease) (The full list of entity types annotated by OGER in this dataset were: organism, sequence, cellular_component, clinical_drug, cell_line, organ/tissue, gene/protein, chemical, cell, disease, molecular_process, biological_process, molecular_function), the matched term (e.g., coronavirus), the preferred-form (e.g., coronavirus), the entity ID and, the origin database (e.g., MeSH diseases, CTD MeSH). OGER has the capacity to return results in different formats, for our purpose TSV was the most convenient (Table 1 for an output example).

The classification task was approached as a multi-class problem for each of the three classification axis; this is, each document can be labeled with at most one topic, one type, and one specialty.

The general approach to use the NERD results consisted in treating each field of OGER output as an alternative representations of the documents and, then, apply different feature extraction over these representations (Fig. 1). To create each representation, all the values of one field of the OGER results were joined as words, with a space as separator. For example, to obtain the *preferred-form* representation of a document, all the value of the preferred-form field of all the identified entities in the document were concatenated. The resulting string is then treated as if it were an additional sentence in the input text and thus, converted using two basic feature extraction strategies: tf-idf and LSA.

The result was 11 features: six corresponding to the tf-idf, LSA and bert-embedding (Due to the BERT model limitation, we use only the first 512 tokens of the article body to build the text embedding) for the document's title and the text representations, and 5 from the NERD extraction: tf-idf of the type, tf-idf and LSA [8,9]

https://doi.org/10.5808/gi.21018

Table 1. Fragment exemplifying an OGER output in TSV format

Type	Start	End	Matched term	Preferred form	Entity ID	Origin
Disease	42	50	COVID-19	COVID-19	C000657245	MeSH supp (Diseases)
Organism	126	137	Coronavirus	Coronavirus	D017934	MeSH desc (Organisms)
Disease	138	145	COVID19	COVID-19	C000657245	MeSH supp (Diseases)
Disease	151	159	COVID-19	COVID-19	C000657245	MeSH supp (Diseases)
Disease	309	317	COVID-19	COVID-19	C000657245	MeSH supp (Diseases)
Disease	360	368	COVID-19	COVID-19	C000657245	MeSH supp (Diseases)
Disease	733	741	COVID-19	COVID-19	C000657245	MeSH supp (Diseases)
Chemical	776	785	Antiviral	Antiviral agent	CHEBI:22587	ChEBI
Chemical	857	866	Antiviral	Antiviral agent	CHEBI:22587	ChEBI
Chemical	907	910	Com	Coenzyme M	CHEBI:17905	ChEBI
Chemical	918	927	Antiviral	Antiviral agent	CHEBI:22587	ChEBI
Clinical_drug	944	955	Chloroquine	Chloroquine	2393	RxNorm
Chemical	944	955	Chloroquine	Chloroquine	D002738	CTD (MESH)

OGER, OntoGene's Biomedical Entity Recogniser; COVID-19, coronavirus disease 2019.

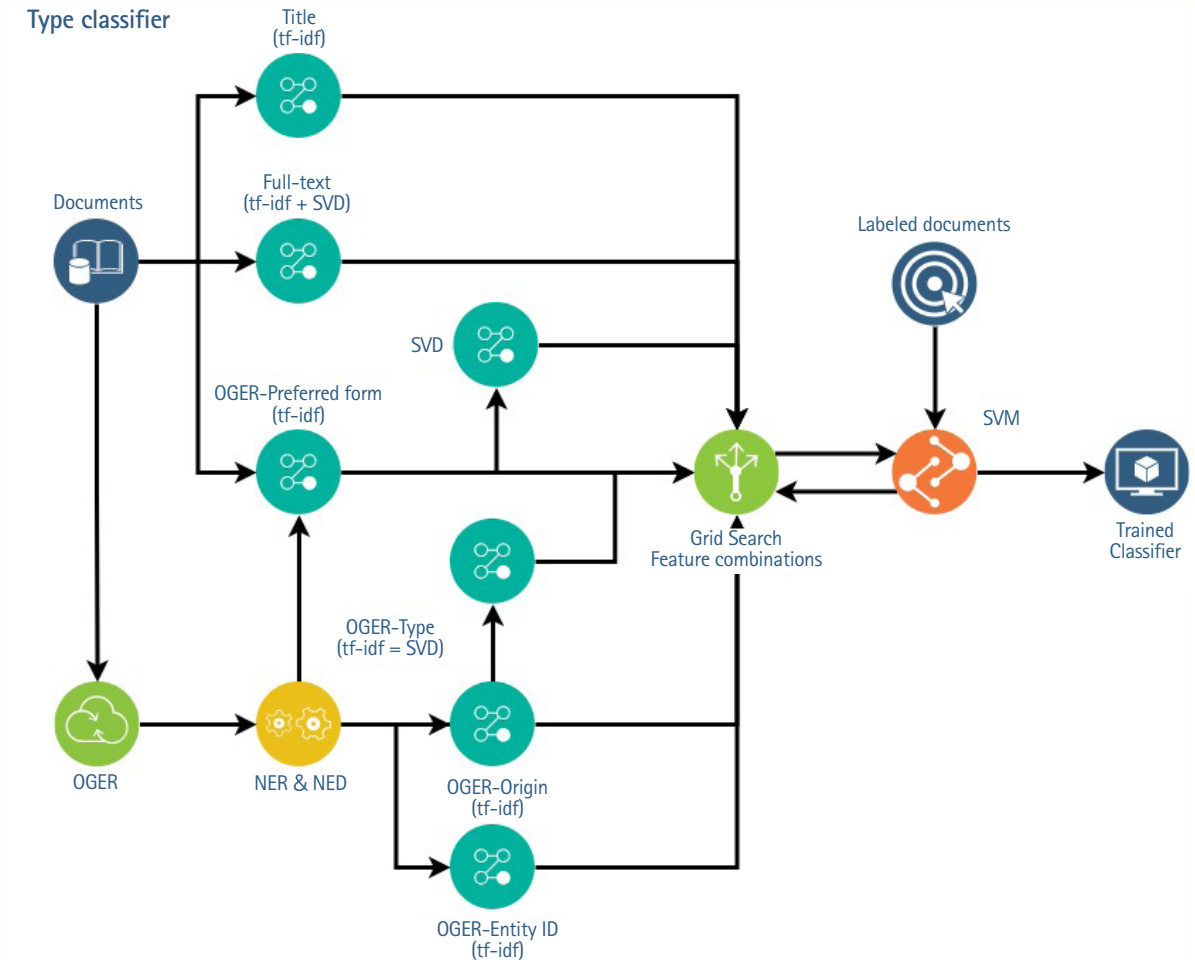

Fig. 1. Schematic representation of the experiment pipeline. At the top half, feature extraction of documents, titles, and full-text was done with tf-idf and LSA techniques, respectively. At the bottom half, documents were first processed with OGER, and then different feature extraction strategies were applied over the extracted NERD fields. Tf-idf was applied to the origin and entity-ID values and LSA to preferred-form. Finally, an exhaustive search of feature combinations with an SVM classifier was performed. tf-idf, Term Frequency-Inverse Document Frequency; LSA, Latent Semantic Analisis; OGER, OntoGene's Biomedical Entity Recogniser; NERD, Named Entity Recognition and Disambiguation; SVM, Support Vector Machine.

of the *preferred-form*, tf-idf of the *entity-id* and, tf-idf of the *origin*. These features were later reduced (independently for each classification axis) through a supervised exhaustive feature selection with a Support Vector Machine (SVM) [10] on a 3-fold cross validation, i.e., through a greedy search evaluating with the SVM classification model all possible combinations of features.

The final validation consisted in training the selected model (the SVM with the reduced features) on a stratified split corresponding to the 80% of the labeled dataset and, test it in the remaining 20%. This was repeated in 30 runs with random train-test split.

One possible limitation of this study is that we performed the search for the best features combination by 3-fold cross-validation in the full dataset. This was done because the dataset was very small, highly unbalanced and some classes had only three examples. Although in the final validation phase training and testing were done in clearly separated subsets of the original data, further experiments with more data are needed to validate the observed benefits.

Dataset

The set of manually labeled documents from the clinical repository was used for these experiments.

It is worth noticing that the labeled dataset consists of 646 articles with a rounded average of 12 words for the title and 3,540 for the body. However not all documents are classified in each of the 3-classification axis: the *Document type* classification (e.g., observational studies, systematic reviews) has 269 examples for the 6 classes; the *Clinical specialties* classification (e.g., immunology, Cardiology) has also 269 examples but for 36 classes; and the *Topic & Subtopic* classification (e.g., epidemiology, diagnosis) counts 383 examples for the 16 topics and 161 examples for the 27 subtopics. Moreover within each classification axis the examples are not uniformly distributed through the classes.

Results

The baseline was selected as the model with the best combination of features extracted from the document title and text, i.e., without the NERD results. In this case, the best baseline model was obtained by using the tf-idf of the title and LSA of the text.

For the Document type classification, the selected features were: tf-idf of the title, LSA of the text and tf-idf of the NE origin. The F1-weighted mean of the 30 validation runs was 0.739 (s = 0.041 [sample standard deviation]) and, by including the NE origin as feature the classifier had a significant 10% gain (t-test with $p < 0.05$) compared to the 0.669 (s = 0.044) F1-weighted score of the baseline (Table 2).

In the specialty classification the best model consisted of the features tf-idf of the title, LSA of the text and tf-idf of the NE type. This combination resulted in a statistically significant 8% gain in the F1-weighted mean compared to the baseline, 0.646 (s = 0.044) versus 0.597 (s = 0.048).

Finally, for the topic classification the best features were tf-idf of the title, LSA of the text and tf-idf of the NER origin. The F1-weighted mean of the 30 runs for this model was 0.599 (s = 0.051) whereas the baseline scored 0.595 (s = 0.049). Although this represented an improvement of 0.8%, it was not statistically significant. Fig. 2 shows the comparison of the results mentioned above.

Discussion and Conclusion

It is worth noticing that the pdf to text conversion was done using a basic approach which produces quite noisy results, i.e., the extracted text includes lines that are not complete sentences due to how they are displayed in the PDF; it also includes statements interrupted by spurious chunks of text like footnotes, page numbers, etc. and other kinds of not cleanly extracted text. This might explain why BERT features played no role, and was detrimental to the OGER processing and the quality of its results.

Interestingly, we found that the named entities' type and origin,

Table 2. Best model results per classification axis

Classification axis	Baseline model	Model with NERD features
Document type	0.669 (s = 0.044)	0.739[a] (s = 0.041)
Specialty	0.597 (s = 0.048)	0.646[a] (s = 0.044)
Topic	0.595 (s = 0.049)	0.599 (s = 0.051)

Values are the average of the F1-weighted mean of 30 runs; s, sample standard deviation.
NERD, Named Entity Recognition and Disambiguation.
[a]The difference is statistically significant.

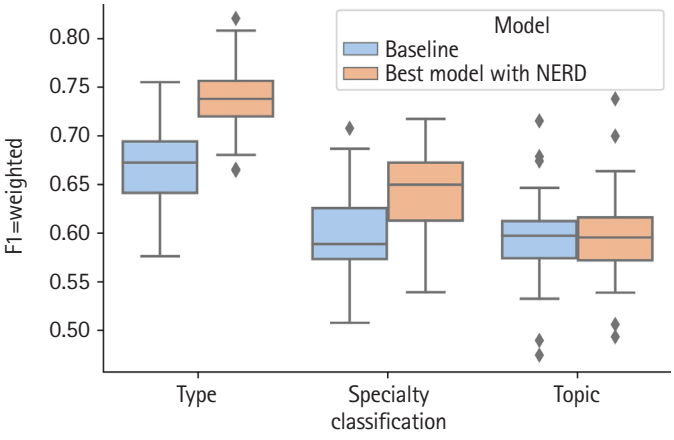

Fig. 2. Comparison of results distributions between baseline and the best model using Named Entity Recognition and Disambiguation (NERD) features for each classification axis.

https://doi.org/10.5808/gi.21018

i.e., more general features, were more informative for the classification than the more specific ones, like the matched term, the preferred-form, and the entity ID. One possible explanation is that due to the granularity of the general features, there are more examples in the training data and, thus, the estimator can better learn the relation between those features and the target classes. The fact that NE's type (e.g., organism, disease) helped classify specialty may unveil an interesting pattern where some biomedical entities are more prevalent in some specialties. On the other hand, the fact that using the NE origin (MeSH diseases, CTD MeSH) improved document type classification is an interesting finding that should be further investigated. These experiments also opened the question of whether using BioBert Embeddings, trained for Biomedical Domain, instead of general Bert Embeddings may help classification.

It is important to stress that these were proof-of-concept experiments, and bearing in mind the methodological limitations due to the small dataset the conclusions here presented are only suggestive, and further experiments with more data are needed to accurately estimate the observed benefits of NER in the classification task.

Finally, it is important to highlight that due to the limited time in the hackathon, the experiments presented here were applied and compared to a baseline and not to the classification strategy that the COVID-19 cooperation group is developing. The next step would be to investigate if similar gains could be obtained when integrating NERD features in that strategy and applied to a larger dataset.

ORCID

Oscar Lithgow-Serrano: https://orcid.org/0000-0003-1995-1669
Joseph Cornelius: https://orcid.org/0000-0002-5427-5005
Vani Kanjirangat: https://orcid.org/0000-0002-2526-1413
Carlos-Francisco Méndez-Cruz: https://orcid.org/0000-0002-2549-1614
Fabio Rinaldi: https://orcid.org/0000-0001-5718-5462

Authors' Contribution

Conceptualization: OLS. Data curation: JC, VK. Formal analysis: OLS. Funding acquisition: FR. Methodology: OLS, FR, CFMC. Writing - original draft: OLS. Writing - review & editing: FR, CFMC, VK, JC.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

We are grateful to the organizer of the Biomedical Linked Annotation Hackathon 2021 for the opportunity to work collaboratively on this project and share it with the other participants.

References

1. Harder T, Sin MA, Bosch-Capblanch X, Coignard B, de Carvalho Gomes H, Duclos P, et al. Towards a framework for evaluating and grading evidence in public health. Health Policy 2015;119:732-736.
2. Prati RC, Batista GE, Monard MC. Class imbalances versus class overlapping: an analysis of a learning system behavior. In: MICAI 2004: Advances in Artificial Intelligence. MICAI 2004. Lecture Notes in Computer Science, Vol. 2972 (Monroy R, Arroyo-Figueroa G, Sucar LE, Sossa H, eds.). Berlin: Springer, 2004. pp. 312-321.
3. Denil M, Trappenberg T. Overlap versus imbalance. In: Advances in Artificial Intelligence. Canadian AI 2010. Lecture Notes in Computer Science, Vol. 6085 (Farzindar A, Keselj V, eds.). Berlin: Springer, 2010. pp. 220-231.
4. Basaldella M, Furrer L, Tasso C, Rinaldi F. Entity recognition in the biomedical domain using a hybrid approach. J Biomed Semantics 2017;8:51.
5. Armour Q. The role of named entities in text classification. M.A.Sc. Thesis. Ottawa: University of Ottawa, 2005.
6. Andelic S, Kondic M, Peric I, Jocic M, Kovacevic A. Text classification based on named entities. In: 7th International Conference on Information Society and Technology, 2017 Mar 12-15, Kopaonik, Serbia. pp. 23-28.
7. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT 2019, 2019 Jun 2-7, Minneapolis, MN, USA. Stroudsburg: Association for Computational Linguistics, 2019. pp. 4171-4186.
8. Deerwester S, Dumais ST, Furnas GW, Ladauer TK, Harshman R. Indexing by latent semantic analysis. J Am Soc Inf Sci 1990;41:391-407.
9. Landauer TK, Dumais ST. A solution to Platos problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol Rev 1997;104:211-240.
10. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273-297.

G&I
Genomics & Informatics

BLAH7
Application note

eISSN 2234-0742
Genomics Inform 2021;19(3):e23
https://doi.org/10.5808/gi.21013

Received: March 17, 2021
Revised: September 3, 2021
Accepted: September 13, 2021

*Corresponding author:
E-mail: xiajingbo.math@gmail.com

LitCovid-AGAC: cellular and molecular level annotation data set based on COVID-19

Sizhuo Ouyang, Yuxing Wang, Kaiyin Zhou, Jingbo Xia*

Hubei Key Lab of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, 430070 Wuhan, China

Currently, coronavirus disease 2019 (COVID-19) literature has been increasing dramatically, and the increased text amount make it possible to perform large scale text mining and knowledge discovery. Therefore, curation of these texts becomes a crucial issue for Bio-medical Natural Language Processing (BioNLP) community, so as to retrieve the important information about the mechanism of COVID-19. PubAnnotation is an aligned annotation system which provides an efficient platform for biological curators to upload their annotations or merge other external annotations. Inspired by the integration among multiple useful COVID-19 annotations, we merged three annotations resources to LitCovid data set, and constructed a cross-annotated corpus, LitCovid-AGAC. This corpus consists of 12 labels including Mutation, Species, Gene, Disease from PubTator, GO, CHEBI from OGER, Var, MPA, CPA, NegReg, PosReg, Reg from AGAC, upon 50,018 COVID-19 abstracts in LitCovid. Contain sufficient abundant information being possible to unveil the hidden knowledge in the pathological mechanism of COVID-19.

Keywords: AGAC, annotation, corpus, knowledge discovery, LitCovid
Availability: AGAC corpus: http://pubannotation.org/projects/AGAC_training; LitCovid-AGAC data set: http://pubannotation.org/projects/LitCovid_AGAC_GENE_OGER.

Introduction

Coronavirus disease 2019 (COVID-19) is an abbreviation for corona virus disease, which caused a pandemic in 2019. People infected with COVID-19 suffers from severe high fever, dyspnea, lung disease and with 0.3%–1.5% chance of death. Due to the severe condition COVID-19 caused, the research upon the disease has been increasing dramatically. As of January 2021, there are over 90,000 related literature published, and make it a huge repository for knowledge discovery. Such a large growth rate makes it difficult for relevant researchers to understand the massive information in time.

Understanding the mechanism of COVID-19 is of importance for containing the virus. Like severe acute respiratory syndrom virus, it enters cells by binding angiotensin-converting enzyme 2 (ACE2) protein on the surface of human cells with S protein. S protein is located in the outermost layer of COVID-19, and exists in the form of trimer. Each monomer contains a receptor binding domain composed of amino acids where S protein binds to ACE2 and infects human cells.

Compared with the whole vision of the COVID-19 mechanism, the above commonsense knowledge is far from sufficiency. For unveiling the mechanism hidden in the huge text data, application of text mining has drawn a good amount of attentions recently. So far, nearly 200 researches have been published in PubMed, which worked on COVID-19 liter-

ature mining. For propelling the COVID-19-oriented text mining researches, NCBI developed a huge public available COVID-19 corpus, LitCovid [1,2], and make it a gold database for knowledge mining.

Fortunately, the Bio-medical Natural Language Processing (BioNLP) community has long focused on fundamental tools development, including bio-medical entity recognition, entity concept normalization, relation extraction, and so forth. For PubMed abstracts and PMC full texts, PubTator [3] efficiently tags and normalizes six types of biological entities, i.e., gene, disease, chemical, mutation, species and cell line.

For example, PubTator is a search database that highlights some keywords in the search results, it's based on the results of PubMed. Pubtator supports six tag types, which are gene, disease, chemical, mutation, species and cell line. The above six kinds of tags are already very useful for unveil hidden mechanism of COVID-19. LitCovid is a reliable corpus which is a collection of texts related to COVID-19. Therefore, when PubTator annotates the LitCovid corpus, the six biological entities in the text will be assigned a corresponding tag. Moreover, the OntoGene's Bio-medical Entity Recognizer (OGER) [4,5] is an important Tagger, which will annotate the following seven bio-medical entities, Disease, Chemical, Sequence, Gene/Protein, Biological_process, Organism and Cell, and these were annotated by using Bio Term Hub (BTH) terminologies. BTH supports the rapid construction of term resources from famous life science databases in a simple standardized format for text mining, and it can label specific concept types such as protein, gene, disease and cell line. However, we use OGER only to add gene ontology (GO) and chemical annotations to our data set.

Considering the need for logical mining, AGAC is good at discovering Regulation relations. Therefore, it is easy to reveal Pathway-like logic. In this research, we release LitCovid-AGAC database. It provides multiple annotations by PubTator, OGER and AGAC.

Methods

AGAC as a corpus for key annotations labeling

The purpose of designing AGAC [6] corpus is to better find the logical lines in the sentence, and designed six tags for this, namely Var, MPA, CPA, PosReg, NegReg, Reg. It took 20 months for 4 annotators to manually annotate and check. AGAC is illuminative to be applied in drug-related knowledge discovery. For example, AGAC was successfully applied in LOF/GOF classification by using tensor decomposition algorithm [7]. As well, it has been adopted as the training data in a competition in the BioNLP open shared task 2019 [8], and applied to extract relevant literature for Alzheimer's disease for the support of gene disease association prediction [9].

AGAC tagger

An AGAC tagger based on the deep neural network was introduced as a baseline method in AGAC track in BioNLP OST 2019. The baseline fully used sophisticated BERT structure and reached sufficient high quality for sequence labeling [7], the F-1 value of which is about 0.5. Such high-quality annotation results indicate that applying AGAC corpus to annotate the text helps to find the convincing logical relationships between biological entities.

PubAnnotation platform for multiple annotations alignment

PubAnnotation [10] is a platform for biologist curator to assemble annotations or annotate their own labels upon interested texts. Till now, there are 45 released projects in PubAnnotation with AGAC included. Co-tagging is possible to carry on automatically via PubAnnotation, as various bio concept taggers, e.g., OGER and PubTator, have already been involved in the system. Co-tagging helps to integrate different annotations and to serve sophisticated knowledge representation. As can be seen from the following example, three resources mentioned above provided different kinds of annotation on a same sentence, which form a complete logical line shown in the figure.

As shown in Fig. 1, TF is the abbreviation of total flavonoid, which has been labeled in the previous article. By combining AGAC labels with other important annotations, we can clearly see the logical lines shown at the bottom right of the figure. The data we uploaded can be downloaded in PubAnnotation in JSON format. The annotation set we released combines the annotation of PubTator, OGER and AGAC, which can be used to mine the logical lines of biological process changes in COVID-19.

It can be seen that different corpora have different annotation focuses. Other corpora mainly label biological concepts and match them to standard data sets. However, AGAC not only focuses on biological concepts, but also focuses on logical lines in sentences. The same biological concept may be given different labels in different contexts, or even will not be labeled. In this way, we can find that some chemicals up regulate or down regulate gene expression in COVID-19.

Automatic annotation pipeline

By integrating the method mentioned above, we performed an automatic annotation pipeline to obtain the LitCovid-AGAC dataset.

Step 1. Data collection: Collect literature data set from LitCovid

https://doi.org/10.5808/gi.21013

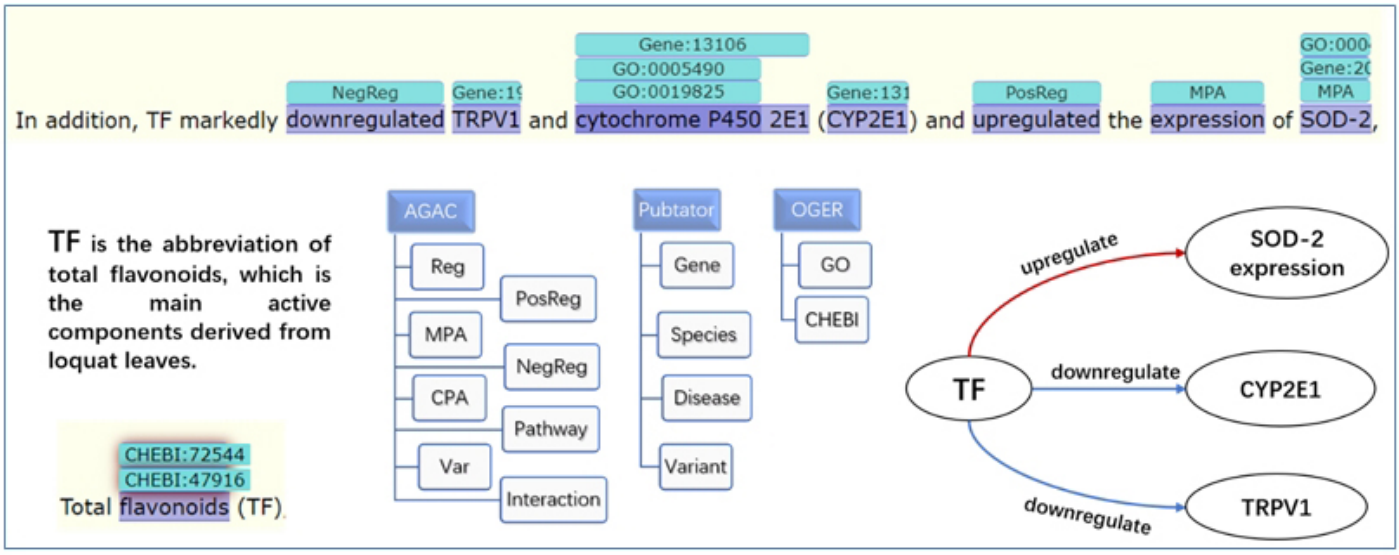

Fig. 1. The knowledge representation based on the LitCovid-AGAC corpus.

[1,2].

Step 2. AGAC annotation: Obtain the AGAC annotations by applying AGAC tagger on literature set.

Step 3. Regulation annotation: Create a regulation dictionary on PubDictionary [10] and automatically annotate the regulation words.

Step 4. PubTator and OGER annotation: Import the annotations from PubTator and OGER by using PubAnnotation.

Results

Statistics of LitCovid-AGAC dataset

LitCovid-AGAC contains 50,018 abstracts from PubMed, and the annotations are from three sources, AGAC, PubTator and OGER. LitCovid-AGAC aims on the regulations of biological process described in COVID-19 literature. Therefore, we applied all the AGAC labels which contains 5 biological concept labels and 3 regulation labels. To enrich the relative annotation, Mutation, Species, Gene, Disease from PubTator and GO, Chemical Entities of Biological Interest (CHEBI) [11] from OGER are included in LitCovid-AGAC dataset. CHEBI includes natural products and synthetic products used to intervene in biological processes, but generally does not include macromolecules encoded by genes. According to the statistics data, the most frequent label is "Disease," which appears 285,135 times, and the least frequent label is "Mutation", which only appears 435 times.

It can be clearly seen that the annotation results of OGER and PubTator are more abundant, on the contrary, the number of AGAC annotations is not in the same order of magnitude as the number of their annotations. It is due to the annotation rules in AGAC that the sentence without the description of regulation is not annotated, so AGAC annotations are less than the annotations from other sources. The more detailed statistics is shown in Table 1.

Table 1. The statistics of LitCovid-AGAC

Name	LitCovid-AGAC
Text type	Title, abstract
Annotation count	AGAC – Var (444), MPA (1,162), CPA (298), NegReg (1,128), PosReg (402), Reg (1,169)
	LitCovid – Mutation (435), Species (152,939), Gene (23,795), Disease (285,135)
	OGER – GO (57,467), CHEBI (111,981)

Knowledge discovery pattern and research paradigm in LitCovid-AGAC dataset

Logical line examples from single sentence in cellular and molecular level

Enriched by PubTator and OGER, the data set contained more complete annotations. For instance, in Fig. 2A, the "Disease" annotation provided by PubTator acts as the cause among the other annotations from AGAC in this sentence, where the "Cell Physiological Activity," lymphocyte, was firstly regulated by the severe acute respiratory syndrom coronavirus 2 (SARS-CoV-2) infection and other two "Cell Physiological Activity," T and B cells and monocytes, were down-regulated subsequently. From the annotations in this sentence, the effect of the SARS-CoV-2 infection in cell level was clearly showed with the sequential order, which could be transformed to a path in a knowledge graph.

Besides, the annotations also unveil the molecule-level biological processes. In Fig. 3, R518W/Q mutations in gene NPC1 inhibited the cholesterol transports and thus resulted the accumulation of

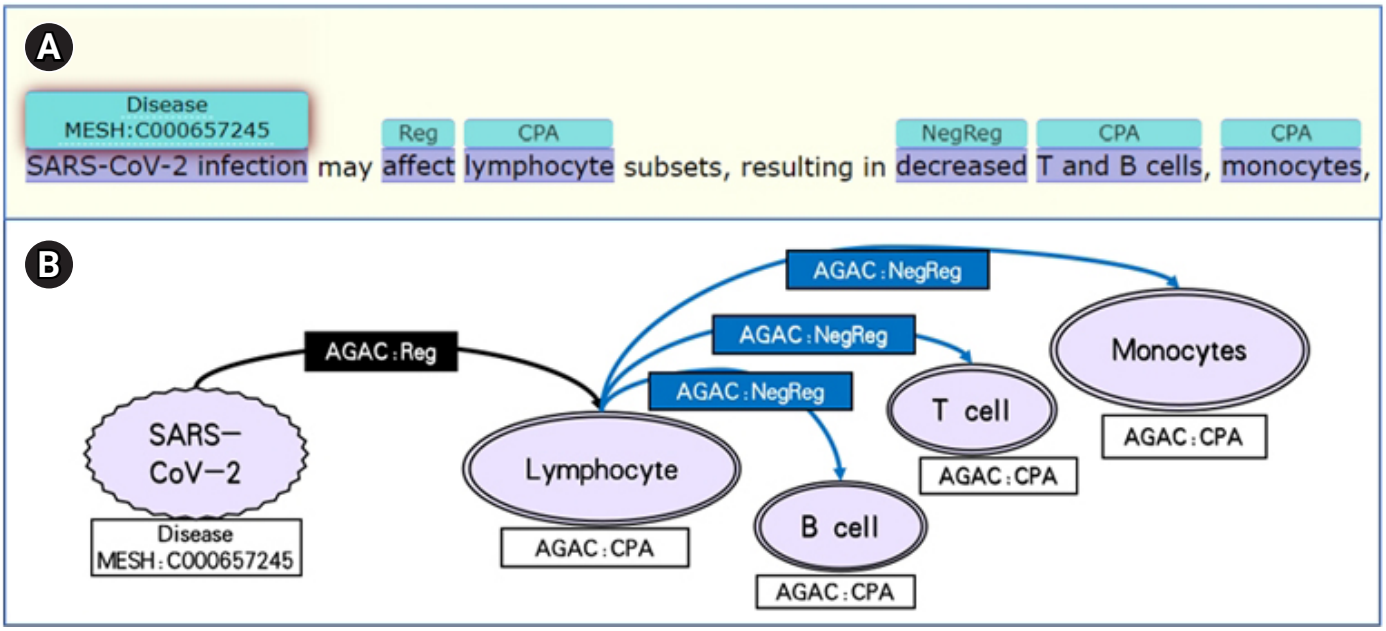

Fig. 2. (A, B) A cellular level annotation example of LitCovid-AGAC data set.

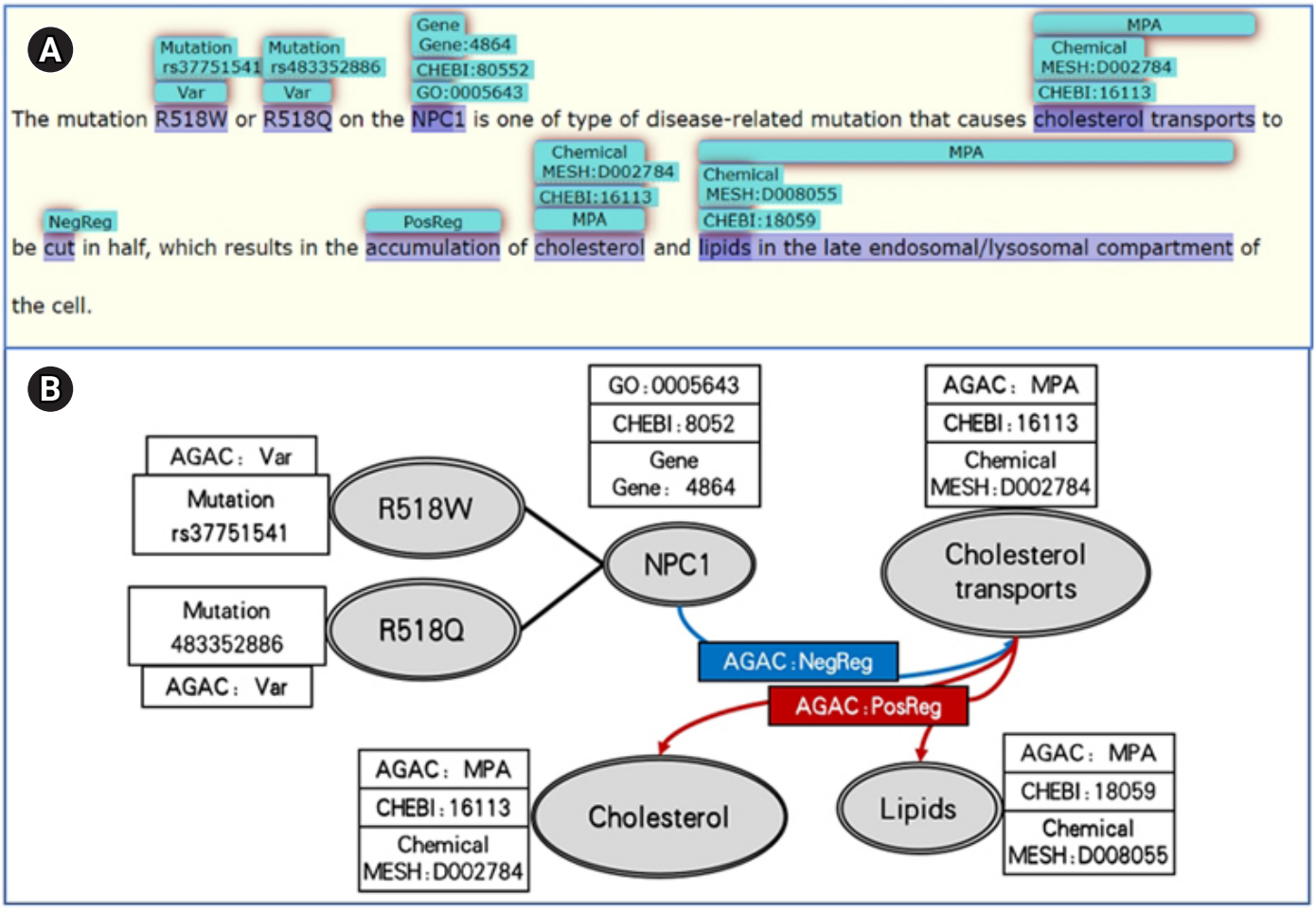

Fig. 3. (A, B) A molecular level annotation example of LitCovid-AGAC data set.

https://doi.org/10.5808/gi.21013

cholesterol and lipids, which are all "Molecular Physiological Activity." In this sentence, AGAC annotations provided the variation, regulation and molecular level processes, while PubTator and OGER provided the gene, variation, chemical and GO [12] annotations with their unique ID which supplemented the information recognition and also provided the normalization on some of the AGAC annotations. GO has three categories, which are biological process, molecular function and cellular component, these terms used to represent all entities and their relationships.

With the annotations in LitCovid-AGAC data set, the genes, diseases, variations and the biological processes in cellular-level and molecular-level are connected by the regulations 4 labels in the

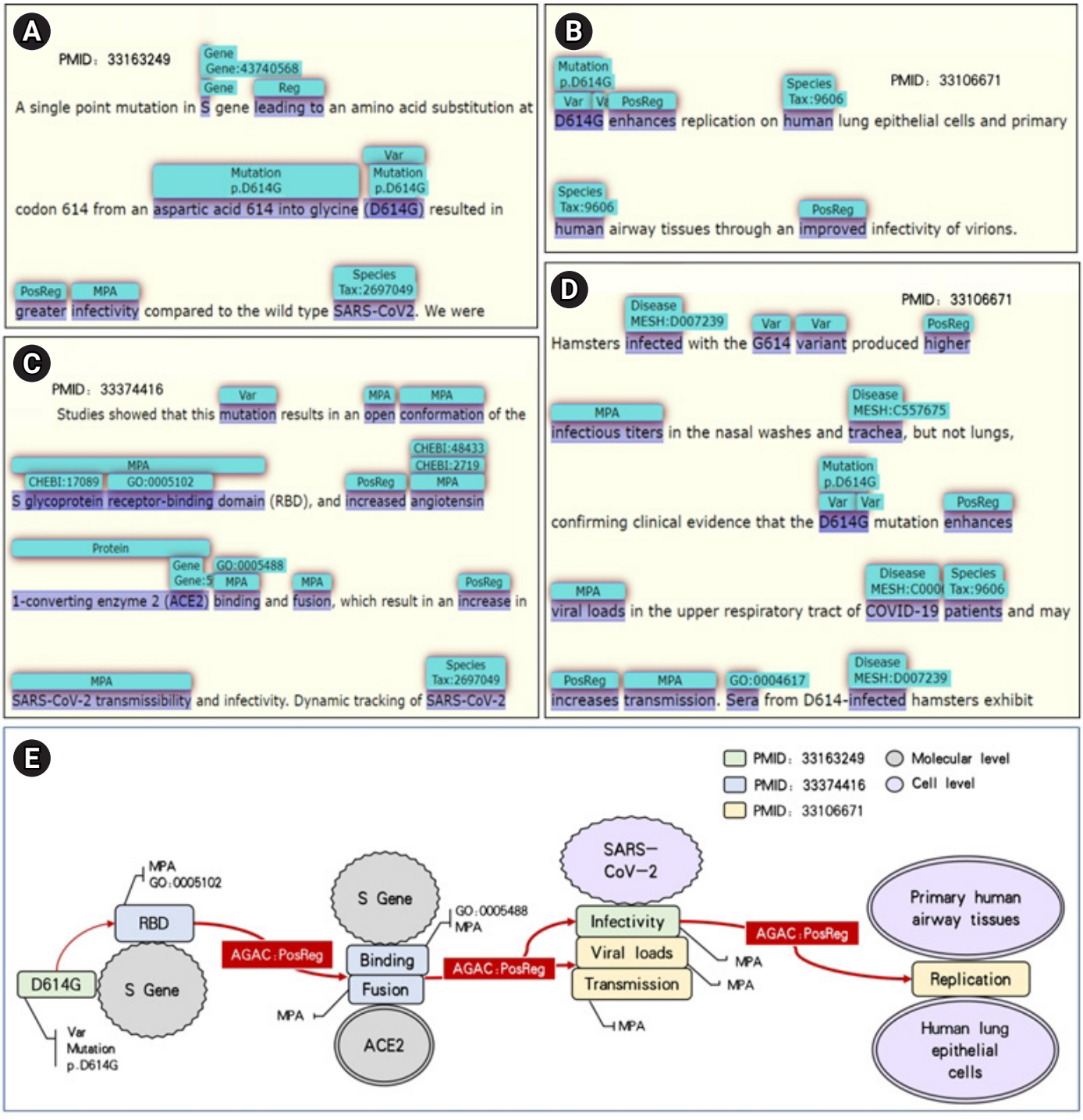

Fig. 4. (A–E) A light logical network inferred from LitCovid-AGAC data set.

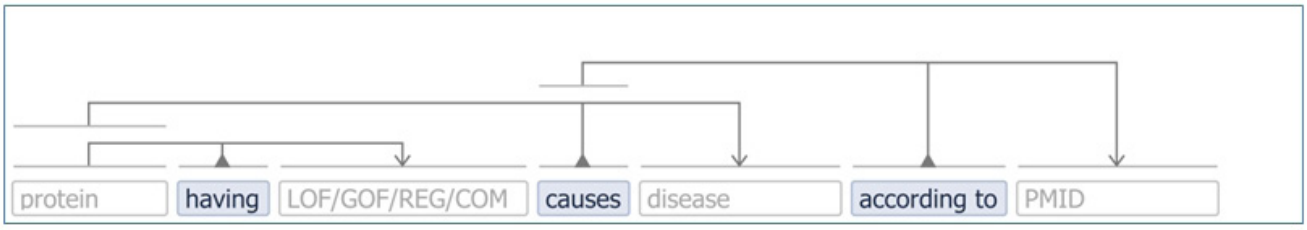

Fig. 5. Visualization semantic structure template.

same sentence. Combining with the semantics information, the sequential order of the regulation events helps to convert them into a directional path which regards regulation label as the edges and the other labels as the nodes. For example, the path in Fig. 2B is a neutral regulation edge from SARS-CoV-2 infection to lymphocytes then three negative regulations to T and B cells and monocytes. The same knowledge pattern is shown in Fig. 3B. The numerous knowledge paths in this data set are able to construct a network with plenty of biological information contained in the COVID-19 literature, which should contribute to the pathological mechanism analysis of COVID-19 and the evolution of this virus.

Combined logical lines from multiple sentences

Combining the annotations in different articles can get a complete logical line. The D614G mutation of spike gene (S gene) in Fig. 4A will lead to greater infectivity of SARS-CoV-2 virus. The information in Fig. 4C and 4D shows that this mutation leads to the open conformation change of S-glycoprotein receptor-binding domain. It also enhances the viral loads of upper respiratory tract and the binding and fusion of ACE2 in patients with COVID-19, which increased the spread of SARS-CoV-2 virus, resulting in the enhancement of the replication of human lung epithelial cells and primary human airway tissues as shown in Fig. 4B.

Combined with the contents of four pictures, we drew the Fig. 4E, which shows the logical lines that are contained in the four examples above. Fig. 4A only shows that D614G mutation will lead to higher infectivity of SARS-CoV-2 virus, but the addition of D614G mutation in Fig. 4C and 4D will lead to the enhancement of ACE2 binding and fusion, which makes SARS-CoV-2 virus produce more virus transmission and viral loads. Therefore, the logical relationship from S gene to ACE2 to SARS-CoV-2 was formed. As the virus infectivity increasing, a series of immune reactions will appear in the patients' body infected with SARS-CoV-2. This information is supplemented in Fig. 4B.

This example reflects not only the information at the molecular level, but also the information at the cellular level, which proves the feasibility of finding and forming a logical line from different texts. Therefore, an idea can be put forward that we can extract the key knowledge from the massive information and form a large logical network when the number of texts is enough. As a result, more hidden information can be discovered and new knowledge can be inferred.

Discussion

As indicated in this research, though single annotation is limited for comprehensive bio-medical knowledge discovery upon the huge literature repository for COVID-19, combination of relevant annotations from different resources makes it possible to bring a rich annotation data set which lead to knowledge with complete semantics.

Furthermore, the suggested knowledge pattern by using LitCovid-AGAC is capable of offering a huge amount of structured logic knowledge, and unveiling the pathological mechanism of COVID-19 in cellular or molecular level.

In addition, it as well makes sense to further curate the obtained results in LitCovid-AGAC, e.g., concept normalization, co-reference, and relation extraction. Meanwhile, it is instructive to visualize the knowledge entry in a syntactic way. The VSM box [13] in Fig. 5 presents a typical knowledge template which carries a type of semantic structure of the information in LitCovid-AGAC. The LOF/GOF/REG/COM can be inferred from the regulation annotations [7], and the pattern in this figure shows the effect of a protein on a disease.

ORCID

Sizhuo Ouyang: https://orcid.org/0000-0001-8335-9868
Yuxing Wang: https://orcid.org/0000-0003-4510-2783
Kaiyin Zhou: https://orcid.org/0000-0002-7314-9776
Jingbo Xia: https://orcid.org/0000-0002-7285-588X

Authors' Contribution

Conceptualization: JX. Data curation: YW, SO, KZ. Formal analysis: SO, YW, JX. Funding acquisition: JX. Methodology: SO, JX, YW. Writing - original draft: SO, JX. Writing - review & editing: JX.

https://doi.org/10.5808/gi.21013

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work is partially funded by the HZAU intramural innovative science funding, grant no. 2662021JC008. We would like to express our gratitude to many instructive discussion among BLAH7 Hackathon (https://blah7.linkedannotation.org/home). Qingyu Chen generously introduced LitCovid and PubTator annotation services. Fabio Rinaldi introduced OGER. Steven Vercruysse kindly offered the knowledge representation template in terms of AGAC mined logic for instructive visualization.

References

1. Chen Q, Allot A, Lu Z. Keep up with the latest coronavirus research. Nature 2020;579:193.
2. Chen Q, Allot A, Lu Z. LitCovid: an open database of COVID-19 literature. Nucleic Acids Res 2021;49:D1534-D1540.
3. Wei CH, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res 2019;47:W587-W593.
4. Furrer L, Jancso A, Colic N, Rinaldi F. OGER++: hybrid multi-type entity recognition. J Cheminform 2019;11:7.
5. Furrer L, Cornelius J, Rinaldi F. Parallel sequence tagging for concept recognition. Preprint at https://arxiv.org/abs/2003.07424 (2020).
6. Wang Y, Zhou K, Kim JD, Cohen KB, Gachloo M, Ren Y, et al. An active gene annotation corpus and its application on anti-epilepsy drug discovery. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2019 Nov 18-21; San Diego, CA, USA. New York: Institute of Electrical and Electronics Engineers, 2019. pp 512-519.
7. Zhou KY, Wang YX, Zhang S, Gachloo M, Kim JD, Luo Q, et al. GOF/LOF knowledge inference with tensor decomposition in support of high order link discovery for gene, mutation and disease. Math Biosci Eng 2019;16:1376-1391.
8. Wang Y, Zhou K, Gachloo M, Xia J. An overview of the active gene annotation corpus and the BioNLP OST 2019 AGAC Track Tasks. Proceedings of the 5th Workshop on BioNLP Open Shared Tasks; 2019 Nov 4; Hong Kong, China. Stroudsburg: Association for Computational Linguistics, 2019. pp 62-71.
9. Zhou K, Wang Y, Bretonnel Cohen K, Kim JD, Ma X, Shen Z, et al. Bridging heterogeneous mutation data to enhance disease gene discovery. Brief Bioinform 2021;22:bbab079.
10. Kim JD, Wang Y, Fujiwara T, Okuda S, Callahan TJ, Cohen KB. Open Agile text mining for bioinformatics: the PubAnnotation ecosystem. Bioinformatics 2019;35:4372-4380.
11. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, et al. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res 2008;36:D344-D350.
12. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25-29.
13. Vercruysse S, Zobolas J, Toure V, Andersen MK, Kuiper M. VSMbox: general-purpose interface for biocuration and knowledge representation. Preprint at https://www.preprints.org/manuscript/202007.0557/v1 (2020).

G&I
Genomics & Informatics

BLAH7
Application note

eISSN 2234-0742
Genomics Inform 2021;19(3):e24
https://doi.org/10.5808/gi.21008

Received: February 26, 2021
Revised: August 4, 2021
Accepted: August 12, 2021

*Corresponding author:
E-mail: mcbarros@fc.ul.pt

**Corresponding author:
E-mail: psruas@fc.ul.pt

***Corresponding author:
E-mail: dfsousa@lasige.di.fc.ul.pt

#These authors contributed equally to this work.

COVID-19 recommender system based on an annotated multilingual corpus

Márcia Barros[1,2#*], Pedro Ruas[1#**], Diana Sousa[1#***], Ali Haider Bangash[3,4], Francisco M. Couto[1]

[1]Large-Scale Informatics Systems Laboratory, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal
[2]Center for Astrophysics and Gravitation, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal
[3]Shifa College of Medicine, Shifa Tameer-e-Millat University, Islamabad 46000, Pakistan
[4]Working Group 3, COST Action EVidence-Based RESearch (EVBRES), Western Norway University of Applied Sciences, Inndalsveien 28, 5063 Bergen, Norway

Tracking the most recent advances in Coronavirus disease 2019 (COVID-19)–related research is essential, given the disease's novelty and its impact on society. However, with the publication pace speeding up, researchers and clinicians require automatic approaches to keep up with the incoming information regarding this disease. A solution to this problem requires the development of text mining pipelines; the efficiency of which strongly depends on the availability of curated corpora. However, there is a lack of COVID-19–related corpora, even more, if considering other languages besides English. This project's main contribution was the annotation of a multilingual parallel corpus and the generation of a recommendation dataset (EN-PT and EN-ES) regarding relevant entities, their relations, and recommendation, providing this resource to the community to improve the text mining research on COVID-19–related literature. This work was developed during the 7th Biomedical Linked Annotation Hackathon (BLAH7).

Keywords: COVID-19, entity extraction, recommendation, relation extraction, text mining
Availability: The code supporting our work and the resulting datasets are publicly available at https://github.com/lasigeBioTM/blah7.

Introduction

Coronavirus disease 2019 (COVID-19) pandemic took the world by surprise due to its impact on global public health. The scientific community, sensing the danger posed by this global health emergency, was quick to join hands in a bid to mitigate its effects. Natural language processing and machine learning (ML) research also focused in the quest of curbing the morbidity and mortality associated with the pandemic, being LitCovid [1] and CORD-19 [2] good examples of this effort. These resources are massive databases of scientific literature generated throughout the world pertinent to the COVID-19 pandemic—the characteristics of its causative organism (severe acute respiratory syndrome coronavirus 2), pathophysiology of the ailment as well as preventive measures that are suggested to be employed.

Recommender systems (RS) are tools for predicting the best items of interest for the users of a system, being mostly based on the past interests of the users. The interests of the users are usually collected through explicit or implicit feedback, for example, using a 5 stars system or the products opened by the users, respectively. The feedback is then used to create recommendation datasets of < user,item,rating > , useful for developing and evaluating

recommendation algorithms. The main approaches used in RS are collaborative-filtering, which uses the similarity between the ratings of the users and its only dependent on the feedback of the users, and content-based, which uses the similarity/relation between the items. RS have been widely used for recommending movies, books, or e-commerce, achieving excellent results. In scientific fields, such as Health and Life Sciences, RS began to be used with the goal of helping health staff and researchers, for example, by recommending drugs to a researcher based on the drugs that she/he already had interest in. The major challenge for RS in scientific fields is the lack of open source recommendation datasets. Some alternatives have been developed, one in particular called LIBRETTI, which uses the scientific literature for creating such datasets [3].

Earlier on, it was realized that a massive resource of literature surely would come in handy while developing management protocols and RS by training ML models. Therefore, efforts have been made to create such pipelines and to fit them onto ML models that allow recommendation [4]. However, since medical literature has its own specific linguistic characteristics and that is fairly more complex than generic text, it was observed that semi-automatic annotation is critical in creating a richer constellation of medical data that can be used for superior training of recommender ML models. Moreover, a large portion of health related text is normally generated in the native language, so text mining tools should also be able to process multilingual corpora.

Therefore, the goals of the present project were to retrieve COVID-19 related documents, to automatically annotate them with entities and relations, generate recommendation datasets of scientific entities, and to manually validate a sample of the obtained annotations. The recommendation datasets are then used to develop new recommendation algorithms in the field of COVID-19.

The contributions of the present work are an automatic pipeline for document retrieval, entity and relation extraction, and recommendation, as well as a set of multilingual parallel datasets (English/Portuguese/Spanish) related with COVID-19 that allows the evaluation of Named Entity Recognition/Linking, Relation Extraction (RE), and Recommendation Systems. We also developed a new recommendation algorithm, called Relation Recommendation Algorithm (RelRA), and conducted preliminary tests with it.

Methodology

Fig. 1 presents the general workflow and the tools used throughout our work.

Document retrieval

The first step was to retrieve COVID-19 related abstracts from PubMed repository using the Bio.Entrez package (https:// biopython.org/docs/1.75/api/Bio.Entrez.html), which is part of Biopython. We used PubMed since it allows the abstract retrieval in more than one language, in our case, we needed English, Spanish, and Portuguese abstracts. Two versions of the dataset were created using different queries: *abstracts_covid_19*, which includes abstracts directly related with COVID-19 and *abstracts_large*, which includes abstracts directly and indirectly related with COVID-19. The queries used are present in Table 1.

Entity extraction

The second step was to extract named entities present in the retrieved documents, more concretely, by performing Named Entity Recognition and Named Entity Linking. This step was accomplished by the Python implementation of MER [5], a dictionary matching system that, given a lexicon with the terms of an ontology or any knowledge base, recognizes entities in text and links them to the respective identifiers. The MER tool is light and efficient, since it does not require neither labelled data for training, as the SOTA supervised approaches usually requires (e.g., BERT) nor extensive time-consuming training. Besides, it works with any given lexicon, even if it is non-English. Consequently, we considered the tool as adequate to achieve our goals in this short-term project. Biomedical entities present in Portuguese, Spanish, and English abstracts were recognized by MER and then linked to the respective DeCS ("Descritores em Ciências da Saúde", https:// decs.bvsalud.org/) term (September 2020 edition). DeCS is multilingual biomedical vocabulary built upon MeSH terminology. It includes versions in several languages, such as Portuguese, Spanish, and English, so it is suitable for our goal of creating a multilingual dataset. Almost all DeCS terms are MeSH terms, however, there exist some specific DeCS terms that do not correspond to any MeSH term (4,378 out of 34,294 terms). Additionally, recognized biomedical entities in English abstracts were linked to the following ontologies (latest edition available at January 2021): Human Disease Ontology (DO, https://disease-ontology.org/), Gene Ontology (GO, http://geneontology.org/), Human Phenotype Ontology (HPO, https://hpo.jax.org/app/), Chemical Elements of Biological Interest (ChEBI) Ontology (https://www.ebi.ac.uk/chebi/), and Coronavirus Infectious Disease Ontology (CIDO, https://github.com/CIDO-ontology/cido).

Relation extraction

In the third step, the relation extraction module performs RE by applying the BiOnt [6] system, built to allow the extraction of relations between biomedical entities supported by ontologies (e.g., HPO and GO). We opted for BiOnt due to its unique use of added

https://doi.org/10.5808/gi.21008

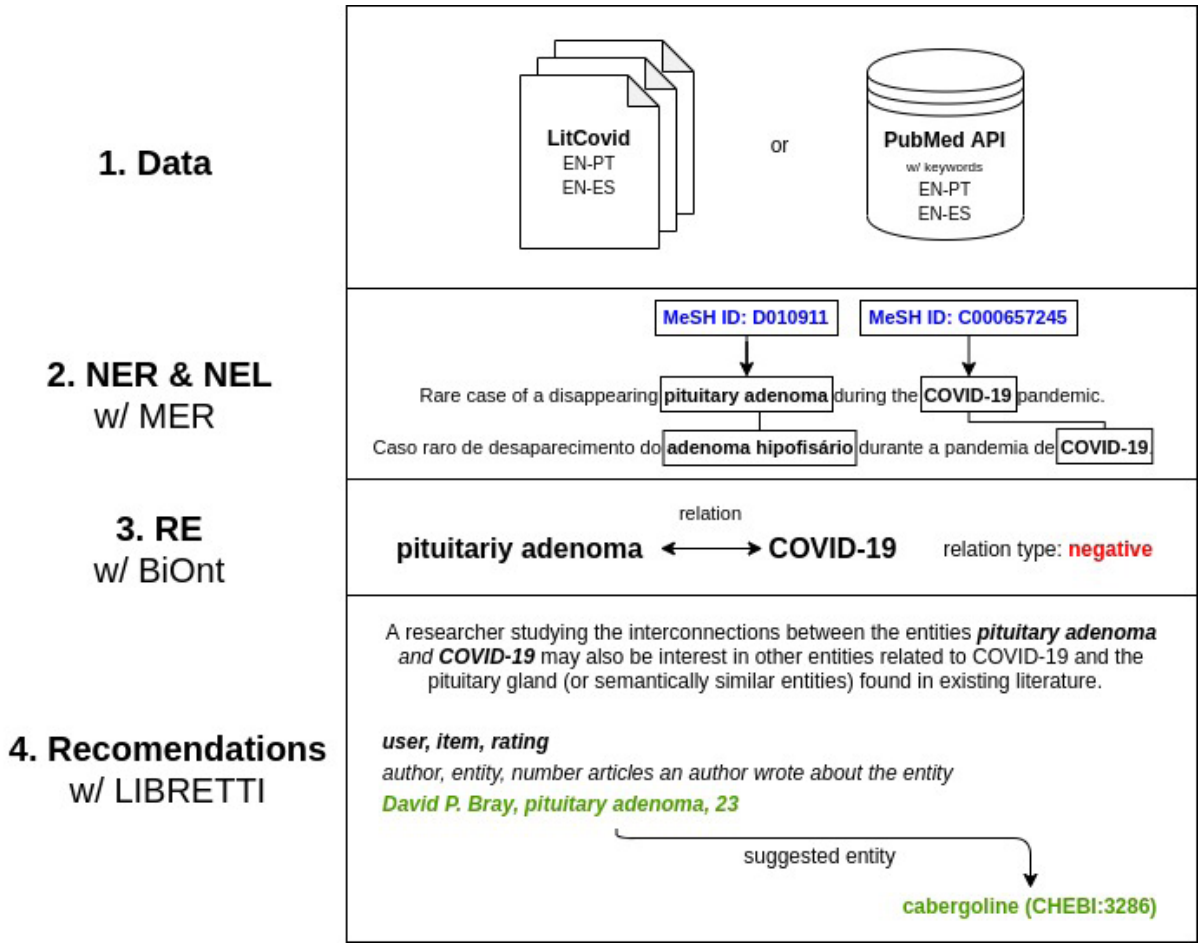

Fig. 1. Pipeline with the tools used at each stage with an example retrieved from article PMID:33220478. NER, Named Entity Recognition; NEL, Named Entity Linking; RE, Relation Extraction; LIBRETTI, LIterature Based RecommEndaTion of scienTific Items; COVID-19, coronavirus disease 2019.

Table 1. Queries used for document retrieval

Set	English query	Portuguese query	Spanish query
abstracts_covid_19	*covid-19 AND English [LANG]*	*AND English [LANG] AND Portuguese [LANG]*	*AND English [LANG] AND Spanish [LANG]*
abstracts_large	*2019 Novel Coronavirus Disease OR 2019 Novel Coronavirus Infection OR 2019-nCoV Disease OR 2019-nCoV Infection OR COVID-19 Pandemic OR COVID-19 Pandemics OR COVID-19 Virus Disease OR COVID-19 Virus OR Infection OR COVID19 OR Coronavirus Disease 2019 OR Coronavirus Disease-19 OR SARS Coronavirus 2 Infection OR SARS-CoV-2 Infection AND English [LANG]*	*2019 Novel Coronavirus Disease OR 2019 Novel Coronavirus Infection OR 2019-nCoV Disease OR 2019-nCoV Infection OR COVID-19 Pandemic OR COVID-19 Pandemics OR COVID-19 Virus Disease OR COVID-19 Virus OR Infection OR COVID19 OR Coronavirus Disease 2019 OR Coronavirus Disease-19 OR SARS Coronavirus 2 Infection OR SARS-CoV-2 Infection AND English [LANG] AND Portuguese [LANG]*	*2019 Novel Coronavirus Disease OR 2019 Novel Coronavirus Infection OR 2019-nCoV Disease OR 2019-nCoV Infection OR COVID-19 Pandemic OR COVID-19 Pandemics OR COVID-19 Virus Disease OR COVID-19 Virus OR Infection OR COVID19 OR Coronavirus Disease 2019 OR Coronavirus Disease-19 OR SARS Coronavirus 2 Infection OR SARS-CoV-2 Infection AND English [LANG] AND Spanish [LANG]*

external knowledge in the form of biomedical ontologies to potentiate RE. Thus, instead of relying on just the training data for the learning process as most RE systems, BiOnt also adds the ancestry information to each entity in a candidate pair by matching it to an ontology term.

In new data, the BiOnt system can identify relations between different and the same type of biomedical entities, such as diseases and human phenotypes, provided we use the pre-trained models trained on available training data. Using the pre-trained models available, we extracted relations between ChEBI and DO entities

and between GO and HPO entities for this project. BiOnt does not make available pre-trained models for all other combinations that we could apply to our entities. We only considered relations in English abstracts since both Portuguese and Spanish abstracts only had annotated MeSH terms, for which we did not have pre-trained models. We did not restrain the relations to sentence-level and instead considered relations within the same abstract. However, since our Portuguese and Spanish abstracts can be linked to their respective English versions, it is also possible to map the extracted relations from English to the corresponding translated abstracts.

Recommendation

Dataset creation

In the fourth step, the goal was to create multilingual recommendation datasets. The datasets were created using a methodology called LIterature Based RecommEndaTion of scienTific Items (LIBRETTI), which consists in developing a standard <user,item,rating> dataset from research articles. The users are the authors from the scientific articles, the items are biomedical entities mentioned in the articles, and the ratings are the number of articles an author wrote about an item [3]. For this work, the input research corpus is the one retrieved in phase 1: Document retrieval, more specifically the abstracts_large collection. The items are biomedical entities recognized in phase 2, i.e., diseases from the DO, gene terms from GO, phenotypes from HPO, and chemical compounds from ChEBI.

RelRA

The primary goal of the RS developed in this work is to recommend entities related to the COVID-19 disease to the researchers. To that end, we developed a new recommender content-based algorithm, based on the relations between the items - RelRA. We developed RelRA during 7th Biomedical Linked Annotation Hackathon (BLAH7), and conducted the first experiments. RelRA is based on the relations between the entities extracted in phase 2. It integrates phase 3: Relation extraction.Consider a user who has already rated some items, we want to know which items are suitable recommendations for this user. The goal of the algorithm is to provide a score to each unrated item in order to rank them. For that, we use the relations between the items, and the score is calculated considering how many relations an unrated item has with the items in the rated list. For this work we used a list of relations created using the method described in phase 3: Relation Extraction.

The RelRa algorithm was evaluated in the datasets EN_PT and EN_ES, created in phase 4. To avoid biases, the list of relations used for evaluating the RelRA algorithm were extracted from a sample of the CORD-19 corpus (nine thousand documents, version from 2020-03-13) with research articles completely different from those used to create the recommendation datasets.

Evaluation

For the evaluation, we tested the RelRa algorithm against a random algorithm. We used a cross-validation strategy, with 80% for the training set and 20% for the test set. For the evaluation the datasets were filtered, thus each user had at least 20 items rated. The evaluation metrics are Precision, Recall, and Mean Reciprocal Ranking (MRR) [7].

Manual validation at BLAH7

For manual validation of the obtained annotations, we randomly selected a sample of 40 English (20) and Portuguese (20) abstracts belonging to *abstracts_covid_19* set with entity annotations and, in the case of the English abstracts, also with relation annotations. During BLAH7, annotations were uploaded to PubAnnotation (http://pubannotation.org/) and 4 participants were responsible for the correction of the existing entity and relation annotations, but also for the addition of new annotations, if deemed necessary.

Results and Discussion

Statistics about the retrieved documents and the dimensions of the recommendations datasets created from each corpus are available in (Table 2). As expected, the abstracts_large datasets have a much higher number of documents, both in PT and ES, than the limited

Table 2. Number of documents in each version of the dataset, and respective dimensions of the recommendation datasets

Set	Languages	Abstracts	nUsers	nItems	nRatings
abstracts_covid_19	EN_PT	80	1,750	1,507	36,614
	EN_ES	53	1,744	669	14,920
abstracts_large	EN_PT	346	1,869	2,403	49,839
	EN_ES	390	1,855	1,036	20,417

EN_PT, created from abstracts in English ("EN") and Portuguese ("PT"); PT, created from abstracts in Portuguese ("PT"); EN_ES, created from abstracts in English ("EN") and Spanish ("ES"); ES, created from abstracts in Spanish ("ES"). nUsers, number of users; nItems, number of items; nRatings, number of ratings.

https://doi.org/10.5808/gi.21008

abstracts_COVID_19. Therefore, the recommendations datasets created from the abstracts_large corpus have as well a higher number of users, items and ratings.

The results of the algorithms RelRA and Random (RAND) for the datasets EN_PT and EN_ES are presented in Fig. 2, respectively, for the evaluation metrics of Precision, Recall and MRR, for the top@5 ranked results.

Fig. 2 shows that RelRa achieved better results for all the evaluation metrics when compared with a random recommendation of the items, for both EN_PT and EN_ES datasets. RelRa seems to have great potential for the recommendation of scientific entities based on their relations, however, the number of documents and relations needs to be improved for further testing.

Fig. 3 shows an example of recommendation. The user represented had interest in six different entities. We have a list of three entities that we wish to know if are suitable for recommendation to this user. Using RelRA, we find the relations between the list of liked items and these unknown items. Severe acute respiratory syndrome (DOID_2945) is related to two items in the liked list, thus it has a score of 2/6. Propyzamide (CHEBI_34935) is related to three entities in the liked list, achieving a score of 3/6. Inflammatory response (GO_0006954) does not have any relation with the items liked by

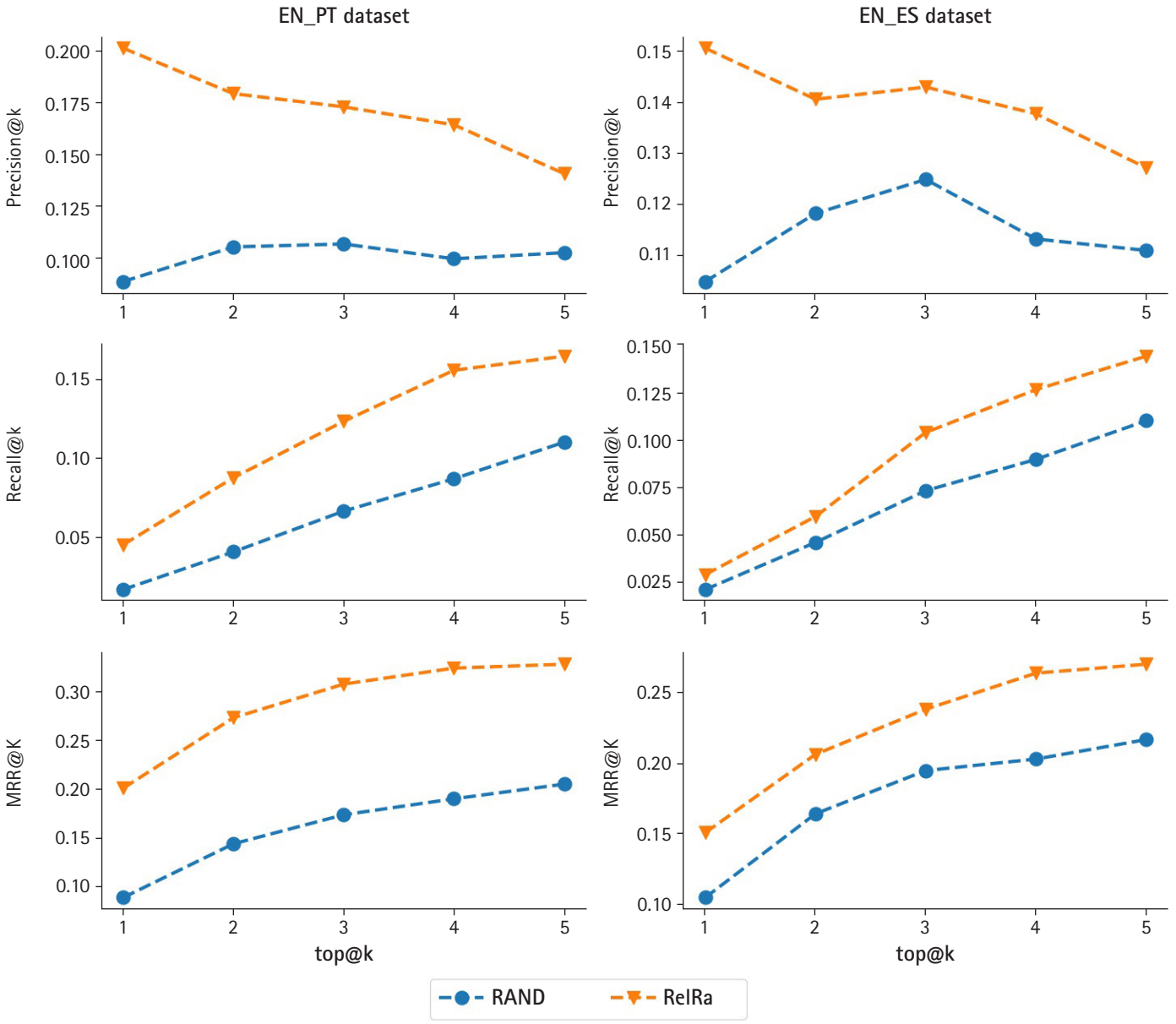

Fig. 2. Results for RelRa vs. RAND, for precision, recall and MRR, for EN_PT and EN_ES datasets. RelRA, Relation Recommendation Algorithm; RAND, Random; MRR, Mean Reciprocal Ranking; EN_PT, created from abstracts in English ("EN") and Portuguese ("PT"); EN_ES, created from abstracts in English ("EN") and Spanish ("ES").

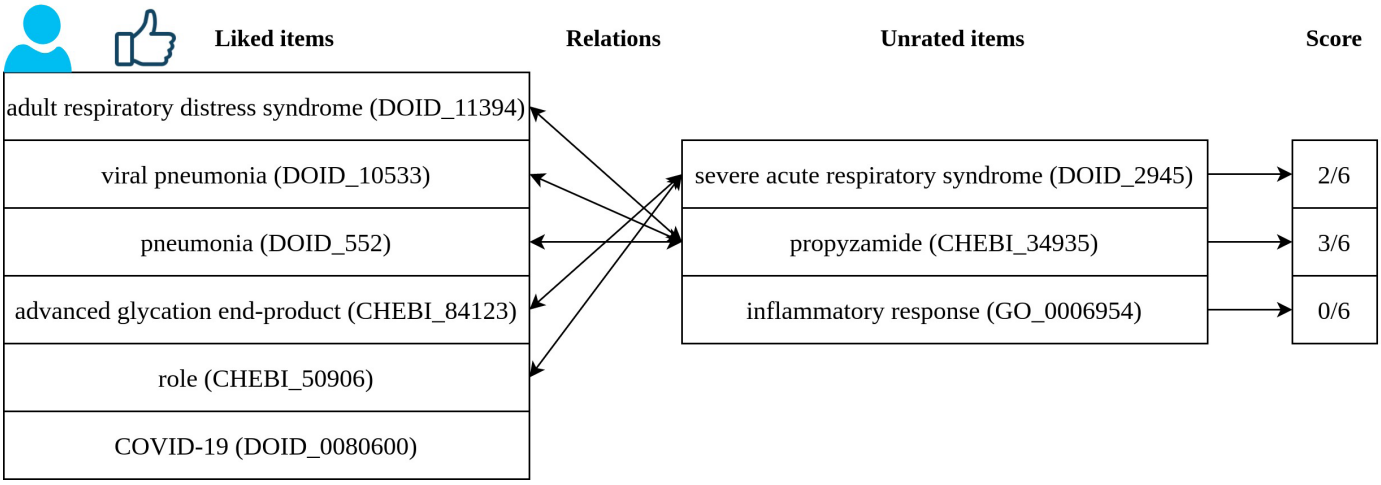

Fig. 3. Example of recommendation using the RelRA algorithm. RelRA, Relation Recommendation Algorithm; COVID-19, coronavirus disease 2019.

this user. Severe acute respiratory syndrome (DOID_2945) and Propyzamide (CHEBI_34935) would be recommended to this user.

These are the first tests with RelRA, which was specially developed during BLAH7.

Manual validation at BLAH7

Table 3 presents the results for the manual validation stage at BLAH7. After the validation process, for the 40 documents, we retrieved more entities, more relations, and with better quality by discarding illy annotated entities and relations. Further, by reaching a consensus between the four annotators, we increased our datasets' quality by adding even more annotations. These datasets are available in the PubAnnotation (http://pubannotation.org/collections/LASIGE:%20Annotating%20a%20multilingual%20COVID-19-related%20corpus%20for%20BLAH7) platform (in their original and consensus format).

Table 3. Final counts for the 40 abstracts sample (20 English and 20 Portuguese), the mean number of each subset for the annotators/curators task, and the final consensus numbers of manual validation

Dataset		Original	Annotated/Curated	Consensus
Portuguese	Entities	245	322	354
English	Entities	493	511	607
	Relations	224	238	250

Conclusion

Our goals for the present project were to retrieve COVID-19 related documents, to automatically annotate them with entities and relations, generate recommendation datasets of scientific entities, and to manually validate a sample of the obtained annotations.

We were able to create an automatic pipeline for document retrieval, entity and relation extraction, and recommendation, as well as a set of multilingual parallel datasets (English/Portuguese/Spanish) related with COVID-19 that allows the evaluation of Named Entity Recognition/Linking, RE, and Recommendation Systems. Further, we partially manually validated our datasets using the PubAnnotation platform.

For future work, the manual validation of the annotations could be improved, more concretely, by leveraging crowdsourcing platforms to recruit a large number of annotators [8]. Besides, due to time constraints, we were not able to manually validate the annotations present in EN_ES datasets during BLAH7, so future work could accomplish this.

ORCID

Márcia Barros: https://orcid.org/0000-0002-9728-9618
Pedro Ruas: https://orcid.org/0000-0002-1293-4199
Diana Sousa: https://orcid.org/0000-0003-0597-9273
Ali Haider Bangash: https://orcid.org/0000-0002-8256-3194
Francisco M. Couto: https://orcid.org/0000-0003-0627-1496

Authors' Contribution

Conceptualization: MB, PR, DS, FMC. Data curation: MB, PR, DS, AHB. Formal analysis: MB, PR, DS, AHB. Funding acquisition: FMC. Methodology: MB, PR, DS, FMC. Writing - original

https://doi.org/10.5808/gi.21008

draft: MB, PR, DS, AHB. Writing - review & editing: MB, PR, DS, AHB, FMC.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work was supported by FCT through funding of Deep Semantic Tagger (DeST) project (ref. PTDC/CCI-BIO/28685/2017) and LASIGE Research Unit (ref. UIDB/00408/2020 and ref. UIDP/00408/2020); and FCT and FSE through funding of PhD Scholarship, ref. 2020.05393.BD, PhD Scholarship, ref. SFRH/BD/128840/2017, and PhD Scholarship, ref. SFRH/BD/145221/2019.

References

1. Chen Q, Allot A, Lu Z. LitCovid: an open database of COVID-19 literature. Nucleic Acids Res 2021;49:D1534-D1540.
2. Lu Wang L, Lo K, Chandrasekhar Y, Reas R, Yang J, Eide D, et al. CORD-19: The Covid-19 Open Research Dataset. Preprint at: https://arxiv.org/abs/2004.10706 (2020).
3. Barros M, Moitinho A, Couto FM. Using research literature to generate datasets of implicit feedback for recommending scientific items. IEEE Access 2019;7:176668-176680.
4. Tworowski D, Gorohovski A, Mukherjee S, Carmi G, Levy E, Detroja R, et al. COVID19 Drug Repository: text-mining the literature in search of putative COVID19 therapeutics. Nucleic Acids Res 2021;49:D1113-D1121.
5. Couto FM, Lamurias A. MER: a shell script and annotation server for minimal named entity recognition and linking. J Cheminform 2018;10:58.
6. Sousa D, Couto FM. BiOnt: deep learning using multiple biomedical ontologies for relation extraction. In: Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science, Vol. 12036 (Jose JM, Yilmaz E, Magalhaes J, Castells P, Ferro N, Silva MJ, et al., eds.). Cham: Springer, 2020. pp. 367-374.
7. Shani G, Gunawardana A. Evaluating recommendation systems. In: Recommender Systems Handbook (Ricci F, Rokach L, Shapira B, Kantor P, eds.). Boston: Springer, 2011. pp. 257-297.
8. Sousa D, Lamurias A, Couto FM. A hybrid approach toward biomedical relation extraction training corpora: combining distant supervision with crowdsourcing. Database (Oxford) 2020; 2020:baaa104.

G&I
Genomics & Informatics

BLAH7
Application note

eISSN 2234-0742
Genomics Inform 2021;19(3):e25
https://doi.org/10.5808/gi.21012

Received: March 17, 2021
Revised: September 1, 2021
Accepted: September 13, 2021

*Corresponding author:
E-mail: atsuko@tcu.ac.jp

#These authors contributed equally to this work.

Constructing Japanese MeSH term dictionaries related to the COVID-19 literature

Atsuko Yamaguchi[1#*], Terue Takatsuki[2#*], Yuka Tateisi[3#*], Felipe Soares[4]

[1]Graduate School of Integrative Science and Engineering, Tokyo City University, Tokyo 158-8557, Japan
[2]Database Center for Life Science, Research Organization of Information and Systems, Kashiwa 277-0871, Japan
[3]National Bioscience Database Center, Japan Science and Technology Agency, Chiyoda, Tokyo 102-8666, Japan
[4]Computer Science Department, The University of Sheffield, Western Bank, Sheffield S10 2TN, UK

The coronavirus disease 2019 (COVID-19) pandemic has led to a flood of research papers and the information has been updated with considerable frequency. For society to derive benefits from this research, it is necessary to promote sharing up-to-date knowledge from these papers. However, because most research papers are written in English, it is difficult for people who are not familiar with English medical terms to obtain knowledge from them. To facilitate sharing knowledge from COVID-19 papers written in English for Japanese speakers, we tried to construct a dictionary with an open license by assigning Japanese terms to MeSH unique identifiers (UIDs) annotated to words in the texts of COVID-19 papers. Using this dictionary, 98.99% of all occurrences of MeSH terms in COVID-19 papers were covered. We also created a curated version of the dictionary and uploaded it to PubDictionary for wider use in the PubAnnotation system.

Keywords: COVID-19, Japanese, medical vocabulary, MeSH
Availability: The dictionaries that we constructed are available at https://github.com/aco-pom/blah7-dic under Creative Commons Attribution-ShareAlike 4.0 International license.

Introduction

An enormous number of research papers about the coronavirus disease 2019 (COVID-19) pandemic has recently been published, and knowledge about COVID-19 is very frequently updated through research articles. As a result of these rapid updates, people's understanding easily becomes outdated. It may be especially difficult for people who are not English speakers to understand research papers about COVID-19 because they include many English medical terms. To share up-to-date knowledge among Japanese speakers from COVID-19 research papers written in English, a Japanese dictionary of medical terms related to COVID-19 is required.

Due to the time constraints of the 7th Biomedical Linked Annotation Hackathon (BLAH7) hackathon, we focused on the Medical Subject Headings (MeSH) unique identifiers (UIDs) annotated in LitCovid [1], which is a curated literature hub providing centralized access to relevant articles in PubMed. Our goal of the hackathon was to construct a dictionary with an open license that has maps of MeSH UIDs in LitCovid and Japanese terms with as many curated terms as possible.

Existing Japanese translations of MeSH terms were not suitable for our purpose. There is a translation of MeSH into Japanese (MeSHJPN) included in the Unified Medical Language Systems (UMLS), provided by the Japan Medical Abstracts Society (JAMAS). The translation is based on the JAMAS Japanese Medical Thesaurus (JJMT). The latest version of JJMT was released in 2019, and corresponds to MeSH 2018, but MeSHJPN is based on the previous version (2014). This means that both dictionaries are older than the onset of the COVID-19 pandemic. In addition, JJMT is not downloadable, and MeSHJPN has a Category 3 License Restriction (https://uts.nlm.nih.gov/uts/license/license-category-help.html#category3) that prohibits the incorporation of the dictionary into any publicly accessible computer-based information systems. Thus, those resources are not suitable for sharing knowledge in an open manner. There are other translations of MeSH into Japanese but, as reported in 5th Biomedical Linked Annotation Hackathon (BLAH5) [2], they have similar problems.

Therefore, we tried to construct a dictionary with an open license by using datasets that are freely available. By using six datasets with open licenses and adding some Japanese terms manually, we constructed a dictionary with a map from MeSH UIDs to Japanese terms.

Methods

As described above, we focused on the MeSH UIDs annotated in LitCovid. We first extracted the MeSH UIDs from LitCovid and sorted them according to the number of occurrences. Next, we obtained pairs of MeSH UIDs and the corresponding Japanese terms using the following six datasets that are freely available.

Wikidata

Wikidata (https://wikidata.org/) is a free and open knowledge base that includes MeSH UIDs and their multilingual labels. Wikidata provides a SPARQL endpoint with a graphical user interface (https://query.wikidata.org/). Using the SPARQL endpoint, we obtained pairs of MeSH UIDs and Japanese labels.

Japan Science and Technology Agency thesaurus headwords from the MeCab user dictionary for science technology terms

The National BioScience Database Center provides dictionaries for the Japanese morphological analyzer MeCab, based on life science-related vocabulary from the Japan Science and Technology Agency (JST) thesaurus, 2015 edition. Although it is not based on the latest version of the JST thesaurus, which is for profit, the user dictionaries based on the 2015 edition are available through a Creative Commons Attribution-Share Alike 4.0 International (CC-BY-SA 4.0). One of the dictionaries consists of terms that can be assigned MeSH UIDs via the Interlinking Ontology of Biological Concepts [3] (https://bioportal.bioontology.org/ontologies/IOBC).

The dictionary can be downloaded from Life Science Database Archive (https://dbarchive.biosciencedbc.jp/en/mecab/ data-2.html). We used the correspondences of Japanese terms and MeSH UIDs from the dictionary.

Human Phenotype Ontology

The Human Phenotype Ontology (HPO) [4] is a standardized vocabulary of phenotypic abnormalities encountered in human disease. HPO terms have links to MeSH UIDs, and Japanese labels for HPO terms are freely available at https:// github.com/ogishima/HPO-japanese. Therefore, we were able to obtain Japanese terms for the MeSH UIDs included in the HPO.

Kyoto Encyclopedia of Genes and Genomes (KEGG) disease and KEGG drug

The KEGG is a collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances [5]. Although the KEGG database is not public, a subset of the databases about medical science, including KEGG Disease and KEGG Drug, is freely available from https://dbarchive.biosciencedbc.jp/en/kegg-medicus/download.html with a CC-BY-SA 4.0 license. These include both English and Japanese labels for the entries. For each MeSH term, if the dictionary included exactly the same English word as a label for an entry, we obtained the Japanese label for the same entry for the MeSH UID.

6th Biomedical Linked Annotation Hackathon (BLAH6) dictionary

Yamada and Tateisi [6] created a dictionary of MeSH UIDs and Japanese terms at the BLAH6 hackathon by using the open-japanese-mesh script (https://github.com/roy29fuku/open-japanese-mesh) from two glossaries: MeSpEn by Barcelona Supercomputing Center (https://temu.bsc.es/mespen/) and MEDUTX (an English-Japanese dictionary by Kitasato University available from Asia-Pacific Association for Machine Translation, https://aamt.info/wp-content/uploads/2019/06/medutx1.05.zip) as reported previously [6]. We obtained the Japanese terms for MeSH IDs from the dictionary.

We obtained pairs of MeSH UIDs and Japanese terms independently for each dataset. Next, we divided the MeSH UIDs into two groups according to whether they had at least 50 occurrences or fewer than 50 occurrences. For the MeSH UIDs in the first group, if two or more different Japanese terms are assigned by the

https://doi.org/10.5808/gi.21012

dictionaries, we selected one Japanese term manually. If no Japanese term was assigned to a MeSH UID in the first group, three native Japanese speakers who have been researchers in the life-sciences domain for more than 20 years mapped the Japanese terms independently and selected one after discussion. Thus, for the first group, every MeSH UID was manually assigned to a Japanese term. The number of terms in the first group for which two or more different Japanese terms were assigned, the number of terms for which one term was assigned and the number of terms for which no term was assigned were 829, 150, and 60, respectively. We termed the dictionary including MeSH UIDs in the first group and Japanese terms assigned manually "curated."

For the MeSH UIDs in the second group, if two or more Japanese terms are assigned, one Japanese term was automatically assigned according to the priority level of the six datasets. The list of the datasets in order of high to low priority was: KEGG Disease, KEGG Drug, JST thesaurus headwords, BLAH6 dictionary, HPO, and Wikidata. The highest priority was given to datasets manually curated by life-sciences experts, as described in the Discussion section. MeSH UIDs were then assigned to Japanese terms according to the priority of the datasets for MeSH UIDs such that at least one term was assigned using the six datasets. By combining the dictionary of the second group with the "curated" dictionary, we constructed a larger dictionary, which we termed "all."

Results

From LitCovid, which is annotated by Pubtator, 8419 MeSH UIDs were obtained. The total number of occurrences of the MeSH UIDs in LitCovid was 989,994. Fig. 1 shows the distribution of the number of occurrences of the MeSH UIDs. The x-axis shows the MeSH UIDs sorted by their occurrences, and the y-axis shows the number of occurrences on a logarithmic scale. As shown in Fig. 1, only a few words appeared very frequently and many words appeared only a few times.

Table 1 shows the number of terms assigned MeSH UIDs, the number of terms for the 8419 MeSH UIDs and the total occurrences of terms for the 989,994 occurrences in LitCovid in the six datasets.

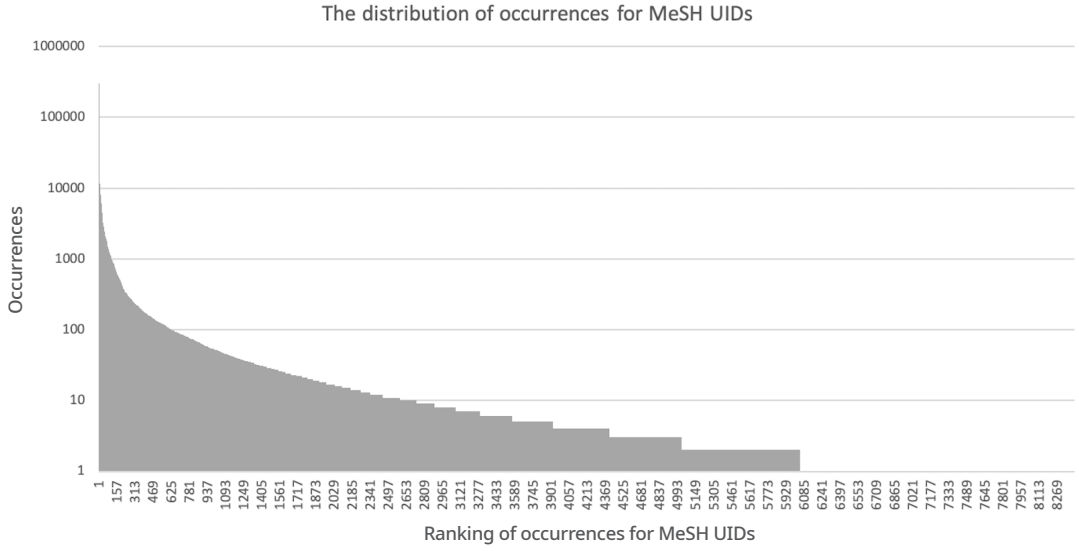

Fig. 1. The distribution of occurrences of MeSH UIDs.

Table 1. Japanese terms for MeSH UIDs obtained from each dataset

Dataset	No. of MeSH terms	Terms in LitCovid	Occurrences in LitCovid
Wikidata	15229	3565	887742
JST thesaurus headwords	15425	2026	353482
HPO	2176	1132	236892
KEGG Disease	2302	961	98616
KEGG Drug	-	1558	88118
BLAH6 dictionary	12771	2811	495494

MeSH, medical subject heading; UID, unique identifier; JST, Japan Science and Technology Agency; HPO, Human Phenotype Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; BLAH6, 6th Biomedical Linked Annotation Hackathon.

There were no entries in KEGG Drug with MeSH UIDs; therefore, we assigned the MeSH UIDs to the Japanese terms that shared KEGG IDs with the English terms that exactly matched the MeSH entry terms, as explained in the Materials and Methods section.

As described in the Materials and Methods section, by using the Japanese terms obtained from datasets, we constructed "curated" and "all" dictionaries. The numbers of MeSH UIDs in the "curated" and "all" dictionaries were 1039 and 5805, respectively. Therefore, 12.34% and 68.95% of 8419 MeSH UIDs appearing in LitCovid were covered by the "curated" and "all" dictionaries. These numbers may seem quite small, especially for the "curated" dictionary. However, because the distribution of occurrences has a long tail, as shown in Fig. 1, and all MeSH UIDs with at least 50 occurrences were included in the "curated" dictionary, the "curated" and "all" dictionaries covered 94.63% and 98.99% of occurrences, respectively.

The "curated" dictionary is expected to be useful for text annotation because all Japanese terms in the dictionary were curated manually. Therefore, we uploaded it to the PubDictionaries system (https://pubdictionaries.org/) for wider use. The "all" dictionary may be useful in cases where more Japanese terms for MeSH UIDs are preferable. Both the "curated" and "all" dictionaries are available at https://github.com/acopom/blah7-dic under Creative Commons Attribution-ShareAlike 4.0 International license.

To check the quality of the dictionaries, we used some abstracts in LitCovid that also had Japanese translations by the authors. For example, the following sentence: "Cerebrovascular disease and vasculitis-related diseases have been reported as systemic complications of coronavirus disease 2019 (COVID-19)," which occurred in the abstract of PubMed ID 33051389, was annotated by PubTator as follows:

D002561: Cerebrovascular disease
D014657: vasculitis
C000657245: coronavirus disease 2019
C000657245: COVID-19

By using the "All" dictionary, we could translate these annotations as follows.

D002561: 脳血管障害
D014657: 脈管炎|血管炎
C000657245: 新型コロナウイルス感染症

The Japanese translation by the author for the sentence was "新型コロナウイルス感染症COVID-19蔓延に伴い，脳血管障害や血管炎関連疾患の合併が報告されるようになった。", where the words in bold correspond to the translations of annotated words. Our dictionary assigned the same translations as the terms used by the author. In another example from PubMed ID 33051386, the sentence "Neuromuscular complications such as cerebrovascular disease, encephalopathy, meningoencephalitis, peripheral neuropathy, and myositis/myopathy have been reported to date." was annotated by PubTator as follows.
D002561: cerebrovascular disease
D001927: encephalopathy
D008590: meningoencephalitis
D010523: peripheral neuropathy
D009220: myositis
D009135: myopathy

By using the "all" dictionary, we could translate these annotations as follows.

D002561: 脳血管障害
D001927: 脳疾患
D008590: 髄膜脳炎
D010523: 末梢神経系疾患
D009220: 筋炎
D009135: 筋疾患

The Japanese translation by the author for the sentence was "神経筋合併症としては，脳血管障害，脳症，髄膜脳炎，末梢神経障害，筋障害などが報告されている。".

In this case, although some terms in the sentence seem different from those of the dictionary, "脳症"/"脳疾患" and "末梢神経障害"/"末梢神経系障害" are synonyms. In addition, "筋障害" in the translation by the author is a broader concept than "筋炎" and "筋疾患" assigned by the dictionary. Therefore, we could confirm the quality of the dictionaries to some extent.

Discussion

The six dictionaries we used have advantages and disadvantages. KEGG and JST contain curated data by experts and are reliable, but expert curation also involves the disadvantage of slow updates. In particular, JST is a dataset from 2015 that does not include new terms related to COVID-19. The same can be said for the BLAH6 dataset, which was based on dictionaries for machine translation in biology and medicine made by domain experts. Another feature of the BLAH6 dataset is that it contains duplicate spellings of the same word (in kanji and hiragana, for example). This is an advantage for use in applications such as assigning MeSH UIDs to Japanese texts, but for English-to-Japanese translations, another step is necessary to choose the standard spelling. In contrast, HPO and

https://doi.org/10.5808/gi.21012

Wikidata are more up-to-date, but not as reliable as other dictionaries. The Japanese terms in HPO are machine translations of the English terms. Wikidata is written by human volunteers who are not guaranteed to be biomedical domain experts.

The MeSH UID for COVID-19 is C000657245 in MeSH 2020, which is also the case in the PubTator data we used for the experiments. However, in MeSH 2021, its status was updated from a Supplementary Concept to a Descriptor, and the UID is now D000086382. As shown by this example, NLM is quick to update MeSH, but it is difficult for official translations to keep up with this pace of MeSH updates. Thus, the methods for the quick and easy translation of terms will be necessary in events such as the current COVID-19 pandemic.

Conclusion

We constructed two Japanese MeSH term dictionaries called "curated" and "all." The numbers of MeSH UIDs in the "curated" and "all" dictionaries are 1039 and 5805, respectively. Although the numbers of MeSH UIDs may seem small, the "curated" and "all" dictionaries covered 94.63% and 98.99% of occurrences in LitCovid, respectively.

As discussed above, methods for quickly constructing a reliable dictionary of Japanese MeSH terms should be considered. To achieve this goal, methods of measuring and guaranteeing the reliability of translations of terms should be presented.

ORCID

Atsuko Yamaguchi: https://orcid.org/0000-0001-7538-5337
Terue Takatsuki: https://orcid.org/0000-0003-0011-764X
Yuka Tateisi: https://orcid.org/0000-0002-3813-5782
Felipe Soares: https://orcid.org/0000-0002-2837-1853

Authors' Contribution

Conceptualization: FS. Data curation: AY, TT, YT. Formal analysis: AY, TT, YT. Methodology: AY, TT, YT, FS. Writing - original draft: AY. Writing - review & editing: AY, TT, YT.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

References

1. Chen Q, Allot A, Lu Z. Keep up with the latest coronavirus research. Nature 2020;579:193.
2. Tateisi Y. Resources for assigning MeSH IDs to Japanese medical terms. Genomics Inform 2019;17:e16.
3. Kushida T, Tateisi Y, Masuda T, Watanabe K, Matsumura K, Kawamura T, et al. Refined JST thesaurus extended with data from other open life science data sources. In: Semantic Technology: JIST2017, Lecture Notes in Computer Science, Vol. 10675 (Wang Z, Turhan AY, Wang K, Zhang X, eds.). Cham: Springer, 2017. pp. 35-48.
4. Kohler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The Human Phenotype Ontology in 2021. Nucleic Acids Res 2021;49:D1207-D1217.
5. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res 2021;49:D545-D551.
6. Yamada R, Tateisi Y. open-japanese-mesh: assigning MeSH UIDs to Japanese medical terms via open Japanese-English glossaries. Genomics Inform 2020;18:e22.

G&I
Genomics & Informatics

BLAH7
Application note

eISSN 2234-0742
Genomics Inform 2021;19(3):e26
https://doi.org/10.5808/gi.21014

Received: March 17, 2021
Revised: September 8, 2021
Accepted: September 10, 2021

*Corresponding author:
E-mail: fs@felipesoares.net

O-JMeSH: creating a bilingual English-Japanese controlled vocabulary of MeSH UIDs through machine translation and mutual information

Felipe Soares[1]*, Yuka Tateisi[2], Terue Takatsuki[3], Atsuko Yamaguchi[4]

[1]Computer Science Department, The University of Sheffield, Western Bank, Sheffield S10 2TN, UK
[2]National Bioscience Database Center, Japan Science and Technology Agency, Tokyo 102-8666, Japan
[3]Database Center for Life Science, Research Organization of Information and Systems, Kashiwa 277-0871, Japan
[4]Database Center for Life Science, Research Organization of Information and Systems, Kashiwa 277-0871, Japan

Previous approaches to create a controlled vocabulary for Japanese have resorted to existing bilingual dictionary and transformation rules to allow such mappings. However, given the possible new terms introduced due to coronavirus disease 2019 (COVID-19) and the emphasis on respiratory and infection-related terms, coverage might not be guaranteed. We propose creating a Japanese bilingual controlled vocabulary based on MeSH terms assigned to COVID-19 related publications in this work. For such, we resorted to manual curation of several bilingual dictionaries and a computational approach based on machine translation of sentences containing such terms and the ranking of possible translations for the individual terms by mutual information. Our results show that we achieved nearly 99% occurrence coverage in LitCovid, while our computational approach presented average accuracy of 63.33% for all terms, and 84.51% for drugs and chemicals.

Keywords: controlled vocabulary, COVID-19, Japanese, multilingualism, natural language processing, translation
Availability: The code and results are available at https://github.com/soares-f/O-JMeSH.

Introduction

References in MEDLINE are indexed according to MeSH terms. MeSH is a controlled vocabulary meta-thesaurus composed of more than 27,000 hierarchically structured descriptors. At higher levels of structure, one can find broad titles (e.g., Diseases - C), while narrow levels contain more specific titles (e.g., Diseases caused by Viruses - C02, Hepatitis A -C02.440.420).

In addition to its use in PubMed, the MeSH vocabulary has been used in a variety of ways in many areas of scientific research, including information retrieval, text mining, citation analysis, education, and bioinformatics research. When applied to information retrieval, MeSH terminology and its indexing results have been used to build visualization tools [1] and to distinguish between homonymous authors [2,3]. Biomedical text mining also makes extensive use of indexing on MeSH terms and has been used in various tasks, such as summarization, document clustering, and syntactic disambiguation [4].

Despite its large usage in natural language processing (NLP) tasks in English, MeSH is translated to a few other languages, such as Spanish, Portuguese, and French. However, some underrepresented languages on biomedical NLP have incomplete or outdated versions of MeSH, which is the case of Japanese. Thus, the ability to generate new open versions of MeSH in other languages, as well as improving the already existing ones, can help foster research in those languages.

In this application note, we propose the creation of a Japanese MeSH by combining different glossaries and exploring automatic translation. As a proof-of-concept, we used the LitCovid dataset [5], focused on coronavirus disease 2019 (COVID-19) research. We used automatic translation of sentences containing terms of interest and term selection via pointwise mutual information.

Methods

Previous approaches to create dictionaries or extract parallel phrases for Japanese have resorted to existing bilingual dictionary and transformation rules to allow such mappings [6], or word alignment [7]. However, given the possible new terms introduced due to COVID-19 and the emphasis on respiratory and infection-related terms, we want to take advantage of the already mapped terms in English and the use of automatic translation.

A straightforward approach would be to directly translate the terms from English to Japanese using commercially available translators, such as Google Translate or Bing. However, given past experiments, this course of action can result in ill-translated terms due to polysemy and lack of context for single tokens [8]. One of the reasons for such behavior is that modern machine translation systems take huge advantage of context in a sentence to make the translation of a single token.

Thus, in our computational approach, we make use of full sentences in English that contain a specific desired MeSH term. After translation to Japanese, the equivalent in the target language will be found. This task can be described as the construction of bilingual dictionaries from parallel data [9].

However, since translating just one sentence for each term might lead to noisy results (that is, one may select a sentence that uses a non-standard or ambiguous translation for a given term), we will look for a set of sentences containing the same English MeSH term and then translate them to Japanese. By collecting multiple sentences, we expect that possible non-standard translations will be given less importance. We will extract the bilingual matching using pointwise positive mutual information (PPMI) [10]. Table 1 shows an example of how the proposed computational approach works considering the term "pulmonary embolism."

The steps for the implementation of the computational method for constructing the bilingual dictionary are as follows:

1. Retrieve the most frequent MeSH terms from LitCovid.
2. For each of the terms from 1, retrieve k sentences in English that contain the given term.
3. Use an MT system to get the translation from English to Japanese of the complete sentence.
4. From the Japanese translation, segment the Japanese tokens using MeCab.
5. From 4, compute all possible $\{1{:}n\}$-grams (e.g. if n = 3, all 1-grams, 2-grams, and 3-grams).
6. Calculate the MeSH-by-ngrams occurrence counts (i.e. the counts for every n-gram for the terms selected in 1).
7. Using PPMI, as in [10], find the most likely Japanese n-gram for a given MeSH.

Results

We selected the MeSH terms in LitCovid appearing at least 50 times, resulting in a total of 1,039 terms. From the selected terms, we found that around 79% could already be found on existing vocabularies.

When using the proposed computational approach of machine translation and pointwise mutual information, we found that it had a precision of 63.33%. Meanwhile, the precision on translating using Google Translate on the isolated terms (without being in a sentence) was of 57.42%. Considering only a subset of MeSH terms

Table 1. Example of the computational approach for the term "pulmonary embolism"

English	MT Japanese
Pulmonary embolism has a high prevalence in COVID-19 patients	肺塞栓症はCOVID-19患者で高い有病率を示す
Pulmonary embolism is shown to increase the risk of death	肺塞栓症は死亡リスクを高めることが示されている
The patient presented bilateral pulmonary embolism	症例は両側性肺塞栓症を呈した。

In the English column, we show three sentences where context is given regarding the term. On the right column, we show the machine translated version for Japanese, with the term identified in bold. In this case, the term "肺塞栓症" is identified as the correct translation without directly inferring that is "pulmonary/lung", 塞栓 as "embolus", and 症 as "illness". However, in this case, due to specialized dictionaries and low polysemy, this term could be directly inferred without requiring context.
COVID-19, coronavirus disease 2019.

https://doi.org/10.5808/gi.21014

representing drug names (validated by KEGG Drugs), the precision increases to 84.51%, which was similar to Google Translate performance of 84.03% on individual terms. We hypothesize that the higher precision is caused by the fact that drug names tend to have less variability (low number of synonyms), thus are easily distinguishable. The overlap of the computational approach with the manual curation of other bilingual sources was 68.43%.

On manual error checking, we found that the computational approach often failed to include the broad term into a specific term. For instance, for most of the cancers, the actual "neoplasm" equivalent in Japanese, "腫瘍", was missing, leaving only the specific organ. The same issue happened for infections, where the actual word "infection" or "viral" was missing on the Japanese part. When considering the nature of MI, which gives less importance to tokens that appear frequently in different groups (terms), this failure is not completely unexpected. As a form of alleviating this, one could create a set of handcrafted rules to pre-filter the candidate terms before using MI to select the most probable one.

Our final bilingual glossary, O-JMeSH, covers approximately 99% of the occurrences in LitCovid, with a coverage of nearly 69% of all MeSH UIDs in the database. Thus, we can see that by combining both manual curation and computational efforts, we can lessen the effort required to map the most frequent terms occurring in COVID-19 related literature to the Japanese language.

ORCID

Felipe Soares: https://orcid.org/0000-0002-2837-1853
Yuka Tateisi: https://orcid.org/0000-0002-3813-5782
Terue Takatsuki: https://orcid.org/0000-0003-0011-764X
Atsuko Yamaguchi: https://orcid.org/0000-0001-7538-5337

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Authors' Contribution

Conceptualization: FS. Data curation: YT , TT, AY. Formal analysis: YT , TT, AY, FS. Funding acquisition: YT , TT, AY. Methodology: FS. Writing - original draft: FS, YT , TT, AY. Writing - review & editing: FS.

Acknowledgments

Felipe Soares would like to acknowledge Google's TensorFlow Research Cloud (TFRC) program as well as AWS Diagnostic Development Initiative (DDI) initiative for providing computational resources. We would also like to acknowledge DeepL for providing access to their API to perform automatic translation.

References

1. Sarkar IN, Schenk R, Miller H, Norton CN. LigerCat: using "MeSH Clouds" from journal, article, or gene citations to facilitate the identification of relevant biomedical literature. AMIA Annu Symp Proc 2009;2009:563-567.
2. Liu W, Islamaj Dogan R, Kim S, Comeau DC, Kim W, Yeganova L, et al. Author name disambiguation for PubMed. J Assoc Inf Sci Technol 2014;65:765-781.
3. Sanyal DK, Bhowmick PK, Das PP. A review of author name disambiguation techniques for the PubMed bibliographic database. J Inf Sci 2019;47:227-254.
4. Jimeno-Yepes AJ, McInnes BT, Aronson AR. Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. BMC Bioinformatics 2011;12:223.
5. Chen Q, Allot A, Lu Z. Keep up with the latest coronavirus research. Nature 2020;579:193.
6. Yamada R, Tatieisi Y. open-japanese-mesh: assigning MeSH UIDs to Japanese medical terms via open Japanese-English glossaries. Genomics Inform 2020;18:e22.
7. Ogawa Y, Nakamura M, Ohno T, Toyama K. Extraction of legal bilingual phrases from the Japanese Official Gazette, English edition. J Inf Telecommun 2018;2:359-373.
8. Soares F, Villegas M, Gonzalez-Agirre A, Krallinger M, Armengol-Estape J. Medical word embeddings for Spanish: development and evaluation. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019 Jun 7, Minneapolis, MN, USA. Stroudsburg: Association for Computational Linguistics, 2019. pp. 124-133.
9. McEwan CJ, Ounis I, Ruthven I. Building bilingual dictionaries from parallel web documents. In: Proceedings of the 24th European Colloquium on Information Retrieval Research, 2002 Mar 25-27, Glasgow, Scotland. Berlin: Springer, 2002. pp. 303-323.
10. Aji S, Kaimal R. Document summarization using positive pointwise mutual information. Int J Comput Sci Inf Technol 2012;4:47-55.

G&I
Genomics & Informatics

BLAH7
Application note

eISSN 2234-0742
Genomics Inform 2021;19(3):e27
https://doi.org/10.5808/gi.21015

Received: March 17, 2021
Revised: July 27, 2021
Accepted: July 27, 2021

*Corresponding author:
E-mail: pierre.larmande@ird.fr

OryzaGP 2021 update: a rice gene and protein dataset for named-entity recognition

Pierre Larmande[1,2*], Yusha Liu[3], Xinzhi Yao[3], Jingbo Xia[3]

[1]DIADE, Univ. Montpellier, IRD, CIRAD, 34394 Montpellier, France
[2]French Institute of Bioinformatics (IFB)—South Green Bioinformatics Platform, Bioversity, CIRAD, INRAE, IRD, Montpellier F-34398, France
[3]Hubei Provincial Key Laboratory of Agricultural Bioinformatics, College of informatics, Huazhong Agricultural University, Wuhan 430070, Hubei Province, China

Due to the rapid evolution of high-throughput technologies, a tremendous amount of data is being produced in the biological domain, which poses a challenging task for information extraction and natural language understanding. Biological named entity recognition (NER) and named entity normalisation (NEN) are two common tasks aiming at identifying and linking biologically important entities such as genes or gene products mentioned in the literature to biological databases. In this paper, we present an updated version of OryzaGP, a gene and protein dataset for rice species created to help natural language processing (NLP) tools in processing NER and NEN tasks. To create the dataset, we selected more than 15,000 abstracts associated with articles previously curated for rice genes. We developed four dictionaries of gene and protein names associated with database identifiers. We used these dictionaries to annotate the dataset. We also annotated the dataset using pretrained NLP models. Finally, we analysed the annotation results and discussed how to improve OryzaGP.

Keywords: biological dataset, gene mention, named entity recognition, natural language processing, *Oryza* species
Availability: OryzaGP is available at https://github.com/pierrelarmande/OryzaGP,while annotations can be visualised and downloaded at http://pubannotation.org/projects/OryzaGP_2021.

Introduction

The past few decades have seen a deluge of information in agronomy. However, a substantial proportion of this information is available in unstructured scientific documents, such as journal articles, reviews, abstracts, and reports. Despite advances in data sciences, innovations in agronomy are still often text-based. One of the challenges is to extract the biological entities and their relationships contained in text fields and scientific papers. Many of these text fields contain molecular mechanisms and phenotypes of interest that are often described by complex expressions associating biological entities linked by specialised semantic relationships (e.g., *"Ehd1 and Hd3a can also be down-regulated by the photoperiodic flowering genes Ghd7 and Hd1"* source PMID: 20566706). To address this issue, the objective is to develop computational tools to extract biological entities and their relationships in order to extract relevant information—here, the entities *Ehd1, Hd3a, Ghd7,* and *Hd1* and the down-regulated relationship. The biomedical field has long experience in developing NLP approaches. The Biocreative [1] and BioNLP conferences [2] have demonstrat-

ed numerous advances in this area achieved through the development of datasets and tools. However, little research has been done on this issue in plant science and, more precisely, in the rice sector. For these reasons, we developed a dedicated dataset for rice named OryzaGP. The first release of OryzaGP was initially published in 2019 during BLAH5. The first version originally gathered relatively few PubMed abstracts and focused on named entity recognition (NER) by providing only entities tagged with gene or protein labels. In this new version, we updated the number of PubMed abstracts and provided both NER and the results of named entity normalisation (NEN) when available. Moreover, we tried to merge several database identifiers coming from different resources under the same name. The next section will describe the procedure of building the OryzaGP dataset and how it was annotated.

Methods

Similarly to the first version, we started by downloading the Oryzabase reference datasets from the Oryzabase [3] web application. Oryzabase provides a manually curated dataset for new rice-related PubMed entries. We filtered out a list of PubMed identifiers that we used to create the OryzaGP_2021 project on PubAnnotation [4]. PubAnnotation [5] is a repository of text annotations related to literature in the life sciences, such as PubMed or PMC articles. It also provides features to create, manage, and access annotations through APIs.

Annotations were conducted through two applications: PubDictionary and HunFlair [6]. PubDictionary is a repository of public dictionaries for the life sciences. It was developed as a model annotation service for PubAnnotation and provides the RESTFul API for dictionary-based text annotation. HunFlair is a NER tagger covering five biomedical entity types. It is integrated into the Flair NLP framework, and it uses a character-level language model pre-trained on roughly 24 million biomedical abstracts and 3 million full texts.

In order to use PubDictionary to annotate OryzaGP, we created several dictionaries of gene/protein entities. We first downloaded the Oryzabase gene dataset, which contains several gene mentions associated with database identifiers. We created the Oryzabase dictionary containing labels, gene names, symbols, synonyms and Oryzabase identifier URIs. Next, we repeated the same process to create the RAPDB [7], MSU [8], and UniProt [9] dictionaries. Additionally, we refined the RAPDB and UniProt dictionaries by adding new entries extracted from the RAPDB gene datasets. All these dictionaries were uploaded to PubDictionary and used to create several annotators. Table 1 shows the size (i.e., the number of entries) of these dictionaries. Finally, we utilized PubAnnotation to run several annotations on OryzaGP using these dictionaries. We merged these annotations in a single project (Fig. 1).

HunFair, which comes with models for genes, proteins, chemicals, diseases, species and cell lines, is an advanced NER tagger for biomedical texts. Compared with other biomedical NER tools, such as GNormPlus [10] and HUNER [11], HunFlair showed better performance on the BioNLP 2013 CG [12] and Plant-Disease corpus [13]. In the OryzaGP project, we imported the HunFlair pre-trained model directly to annotate the abstracts in OryzaGP. HunFlair annotated each abstract with genes, proteins, chemicals, diseases, and species, and converted the JSON results into a format that met the requirements of the PubAnnotation platform. All annotations created by HunFlair were prefixed with *hunflair:NA* plus the entity type (e.g., gene, disease, cell line, chemical, and species).

Results

Compared to the first version of OryzaGP, this updated version was significantly improved. Table 2 compares basic statistics on both versions. The number of articles was increased from 10,000 to 15,000, and consequently the number of sentences and words increased as well. The number of annotations also increased. In the first version, the annotations were produced with an improved Bi-LSTM-CRF model from [14,15] previous research [12,13]. Around 29,000 annotations were found. In this current version, we used multiple annotators to achieve this goal (see the Materials and Methods section)

Table 1. Description of the dictionaries

Name	Size
OryzaGeneName_Oryzabase	175,158
OryzaGeneName_RAPDB	110,539
OryzaGeneName_MSU	112,309
OryzaGeneName_UniProt	66,934

Table 2. Description of the dataset

Name	OryzaGP	OryzaGP 2021
Text genre	Article	Article
Text type	Abstract & Title	Abstract & Title
Entity type	Gene, Protein	Gene, Protein
No. of articles	10,400	15,041
No. of sentences	75,096	150,604
No. of words	2,697,726	4,101,648
No. of annotations	29,098	1,064,353
No. of gene mentions	None	677,938

The number of annotations corresponds to the total annotations detailed in Table 3. The number of gene mentions was calculated from the fourth PubAnnotation (oryzabase.gene, rapdb.gene, uniprot, msu.gene) results because the corresponding dictionaries contained the URIs of the entities.

https://doi.org/10.5808/gi.21015

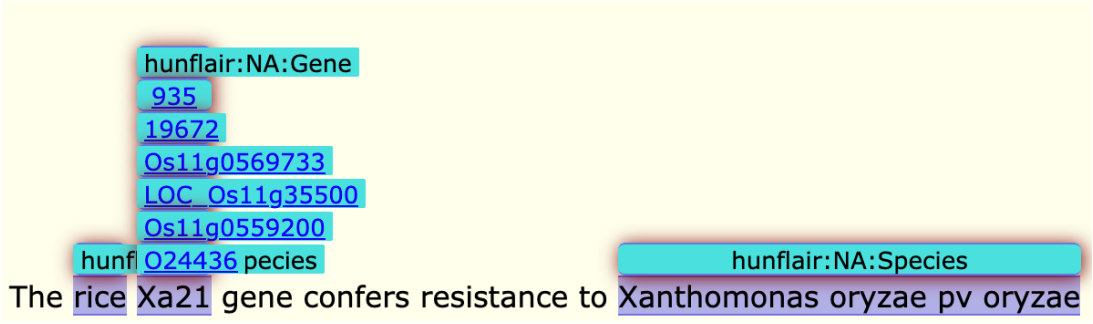

Fig. 1. Example of merged annotations with TexTAE tool.

Table 3. Description of the annotations

Annotation type	No. of annotations
hunflair:NA:CellLine	5,195
hunflair:NA:Chemical	86,770
hunflair:NA:Disease	12,369
hunflair:NA:Gene	171,761
hunflair:NA:Species	110,320
PubAnnotation.oryzabase.gene	189,081
PubAnnotation.rapdb.gene	175,285
PubAnnotation.uniprot	140,354
PubAnnotation.msu.gene	173,218
Total	1,064,353

and obtained about 1 million annotations (Table 3). The annotations were merged into the single project. Fig. 1 shows an example of merged annotations with the TextAE editor from PubAnnotation. We can see tagged entities with a class label and other entities tagged with database identifiers (i.e., NEN). We obtained NEN results in 64% of cases, which means that nearly two-thirds of the annotations are linked with a database identifier.

To our knowledge, OryzaGP is the first dataset created for genes and proteins in rice species. It can help to better train NLP tools to recognize rice-related biological entities. Moreover, this new version contains a large number of normalized genes and proteins. However, manual checking of these annotations revealed some false positives. For this iteration of the project, it was not possible to develop a strategy to automatically evaluate the rate of false-positive and false-negative annotations. This remains a task for future work.

Future work

Our future work will first focus on identifying false positives and negatives to improve annotations. We manually observed that false positives often occurred with gene and protein full names. Some annotations did not match the whole sequence of words. Our hypothesis is that there often exist co-occurrences of full names and symbols in the same sentence or abstract. Thus, we will analyze and classify these co-occurrences.

Next, we plan to normalise the annotations done by HunFlair. Some are already merged with NEN, but some are not. We plan to analyse these annotations, especially those standing for gene symbols, and set up a strategy to normalise them.

Finally, we are interested in adding new annotation types such as plant organs or plant traits. Thus, we will create dictionaries and train NLP tools to achieve this goal.

ORCID

Pierre Larmande: https://orcid.org/0000-0002-2923-9790
Yusha Liu: https://orcid.org/0000-0002-1596-8802
Xinzhi Yao: https://orcid.org/0000-0001-6795-2653
Jingbo Xia: https://orcid.org/0000-0002-7285-588X

Authors' Contribution

Conceptualization: PL. Formal analysis: PL, YL. Methodology: PL, YL, XY. Writing – original draft: PL, YL. Writing – review & editing: PL, JX.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work was supported by IRD, UMR DIADE, and CGIAR CRP RICE.

References

1. Morgan AA, Lu Z, Wang X, Cohen AM, Fluck J, Ruch P, et al. Overview of BioCreative II gene normalization. Genome Biol 2008;9 Suppl 2:S3.

2. Kim JD, Nguyen N, Wang Y, Tsujii J, Takagi T, Yonezawa A. The genia event and protein coreference tasks of the BioNLP Shared Task 2011. BMC Bioinformatics 2012;13 Suppl 11:S1.
3. Kurata N, Yamazaki Y. Oryzabase: an integrated biological and genome information database for rice. Plant Physiol 2006; 140:12-17.
4. PubAnnotation. Kashiwa: DBCLS (Database Center for Life Science), 2021. Accessed 2021 Mar 9. Available from: http://pubannotation.org/projects/OryzaGP_2021.
5. Kim JD, Wang Y. PubAnnotation: a persistent and shareable corpus and annotation repository. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, 2012 Jun 8, Montreal, Canada. Stroudsburg: Association for Computational Linguistics, 2012. pp. 202-205.
6. Weber L, Sanger M, Munchmeyer J, Habibi M, Leser MU, Akbik A. HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. Preprint at https://arxiv.org/abs/2008.07347 (2020).
7. Sakai H, Lee SS, Tanaka T, Numa H, Kim J, Kawahara Y, et al. Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. Plant Cell Physiol 2013; 54:e6.
8. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, et al. The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res 2007;35:D883-D887.
9. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47:D506-D515.
10. Wei CH, Kao HY, Lu Z. GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. Biomed Res Int 2015;2015:918710.
11. Weber L, Munchmeyer J, Rocktaschel T, Habibi M, Leser U. HUNER: improving biomedical NER with pretraining. Bioinformatics 2020;36:295-302.
12. Pyysalo S, Ohta T, Rak R, Rowley A, Chun HW, Jung SJ, et al. Overview of the cancer genetics and pathway curation tasks of BioNLP Shared Task 2013. BMC Bioinformatics 2015;16 Suppl 10:S2.
13. Kim B, Choi W, Lee H. A corpus of plant-disease relations in the biomedical domain. PLoS One 2019;14:e0221582.
14. Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics 2017;33:i37-i48.
15. Do H, Than K, Larmande P. Evaluating named-entity recognition approaches in plant molecular biology. Preprint at https://doi.org/10.1101/360966 (2018).

https://doi.org/10.5808/gi.21015

G&I
Genomics & Informatics

BLAH7
Opinion

eISSN 2234-0742
Genomics Inform 2021;19(3):e28
https://doi.org/10.5808/gi.21016

Received: March 17, 2021
Revised: August 29, 2021
Accepted: August 29, 2021

*Corresponding author:
E-mail: edohi@bri.niigata-u.ac.jp

Visualizing the phenotype diversity: a case study of Alexander disease

Eisuke Dohi[1*], Ali Haider Bangash[2,3]

[1]Department of Neuroscience of Disease, Brain Research Institute, Niigata University, Niigata 951-8122, Japan
[2]Shifa College of Medicine, Shifa Tameer-e-Millat University, Islamabad 46000, Pakistan
[3]COST Action EVidence-Based RESearch (EVBRES), Western Norway University of Applied Sciences, Bergen 5063, Norway

Since only a small number of patients have a rare disease, it is difficult to identify all of the features of these diseases. This is especially true for patients uncommonly presenting with rare diseases. It can also be difficult for the patient, their families, and even clinicians to know which one of a number of disease phenotypes the patient is exhibiting. To address this issue, during Biomedical Linked Annotation Hackathon 7 (BLAH7), we tried to extract Alexander disease patient data in Portable Document Format. We then visualized the phenotypic diversity of those Alexander disease patients with uncommon presentations. This led to us identifying several issues that we need to overcome in our future work.

Keywords: annotation, biomedical text mining, case reports, phenotype, rare diseases, symptoms

Introduction

Since only a relatively small number of patients have a rare disease, not all clinicians experience such patients, and not many experts experience a significant number of these patients. Such expert clinicians can experience so few of these patients that they cannot reach a comprehensive understanding of each rare disease. We clinicians, then, continue to further our knowledge through learning from the textbooks and previous publications. In doing so, we tend to learn only about the typical phenotype or the most well-known subtypes of a rare disease. In the real-life clinical setting, however, not every patient presents with typical symptoms, and their phenotypes are often diverse. Seeing patients with uncommon presentations of rare diseases is one of the reasons why we clinicians misdiagnose. This is a challenging issue, one in which we must make sure we do not overlook such patients, especially those whose delay of diagnosis could result in negative prognoses and even their lives being threatened.

On-site Issues for Clinicians: Interpreting the Symptoms of Patients with Rare Diseases

Recent advances in diagnostic support tools allow us to develop better differential diagnosis lists based on a patient's symptoms [1]. For example, PubCaseFinder includes information on different phenotypes that are associated with a disease. These have been extracted from the titles and abstracts of entire case reports found in PubMed [2]. In addition to these support tools, the cost of whole-genome sequencing is getting progressively cheaper. With this in mind, in the near future, we may encounter more patients than at present who

show uncommon presentations with genetic mutations. Even now, in the case reports, we can find uncommon presentations in both common and uncommon diseases. Also, in the clinical practice, we clinicians sometimes find difficulty in interpreting the genetic results to determine whether these symptoms are due to genetic mutations or not. In rare diseases, in particular, clinicians often do not know enough about the wide diversity of the disease phenotypes.

Attempt to Visualize the Phenotype Diversity in Alexander Disease

To address this issue, we tried to visualize the phenotype diversity in Alexander disease. The reasons why we chose Alexander disease are as follows: 1, The majority of cases of Alexander disease are caused by a genetic mutation of the *GFAP* gene [3]; 2, There are a manageable number of reported Alexander disease cases (less than 1,000); 3, There is a diversity in the age of onset, severity, and combination of symptoms; 4, There is an established feature-based disease classification for Alexander disease; and 5, Genotype-phenotype correlations have been examined [4]. Based on the established previous knowledge, we planned to validate, compare, and analyze the results of our method.

The outline of our plan to visualize the phenotype diversity in Alexander disease is shown in the accompanying figure. Medical Subject Headings 2021 (MeSH) Browser was used to search for the primary descriptors identifying Alexander disease in PubMed-indexed articles [5]. The MeSH term "Alexander Disease" was retrieved by an advanced search of PubMed [6]. We filtered the search results with "case reports" and obtained 139 case reports (as of 19 January 2021). Of the 139 case reports, 116 PDF files were available. We downloaded these and converted the entire text of all PDF files to text data with Apowersoft PDF Converter [7]. After ragged alignments were manually corrected, we annotated the extracted text data with Pubannotation [8]. For the annotation, the Human Phenotype Ontology (HPO) dictionaries on Pubdictionaries [9] were applied. Due to several problems, we could not proceed further in the process using automated means. So, we changed our direction toward listing the problems that we need to overcome.

Discussion and Future Direction

Regarding the collecting of case reports of Alexander disease patients, we could not access the total number of reported cases (more than 500) [4]. One of the reasons for this was that we could not extract patient data from a particular type of publication, namely case series. Case series usually provide patient data in table format, and issues occurred when we extracted patient data from tables in PDF form. In step 1 of the figure, data extraction from PDF has several issues, but novel methods and algorithms [10] are being developed. We need to optimize our process in order to make automated data extraction possible. In step 2 (annotation), it was hoped that by using Pubdictionaries along with the HPO we would be able to annotate the technical terms related to the disease. However, a number of technical words, especially those describing neurological and psychiatric symptoms, were not annotated. It will be necessary to improve the dictionary with the help of expert clinicians from each area of expertise. Interestingly, we found that clinicians sometimes used non-technical terms to describe patients, even in scientific papers, and these were not annotated. For example, the term "occasional fall" is used to describe "unsteadiness," and such words sometimes do not show up in the medical dictionaries. Also, the majority of symptoms should be interpreted in a context-dependent manner. For example, in "He became verbally abusive, introverted and aggressive, and had little insight into his condition," the underlined part of the sentence describes the disease progression and severity. Symptoms, unlike other objective clinical parameters such as laboratory or radiographic tests, tend to be described in subjective and descriptive manners. We need, then, to apply natural language processing and develop more powerful dictionaries with the help of experienced clinicians. In step 3, we could not find an adequate tool for extracting "individual patient data." Since paragraphs and locations of patient data are not structured between each paper and each journal, an algorithm and novel dictionaries would be required to extract and restructure "individual patient data." We are currently developing such tools and hope to be soon able to extract and restructure this data. We had not reached this point in step 4. We noticed, however, that if we try to collect all of the patient features, then the number of features might be relatively large compared with the number of patients, especially for rare diseases. Although symptoms may be the same, the number of features would be multiplied if we try to include the severity of each symptom. The same thing would happen if we decided to include the time course of newly emerged symptoms and the progression of each symptom along with the disease progression. Through steps 1 to 4, we will successfully establish the "symptom-DataFrame," which will be indexed with individual Alexander disease patients (rows) and their symptoms (columns) (Fig. 1). The columns could be expanded with other data such as lab data, image data, and any individual clinical information. In step 5, we will attempt to classify the individual patient with the disease. We can get the patient's DataFrame using the established patient's classification. Alternatively, we can apply a machine learning approaches to classify the relevant patient features. We will explore this further once we develop the "symptom-DataFrame."

https://doi.org/10.5808/gi.21016

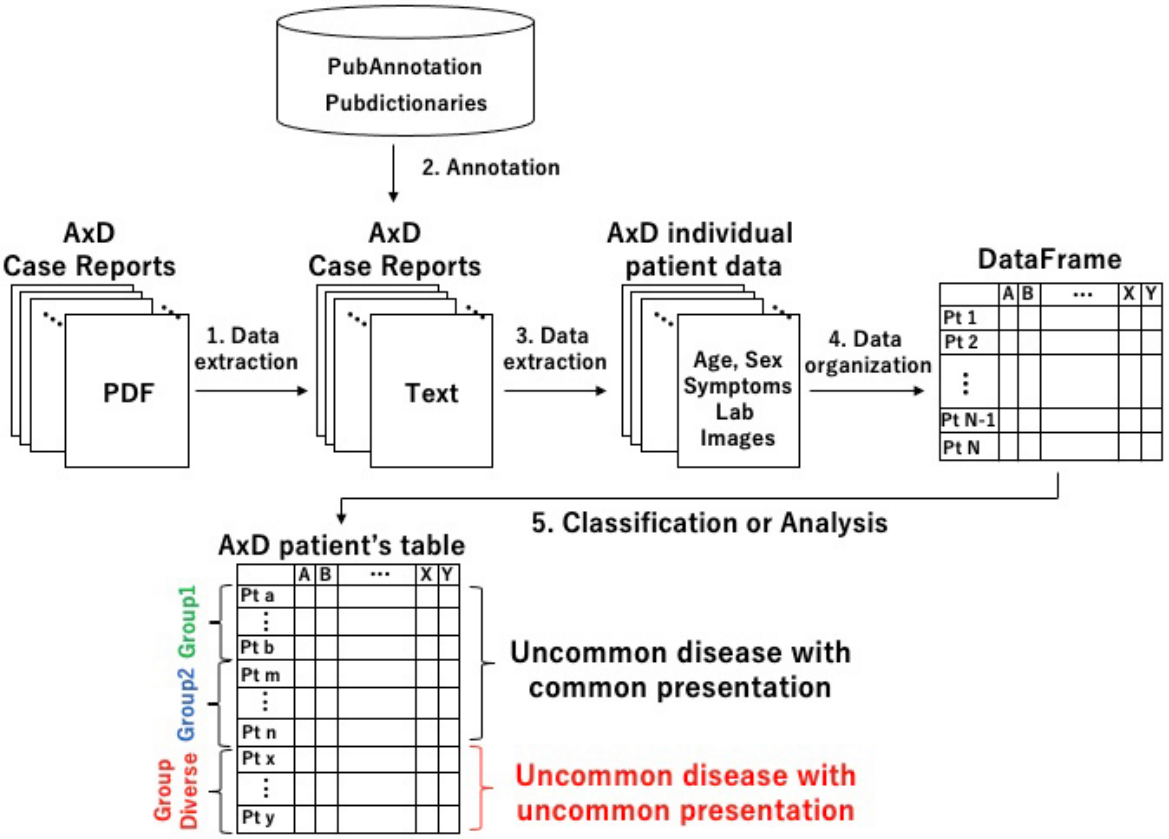

Fig. 1. Workflow of individual patient data extraction from case report PDF for the visualization of uncommon disease with uncommon presentations. Step 1: Data extraction from PDF. Step 2: Annotation of data extraction with PubAnnotation and Pubdictionaries. Step 3: Data extraction from annotated text data as individual patient data. Step 4: Generation of the DataFrame with individual patient and clinical data (such as symptoms, lab images, and genetic mutation). Step 5: Classification of the patients based on their combination of phenotypes. AxD, Alexander disease.

As mentioned above, there are problems that need to be overcome. However, the solving of several problems is already ongoing using multiple approaches, including the testing of a suitable PDF data extraction method [10], the development of new dictionaries, the application of validated natural language processing [11], testing the effect of limiting the relevant patient features, and application of machine learning to explore the relevant patient features. In addition to these, while the structured recording and storing of patient data for all diseases will need to be carried out consistently in the future, we need to seriously think about a more flexible way of doing this, as the sheer volume of data from individual patients is increasing massively in this era of big data. With this in mind, we would like to emphasize that seamless collaboration between informaticians and clinicians will become increasingly important to develop such "useful tools" and get clinicians to use them. We thoroughly enjoyed this attempt during Biomedical Linked Annotation Hackathon 7 (BLAH7), and we hope many more clinicians will be interested in this field and join with informaticians to develop and apply "useful tools." Once we develop this automated kind of system, we will be able to analyze patient data from rare diseases that have not been well investigated. It will also, of course, be helpful to understand the individual patients not only for clinicians but for the patients themselves and their families.

ORCID

Eisuke Dohi: https://orcid.org/0000-0002-5365-4900
Ali Haider Bangash: https://orcid.org/0000-0002-8256-3194

Authors' Contribution

Conceptualization: ED. Data curation: ED. Formal analysis: ED. Methodology: ED, AHB. Writing - original draft: ED, AHB. Writing - review & editing: ED.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

References

1. Faviez C, Chen X, Garcelon N, Neuraz A, Knebelmann B, Salomon R, et al. Diagnosis support systems for rare diseases: a scoping review. Orphanet J Rare Dis 2020;15:94.
2. Fujiwara T, Yamamoto Y, Kim JD, Buske O, Takagi T. PubCaseFinder: a case-report-based, phenotype-driven differential-diangnosis system for rare diseases. Am J Hum Genet 2018;103:389-399.
3. Barkovich AJ, Messing A. Alexander disease: not just a leukodystrophy anymore. Neurology 2006;66:468-469.
4. Srivastava S, Waldman A, Naidu S. Alexander disease. GeneReviews. Seattle: University of Washington, 1993-2021. Accessed 2021 Jan 20. Available from: http://www.ncbi.nlm.nih.gov/books/NBK1172/.
5. MeSH Browser. Bethesda: National Library of Medicine, 2021. Accessed 2021 Mar 11. Available from: https:// meshb.nlm.nih.gov/search.
6. Advanced Search Results - PubMed. Bethesda: National Library of Medicine, 2021. Accessed 2021 Mar 11. Available from: https://pubmed.ncbi.nlm.nih.gov/advanced.
7. Apowersoft PDF converter. Hong Kong: Wangxu Ltd., 2021.
8. PubAnnotation. Bethesda: Database Center for Life Science, 2021. Accessed 2021 Mar 11. Available from: https://pubannotation.org/.
9. Pubdictionaries. Kashiwa: Database Center for Life Science, 2021. Accessed 2021 Mar 11. Available from: http:// pubdictionaries.org/.
10. Bast H, Korzen C. A benchmark and evaluation for text extraction from PDF. In: ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2017 Jun 19-23, Toronto, ON, Canada.
11. Funk C, Baumgartner W Jr, Garcia B, Roeder C, Bada M, Cohen KB, et al. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. BMC Bioinformatics 2014;15:59.

https://doi.org/10.5808/gi.21016

G&I
Genomics & Informatics

Original article

eISSN 2234-0742
Genomics Inform 2021;19(3):e29
https://doi.org/10.5808/gi.21034

Received: June 14, 2021
Revised: July 16, 2021
Accepted: July 22, 2021

*Corresponding author:
E-mail: reddyg.bp@icmr.gov.in

**Corresponding author:
E-mail: drdirisala@gmail.com

Molecular insights into the role of genetic determinants of congenital hypothyroidism

Yedukondalu Kollati[1], Radha Rama Devi Akella[2,3], Shaik Mohammad Naushad[3], Rajesh K. Patel[4], G. Bhanuprakash Reddy[5*], Vijaya R. Dirisala[1**]

[1]Department of Biotechnology, Vignan's University, Vadlamudi, Guntur, Andhra Pradesh 522213, India
[2]Department of Genetics, Rainbow Children's Hospital, Banjara Hills, Hyderabad, Telangana 500009, India
[3]Department of Biochemical Genetics and Pharmacogenomics, Sandor Speciality Diagnostics Pvt. Ltd, Banjara Hills, Hyderabad, Telangana 500034, India
[4]Department of Genetics, Genetic Group of Gujarat Diagnostic Centre, Mehsana, Gujarat 384002, India
[5]Biochemistry Division, National Institute of Nutrition, Hyderabad, Telangana 500007, India

In our previous studies, we have demonstrated the association of certain variants of the thyroid-stimulating hormone receptor (*TSHR*), thyroid peroxidase (*TPO*), and thyroglobulin (*TG*) genes with congenital hypothyroidism. Herein, we explored the mechanistic basis for this association using different *in silico* tools. The mRNA 3′-untranslated region (3′-UTR) plays key roles in gene expression at the post-transcriptional level. In *TSHR* variants (rs2268477, rs7144481, and rs17630128), the binding affinity of microRNAs (miRs) (hsa-miR-154-5p, hsa-miR-376a-2-5p, hsa-miR-3935, hsa-miR-4280, and hsa-miR-6858-3p) to the 3′-UTR is disrupted, affecting post-transcriptional gene regulation. TPO and TG are the two key proteins necessary for the biosynthesis of thyroid hormones in the presence of iodide and H_2O_2. Reduced stability of these proteins leads to aberrant biosynthesis of thyroid hormones. Compared to the wild-type TPO protein, the p.S398T variant was found to exhibit less stability and significant rearrangements of intra-atomic bonds affecting the stoichiometry and substrate binding (binding energies, ΔG of wild-type vs. mutant: −15 vs. −13.8 kcal/mol; and dissociation constant, K_d of wild-type vs. mutant: $7.2E^{-12}$ vs. $7.0E^{-11}$ M). The missense mutations p.G653D and p.R1999W on the TG protein showed altered ΔG (0.24 kcal/mol and 0.79 kcal/mol, respectively). In conclusion, an *in silico* analysis of *TSHR* genetic variants in the 3′-UTR showed that they alter the binding affinities of different miRs. The TPO protein structure and mutant protein complex (p.S398T) are less stable, with potentially deleterious effects. A structural and energy analysis showed that TG mutations (p.G653D and p.R1999W) reduce the stability of the TG protein and affect its structure-functional relationship.

Keywords: congenital hypothyroidism, miR, *TG*, *TPO*, *TSHR*, 3′-UTR

Introduction

Congenital hypothyroidism (CH) is one of the most common endocrine disorders, reported to occur in 1 in 3,000 to 4,000 newborns worldwide [1,2], and 1 in 1,100 in India [3]. On a worldwide basis, CH frequently results from iodine deficiency. Otherwise, CH is commonly caused by thyroid gland development defects, which can lead to thyroid dys-

genesis (80%–85%) [1]. The majority of these cases involve thyroid dysgenesis [4], agenesis (35%–40%), ectopic tissue (30%–45%), or hypoplasia (5%) [1]. Thyroid dysgenesis is caused by genes (thyroid transcription factor-1 [*TTF-1*], thyroid transcription factor-2 [*TTF-2*], and paired box 8 [*PAX-8*]) associated with syndromic CH and those causing non-syndromic CH (thyroid-stimulating hormone receptor [*TSHR*]) [5]. The remaining 15%–20% of cases are due to hereditary defects in the genes involved in the intermediary steps of biosynthesis in the thyroid, leading to dyshormonogenesis [6]. Thyroid dyshormonogenesis is associated with multiple genetic defects, including dual oxidases (*DUOXs*; *DUOX1* and *DUOX2*), and its maturation factors (*DUOXA1* and *DUOXA2*), thyroid peroxidase (*TPO*), thyroglobulin (*TG*), dehalogenase 1 (*DEHAL1*) and solute carrier families: 26 (*SLC26A4* or *PDS*) and 5 (*SLC5A5* or *NIS*) [1,4]. The synthesis of thyroid hormones (T_4 and T_3) is affected in 20% of all cases involving inborn genetic errors in the enzymatic cascade, which is defined as thyroid dyshormonogenesis [1]. Most often, these defects appear to be transmitted in an autosomal recessive manner [6], but autosomal dominant inheritance has also been reported [7].

Splicing is dependent on the exact identification of exons, which are perfectly recognized within pre-mRNAs. The presence of 5′ and 3′ splice sites and the branch points may not be sufficient to define intron-exon boundaries. The exonic elements are represented by exonic splicing enhancers (ESEs), where SR proteins bind and play a pivotal role in spliceosome assembly. Sequences that act as exonic splicing silencers (ESSs) bind to negative regulators, which belong to the heterogeneous nuclear ribonucleoprotein family. Both ESEs and ESSs appear to play an instrumental role in the regulation of alternative splicing events, other than the sequences that may play a pertinent role in the definition of constitutive exons [8].

Furthermore, microRNAs (miRs), which are 22-23 nucleotides in length, bind to the 3′-untranslated region (3′-UTR) and regulate mRNAs post-transcriptionally, either by facilitating mRNA degradation or by inhibiting mRNA transcription [9]. More than 30% of genes encoding for proteins are regulated by miRs [10]. Any genetic variation in the 3′-UTR may interfere with miR binding to its target, thereby influencing the expression of the targeted gene [11].

In our previous study, we investigated the effects of *TSHR*, *TPO*, and *TG* genetic variants in CH and identified 22 variants [12]. Three of these 22 variants (p.S398T in *TPO* and p.G653D and p.R1999W in *TG*) were predicted to be deleterious [12]. In this study, we aimed to perform an *in silico* analysis to achieve a better understanding of polymorphic variants in the 3′-UTR of the *TSHR* gene, which interferes with the binding affinities of miRs that might inhibit gene expression. Further, we analyzed mutant protein structure-function relationships through molecular modeling and interaction studies for the p.S398T mutation of TPO. In addition, molecular modeling, mutation analyses, and thermodynamic energy calculations for variants of p.G653D and p.R1999W in the *TG* gene were also performed.

Methods

As part of a newborn screening (NBS) program for CH, we analyzed 49,432 newborns, of whom 1,099 were screened with negative findings, along with 45 confirmed cases of CH. The institutional ethical committee of Rainbow Children's Hospital, located in Hyderabad, India, approved the study protocol (RCH-BH/066/02-2018). Informed consent was obtained from the parents or guardians of all neonates [12].

In silico analysis of SNP-miR interactions

Four web-based tools—PolymiRTS (http://compbio.uthsc.edu/miRSNP/) [13], miRDB (http://www.mirdb.org/cgi-bin/ search.cgi) [14], TargetScan (http://www.targetscan.org) [15], and STarMir (http://sfold.wadsworth.org) [16]—were used to ascertain whether the identified single-nucleotide polymorphisms (SNPs) in the 3′-UTR region interfered with miR binding.

Molecular modeling and protein-protein docking of the TPO protein

To further analyze the predicted role of SNPs in the protein interactions, molecular modeling and interaction studies were carried out. The Uniprot IDs of TPO and DUOX1 are P07202 and Q9NRD9, respectively. As the structures of TPO and DUOX1 are not crystallized, model prediction for the functional domains of the proteins was carried using the I-Tasser server [17]. The TPO domain (residues 167-734) of human TPO was modeled, as was the similarly interacting protein of the human dual peroxidase domain of DUOX. The 398 Ser-Thr mutant of TPO was modeled and energy-minimized using Chimera [18], and energy minimization was carried out using the steepest descent method. The modeled structures were docked using the ClusPro server [19] and the probable interactions were predicted using the PIC webserver [20]. The protein-protein interaction energies were calculated using the PRODIGY server [21].

Molecular modeling of the wild-type and mutant TG protein

To understand the effect of SNPs such as p.G653D and p.R1999W on the TG protein, we performed molecular modeling, mutation analyses, and thermodynamic energy calculations using the SAAM-BE-3D server [22]. The crystal structure of human TG was taken from the protein database (PDB ID: 6SCJ). The crystal structure

https://doi.org/10.5808/gi.21034

was solved using electron microscopy with a resolution of 3.60 Å [23]. Mutations such as p.G653D and p.R1999W were inserted *in silico* into the wild-type TG protein using the 'Mutagenesis' wizard of the 'PyMol' software [24]. These modeled structures of mutant TG proteins were further used to understand the effects of the mutations on protein structural stability and its intramolecular interactions. The effect of mutations such as p.G653D and p.R1999W on TG protein stability were checked using the SAAMBE-3D server [22], which predicts the effects of mutations on protein stability. In addition, we used the DynaMut server [25] to understand the effects of mutations on various types of interactions, such as van der Waals, weak polar van der Waals, polar proximal, amide-amide interactions, and so on.

Results and Discussion

In our previous studies, we established the reference intervals for thyroid-stimulating hormone (TSH) through NBS data, and the specific genotype-phenotype correlations were exhibited in confirmed CH cases with *TSHR*, *TPO*, and *TG* gene variants [12]. Previously, we published an *in silico* analysis on p.D727E in the *TSHR* gene, which might control the signal transduction (cAMP-mediated) pathway, consequently contributing to the pathophysiology of CH [26]. In this study, we specifically focused on an *in silico* analysis of three variants: p.S398T in TPO, and p.G653D and p.R1999W in TG.

In a study we reported earlier, we identified eight intronic variants (g. IVS 01+63 G > C, g.IVS 06-69 C > T, g.IVS 06+13 A > G, g.IVS 09+58 T > G in the *TSHR* gene, g.IVS 11+20 G > A, g.IVS 13+128 C > T, g.IVS 14-37 G > A, g.IVS 14-19 G > C in the *TPO* gene) [12]. The interpretation of the intronic variants is that g.IVS 01+63 G > C, g.IVS 06-69 C > T, g.IVS 09+58 T > C are involved in the creation of an intronic ESE site and g.IVS 06+13 A > G is an alteration of an intronic ESS site. The intronic variant g.IVS 11+20 G > A predicted a signal wherein an ESS site is broken and a new ESS site is created; this is interpreted as involving an alteration of an intronic ESS site and creation of an intronic ESE site. These results may not have an impact on splicing; no significant splicing motif alterations were detected for the remaining SNPs, which probably have no impact on the splicing mechanism (http://www.umd.be/HSF3/HSF.shtml).

In silico analysis of SNP-miR interaction

The four SNPs in the 3′-UTR of the *TSHR* gene (rs2268477, rs373305430, rs7144481, and rs17630128) were predicted to alter the binding of 10 common miRs as per the data obtained from four different databases (PolymiRTS, miRDB, TargetScan, and STarMir). As shown in Table 1, the presence of the rs2268477 polymorphic variant destroys the binding site for hsa-miR-154-5p. The rs373305430, rs7144481, and rs17630128 polymorphic variants alter the binding affinities of hsa-miR-1237-5p, hsa-miR-4488, hsa-miR-4697-5p, hsa-miR-6846-5p, hsa-miR-6848-5p, hsa-miR-376a-2-5p, hsa-miR-3935, hsa-miR-4280, and hsa-miR-6858-3p. The 3′-UTR SNPs-miR interaction hybrid diagrams are shown in Supplementary Fig. 1.

The rs2268477, rs373305430, rs7144481, and rs17630128 SNPs were localized in the 3′-UTR region of the *TSHR* gene and hence associated with binding of the 10 different miRs. These SNPs in the miR target site on the 3′-UTR may affect the binding efficacy of miR. SNPs may alter target gene expression to affect post-transcriptional processing and polyadenylation, and they may even contribute to CH. Among the risk variants, rs2268477 was found to destroy the binding site of hsa-miR-154-5p. In papillary thyroid carcinoma, hsa-miR-154 was reported to be downregulated [27], which substantiates the role of miR-154 in thyroid development. The rs7144481 variant was found to alter binding affinities of hsa-miR-376a-2-5p, hsa-miR-3935 and miR-4280. Long non-coding RNA

Table 1. *In silico* studies revealing TSHR SNP-miRNA interactions

Gene	SNP	miRNA	Seed match	Wild	ΔG (kcal/mol)	Mutant	ΔG (kcal/mol)
TSHR	rs2268477	hsa-miR-154-5p	7-mer	uauUAA**C**CUAa	−22.0	Disruption of binding site	
	rs373305430	hsa-miR-1237-5p	6-mer	auu**G**CCCCCa	−23.4	auu**U**CCCCCa	−17.4
		hsa-miR-4488	6-mer	auu**G**CCCCCa	−25.8	auu**U**CCCCCa	−15.9
		hsa-miR-4697-5p	6-mer	auu**G**CCCCCa	−29.8	auu**U**CCCCCa	−22.8
		hsa-miR-6846-5p	6-mer	auu**G**CCCCCa	−29.6	auu**U**CCCCCa	−17.3
		hsa-miR-6848-5p	6-mer	auu**G**CCCCCa	−26.3	auu**U**CCCCCaa	−22.0
	rs7144481	hsa-miR-376a-2-5p	8-mer	auAAUCUA**C**A	−17.5	auAAUCUA**U**A	−16.1
		hsa-miR-3935	7-mer	auaAUCUA**C**Ac	−18.7	auaAUCUA**U**Ac	−14.0
		hsa-miR-4280	7-mer	aauCUA**C**ACUa	−19.6	aauCUA**U**ACUa	−17.7
	rs17630128	hsa-miR-6858-3p	7-mer	cacG**U**UGGCUc	−20.2	cacG**C**UGGCUc	−22.6

In the table given above, capital letters of the miR site indicates the miRNA binding site of the 3'-UTR. Bold letters indicate SNPs identified in our study.
TSHR, thyroid stimulating hormone receptor; SNP, single-nucleotide polymorphism; miRNA, microRNA; ΔG, binding energy; mer, nucleotide pairing.

LINC00488, which is reported to be highly expressed in thyroid cancer cell lines, directly binds to miR-376a and downregulates its expression [28]. The rs17630128 variant was found to alter the binding affinity of hsa-miR-6858-3p.

The binding of TSH to TSHR induces angiogenesis by modulating vascular endothelial growth factor expression through cAMP-mammalian target of rapamycin signaling [29]. In endurance athletes, the frequency of rs7144481 C-allele (wild-type allele) was reported to be higher than in controls, contributing to a high metabolic rate and better aerobic performance [30]. Through the regulation of gene expression, the rs7144481 SNP may decrease angiogenesis and/or the metabolic rate. Campo and his group found that rare allele carriers of rs7144481 in *TSHR* were at an increased risk of well-differentiated thyroid cancer [31].

In silico predictions of function of the c. 1284 G>C (p.S398T) mutation in the TPO protein

The docked structures of the wild-type and mutant structures with DUOX were saved for further structure-based analysis. The predicted wild-type and mutant structures were aligned using chimera to analyze the structural variation caused by the mutation. Various interactions between the wild-type and the mutant protein are tabulated and provided in Supplementary Table 1. The binding energies and dissociation constants for the wild-type and mutant complexes were also analyzed to reveal the functional alterations induced by the SNPs (Fig. 1). The TPO structures show a fluctuation in the root-mean-square deviation (RMSD) of 0.152 Å and the interacting DUOX shows an RMSD of 1.091 Å. This shows that the SNP induces more structural variations in the binding protein DUOX than in TPO, which harbors the SNP. These structural variations induced in the binding complex alter the binding pattern which causes major variations in functional aspects of *TPO*. The mutant induces novel non-bonded hydrophobic interactions by the Ala263 residue, while there is a steep fall in the numbers of hydrophobic interactions by Trp200 and Leu267. A similar fall in the numbers of hydrogen-bonds involving main chain–main chain, main chain–side chain, and side chain–side chain patterns are observed. A steep fall in the number of ionic interactions by Arg198 with DUOX is observed in the mutant protein complex, whereas the aromatic-aromatic interactions and cation-pi interactions are retained. The complete list of the molecular interactions between the TPO wild-type and mutant proteins and the DUOX protein are given in Supplementary Table 1. These altered interactions further decrease the binding energies. The wild-type complex showed a ΔG of –15 kcal/mol and a dissociation constant (K_d) of $7.2E^{-12}$ M, and the mutant showed a ΔG of –13.8 kcal/mol and a K_d of $7.0E^{-11}$ M. These binding energies and dissociation constants further show that the mutant complex is less stable and dissociates more easily than the wild-type protein. The *in silico* analysis of p.S398T revealed that the mutant protein complex is less stable and it may be deleterious.

TPO and TG play a pivotal role in the biosynthesis of thyroid hormones by supplying hydrogen peroxide (H_2O_2) and by serving as an iodine acceptor [32]. TPO protein activity hinges on the

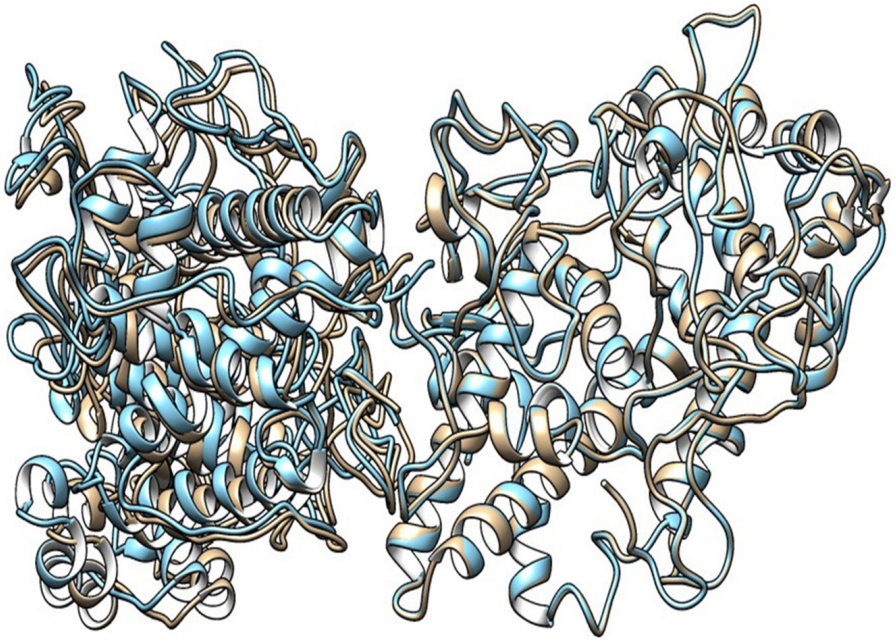

Fig. 1. Alignment showing the wild-type thyroid peroxidase (TPO) and mutant TPO with dual oxidase. The wild-type complex is shown in blue and the mutant complex is shown as mutant TPO. Deviation in the complex is shown by the non-alignment of the structures.

https://doi.org/10.5808/gi.21034

proper folding and insertion of the membrane, as well as a complete catalytic site with the heme-binding region, which is encoded by exons 8, 9, and 10 [33]. *TPO* gene-inactivating mutations lead to iodide organification defect (IOD), which is either partial (PIOD) or total (TIOD) depending on the mutation type and position. IOD is diagnosed by a positive perchlorate discharge test (PDT) [34]. Turkkahraman et al. [35] performed PDT after the fourth week of the LT_4 period. They found that one patient (variant-p.S398T) had PIOD (24.4%) with positive PDT (normal range, < 10%). They suggested that PIOD should be considered in infants with permanent subclinical hypothyroidism [35].

TPO c.1284G > C (p.S398T) is located in the eighth exon of *TPO*, which is part of the heme-binding catalytic site of *TPO*. Hence, the decreased interaction stability results in impaired binding to heme and affects its interaction with iodide and TG [36]. Guriaet al. demonstrated that p.A373S, p.S398T, and p.T725P had damaging effects on *TPO* mRNA expression and protein activity [37]. Furthermore, they measured wild-type and mutant enzyme activity using an iodide and guaiacol assay and confirmed that the p.A373S and p.T725P mutants were more damaging than the p.S398T mutant [37]. Begum et al. performed an analysis of quantum mechanics/molecular mechanics and molecular dynamics on p.A373S, p.S398T, and p.T725P. This molecular docking study showed that the full-length TPO mutant p.S398T structure interacted with all the crucial amino acids in the catalytic site of the TPO protein. Begum et al. [38] concluded that the mutant variants p.A373S, p.S398T, and p.T725P were involved in Bangladeshi patients with thyroid dyshormonogenesis, and their molecular docking– based study showed that the three mutant variants had damaging effects on the activity of the TPO protein. In our study, the protein structure of TPO, along with that of *DUOX1*, was crystallized, and the mutant protein complex was predicted to be less stable, which may have an effect on the function of the protein. The mutant protein p.S398T was analyzed for binding energies and dissociation constants for wild-type versus mutant complexes, with results of ΔG –15 vs. –13.8 kcal/mol and K_d $7.2E^{-12}$ vs. $7.0E^{-11}$ M, respectively. The RMSD of the TPO protein structure is high (1.091 Å) when it interacted with the DUOX protein, which illuminates the fact that the variant induced more structural variations in the binding protein DUOX than in TPO, which harbored the variant.

In silico predictions of function of c.1999 G>A (p.G653D) and c.6036 C>T (p.R1999W) in the TG protein

No three-dimensional structures are available for any TG regions [39]. Three-dimensional structural folding of a long-chain amino acid sequence was seen as a complex problem in the past [40]. Due to high molecular weight of TG and the lack of its crystal structure, *in silico* studies on TG have not been conducted to date. The recent elucidation of its crystal structure (PDB ID: 6SCJ) [23] facilitated the *in silico* exploration of *TG* variants. The predicted wild-type and mutant (p.G653D and p.R1999W) structures of TG proteins were aligned using the PyMol software [24] to analyze the effects of structural variation caused by the mutation in the surrounding residues within 4 Å.

Analysis of the effect of the p.G653D mutation

The analysis of the structure of the p.G653D mutant of TG showed that the mutation of Gly653 to Asp653 formed hydrogen-bonding interactions with the surrounding residue Ser990 (2.4 Å) as shown in Fig. 2, resulting in altered TG protein structure (Supplementary Video 1). The SAAMBE-3D prediction showed that the mutation destabilized the TG protein, with a ΔG value of 0.24 kcal/mol.

Next, the effect of the mutation on the adjacent residues was investigated using the DynaMut server [25]. The mutant Asp653 formed a weak polar van der Waals bond with the residues Gly555 and Ser885, polar proximal interactions with Gln576, and amide-amide interactions with Ser885, as shown in Fig. 3.

Analysis of the effect of the p.R1999W mutation

Next, the analysis of the mutant (p.R1999W) TG structures showed that the mutation of Arg1999 to Trp1999 formed hydrogen-bonding interactions with the surrounding residues, such as Val1955 (2.7 Å), Asn1980 (2.4 Å), and Thr2026 (2.7 Å) shown in Fig. 4 (Supplementary Video 2). This shows that the mutation of p.R1999W resulted in bonding interactions with the surrounding residues, which may affect the structure-function relationship of the TG protein. The effect of the p.R1999W mutation on protein stability was further assessed using the SAAMBE-3D server. The SAAMBE-3D prediction showed that the mutation destabilized the *TG* protein and the ΔG value was 0.79 kcal/mol.

The analysis of bonding and non-bonding interactions shows that the p.R1999W mutation results in non-bonding interactions with the residues; for instance, Trp1999 shows van der Waals clashes with Glu1820 and Thr1845, hydrogen-bond and van der Waals clashes with Glu1818, hydrophobic proximal clashes with Val1805 and Leu1780, hydrogen-bond proximal interactions with Ala1825, van der Waals clashes with Asn1803, and carbon-pi interactions with Val1805 (Fig. 5).

These altered interactions due to the p.G653D and p.R1999W mutations decrease the binding energies; here, the mutation of p.R1999W shows a more destabilizing effect on the TG protein due to more hydrogen-bonding interactions (Fig. 4), van der Waals and carbon-pi interactions (Fig. 5), and a higher binding energy (ΔG = 0.79 kcal/mol) compared to the p.G653D mutation (ΔG = 0.24

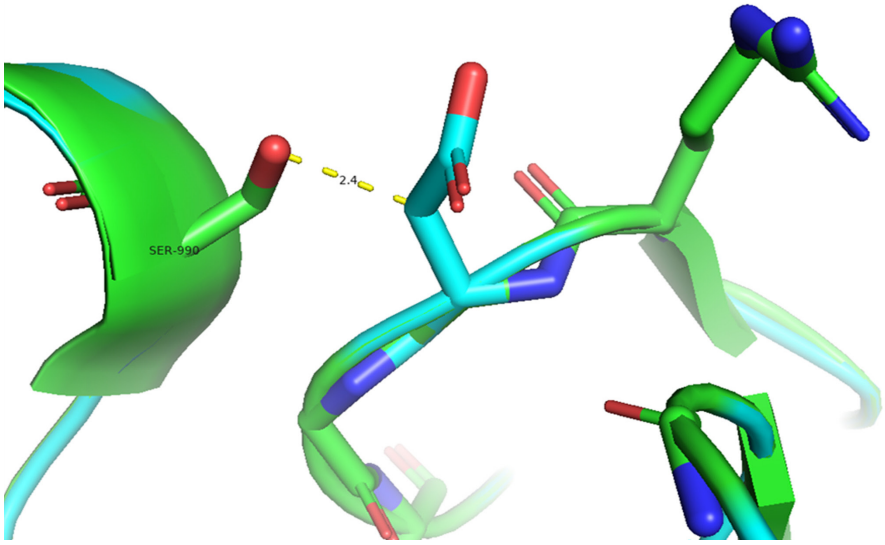

Fig. 2. An *in silico* analysis shows overlapping images of the wild-type (green) and mutant G653D (cyan) thyroglobulin protein. Here, the mutant residue Asp653 is shown in a stick model with hydrogen-bonding interactions with the Ser990 residue.

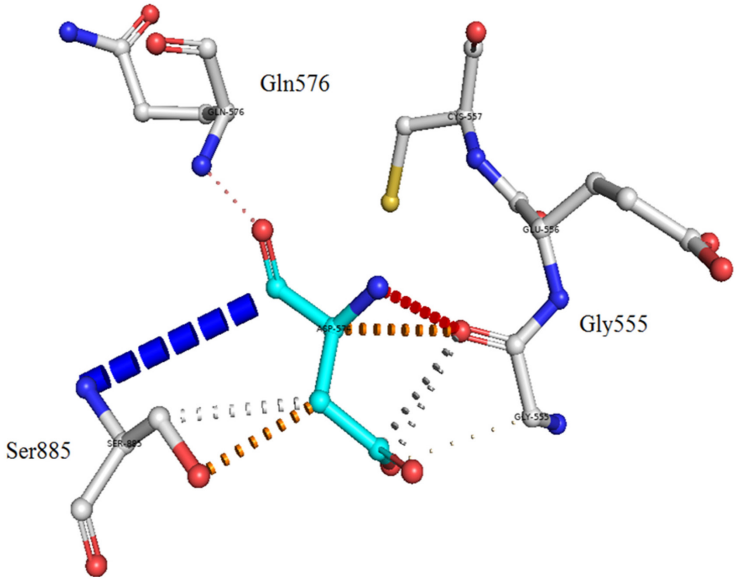

Fig. 3. An *in silico* analysis shows the mutant residue Asp653 colored in cyan, which is also represented as sticks alongside with the surrounding residues colored in white, which are involved in any type of interactions. The mutant thyroglobulin protein shows the weak polar van der Waals clashes with the residues Gly555 and Ser885, polar proximal interactions with Gln576, and amide-amide interactions with Ser885.

kcal/mol). These structural and energetic analyses show that the mutants are less stable and affect the structure-functional relationship of the human TG protein.

Aljouie et al. studied synonymous and nonsynonymous variants such as rs76672487 in the *ABCC2* gene and rs2069548 (p.G653D) in the *TG* gene. According to the Human Protein Atlas, these two genes are cancer-related genes. The variant p.G653D was ranked fourth of the selected variants using the chi-square test from the top SNPs in the glioblastoma multiforme +1000 genome database. The nearby tissue was enriched in this rare variant. This variant affects cancer susceptibility by suppressing mRNA expression (transcription and translation) [41]. Autoimmune thyroid diseases, including hyperthyroidism/Graves' disease and autoimmune hypothyroidism/Hashimoto's thyroiditis, are complex diseases caused by a malfunction in immune tolerance to self-thyroid antigens, such as TSHR, TPO, and TG [42]. Pyun et al. [43] identified epistasis between two polymorphic variants of the *HSD17B4* and *TG* genes. One variant, p.R1999W (c.5995 C > T), in the *TG* gene was tested

https://doi.org/10.5808/gi.21034

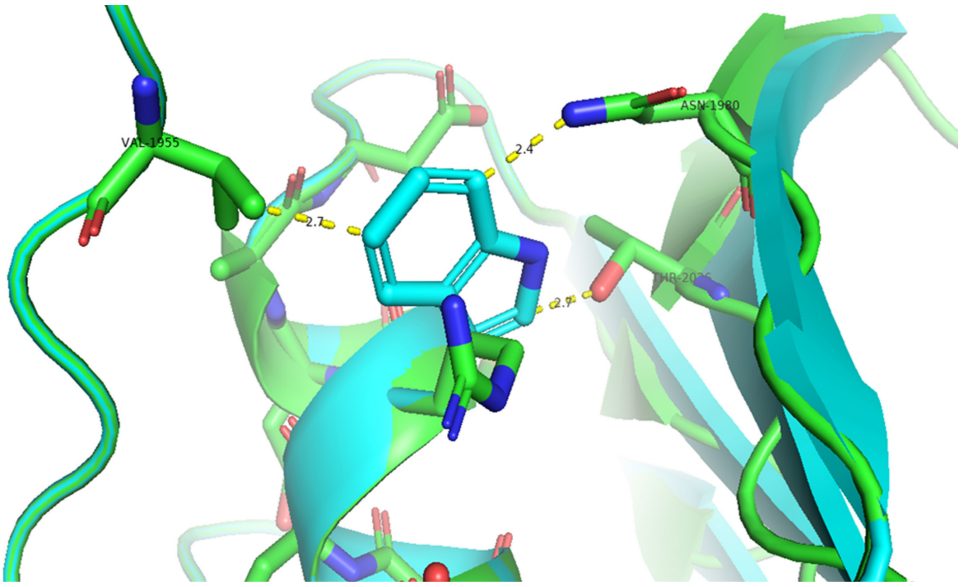

Fig. 4. An *in silico* analysis shows the overlapped images of wild-type (green) and mutant R1999W (cyan) thyroglobulin protein. The wild-type Arg1999 is shown in green and the mutant Trp1999 is shown in cyan. The mutant Trp1999 forms bonding interactions with Val1955, Asn1980, and Thr2026.

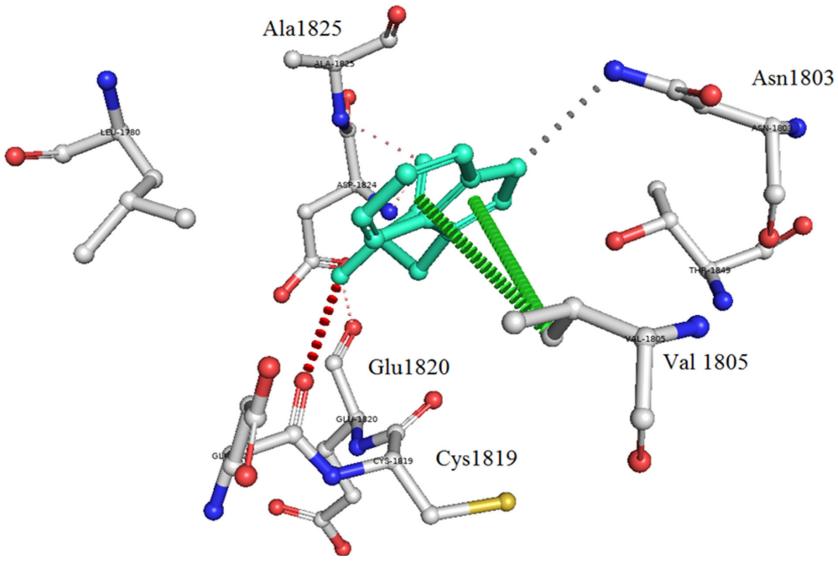

Fig. 5. An *in silico* analysis shows the mutant p.R1999W residue colored in light-green and also represented as sticks alongside with the surrounding residues colored in white, which are involved in any type of interactions. Here, Trp1999 shows van der Waals clashes with Glu1820 and Thr1845, hydrogen-bond and van der Waals clashes with Glu1818, hydrophobic proximal clashes with Val1805 and Leu1780, hydrogen-bond proximal interactions with Ala1825, van der Waals clashes with Asn1803, and carbon-pi interactions with Val1805.

for epistasis with the *HSD17B4* variant p.T687I (c.2060 C > T). The combined effect of these variants was significantly associated with premature ovarian failure, although these variants alone showed no significant association. Ban et al. [44] performed case-control association studies for 14 discovered *TG* variants in 285 autoimmune thyroid disease patients and 150 controls. One variant cluster (p.S753A & p.P797P in exon 10 and p.M1027V in exon 12) [44] and the exon 33 variant p.R1999W showed significant associations with patients suffering from autoimmune thyroid disease [44,45] in the United States [44], but were not associated with patients in the United Kingdom suffering from the same disease [42]. The combination of these variants conferred susceptibility to autoimmune thyroid disease. Ban et al. [44] analyzed gene-gene interactions between p.R1999W in the *TG* gene and *HLA-*

DR3. The variant (p.R1999W) showed promising results for interaction with *HLA-DR3* in conferring susceptibility to Graves' disease (odds ratio, 6.1) [44].

Although there are no reports depicting the direct association of p.G653D and p.R1999W variants of the TG protein with CH, several recent studies have illuminated the role of TG variants in the etiology of CH. These studies illustrate TG as one of the primary candidate genes to be evaluated in CH patients [46-49]. In our study, we investigated the effect of mutations such as p.G653D and p.R1999W, which were analyzed using computational structural biology methods, including mutant model building for the TG protein and structural analysis of interaction networks such as hydrogen-bonding. A further analysis of binding energies using the SAAMBE-3D server demonstrated that the p.G653D and p.R1999W mutations showed a protein destabilizing effect, which revealed that these SNPs may be deleterious and affect the TG protein structure-function relationship.

In conclusion, *in silico* studies revealed that SNPs in the 3′-UTR region altered the binding affinity of various miRs, thus influencing the expression of thyroid-associated genes. In this study, analyses of the computational protein-protein interactions and the binding energies of the p.S398T mutation in the *TPO* gene showed that the mutant protein complex was less stable than the wild-type complex, implying that this SNP may be deleterious. The altered variants p.G653D and p.R199W decreased the binding energies and contributed to a destabilizing effect on the TG protein.

ORCID

Yedukondalu Kollati: https://orcid.org/0000-0002-1851-3213
Radha Rama Devi Akella: https://orcid.org/0000-0002-4108-6213
Shaik Mohammad Naushad: https://orcid.org/0000-0001-8952-9581
Rajesh K. Patel: https://orcid.org/0000-0002-7210-6168
G. Bhanuprakash Reddy: https://orcid.org/0000-0003-4787-3944
Vijaya R. Dirisala: https://orcid.org/0000-0002-9433-2780

Authors' Contribution

Conceptualization: VRD, SMN, RRDA, GBR. Data curation: YK, RRDA. Formal analysis: YK, RKP. Funding acquisition: VRD. Methodology: YK, SMN, VRD. Writing - original draft: YK, SMN. Writing - review & editing: RRDA, RKP, GBR, VRD.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work was partly supported by a grant from DST-SERB, Government of India (ECR/2016/00304). The authors specially thank Dr. Pasumarthi NBS Srinivas, Dr. Hari Krishna K, Dr. Bajarang Vasant Kumbhar, Anusha Puvvada, Uma Maheshwar P for their support during the investigation.

Supplementary Materials

Supplementary data can be found with this article online at http://www.genominfo.org.

References

1. Agrawal P, Philip R, Saran S, Gutch M, Razi MS, Agroiya P, et al. Congenital hypothyroidism. Indian J Endocrinol Metab 2015;19:221-227.
2. Ahmad N, Irfan A, Al Saedi SA. Congenital hypothyroidism: screening, diagnosis, management, and outcome. J Clin Neonatol 2017;6:64-70.
3. ICMR Task Force on Inherited Metabolic Disorders. Newborn screening for congenital hypothyroidism and congenital adrenal hyperplasia. Indian J Pediatr 2018;85:935-940.
4. Kollati Y, Ambati RR, Reddy PN, Kumar NS, Patel RK, Dirisala VR. Congenital hypothyroidism: facts, facets and therapy. Curr Pharm Des 2017;23:2308-2313.
5. Ramesh BG, Bhargav PR, Rajesh BG, Devi NV, Vijayaraghavan R, Varma BA. Genotype-phenotype correlations of dyshormonogenetic goiter in children and adolescents from South India. Indian J Endocrinol Metab 2016;20:816-824.
6. Kota SK, Modi K, Kumaresan K. Elevated thyroid stimulating hormone in a neonate: drug induced or disease? Indian J Endocrinol Metab 2011;15(Suppl 2):S138-S140.
7. Lee CC, Harun F, Jalaludin MY, Heh CH, Othman R, Kang IN, et al. Variable clinical phenotypes in a family with homozygous c.1159G > A mutation in the thyroid peroxidase gene. Horm Res Paediatr 2014;81:356-360.
8. Sironi M, Menozzi G, Riva L, Cagliani R, Comi GP, Bresolin N, et al. Silencer elements as possible inhibitors of pseudoexon splicing. Nucleic Acids Res 2004;32:1783-1791.
9. Dhanoa JK, Sethi RS, Verma R, Arora JS, Mukhopadhyay CS. Long non-coding RNA: its evolutionary relics and biological implications in mammals: a review. J Anim Sci Technol 2018;60:25.
10. Christopher AF, Kaur RP, Kaur G, Kaur A, Gupta V, Bansal P. MicroRNA therapeutics: discovering novel targets and developing specific therapy. Perspect Clin Res 2016;7:68-74.

https://doi.org/10.5808/gi.21034

11. Popp NA, Yu D, Green B, Chew EY, Ning B, Chan CC, et al. Functional single nucleotide polymorphism in IL-17A 3' untranslated region is targeted by miR-4480 in vitro and may be associated with age-related macular degeneration. Environ Mol Mutagen 2016;57:58-64.
12. Kollati Y, Akella RR, Naushad SM, Borkar D, Thalla M, Nagalingam S, et al. Newborn screening and single nucleotide variation profiling of *TSHR*, *TPO*, *TG* and *DUOX2* candidate genes for congenital hypothyroidism. Mol Biol Rep 2020;47:7467-7475.
13. Bhattacharya A, Ziebarth JD, Cui Y. PolymiRTS Database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. Nucleic Acids Res 2014;42:D86-D91.
14. Liu W, Wang X. Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. Genome Biol 2019;20:18.
15. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. Elife 2015;4:e05005.
16. Kanoria S, Rennie W, Liu C, Carmack CS, Lu J, Ding Y. STarMir tools for prediction of microRNA binding sites. Methods Mol Biol 2016;1490:73-82.
17. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nat Methods 2015;12:7-8.
18. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF chimera: a visualization system for exploratory research and analysis. J Comput Chem 2004;25:1605-1612.
19. Kozakov D, Hall DR, Xia B, Porter KA, Padhorny D, Yueh C, et al. The ClusPro web server for protein-protein docking. Nat Protoc 2017;12:255-278.
20. Tina KG, Bhadra R, Srinivasan N. PIC: protein interactions calculator. Nucleic Acids Res 2007;35:W473-W476.
21. Vangone A, Bonvin AM. Contacts-based prediction of binding affinity in protein-protein complexes. Elife 2015;4:e07454.
22. Pahari S, Li G, Murthy AK, Liang S, Fragoza R, Yu H, et al. SAAMBE-3D: Predicting Effect of Mutations on Protein-Protein Interactions. Int J Mol Sci 2020;21:2563.
23. Coscia F, Taler-Vercic A, Chang VT, Sinn L, O'Reilly FJ, Izore T, et al. The structure of human thyroglobulin. Nature 2020;578:627-630.
24. Delano WL. The PyMOL molecular graphics system, version 1.8. New York: Schrodinger, 2002.
25. Rodrigues CH, Pires DE, Ascher DB. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res 2018;46:W350-W355.
26. Kollati Y, Akella RR, Naushad SM, Thalla M, Reddy GB, Dirisala VR. The rs1991517 polymorphism is a genetic risk factor for congenital hypothyroidism. 3 Biotech 2020;10:285.
27. Marini F, Luzi E, Brandi ML. MicroRNA role in thyroid cancer development. J Thyroid Res 2011;2011:407123.
28. Xie F, Li L, Luo Y, Chen R, Mei J. Long non-coding RNA LINC00488 facilitates thyroid cancer cell progression through miR-376a-3p/PON2. Biosci Rep 2021;41:BSR20201603.
29. Balzan S, Del Carratore R, Nicolini G, Beffy P, Lubrano V, Forini F, et al. Proangiogenic effect of TSH in human microvascular endothelial cells through its membrane receptor. J Clin Endocrinol Metab 2012;97:1763-1770.
30. Ahmetov I, Kulemin N, Popov D, Naumov V, Akimov E, Bravy Y, et al. Genome-wide association study identifies three novel genetic markers associated with elite endurance performance. Biol Sport 2015;32:3-9.
31. Campo C, Kohler A, Figlioli G, Elisei R, Romei C, Cipollini M, et al. Inherited variants in genes somatically mutated in thyroid cancer. PLoS One 2017;12:e0174995.
32. Targovnik HM, Citterio CE, Rivolta CM. Iodide handling disorders (NIS, TPO, TG, IYD). Best Pract Res Clin Endocrinol Metab 2017;31:195-212.
33. Deladoey J, Pfarr N, Vuissoz JM, Parma J, Vassart G, Biesterfeld S, et al. Pseudodominant inheritance of goitrous congenital hypothyroidism caused by *TPO* mutations: molecular and *in silico* studies. J Clin Endocrinol Metab 2008;93:627-633.
34. Turkkahraman D, Alper OM, Pehlivanoglu S, Aydin F, Yildiz A, Luleci G, et al. Analysis of *TPO* gene in Turkish children with iodide organification defect: identification of a novel mutation. Endocrine 2010;37:124-128.
35. Turkkahraman D, Alper OM, Aydin F, Yildiz A, Pehlivanoglu S, Luleci G, et al. Final diagnosis in children with subclinical hypothyroidism and mutation analysis of the thyroid peroxidase gene (*TPO*). J Pediatr Endocrinol Metab 2009;22:845-851.
36. Rivolta CM, Moya CM, Esperante SA, Gutnisky VJ, Varela V, Targovnik HM. The thyroid as a model for molecular mechanisms in genetic diseases. Medicina (B Aires) 2005;65:257-267.
37. Guria S, Bankura B, Balmiki N, Pattanayak AK, Das TK, Sinha A, et al. Functional analysis of thyroid peroxidase gene mutations detected in patients with thyroid dyshormonogenesis. Int J Endocrinol 2014;2014:390121.
38. Begum MN, Islam MT, Hossain SR, Bhuyan GS, Halim MA, Shahriar I, et al. Mutation sectrum in TPO gene of Bangladeshi patients with thyroid dyshormonogenesis and analysis of the effects of different mutations on the structural features and functions of TPO protein through in silico approach. Biomed Res Int 2019;2019:9218903.

39. Targovnik HM, Citterio CE, Rivolta CM. Thyroglobulin gene mutations in congenital hypothyroidism. Horm Res Paediatr 2011;75:311-321.
40. Haddad Y, Adam V, Heger Z. Ten quick tips for homology modeling of high-resolution protein 3D structures. PLoS Comput Biol 2020;16:e1007449.
41. Aljouie A, Schatz M, Roshan U. Machine learning based prediction of gliomas with germline mutations obtained from whole exome sequences from TCGA and 1000 Genomes Project. In: The 3rd International Conference on Intelligent Computing in Data Sciences (ICDS), 2019 Oct 28-30, Marrakech, Morocco.
42. Collins JE, Heward JM, Howson JM, Foxall H, Carr-Smith J, Franklyn JA, et al. Common allelic variants of exons 10, 12, and 33 of the thyroglobulin gene are not associated with autoimmune thyroid disease in the United Kingdom. J Clin Endocrinol Metab 2004;89:6336-6339.
43. Pyun JA, Kim S, Cha DH, Ko JJ, Kwack K. Epistasis between the HSD17B4 and TG polymorphisms is associated with premature ovarian failure. Fertil Steril 2012;97:968-973.
44. Ban Y, Greenberg DA, Concepcion E, Skrabanek L, Villanueva R, Tomer Y. Amino acid substitutions in the thyroglobulin gene are associated with susceptibility to human and murine autoimmune thyroid disease. Proc Natl Acad Sci U S A 2003;100:15119-15124.
45. Zhang ML, Zhang DM, Wang CE, Chen XL, Liu FZ, Yang JX. Association between thyroglobulin polymorphisms and autoimmune thyroid disease: a systematic review and meta-analysis of case-control studies. Genes Immun 2019;20:484-492.
46. Raef H, Al-Rijjal R, Al-Shehri S, Zou M, Al-Mana H, Baitei EY, et al. Biallelic p.R2223H mutation in the thyroglobulin gene causes thyroglobulin retention and severe hypothyroidism with subsequent development of thyroid carcinoma. J Clin Endocrinol Metab 2010;95:1000-1006.
47. Hu X, Chen R, Fu C, Fan X, Wang J, Qian J, et al. Thyroglobulin gene mutations in Chinese patients with congenital hypothyroidism. Mol Cell Endocrinol 2016;423:60-66.
48. Santos-Silva R, Rosario M, Grangeia A, Costa C, Castro-Correia C, Alonso I, et al. Genetic analyses in a cohort of Portuguese pediatric patients with congenital hypothyroidism. J Pediatr Endocrinol Metab 2019;32:1265-1273.
49. Tanaka T, Aoyama K, Suzuki A, Saitoh S, Mizuno H. Clinical and genetic investigation of 136 Japanese patients with congenital hypothyroidism. J Pediatr Endocrinol Metab 2020;33:691-701.

https://doi.org/10.5808/gi.21034

G&I
Genomics & Informatics

Original article

eISSN 2234-0742
Genomics Inform 2021;19(3):e30
https://doi.org/10.5808/gi.21048

Received: August 19, 2021
Revised: September 15, 2021
Accepted: September 17, 2021

*Corresponding author:
E-mail: yejun@catholic.ac.kr

Rapid and sensitive detection of *Salmonella* species targeting the *hilA* gene using a loop-mediated isothermal amplification assay

Jiyon Chu[1,2,3,4], Juyoun Shin[5], Shinseok Kang[6], Sun Shin[2,3,4], Yeun-Jun Chung[1,2,3,4*]

[1]Department of Biomedicine & Health Sciences, Graduate School, The Catholic University of Korea, Seoul 06591, Korea
[2]Department of Microbiology, College of Medicine, The Catholic University of Korea, Seoul 06591, Korea
[3]Precision Medicine Research Center, College of Medicine, The Catholic University of Korea, Seoul 06591, Korea
[4]Integrated Research Center for Genome Polymorphism, College of Medicine, The Catholic University of Korea, College of Medicine, Seoul 06591, Korea
[5]ConnectaGen, Hanam 12918, Korea
[6]Chungbuk Veterinary Services Laboratory, Chungju 27336, Korea

Salmonella species are among the major pathogens that cause foodborne illness outbreaks. In this study, we aimed to develop a loop-mediated isothermal amplification (LAMP) assay for the rapid and sensitive detection of *Salmonella* species. We designed LAMP primers targeting the *hilA* gene as a universal marker of *Salmonella* species. A total of seven *Salmonella* species strains and 11 non-*Salmonella* pathogen strains from eight different genera were used in this study. All *Salmonella* strains showed positive amplification signals with the *Salmonella* LAMP assay; however, there was no non-specific amplification signal for the non-*Salmonella* strains. The detection limit was 100 femtograms (20 copies per reaction), which was ~1,000 times more sensitive than the detection limits of the conventional polymerase chain reaction (PCR) assay (100 pg). The reaction time for a positive amplification signal was less than 20 minutes, which was less than one-third the time taken while using conventional PCR. In conclusion, our *Salmonella* LAMP assay accurately detected *Salmonella* species with a higher degree of sensitivity and greater rapidity than the conventional PCR assay, and it may be suitable for point-of-care testing in the field.

Keywords: *hilA* gene, loop-mediated isothermal amplification, point-of-care testin, *Salmonella*

Introduction

Salmonella species are among the major pathogens that cause foodborne infections. Indeed, *Salmonella* species are thought to be the main factor contributing to diarrheal disease-associated deaths [1]. According to the World Health Organization, *Salmonella* accounts for approximately 9% of diarrheal illnesses and it costs 2.7 billion US dollars to care for these patients in the United States per year [2,3]. Especially, since food production is evolving to reflect demand for raw and lightly cooked foods, foodborne pathogens including *Salmonella* have continued to spread all over the world [4,5].

To detect *Salmonella* at the early stage of infection, a sensitive and reliable method is crucial. Various methods have been developed for *Salmonella* detection using molecular tools such as direct polymerase chain reaction (PCR) and multiplex quantitative PCR [6]. A PCR-based assay is currently the most commonly used tool for molecular screening of *Salmonella* [7]. However, PCR methods require specialized equipment that must be optimized for the amplification of target genes from samples, along with the extraction and purification steps carried out by a well-trained individual. Several technical issues such as long reaction times (about 2 h) and the requirement for well-purified nucleic acid are drawbacks for clinical application in many real-world settings. The lack of hardware resources in the field constitutes a critical barrier to the detection of pathogens directly from clinical samples [8]. Taken together, the limitations of conventional PCR-based methods make these assays difficult to use for point-of-care testing (POCT). POCT is a new concept in the laboratory-medicine discipline that enables screening for pathogens at the site where the infection occurs without transporting the test materials to well-equipped laboratories; by obviating the need for transport, POCT would be helpful to reduce the time necessary for clinical decision-making [9,10]. Therefore, POCT requires fast reaction time and simple equipment.

Loop-mediated isothermal amplification (LAMP) technology, an isothermal nucleic acid amplification method developed by Notomi et al. [11], can amplify target nucleic acids at a consistent temperature without changing the temperature for PCR cycling. The LAMP assay has been proven to be an effective tool for POCT in the field to test for infectious diseases [12-14]. There are several advantages of the LAMP technology. First, due to isothermal amplification, LAMP does not need a thermal cycler; therefore, this technology is ideal for POCT in the field. Second, the reaction time is much shorter than that of PCR, which is advantageous for rapid testing of highly contagious infections and preventing the spread of infections at an early stage. Third, the detection sensitivity of LAMP is higher than that of conventional PCR due to the loop-mediated amplification strategy. Fourth, LAMP is relatively less sensitive to PCR inhibitors, which can contaminate samples from sources such as feces; therefore, LAMP would be advantageous for minimizing the sample preparation procedure. During the LAMP reaction, the inner primers anneal by Watson-Crick complementarity to a region within the target and a hybridized loop structure is generated by strand invasion, which allows more efficient amplification for the synthesis of large amounts of DNA in a short time [11].

Several studies have reported LAMP assays targeting *Salmonella* in food [15,16]. The invA gene has been the most frequent target for *Salmonella* detection LAMP assays [15]. In this study, we aimed to develop a real-time LAMP assay targeting the *hilA* gene for universal detection of *Salmonella* species. We also validated the LAMP assay with seven clinically important *Salmonella* pathogens.

Methods

Bacterial strains

In this study, seven major *Salmonella* species were used. Six species (*Salmonella* Typhi, *Salmonella* Typhimurium, *Salmonella enteritidis*, *Salmonella* Paratyphi A, *Salmonella* Paratyphi B, and *Salmonella* Infantis) were collected from the Korea National Institute of Health (KNIH), Republic of Korea. *Salmonella enterica* was collected from Chungbuk Veterinary Service Laboratory, Chungju, Republic of Korea. For the specificity test, we used 11 non-*Salmonella* pathogens: seven Gram-negative and four Gram-positive bacteria (Table 1).

DNA extraction and template DNA preparation

All *Salmonella* strains (n = 7) were cultivated for 24 h at 37°C on sheep blood agar (Bandio, Pocheon, Korea). Non-*Salmonella* strains (n = 11) were cultivated for 36 hours at 37°C on Difco Luria-Bertani agar (BD, Franklin Lakes, NJ, USA). DNA extraction was done via QlAamp DNA Mini kit (Qiagen, Germantown, MD, USA) according to the manufacturer's instructions using the proto-

Table 1. Strains used for testing the *Salmonella* LAMP assay based on the *hilA* gene

Species		Strain No.
Salmonella strains		
1	*Salmonella* Typhi	NCCP 10820
2	*Salmonella* Typhimurium	NCCP 16207
3	*Salmonella* Enteritidis	NCCP 14547
4	*Salmonella enterica*	CVSL 0029
5	*Salmonella* Paratyphi A	NCCP 14759
6	*Salmonella* Paratyphi B	NCCP 12204
7	*Salmonella* infantis	NCCP 12233
Non-*Salmonella* strains		
1	*Pseudomonas aeruginosa*	NCCP 14781
2	*Klebsiella pneumoniae*	NCCP 14764
3	*Enterobacter aerogenes*	NCCP 14761
4	*Escherichia coli*	NCCP 14538
5	*Acinetobacter baumannii*	NCCP 14782
6	*Shigella flexneri*	NCCP 14744
7	*Shigella sonnei*	NCCP 14773
8	*Streptococcus pneumoniae*	NCCP 15898
9	*Streptococcus pyogenes*	NCCP 14783
10	*Staphylococcus epidermidis*	NCCP 14768
11	*Streptococcus salivarius*	NCCP 16179

LAMP, loop-mediated isothermal amplification.

https://doi.org/10.5808/gi.21048

col. For Gram-negative bacteria, colonies on the freshly cultured bacteria were added to the ATL buffer (180 μL) containing proteinase K (20 μL; 20 mg/mL) and suspension. The sample was then mixed by vortexing and incubated at 56°C until the bacteria were completely lysed. Next, 200 μL of lysis buffer was added to each sample and incubated at 70°C for 10 min. Then, vortexing for 15 seconds was performed right after adding 200 μL of ethanol. The mixture was transferred to the QIAamp Mini Spin column and centrifuge at 6,000 ×g for 1 min. The genomic DNA in the spin column was washed using 500 μL of AW1 buffer and centrifuged at 6,000 ×g for 1 min. Next, 500 μL of AW2 buffer was added and centrifuged at 20,000 ×g for 3 min. For the elution process, 100 μL of distilled water was added and centrifuged at 6,000 ×g. The genomic DNA was measured using Qubit 3.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) and the DNA was stored at −20°C for further study. Gram-positive bacteria were added to 180 μL of lysis buffer containing proteinase K (20 μL; 20 mg/mL) and incubated for 30 min at 37°C instead of using the ATL buffer. The other steps were the same as in the protocol for Gram-negative bacteria.

Primer design for the *Salmonella* LAMP assay

Primer Explorer V4 (http://primerexplorer.jp/elamp4.0.0/index.html) was used to design *Salmonella* species LAMP primers. The three sets of primers included a forward outer primer (F3), a backward outer primer (B3), a forward inner primer (FIP), a backward inner primer (BIP), and two loop primers: a forward loop primer (LF) and a backward loop primer (LB). All sets of primers were then validated by the BLAST program. Of the three sets of LAMP primers, the set that demonstrated the best amplification performance was selected as the *Salmonella* LAMP primer set (Fig. 1).

LAMP assay

The LAMP reaction was carried as described elsewhere with some modifications [17]. In brief, a 20 μL reaction mixture was prepared that contained 1 μL of target genomic DNA along with 1.6 μM each of the primers FIP and BIP, 0.2 μM each of F3 and B3, 0.4 μM of LF and LB. 4U of the Bst 2.0 DNA polymerase (New England Biolabs, Ipswich, MA, USA), 8 mM of $MgSO_4$ (New England Biolabs), 1.5 mM of dNTPs (Thermo Fisher Scientific), 1× isothermal amplification buffer (New England Biolabs), and 1.25 M of N-methyl formamide (NMF) and isobutylamide (IBA). Furthermore, 2 μM of SYTO 9 (Thermo Fisher Scientific) was added to enhance fluorescence in the presence of DNA in the real-time assay. The amplification reaction was performed at 60°C for 45 min and terminated by heating at 80°C for 3 min using the CFX96 Touch Real-Time PCR Detection System (Bio-Rad Laboratories, Hercules, CA, USA). The results were shown as a graph on the monitor of real-time analysis software (BioRad CFX Manager, Bio-Rad Laboratories). The fluorescence curve was captured from the BioRad CFX Manager graph.

Optimization of the LAMP reaction and evaluation of the detection sensitivity and specificity

To optimize the amplification conditions, the LAMP reaction was performed at different temperatures (60°C–65°C). To evaluate the detection sensitivity of the LAMP assays, we tested the detection limit using *S. enterica* DNA, which was 10-fold serially diluted from

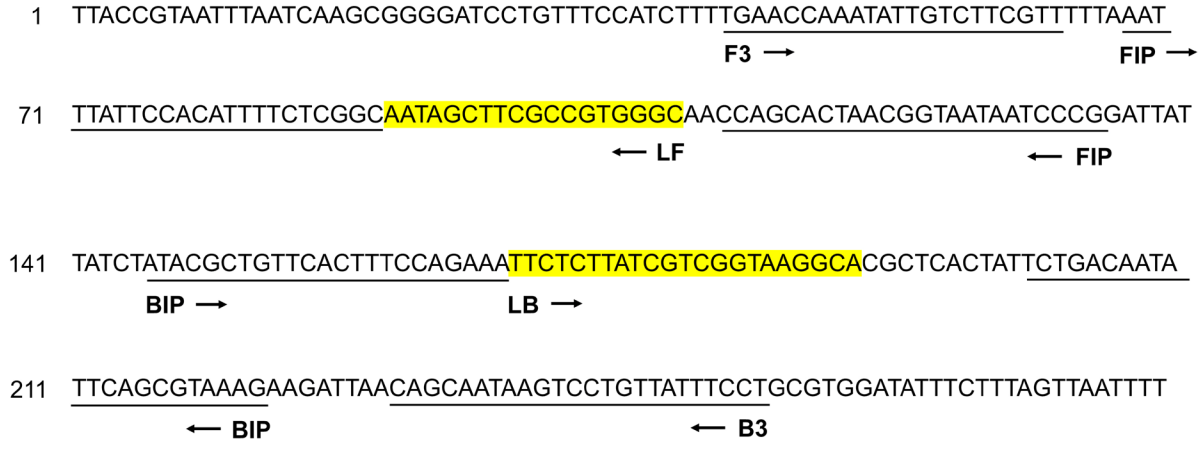

Fig. 1. Primer design for the *Salmonella* loop-mediated isothermal amplification assay. The positions of six primers (F3, B3, FIP, BIP, LB, and LF) were aligned on the partial nucleotide sequence of the *hilA* gene (GeneBank accession No. CP075108). The forward/backward outer primers (F3/B3) and forward/backward inner primers (FIP/BIP) are underlined. The forward/backward loop primers (LF/LB) are highlighted in yellow. Arrowheads indicate the direction of the primers.

1 ng to 1 ag and used in the LAMP reaction. To test for non-specific amplification of non-*Salmonella* pathogens, we conducted the LAMP assay with seven non-*Salmonella* Gram-negative and four Gram-positive bacteria.

Results

Design of LAMP primers

Six LAMP primers were designed targeting the *hilA* gene as a universal marker of *Salmonella* species: two outer primers (F3 and B3), two inner primers (FIP and BIP), and two loop primers (LF and LB) (Fig. 1). Validation of the primer set was performed using *S. enterica* DNA under standard LAMP conditions (60°C for 45 min) (Fig. 2). We performed the LAMP reaction with four different combinations of primers. Set 1 involved LAMP with two inner and two outer primers but without a loop primer, set 2 used the set 1 primers with one loop primer (LB), set 3 used the set 1 primers with another loop primer (LF), and set 4 used all six primers (Fig. 2). As expected, LAMP without the two loop primers showed the slowest amplification reaction. LAMP with one loop primer showed a faster amplification reaction than LAMP without any loop primer. LAMP with all six primers, including two loop primers, showed the fastest amplification reaction, in which a fluorescent signal appeared just 12 min after starting the reaction. None of the negative controls showed a positive signal during the reaction. Therefore, we decided to include all six primers in our *Salmonella* species LAMP system.

Optimization of LAMP reaction conditions

To determine the best LAMP conditions, we compared three different reaction conditions: 60°C, 63°C, and 65°C for 45 min (Fig. 3). Consistent with the results in Fig. 2, LAMP with all six primers including two loop primers showed the fastest amplification reaction regardless of reaction temperature. Regarding the reaction conditions, a reaction temperature of 60°C showed the best amplification performance. Integrating these, we set the conditions for our *Salmonella* LAMP assay as using all six primers at 60°C for 45 min.

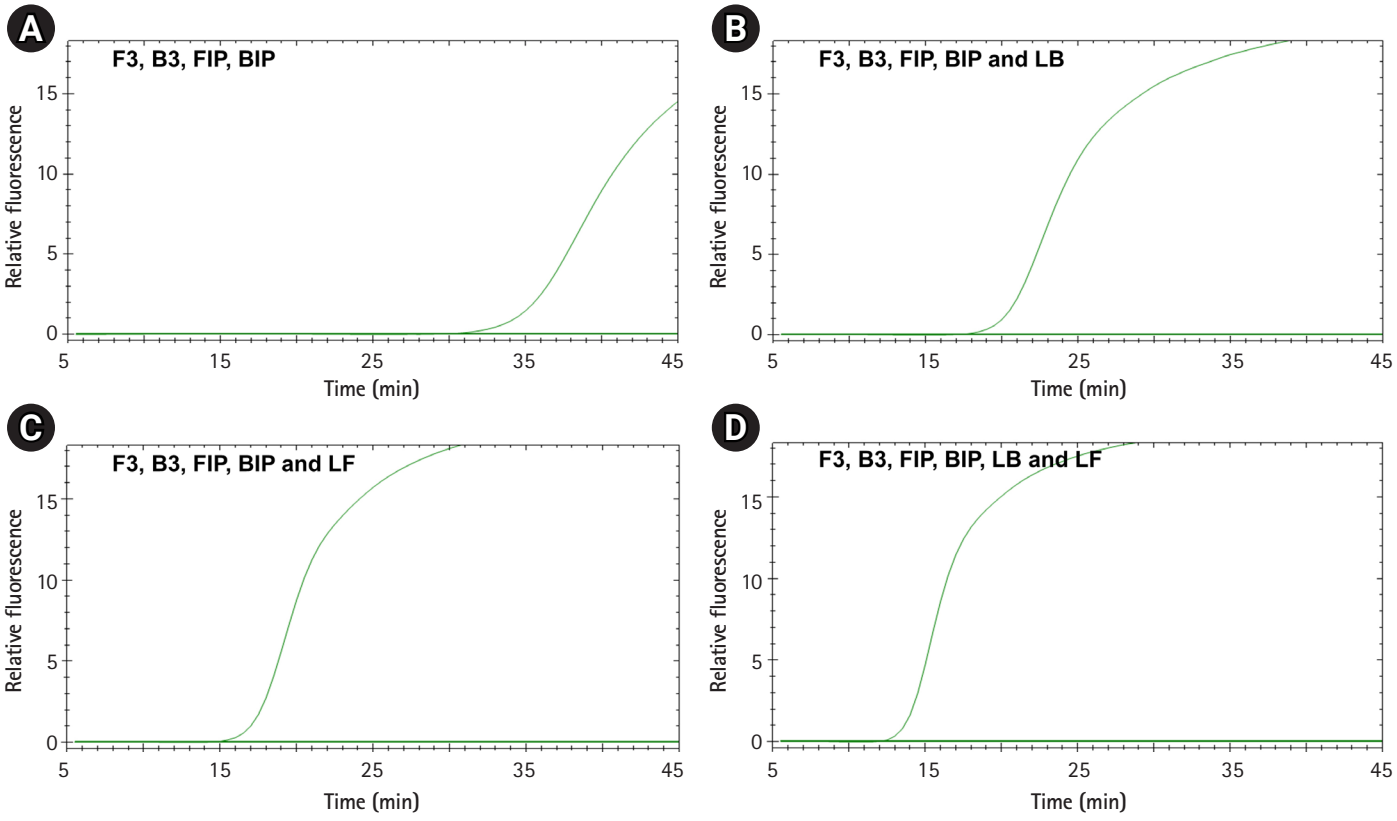

Fig. 2. Optimization of the combinations of primers. Loop-mediated isothermal amplification (LAMP) reactions were performed with *Salmonella enterica* DNA with four different combinations of primers. (A) LAMP with two inner and two outer primers. (B) LAMP with the combination-A primers and one loop primer (LB). (C) LAMP with the combination-A primers and another loop primer (LF). (D) LAMP with all six primers. The x-axis represents the time needed for the LAMP reaction using a real-time PCR cycler; the y-axis represents the relative fluorescence signal. The flat amplification curve was found in the negative control.

https://doi.org/10.5808/gi.21048

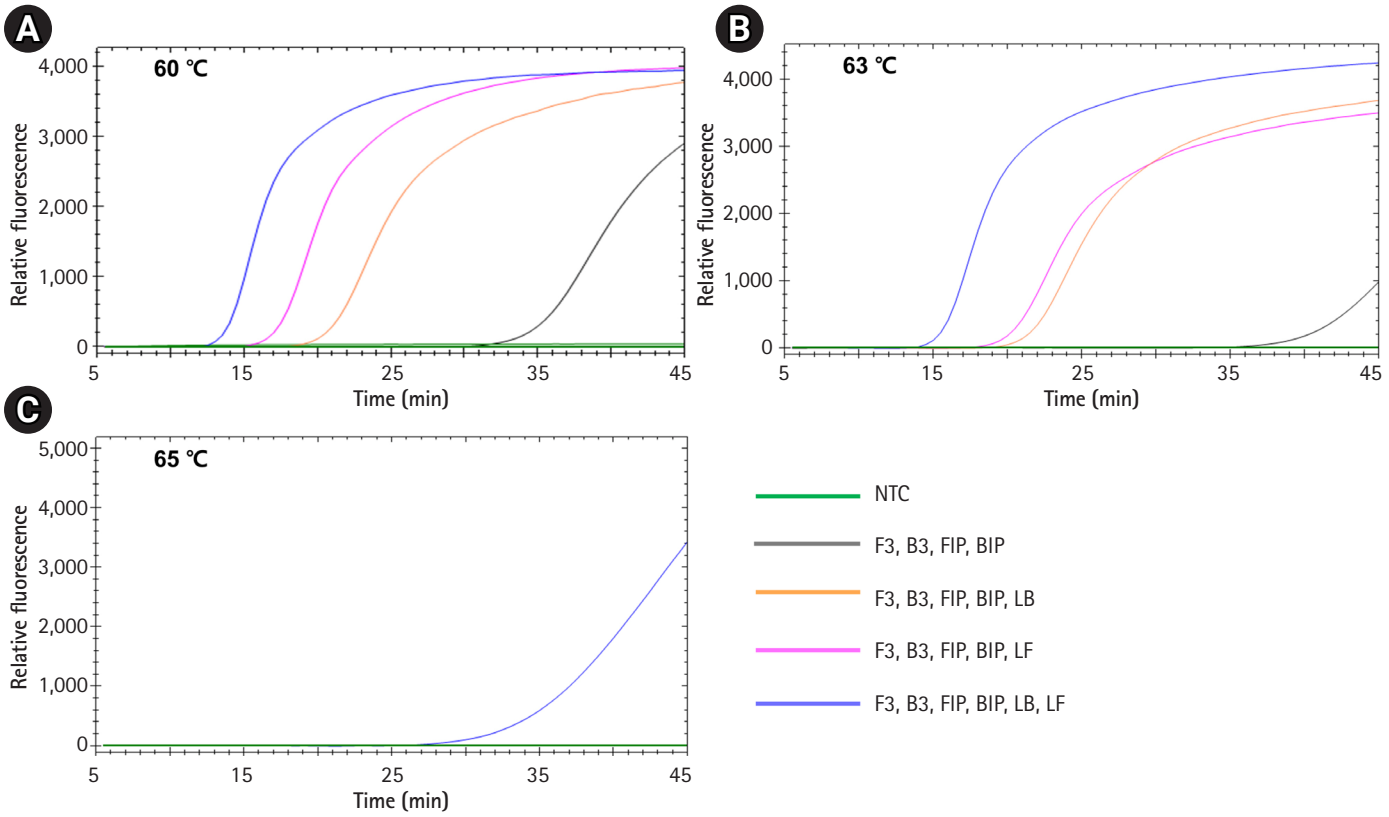

Fig. 3. Optimization of loop-mediated isothermal amplification (LAMP) reaction conditions. LAMP reactions were performed with *Salmonella enterica* DNA under three different reaction conditions: 60°C (A), 63°C (B), and 65°C (C). The combinations of primers were the same as shown in Fig. 2. NTC, negative control.

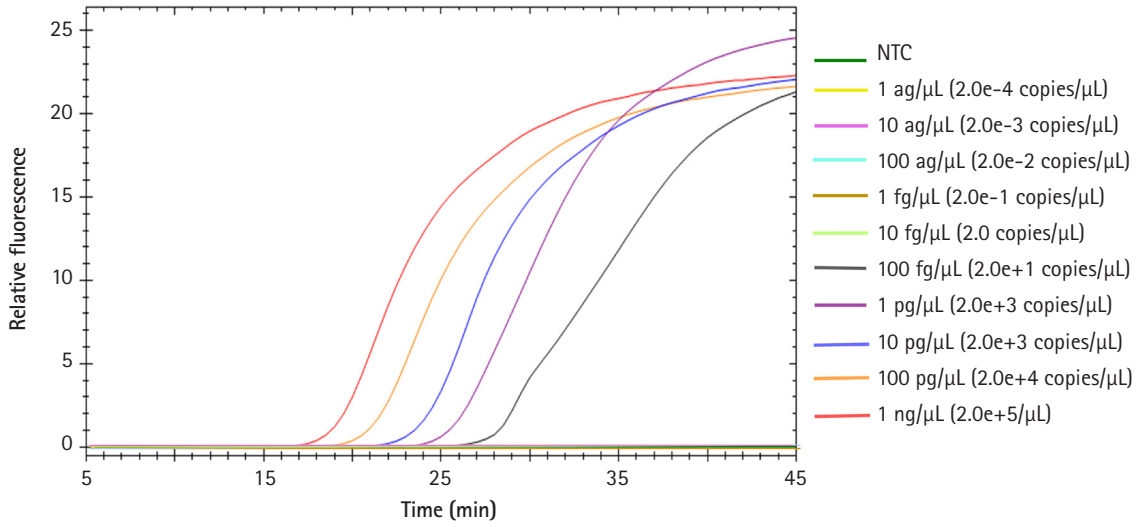

Fig. 4. Sensitivity of the *Salmonella* species loop-mediated isothermal amplification (LAMP) assay. The template DNA (*Salmonella enterica* DNA) was 10-fold serially diluted (1 ng/μL to 1 ag/μL) and tested with our *Salmonella* LAMP assay. NTC, negative control.

Detection sensitivity of the LAMP assay

To test the detection sensitivity of our *Salmonella* LAMP assay, we observed the limit of detection (LOD) of this assay. *Salmonella enterica* DNA was 10-fold serially diluted from 1 ng to 1 ag and used in the LAMP reaction (Fig. 4). The LOD of the *Salmonella* LAMP assay was 100 fg which corresponds to 20 copies of *Salmonella* DNA. The time for the fluorescent signal to appear from 1 ng of template DNA was less than 20 min. Even in the case of 1 pg of

template DNA, the amplification signal appeared within 25 min.

Validating the universal detection of *Salmonella* species and specificity

To validate whether our universal *Salmonella* LAMP assay can detect diverse *Salmonella* species, we prepared seven clinically important *Salmonella* species (*S.* Typhi, *S.* Typhimurium, *S.* Enteritidis, *S. enterica*, *S.* Paratyphi A, *S.* Paratyphi B, *S.* Infantis) and used them with this assay (Fig. 5). All seven *Salmonella* species showed specific amplification signals at around 15 min except *S.* Paratyphi A. In contrast, the negative control did not show any amplification signal. To verify the *Salmonella*-specific detection, we also conducted our *Salmonella* LAMP assay with 11 non-*Salmonella* pathogens (seven Gram-negative and four Gram-positive bacteria) (Table 1). None of the non-*Salmonella* pathogens showed any amplification signal with our *Salmonella* LAMP assay (data not shown).

Discussion

In this study, to develop a real-time LAMP assay for the rapid and sensitive detection of *Salmonella* species, we designed a LAMP primer set targeting the *hilA* gene. The *hilA* gene, a member of the transcriptional regulator genes encoded in the *Salmonella* pathogenicity island 1 (SPI1), plays an important role in the pathogenesis of *Salmonella* infection by activating the expression of SPI1 [18,19]. SPI1 is one of the key virulence factors for *Salmonella* infection and invasion. The *hilA* gene is known to be *Salmonella* species-specific and absent in other Gram-negative bacteria [20]. Therefore, this gene has been targeted for detecting *Salmonella* infection by PCR [21]. However, no study has yet developing a LAMP assay targeting the *hilA* gene for the detection of *Salmonella* species.

In the LAMP assay, a high level of amplification efficiency can be achieved under isothermal conditions due to strand displacement by the Bst DNA polymerase enzyme [11]. However, the possibility of non-specific amplification is also high due to the high level of amplification. To minimize this possibility, we added NMF and IBA as described previously [17]. In addition, we optimized the LAMP reaction conditions to facilitate the efficient and reliable detection of *Salmonella* species.

When we checked the detection sensitivity of our *Salmonella* LAMP assay using *S. enterica* DNA, the LOD was 100 fg, which corresponds to just 20 copies of *Salmonella* DNA. In the previous PCR assay targeting the *hilA* gene using *S. typhimurium* DNA, the LOD was 100 pg [21]. Therefore, although the *Salmonella* strains used for the sensitivity test were different, our *Salmonella* LAMP assay is approximately 1,000 times more sensitive than the PCR-based assay to detect *Salmonella* species. Of note, when we tested whether our universal *Salmonella* LAMP assay can detect diverse *Salmonella* species, all seven *Salmonella* species (including *S.* Typhimurium) showed consistent detection performance, providing further support that our *Salmonella* LAMP assay is much more sensitive than the PCR assay. This result also suggests that this *Salmonella* LAMP assay is universally applicable for the detection of diverse clinically important *Salmonella* species.

When we tested the assay using 11 non-*Salmonella* pathogens, none of them showed any amplification signals with our *Salmonella*

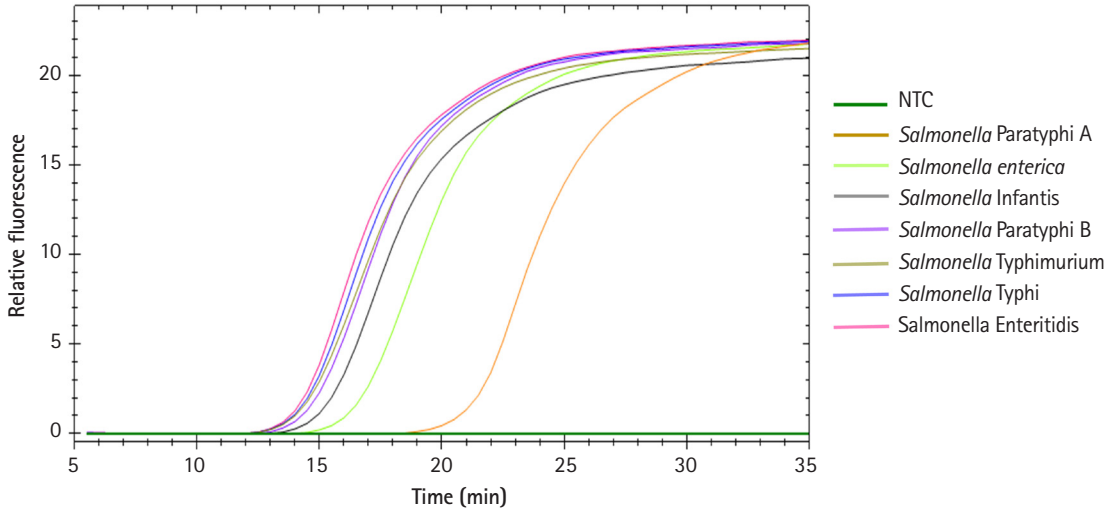

Fig. 5. Universal detection of *Salmonella* species using our loop-mediated isothermal amplification (LAMP) assay. DNA extracted from seven clinically important *Salmonella* species (*Salmonella* Typhi, *Salmonella* Typhimurium, *Salmonella* Enteritidis, *Salmonella enterica*, *Salmonella* Paratyphi A, *Salmonella* Paratyphi B, *Salmonella* Infantis) was used in our *Salmonella* LAMP assay. NTC, negative control.

https://doi.org/10.5808/gi.21048

LAMP assay, suggesting high *Salmonella* species specificity of our LAMP assay. According to the BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) analysis, none of the 11 non-*Salmonella* pathogens showed similar sequences to the *hilA* gene (data not shown). Of particular note, a previous study based on the conventional PCR method targeting the *hilA* gene took approximately 1 hour and 45 minutes [21]. However, all *Salmonella* species tested in this study showed specific amplification signals at around 15 minutes except S. Paratyphi A, demonstrating that our *Salmonella* LAMP assay would be ideal for POCT.

One of the fundamental limitations of conventional PCR is that PCR is easily inhibited by substances in the sample. LAMP technology can overcome this drawback of PCR, as LAMP is relatively less sensitive to PCR inhibitors, which can contaminate samples from sources such as feces. Therefore, LAMP would be advantageous to minimize the sample preparation procedure.

In conclusion, our *Salmonella* LAMP assay targeting the *hilA* gene demonstrated a high level of sensitivity and specificity. Considering the higher sensitivity of our assay than conventional PCR, its rapid reaction time, and its lower sensitivity to potential PCR inhibitors, our assay could be a useful POCT tool for *Salmonella* species.

ORCID

Jiyon Chu: https://orcid.org/0000-0003-4349-009X
Juyoun Shin: https://orcid.org/0000-0001-9207-7053
Shinseok Kang: https://orcid.org/0000-0002-6759-9324
Sun Shin: https://orcid.org/0000-0002-2146-4336
Yeun-Jun Chung: http://orcid.org/0000-0002-6943-5948

Authors' Contribution

Conceptualization: JC, JS, YJC. Data curation: JC, JS, SS, YJC. Formal analysis: JC, JS. Funding acquisition: YJC. Methodology: JC, JS, SS, SK. Writing - original draft: JC. Writing - review & editing: YJC.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work was supported by a grant from the Korea Health Industry Development Institute (KHIDI) (HI21C0561) and a grant from the National Research Foundation of Korea (2017M3C9A6047615). We thank KREONET (Korea Research Environment Open NETwork) and KISTI (Korea Institute of Science and Technology Information) for allowing us to use their network infrastructure.

References

1. Tauxe RV, Doyle MP, Kuchenmuller T, Schlundt J, Stein CE. Evolving public health approaches to the global challenge of foodborne infections. Int J Food Microbiol 2010;139 Suppl 1:S16-28.
2. Kokkinos PA, Ziros PG, Bellou M, Vantarakis A. Loop-mediated isothermal amplification (LAMP) for the detection of *Salmonella* in food. Food Anal Methods 2014;7:512-526.
3. Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson MA, Roy SL, et al. Foodborne illness acquired in the United States: major pathogens. Emerg Infect Dis 2011;17:7-15.
4. Besser JM. *Salmonella* epidemiology: a whirlwind of change. Food Microbiol 2018;71:55-59.
5. Zhuang L, Gong J, Li Q, Zhu C, Yu Y, Dou X, et al. Detection of *Salmonella* spp. by a loop-mediated isothermal amplification (LAMP) method targeting bcfD gene. Lett Appl Microbiol 2014;59:658-664.
6. Heymans R, Vila A, van Heerwaarden CA, Jansen CC, Castelijn GA, van der Voort M, et al. Rapid detection and differentiation of *Salmonella* species, *Salmonella* Typhimurium and *Salmonella* Enteritidis by multiplex quantitative PCR. PLoS One 2018;13:e0206316.
7. Chen G, Chen R, Ding S, Li M, Wang J, Zou J, et al. Recombinase assisted loop-mediated isothermal DNA amplification. Analyst 2020;145:440-444.
8. Nzelu CO, Kato H, Peters NC. Loop-mediated isothermal amplification (LAMP): an advanced molecular point-of-care technique for the detection of Leishmania infection. PLoS Negl Trop Dis 2019;13:e0007698.
9. Luppa PB, Muller C, Schlichtiger A, Schlebusch H. Point-of-care testing (POCT): current techniques and future perspectives. Trends Analyt Chem 2011;30:887-898.
10. Niemz A, Ferguson TM, Boyle DS. Point-of-care nucleic acid testing for infectious diseases. Trends Biotechnol 2011;29:240-250.
11. Notomi T, Okayama H, Masubuchi H, Yonekawa T, Watanabe K, Amino N, et al. Loop-mediated isothermal amplification of DNA. Nucleic Acids Res 2000;28:E63.
12. Srividya A, Maiti B, Chakraborty A, Chakraborty G. Loop mediated isothermal amplification: a promising tool for screening genetic mutations. Mol Diagn Ther 2019;23:723-733.
13. Zhao Y, Chen F, Li Q, Wang L, Fan C. Isothermal amplification of nucleic acids. Chem Rev 2015;115:12491-12545.
14. Kashir J, Yaqinuddin A. Loop mediated isothermal amplification

(LAMP) assays as a rapid diagnostic for COVID-19. Med Hypotheses 2020;141:109786.
15. Yang Q, Domesle KJ, Ge B. Loop-mediated isothermal amplification for *Salmonella* detection in food and feed: current applications and future directions. Foodborne Pathog Dis 2018;15:309-331.
16. Hara-Kudo Y, Yoshino M, Kojima T, Ikedo M. Loop-mediated isothermal amplification for the rapid detection of *Salmonella*. FEMS Microbiol Lett 2005;253:155-161.
17. Zhang S, Shin J, Shin S, Chung YJ. Development of reverse transcription loop-mediated isothermal amplification assays for point-of-care testing of avian influenza virus subtype H5 and H9. Genomics Inform 2020;18:e40.
18. Fahlen TF, Mathur N, Jones BD. Identification and characterization of mutants with increased expression of hilA, the invasion gene transcriptional activator of *Salmonella* typhimurium. FEMS Immunol Med Microbiol 2000;28:25-35.
19. Schechter LM, Lee CA. AraC/XylS family members, HilC and HilD, directly bind and derepress the *Salmonella* typhimurium hilA promoter. Mol Microbiol 2001;40:1289-1299.
20. Bajaj V, Hwang C, Lee CA. hilA is a novel ompR/toxR family member that activates the expression of *Salmonella* Typhimurium invasion genes. Mol Microbiol 1995;18:715-727.
21. Pathmanathan SG, Cardona-Castro N, Sanchez-Jimenez MM, Correa-Ochoa MM, Puthucheary SD, Thong KL. Simple and rapid detection of *Salmonella* strains by direct PCR amplification of the *hilA* gene. J Med Microbiol 2003;52:773-776.

https://doi.org/10.5808/gi.21048

G&I
Genomics & Informatics

Original article

eISSN 2234-0742
Genomics Inform 2021;19(3):e31
https://doi.org/10.5808/gi.21037

Received: July 9, 2021
Revised: September 6, 2021
Accepted: September 12, 2021

*Corresponding author:
E-mail: sp.medbiochemcu@gmail.com

Comparative genome characterization of *Leptospira interrogans* from mild and severe leptospirosis patients

Songtham Anuntakarun[1], Vorthon Sawaswong[1], Rungrat Jitvaropas[2], Kesmanee Praianantathavorn[3], Witthaya Poomipak[4], Yupin Suputtamongkol[5], Chintana Chirathaworn[6], Sunchai Payungporn[3,7*]

[1]Program in Bioinformatics and Computational Biology, Graduate School, Chulalongkorn University, Bangkok 10330, Thailand
[2]Division of Biochemistry, Department of Preclinical Science, Faculty of Medicine, Thammasat University, Pathum Thani 12120, Thailand
[3]Department of Biochemistry, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand
[4]Research Affairs, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand
[5]Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand
[6]Department of Microbiology, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand
[7]Research Unit of Systems Microbiology, Chulalongkorn University, Bangkok 10330, Thailand

Leptospirosis is a zoonotic disease caused by spirochetes from the genus *Leptospira*. In Thailand, *Leptospira interrogans* is a major cause of leptospirosis. Leptospirosis patients present with a wide range of clinical manifestations from asymptomatic, mild infections to severe illness involving organ failure. For better understanding the difference between *Leptospira* isolates causing mild and severe leptospirosis, illumina sequencing was used to sequence genomic DNA in both serotypes. DNA of *Leptospira* isolated from two patients, one with mild and another with severe symptoms, were included in this study. The paired-end reads were removed adapters and trimmed with Q30 score using Trimmomatic. Trimmed reads were constructed to contigs and scaffolds using SPAdes. Cross-contamination of scaffolds was evaluated by ContEst16s. Prokka tool for bacterial annotation was used to annotate sequences from both *Leptospira* isolates. Predicted amino acid sequences from Prokka were searched in EggNOG and David gene ontology database to characterize gene ontology. In addition, *Leptospira* from mild and severe patients, that passed the criteria e-value < $10e^{-5}$ from blastP against virulence factor database, were used to analyze with Venn diagram. From this study, we found 13 and 12 genes that were unique in the isolates from mild and severe patients, respectively. The 12 genes in the severe isolate might be virulence factor genes that affect disease severity. However, these genes should be validated in further study.

Keywords: genome annotation, leptospirosis, *Leptospira interrogans*, virulence factor genes

Introduction

Leptospirosis is a worldwide zoonotic disease that influences humans and animals worldwide [1]. It is a zoonosis caused by bacteria in the genus *Leptospira*. *Leptospira* can be clustered in three groups including pathogenic, intermediate pathogenic and saprophytic

groups. The various clinical manifestations are caused by the pathogenic and intermediate groups, while the saprophytic group does not cause the disease in humans or animals [2]. Human leptospirosis can be acquired by contact with the urine of infected animals or soil and water contaminated with *Leptospira* [1]. There are two chromosomes in the *Leptospira* species with a cumulative length ranging from 3.9 to 4.6 Mb. This variability in the genome length confers the bacteria with an ability to live within diverse environments and adapt to a wide range of hosts [3]. Approximately 60% of the functional genes that affect the unique pathogenic mechanisms caused by *Leptospira* are unknown [4].

In 2017, the 100K Pathogen Genome Project was established with internationalization coprojects by many countries, including China, South Korea, and Mexico. This project provides various pathogen draft genomes from many areas, and which include human and animal diseases, food, environmental reservoirs of those pathogens and wildlife. Several species such as *Campylobacter*, *Shigella*, *Salmonella*, *Listeria*, *Helicobacter*, and *Vibrio* are currently involved in the project [5]. Virulence genes code for virulence factors that are essential for successful infection and pathogenesis, such as invasion, colonization, adaptation in host environments, immune evasion and tissue damage. Comparison of genomes from microorganisms causing the variety of symptoms provides insight into the mechanisms of microbial infection and pathogenesis. The virulence factor database (VFDB) [6] provides up-to-date information of virulence factor genes from various bacterial pathogens.

In this study, we compared the genomes of *Leptospira* isolated in Thailand from both mild and severe leptospirosis patients. The data provide insight into the genomic characteristics of *Leptospira interrogans*. In addition, virulence factor genes were analyzed using bioinformatics approaches. This research provides information for therapeutic and vaccine development for leptospirosis.

Methods

Isolation of *Leptospira*

Leptospira isolated from human patients in this study were obtained from the Department of Medicine, Faculty of Medicine, Siriraj Hospital, Mahidol University, Bangkok, Thailand. The protocol was approved by the Ethical Committee of the Ministry of Public Health, Royal Government of Thailand. One isolate was from a mild leptospirosis patient, while the other was from a patient presenting with a severe clinical manifestation. Leptospirosis was laboratory confirmed by detecting IgM antibody to *Leptopsira* by indirect immunofluorescent assay and PCR for *lipL32* gene detection. Briefly, the mild case (TH_mild) was a 25-year-old male, admitted to Loei Hospital on 21 August 2001. He presented with three days of fever, headache and myalgia. *Leptospira* detected from his blood culture was identified as Serogroup Pyrogenase. The severe case (TH_severe) was a 59-year-old male admitted to Nakhon Ratchasima Hospital on 2 July 2012. He presented with septic shock and died within 48 h of admission. He had a history of 3 days of fever and developed hypotension, jaundice, acute renal failure and upper gastrointestinal hemorrhage. He had no hemoptysis or acute respiratory distress syndrome.

Library preparation

DNA was extracted from the leptospires grown in EMJH medium using QIAamp DNA mini kit (Qiagen, Valencia, CA, USA) according to the manufacturer's instructions. In the fragmentation step, a Covaris M220 focused-ultrasonicator (Covaris, Brighton, UK), with 20% duty factor, 50 unit of peak incident power (W), and 200 cycles per burst for 150 s, was used to fragment 1 μg of DNA. In the DNA library preparation, the fragmented DNA was prepared based on the TruSeq DNA LT Sample Prep Kit (Illumina, San Diego, CA, USA) following the manufacturer's instructions. Then, AMPure XP beads (Beckman Coulter, Danvers, MA, USA) was used to perform clean up and size selection of the DNA library. The concentration of the DNA library was measured using the KAPA Library Quantification Kit (Kapa Biosystems, Wilmington, MA, USA). The DNA library was diluted to 6 pM. Finally, the diluted DNA library was paired-end sequenced (2 × 150 bp) with the MiSeq platform (Illumina), using MiSeq Reagent Kits V2 (300 cycles) according to the standard protocol.

Quality filter and genome assembly

MiSeq was used to sequence the mild and severe strains of *Leptospira* isolated from the Thai patients. Trimmomatic-0.38 [7] was used to trim and remove low quality reads using default parameter. *De novo* assembly was performed in both strains using SPAdes-3.13.0 [8]. All scaffolds were checked for contamination of 16S rRNA using the ContEST16s database [9]. The Artermis comparison tool (ACT) [10] was used to perform alignment of assembled sequences to a reference genome using *L. interrogans* serovar Lai 56601 as a reference. The DNA sequences were deposited in the Sequence Read Archive data of NCBI server (BioProject PRJNA716760).

Gene prediction and functional annotation

In the gene prediction step, Prokka 1.13.3 [11] was used to predict genes in the mild and severe *Leptospira* genome. Putative protein coding sequences from Prokka were performed in the functional annotation. The integration of annotation data from the EggNOG database version 1.0.3 [12] and the David gene ontology (GO) database [13] represent the function of predicted genes including the

https://doi.org/10.5808/gi.21037

cluster of orthologous groups of proteins (COGs), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway [14], and GO annotation.

Prediction of virulence factor gene

The putative protein coding sequences were searched using blastP with the VFDB. The criteria for the determination of candidate virulence sequences was based on an e-value of $10e^{-5}$. Venn diagram analysis was used to find unique candidate virulence sequences in a specific strain. Lipoprotein prediction in gram-negative bacteria was performed using LipoP 1.0 [15].

Identification of phages in mild and severe *Leptospira* genomes

PHASTER (PHAge Search Tool Enhanced Release) [16] was performed to identify phages in both the mild and severe genomes.

Results

Genome characteristics of mild and severe strain

There was a total of 5,439,790 and 2,162,355 reads with 150 bp paired-end library using mean Phred score (Q) > 30 in mild and severe strain, respectively. The number of scaffolds more than 500 bp are 165 in the mild strain and 309 in the severe strain. The overview of fastq and *de novo* data assembly of mild and severe strains is shown in Table 1. After merging and ordering scaffolds with ACT, there are 3,947 and 297 predicted genes in the final assembly of chromosome 1 (4.70 Mb) and chromosome 2 (0.36 Mb), respectively. In the severe strain, there are 4,373 and 236 predicted genes in the final assembly of chromosome 1 (5.14 Mb) and chromosome 2 (0.37 Mb), respectively. The large variations of the CG content regions in the genome may be caused by being over- or under-fragmented during the library construction. The percentage of GC content in *Leptospira interogans* ranges from 35%–41% [17]. The mild genome had an average GC content of 35%, and the severe genome had an average GC content of 37%.

From COGs analysis of mild and severe strains, the top three categories included function unknown, membrane/envelope biogenesis and signal transduction mechanisms, as indicated in Fig. 1. For the KEGG pathway analysis, the top three pathways included metabolic pathways, biosynthesis of amino acids, and 2-oxocarboxylic metabolism acid, as shown in Supplementary Fig. 1. Functional annotation is the process of collecting information about the function of genes. The GO system [18] was used in this study. There are three distinct categories in GO, namely molecular function, cellular component and biological process. The results of GO analysis given in Supplementary Figs. 2–4 show that the top three molecular functions are sigma factor activity, magnesium ion binding, and structural constituent of ribosome. The top three cellular components are cytoplasm, ribosome, and large ribosomal subunit. The top three biological processes are DNA-templated transcription/initiation, translation, and peptidoglycan biosynthetic process. There is no significant difference between mild and severe strains from COGs, KEGG pathway and GO analysis.

Table 1. Characteristics of mild and severe data and *de novo* assembly

Feature	Mild	Severe
Length (bp)	150	150
Raw reads	5,989,479	2,590,133
Q30 reads	5,439,790	2,162,355
No. of scaffolds	619	1,210
No. of scaffolds (> 500 bp)	165	309
N50	97,013	185,969

Putative virulence factor analysis

A total of 4,244 and 4,699 predicted genes in mild and severe strains, respectively from Prokka were used to identify virulence factor gene with VFDB. The 162 and 161 virulence factor genes were found in mild and severe stains, respectively using blastP with an e-value < $10e^{-5}$. Venn diagram analysis was used to compare virulence factor genes between mild and severe strains. Fig. 2A shows that 12 genes and 10 genes, respectively, of chromosome 1 were found in only the mild strain and only the severe strain. In chromosome 2, one gene was found in the mild strain only and two genes were found in the severe strain only (Fig. 2B). The gene lists that were discovered in only the mild strain included *AfaG-VII, neuA/flmD, rhmA, dapH, yhbX, murB, ahpC, flhB, LA_3103, nuc, PS_PT04340, ipaH2.5*, and *rfaK*". Meanwhile, the gene lists found in only the severe strain consist of *mntB, iga, flgG, proC, kdnB, neuA_1, neuA_2, pyrB, C8J_1334, rfbB, gtf1,* and *hemB*. The description of virulence factor genes is shown in Tables 2 and 3. In Fig. 2C, the regions of virulence factor genes were mapped into chromosomes of mild and severe strains. There are many different regions of virulence factor genes found in mild and severe strains, especially in chromosome 1. In chromosome 2 of the severe strain, the group of virulence factor genes were located in the range of 4.8-5.2 Mb. In addition, nearby virulence factor genes might exhibit co-expression or regulation. However, nearby virulence factor genes will be studied further.

Phage analysis

For phage investigation, prophage sequences in mild and severe strain genomes were identified and annotated using PHASTER. Prophages play an important role in the evolution of the bacterial

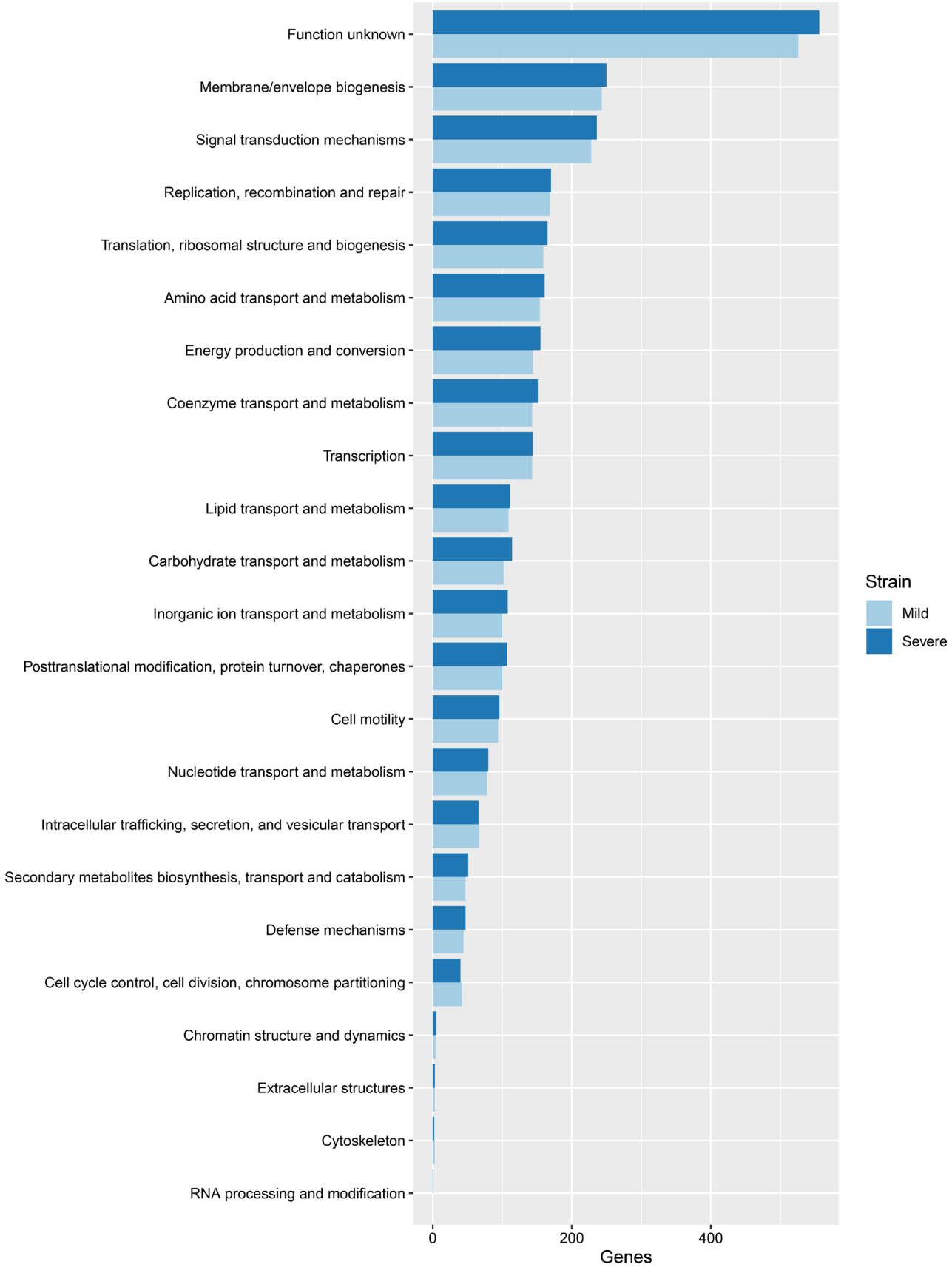

Fig. 1. Comparison of clusters of orthologous groups of proteins between mild and severe strains.

https://doi.org/10.5808/gi.21037

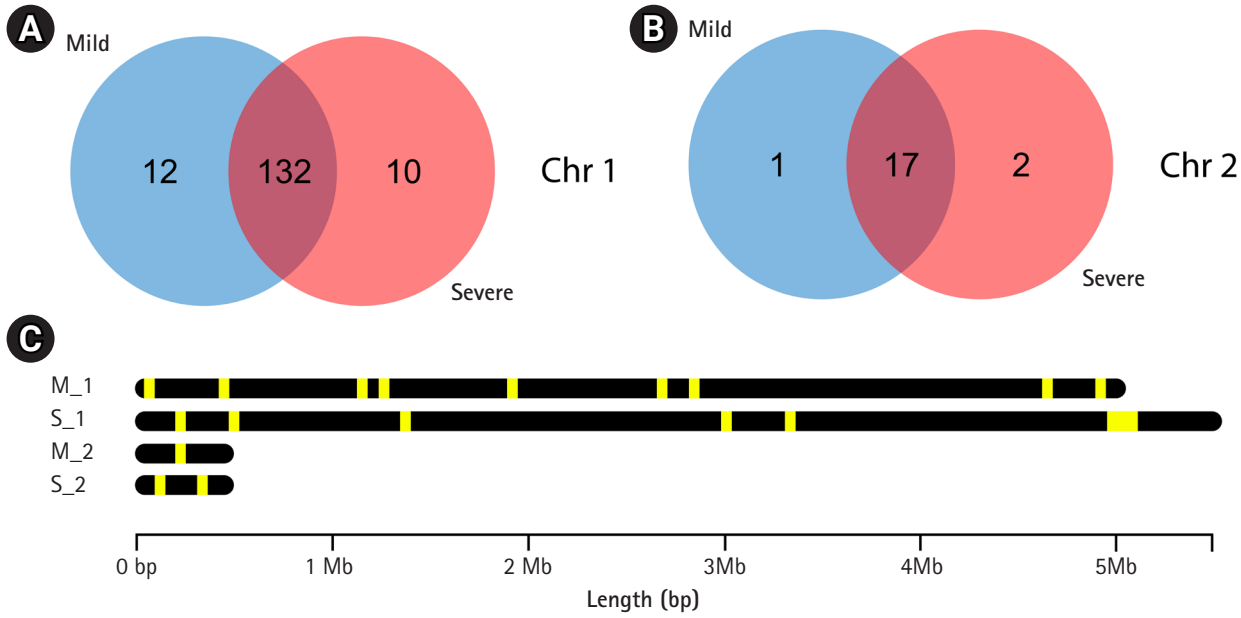

Fig. 2. Comparison of virulence factor genes between mild and severe strains. (A) Venn diagram analysis between mild and severe strains in chromosome 1. (B) Venn diagram analysis between mild and severe strains in chromosome 2. (C) Comparison region of predicted virulence factor genes in each chromosome of both mild and severe strains (M_1: chromosome 1 in mild strain, M_2: chromosome 2 in mild strain, S_1: chromosome 1 in severe strain and S_2 chromosome 2 in severe strain; Yellow stripe in the black bar: region of virulence factor genes).

Table 2. Description of predicted virulence factor genes in mild strains

Gene	Description	TH_mild	FMAS_KW1	FMAS_KW2	FMAS_AW1
AfaG-VII	Afimbrial adhesin	√	X	X	X
neuA/flmD	CMP-N-acetylneuraminic acid synthetase	√	X	X	X
rhmA	2-Keto-3-deoxy-L-rhamnonate aldolase	√	X	X	X
dapH	2,3,4,5-Tetrahydropyridine-2,6-dicarboxylate N-acetyltransferase	√	X	X	X
yhbX	Outer membrane protein YhbX	√	√	√	√
murB	UDP-N-acetylenolpyruvoylglucosamine reductase	√	√	√	√
ahpC	Alkyl hydroperoxide reductase C	√	√	√	√
flhB	Flagellar biosynthetic protein FlhB	√	√	√	√
LA_3103	Fibronectin-binding protein	√	√	√	√
nuc	Thermonuclease	√	X	X	X
PS_PTO4340	Insecticidal toxin protein, putative	√	X	X	X
ipaH2.5	Invasion plasmid antigen	√	X	X	X
rfaK	Alpha 1,2 N-acetylglucosamine transferase	√	X	X	X

host and are commonly found in the bacterial genome [19]. In our results, there is no phage in either mild and severe genomes. However, the size ranges of incomplete phages from 6.9–11.3 kb were detected in both strains. PHAGE_Synech_S_CAM7_NC_031927, PHAGE_Sphing_PAU_NC_019521, PHAGE_Synech_ACG_2014b_NC_027130, PHAGE_Bacill_Finn_NC_020480, PHAGE_Psychr_pOW20_A_NC_020841 and PHAGE_Shigel_Sf6_NC_005344 were found in the mild genome. Moreover, PHAGE_Acinet_Acj9_NC_014663, PHAGE_Bacill_SP_15_NC_031245, PHAGE_Synech_S_CAM7_NC_031927, PHAGE_Sphing_PAU_NC_019521 and PHAGE_Synech_ACG_2014f_NC_026927 were found in the severe genome. Almost all of the incomplete prophages were similar to other *Leptospira* species that contained incomplete phages with sizes ranging from 4.1 to 13.8 kb [20]. However, PHAGE_Acinet_Acj9_NC_014663 which was found in the severe strain, is the one multiple-drug resistant species [21].

Plasmid analysis

Additional investigation of plasmids in the TH_mild and TH_severe strains isolated form Thai patients found that both strains con-

Table 3. Description of predicted virulence factor genes in severe strains

Gene	Description	TH_severe	Taganrog-2018	SK-1
mntB	Manganese transport system membrane protein MntB	√	X	X
iga	IgA-specific serine endopeptidase	√	X	X
flgG	Flagellar basal-body rod protein FlgG	√	√	√
proC	Pyrroline-5-carboxylate reductase	√	√	√
kdnB	3-Deoxy-alpha-D-manno-octulosonate 8-oxidase	√	X	X
neuA_1	N-Acylneuraminate cytidylyltransferase	√	√	√
neuA_2	CMP-N,N′-diacetyllegionaminic acid synthase	√	√	√
pyrB	Aspartate carbamoyltransferase catalytic subunit	√	√	√
C8J_1334	Hypothetical protein	√	X	X
rfbB	dTDP-glucose 4,6-dehydratase	√	√	√
gtf1	Glycosyltransferase Gtf1	√	X	√
hemB	Delta-aminolevulinic acid dehydratase	√	X	√

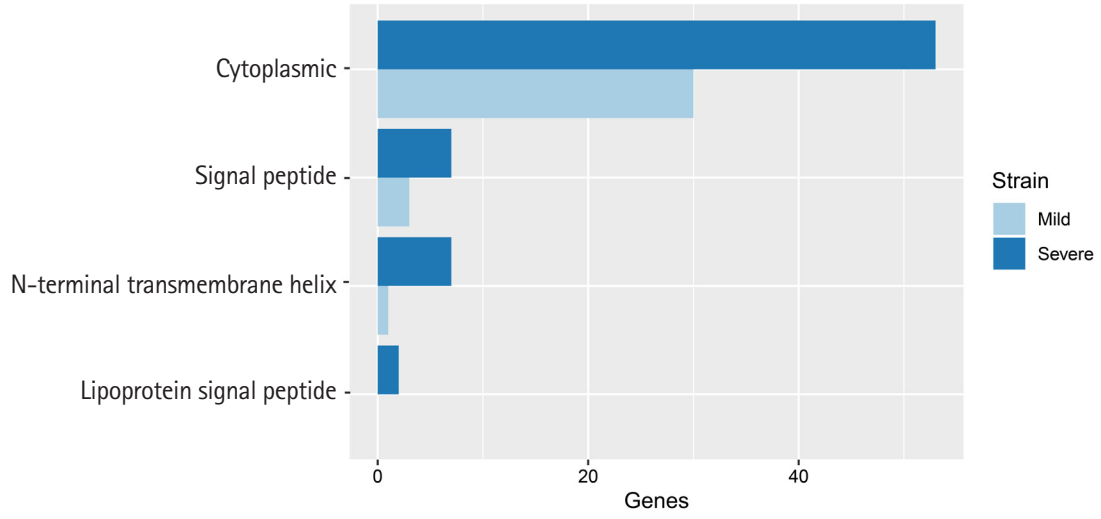

Fig. 3. Comparison of lipoprotein predicted genes between mild and severe strains. The class of prediction from LipoP 1.0 was separated into four groups including cytoplasmic, signal peptide, N-terminal transmembrane helix, and lipoprotein signal peptide.

tained *L. interrogans* serovar Canicola strain Gui44 plasmids (pGui1 and pGui2), *L. interrogans* serovar Linhai str. 56609 plasmids (lcp1 and lcp2) and *L. interrogans* serovar Manilae strain UP-MMC-NI-ID LP plasmid pLIMLP1. Interestingly, the *L. borgpetersenii* serovar Ballum strain 56604 plasmid lbp2 was found only in the TH_severe strain, implying that this plasmid might be associated with the pathogenesis or severity of *Leptospira*.

Lipoprotein analysis

Lipoproteins of bacteria are a set of membrane proteins. There are many functions in the role of pathogenesis and host-pathogen interaction, especially the functions of surface adhesion and initiation of inflammatory processes through translocation of virulence factors in the host cytoplasm [22]. In our study, we used 32 and 67 unique genes in mild and severe strains, respectively, from eggNOG annotation to predict lipoprotein signal peptide using LipoP 1.0. This software can discriminate between lipoprotein and other signal peptides. The prediction was separated into four groups, including cytoplasmic, signal peptide, N-terminal transmembrane helix and lipoprotein signal peptide. In addition, this result in Fig. 3 showed that a protein sequence was assigned to a lipoprotein signal peptide found in the severe strain only.

Discussion

LipoP1.0 predicts lipoproteins and discriminates between lipoprotein signal peptides and other signal peptides in gram-negative bacteria using a Hidden Markov model (HMM). They report that the

https://doi.org/10.5808/gi.21037

accuracy performance of prediction in gram-negative bacteria is 96.8%. Another lipoprotein prediction is called LIPOPREDICT which predicts signal peptides using a support vector machine [23]. The accuracy of this tool is 97%. Support vector machine has a similar performance to HMM. We would like to use LIPOPREDICT to predict lipoproteins in our genomes. Unfortunately, LIPOPREDICT is not available so far.

The genome characteristics of mild and severe strains in this study were compared with the *L. interrogans* genomes previously reported from Russia (strain Taganrog-2018) [24], Sri Lanka (strain FMAS_KW1, FMAS_KW2, and FMAS_AW1) [25], and Saint Kitts (strain SK-1) [26]. The *Leptospira* strains from Russia and Saint Kitts were classified as severe strains. The result of genome characteristics comparison was represented in Supplementary Table 1. In addition, the virulence factor genes were compared among our strains and other strains as shown in Tables 2 and 3. The result revealed that *yhbX, murB, ahpC, flhB*, and *LA_3103* genes were found in *Leptospira interrogans* stains FMAS_KW1, FMAS_KW2, and FMAS_AW1 similar to those found in our mild strain. Moreover, *flgG, proC, neuA_1, neuA_2, pyrB* and *rfbB* genes were also found in *Leptospira interrogans* strains in this study, Taganrog-2018 and SK-1 isolated from severe cases. However, *mntB, iga, kdnB*, and *C8J_1334* genes were found only in our severe strain.

IgA-specific serine endopeptidase or IgA protease is secreted by gram-negative bacteria. This enzyme plays an important role in human antibodies. They can specifically cleave IgA, which provides an antibody for defending the mucosal surface [27]. The inactivation of IgA protease might have the potential to reduce bacterial colonization on mucosal surfaces [28]. Aminoglycosides are broad-spectrum antibiotics that are used in gram-negative and gram-positive organisms [29]. Many reports showed that *Leptospira* are sensitive to aminoglycosides [30,31]. dTDP-glucose-4,6-dehydratase genes were related in a gene cluster in an aminoglycoside antibiotics producer [32].

In bacteria, metal ions play an important role in survival in their host environment. Bacteria which cannot maintain proper homeostasis of metals are less virulent [33]. In many biological processes metal ions are needed as metalloprotein materials, which function as enzyme cofactors or structural elements. Manganese (Mn) is one important example. Many bacteria require manganese with eukaryotic host cells to form pathogenic or symbiotic interactions [34]. Currently, there is evidence that the invading microbe uses Mn as the main micronutrient to avoid the effects of host-mediated oxidative stress and thus plays a significant role in the human host's tolerance to pathogenic bacteria [35]. In our study, we found manganese transport system membrane protein MntB (mntB) in the severe *Leptospira* strain. This gene encodes transmembrane protein. The *mntB* gene is part of the ABC transporter system for manganese that mediates the movement of various substrates from microbes to humans across different biological membranes [36]. The lack of the *mntB* gene might affect the homeostasis of metal in bacteria that are less virulent.

The flagellum consists of three main sections, including a flagellar filament, a hook complex, and a basal body in both gram-negative and gram-positive bacteria. There are many genes related to flagellar biosynthetic protein such as *flhA, flhB* [37,38]. The results showed that *flhB* was found in the mild strain. This result came from blastP with a VFDB. However, *flhB* was also found in the severe strain from Prokka annotation. In this case, some genes in the mild strain are similar to the *flhB* gene in other species of bacteria in the VFDB.

In this study, two strains of *Leptospira* spp. isolated from mild and severe Thai patients were compared. Our analysis showed 3,947 and 297 predicted genes in the final assembly of chromosome 1 (4.70 Mb) and chromosome 2 (0.36 Mb), respectively, in the mild strain. In addition, there are 4,373 and 236 predicted genes in the final assembly of chromosome 1 (5.14 Mb) and chromosome 2 (0.37 Mb), respectively, in the severe strain. The difference of virulence factor genes was found in both strains. Our results focus on predicting virulence factor genes in the severe strain that is not found in the mild strain. The virulence factor genes in the severe strain are only related to host immune response, and survival in the host environment might be the vital virulence factor genes. However, these genes should be validated in further study.

ORCID

Songtham Anuntakarun: https://orcid.org/0000-0002-6849-0523
Vorthon Sawaswong: https://orcid.org/0000-0003-2805-6690
Rungrat Jitvaropas: https://orcid.org/0000-0001-7555-0048
Kesmanee Praianantathavorn: https://orcid.org/0000-0002-5368-3015
Witthaya Poomipak: https://orcid.org/0000-0002-3282-7219
Yupin Suputtamongkol: https://orcid.org/0000-0001-7324-1698
Chintana Chirathaworn: https://orcid.org/0000-0002-2131-1815
Sunchai Payungporn: https://orcid.org/0000-0003-2668-110X

Authors' Contribution

Conceptualization: SP, SA. Data curation: CC, YS. Formal analysis: SA, VS. Funding acquisition: SP. Methodology: SA, KP, WP. Writing - original draft: SA. Writing - review & editing: SP, CC, RJ.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

The authors would like to acknowledge the supports from Graduate School, Faculty of Science and Faculty of Medicine, Chulalongkorn University (the 100th Anniversary Chulalongkorn University Fund for Doctoral Scholarship; the 90th Anniversary of Chulalongkorn University Ratchadaphiseksomphot Endowment Fund).

Supplementary Materials

Supplementary data can be found with this article online at http://www.genominfo.org.

References

1. Bharti AR, Nally JE, Ricaldi JN, Matthias MA, Diaz MM, Lovett MA, et al. Leptospirosis: a zoonotic disease of global importance. Lancet Infect Dis 2003;3:757-771.
2. Adler B, de la Pena Moctezuma A. *Leptospira* and leptospirosis. Vet Microbiol 2010;140:287-296.
3. Picardeau M, Bulach DM, Bouchier C, Zuerner RL, Zidane N, Wilson PJ, et al. Genome sequence of the saprophyte *Leptospira biflexa* provides insights into the evolution of *Leptospira* and the pathogenesis of leptospirosis. PLoS One 2008;3:e1607.
4. Xu Y, Zhu Y, Wang Y, Chang YF, Zhang Y, Jiang X, et al. Whole genome sequencing revealed host adaptation-focused genomic plasticity of pathogenic Leptospira. Sci Rep 2016;6:20020.
5. Weimer BC. 100K Pathogen Genome Project. Genome Announc 2017;5:e00594-17.
6. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, et al. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res 2005;33:D325-D328.
7. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114-2120.
8. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19:455-477.
9. Lee I, Chalita M, Ha SM, Na SI, Yoon SH, Chun J. ContEst16S: an algorithm that identifies contaminated prokaryotic genomes using 16S RNA gene sequences. Int J Syst Evol Microbiol 2017;67:2053-2057.
10. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. ACT: the Artemis Comparison Tool. Bioinformatics 2005;21:3422-3423.
11. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;30:2068-2069.
12. Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 2019;47:D309-D314.
13. Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 2007;8:R183.
14. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017;45:D353-D361.
15. Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H, Krogh A. Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci 2003;12:1652-1662.
16. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res 2016;44:W16-W21.
17. Farrar J, Hotez P, Junghanss T, Kang G, Lalloo D, White N, et al. Manson's Tropical Diseases. Oxford: Saunders, 2013.
18. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nat Genet 2000;25:25-29.
19. Fortier LC, Sekulovic O. Importance of prophages to evolution and virulence of bacterial pathogens. Virulence 2013;4:354-365.
20. Kurilung A, Keeratipusana C, Suriyaphol P, Hampson DJ, Prapasarakul N. Correction to: Genomic analysis of Leptospira interrogans serovar Paidjan and Dadas isolates from carrier dogs and comparative genomic analysis to detect genes under positive selection. BMC Genomics 2019;20:246.
21. Turner D, Ackermann HW, Kropinski AM, Lavigne R, Sutton JM, Reynolds DM. Comparative analysis of 37 Acinetobacter bacteriophages. Viruses 2017;10:5.
22. Kovacs-Simon A, Titball RW, Michell SL. Lipoproteins of bacterial pathogens. Infect Immun 2011;79:548-561.
23. Kumari SR, Kadam K, Badwaik R, Jayaraman VK. LIPOPREDICT: bacterial lipoprotein prediction server. Bioinformation 2012;8:394-398.
24. Samoilov AE, Stoyanova NA, Tokarevich NK, Evengard B, Zueva EV, Panferova YA, et al. Lethal outcome of leptospirosis in southern Russia: characterization of Leptospira interrogans isolated from a deceased teenager. Int J Environ Res Public Health 2020;17:4238.
25. Senevirathna I, Jayasundara D, Lefler JP, Chaiboonm KL, War-

https://doi.org/10.5808/gi.21037

nasekara J, Agampodi S, et al. Complete genome sequence of *Leptospira interrogans* strains FMAS_KW1, FMAS_KW2 and FMAS_AW1 isolated from leptospirosis patients from Karawanalla and Awissawella, Sri Lanka. J Genomics 2020;8:49-52.
26. Llanes A, Prakoso D, Restrepo CM, Rajeev S. Complete genome sequence of a virulent *Leptospira interrogans* serovar Copenhageni strain, assembled with a combination of nanopore and Illumina reads. Microbiol Resour Announc 2020;9:e00200-20.
27. Plaut AG. The IgA1 proteases of pathogenic bacteria. Annu Rev Microbiol 1983;37:603-622.
28. Mistry D, Stockley RA. IgA1 protease. Int J Biochem Cell Biol 2006;38:1244-1248.
29. Krause KM, Serio AW, Kane TR, Connolly LE. Aminoglycosides: an overview. Cold Spring Harb Perspect Med 2016;6:a027029.
30. Faine S, Adler B, Bolin C, Perolat P. "*Leptospira*" and leptospirosis. 2nd ed. Melbourne: MediSci, 1999.
31. Kobayashi Y. Clinical observation and treatment of leptospirosis. J Infect Chemother 2001;7:59-68.
32. Du Y, Li T, Wang YG, Xia H. Identification and functional analysis of dTDP-glucose-4,6-dehydratase gene and its linked gene cluster in an aminoglycoside antibiotics producer of *Streptomyces tenebrarius* H6. Curr Microbiol 2004;49:99-107.
33. Hood MI, Skaar EP. Nutritional immunity: transition metals at the pathogen-host interface. Nat Rev Microbiol 2012;10:525-537.
34. Zeinert R, Martinez E, Schmitz J, Senn K, Usman B, Anantharaman V, et al. Structure-function analysis of manganese exporter proteins across bacteria. J Biol Chem 2018;293:5715-5730.
35. Lisher JP, Giedroc DP. Manganese acquisition and homeostasis at the host-pathogen interface. Front Cell Infect Microbiol 2013;3:91.
36. Saier MH Jr. Molecular phylogeny as a basis for the classification of transport proteins from bacteria, archaea and eukarya. Adv Microb Physiol 1998;40:81-136.
37. Lambert A, Picardeau M, Haake DA, Sermswan RW, Srikram A, Adler B, et al. FlaA proteins in Leptospira interrogans are essential for motility and virulence but are not required for formation of the flagellum sheath. Infect Immun 2012;80:2019-2025.
38. Cheng C, Wang H, Ma T, Han X, Yang Y, Sun J, et al. Flagellar basal body structural proteins FlhB, FliM, and FliY are required for flagellar-associated protein expression in *Listeria monocytogenes*. Front Microbiol 2018;9:208.

G&I
Genomics & Informatics

Original article

eISSN 2234-0742
Genomics Inform 2021;19(3):e32
https://doi.org/10.5808/gi.21039

Received: July 29, 2021
Revised: August 27, 2021
Accepted: September 6, 2021

*Corresponding author:
E-mail: furim@hanmail.net

#These authors contributed equally to this work.

Draft genome of *Semisulcospira libertina*, a species of freshwater snail

Jeong-An Gim[1#], Kyung-Wan Baek[2,3#], Young-Sool Hah[4], Ho Jin Choo[5], Ji-Seok Kim[2], Jun-Il Yoo[3*]

[1]Medical Science Research Center, Korea University Guro Hospital, Korea University College of Medicine, Seoul 08308, Korea
[2]Department of Physical Education, Gyeongsang National University, Jinju 52727, Korea
[3]Department of Orthopaedic Surgery, Gyeongsang National University Hospital, Jinju 52727, Korea
[4]Biomedical Research Institute, Gyeongsang National University Hospital, Jinju 52727, Korea
[5]South Korea 4H Association, Seoul 05269, Korea

Semisulcospira libertina, a species of freshwater snail, is widespread in East Asia. It is important as a food source. Additionally, it is a vector of clonorchiasis, paragonimiasis, metagonimiasis, and other parasites. Although *S. libertina* has ecological, commercial, and clinical importance, its whole-genome has not been reported yet. Here, we revealed the genome of *S. libertina* through *de novo* assembly. We assembled the whole-genome of *S. libertina* and determined its transcriptome for the first time using Illumina NovaSeq 6000 platform. According to the *k*-mer analysis, the genome size of *S. libertina* was estimated to be 3.04 Gb. Using RepeatMasker, a total of 53.68% of repeats were identified in the genome assembly. Genome data of *S. libertina* reported in this study will be useful for identification and conservation of *S. libertina* in East Asia.

Keywords: *de novo* assembly, draft genome, *Semisulcospisa libertina*

Introduction

As a species of freshwater snail, *Semisulcospira libertina* is widespread in East Asia and it is an important food source. It is also a vector of clonorchiasis, paragonimiasis, metagonimiasis, and other parasites. It inhabits clean running waters or pools such as drainage ditches, slow flowing rivers, rice paddies, and streams. The phylogeography of *S. libertina* in Taiwan has been revealed in two studies [1,2] by its mitochondrial cytochrome c oxidase subunit I (COI) sequences. *S. libertina* belongs to genus *Semisulcospira*, a well-known group of freshwater snails. *S. libertina* can be readily identified by its nuclear seqeunce (28S ribosomal RNA) and mitochondrial sequence (16S ribosomal RNA) [3]. In genus *Semisulcospira*, mitochondrial genomes of *S. libertina* [4], *S. coreana* [5], and *S. gottsei* [6] have been reported. In Gastropoda, mitochondrial genome studies have been performed to classify species until now, as well as genomes were revealed in some species. The genome of *Biomphalaria glabrata*, a freshwater snail, has been reported [7]. Genomes of owl limpet (*Lottia gigantea*) [8] and abalones (*Haliotis discus hannai*) [9] have also been revealed. However, no study has reported whole-genome of *Semisulcospira* genus. A draft genome of *Radix auricularia* (big-ear Radix) [10] and a genome of *Conus tribblei* [11] are cases of genome sequencing in Gastropoda.

S. libertina has ecological, commercial, and clinical importance [12,13], thus whole-genome data of *S. libertina* could be of great help in many ways. In this study, we sequenced

the whole-genome and transcriptome of *S. libertina* for the first time using Illumina NovaSeq 6000 platform. To enhance the accuracy of gene prediction, we integrated *S. libertina* transcriptome data with gene set annotation for the assembled genome. Our genomic data could provide basic knowledge for understanding genomic features of *S. libertina*. They could be used for further comparative, systemic, and functional genomic studies of freshwater snails.

Methods

Sample collection and nucleic acid extraction

Specimens of healthy *S. libertina* were collected from the upstream of Bukhan River basin, South Korea (37°47'32.3"N, 127°31'49.8"E) in June 2019. Morphometric characteristics such as shell length (20–30 mm) and weight (5–6 g) of collected *S. libertina* samples were determined. The samples were stored in a −80°C freezer. Freshest individuals (five for DNA and five for RNA) with the best DNA or RNA quality were studied. Genomic DNAs were extracted from muscle tissues using DNeasy Blood & Tissue Kits (Qiagen, Hilden, Germany). RNAs were extracted using Trizol reagent (Invitrogen, Carlsbad, CA, USA). The quality of RNA was comfirmed based on 28S/18S ratio. RNA integrity number (RIN) of the extracted RNA was determiend using a Tecan F-200 and an Agilent Bioanalyzer 2100 system (Agilent, Santa Clara, CA, USA). All RNAs extracted from samples had RIN values of 6.5–7.0. One sample with the highest quality among five DNA or RNA samples was used for sequencing.

Sequencing library construction

To construct sequencing library, high molecular weight genomic DNAs were sheared to ~500 bp using a Covaris S2 Ultrasonicator system. All DNA libraries for sequencing were constructed following Illumina's instruction. To check the quality of the library constructed, the size of the library was detemrined with a 2200 TapeStation (Agilent). Normalized libraries were diluted with hybridization buffer. Clusters of each library were then made with a cBot system and a HiSeq Rapid Duo cBot Sample Loading Kit (Illumina, San Diego, CA, USA). Pair-end libraries were prepared following the manufacturer's guideline (Illumina). Final library products were sequenced on an Illumina NovaSeq 6000 platform using HiSeq Rapid Paired End Cluster Kit v2 and SBS Kit V2 for 100 PE sequencing (Illumina). Raw fastq sequences are available under BioProject ID PRJNA659426.

Filtering raw sequences for *de novo* assembly

To maintain quality of sequences, raw reads were filtered to remove the following: (1) reads presented with letter N (ambiguous bases) or poly-A motif; (2) reads with low-quality bases (below base quality 7) from the 549 bp insert size library; (3) reads with adapter contamination; (4) reads with small sizes of inserts in which read 1 and read 2 overlapped for more than 10 bp (only 10% mismatch allowed); (5) PCR duplicates (reads were considered duplicates when read 1 and read 2 of two pair-end reads were identical).

De novo assembly of the *S. libertina* genome

K-mer size of 17-bp was estimated using SOAPec v2.01, and the best k was 77. The genome size was calculated using the following formula: genome size = total number of *k*-mer/*k*-mer depth. The size of the *S. libertina* genome was estimated to be 3.04 Gb. The genome was then assembled using qualified reads from the pair-end libraries. *De novo* assembly invovled contig construction followed by scaffolding and gap closure. In the step of contig construction, a short insert library (429 bp) was used to construct a de Brujin graph using SOAPdenovo v2.04 with default parameters [14]. All erroneous data derived from clip tips, bubbles, and connection with low coverage were eliminated. All qualified reads were then realigned with contig sequences. Reads were mapped with bowtie2 v2.2.5 using end-to-end mode and default options. Mapping was perforemd with samtools v1.2.1 and bedtools v2.26. We used benchmarking universal single-copy orthologs software (BUSCO; v2.0) to assess the genome completeness [15].

Identification of repeat sequences

To identify repeat sequences in the genome of *S. libertina*, the follwoing two approaches were applied: (1) a homology-based approach; and (2) a *de novo*-based approach. Identification of homology-based repeat sequences was performed with RepeatMasker (v4.0.9) using Repbase libraries (2019, volume 19, issue 1) containing identified repeat sequences [16].

Identification of *de novo*-based repeat sequences was then finished with RepeatModeler v1.0.8 [16]. Simple sequence repeats (SSRs) were identified using perl script of SSR identification tool (SSRIT; ftp://ftp.gramene.org/pub/gramene/ archives/software/scripts/ssr.pl). SSR target primer pairs were designed with flanking sequences of SSR using Primer 3 program (v0.4.0) [17]. These primers met the following criteria: having GC content > 50%, annealing temperature range at 55–62°C, and primer length of 18–26 bp in size.

Prediction of noncoding RNAs

From *de novo* assembled *S. libertina* genome, four types of noncoding RNA (ncRNA; miRNAs, tRNAs, rRNAs, and snRNAs) were identified by searching databases as follows, tRNAscan-SE with default setting was applied to search for definite tRNA positions [18]. To detect snRNAs and miRNAs, INFERNAL v1.1.1 was used to search for putative sequences with Rfam database (release 9.1)

https://doi.org/10.5808/gi.21039

[19]. For rRNA predictions in the *S. libertina* genome, BLAST (v2.2.29+) homology search was performed [20].

Transcriptome sequencing

For RNA sequencing, cDNA libraries were constructed. mRNA was enriched with oligo-dT attached magnetic beads from total RNA (2 mg). Purified mRNAs were sheared into short fragments and synthesized into double-stranded cDNAs by reverse-transcription immediately. Synthesized cDNAs were subjected to end-repair, poly-A addition, and ligations with adaptors provided by a TruSeqRNA sample prep Kit (Illumina). Modified mRNA fragments were separated on bluepippin 2% agarose gel cassette. Suitable fragments were automatically purified and used as templates for PCR amplification. Final products were 400–500 bp in length and evaluated with an Agilent High Sensitivity DNA Kit (Agilent) on an Agilent Bioanalyzer 2100 system. Subsequently, the constructed libraries were sequenced using an Illumina HiSeq 2500 sequencer (Illumina). All processes were conducted by TheragenETEX Bio Institute (Suwon, Korea).

Gene prediction and annotation

For the annotation of *S. libertina* genome, a combination of evidence-based gene prediction (RNA-sequencing [RNA-seq] and proteins) and *ab initio* gene prediction was used. First, transcript alignment was performed with STAR v2.7.0a using a set of gene model annotations [21]. From RNA-seq data, clean reads with average quality scores of higher than Q30 were aligned from all libraries and used for gene prediction using GeneMark-ET v4.29 [22]. Next, homologous proteins of other species were aligned to the genome using TBlastN v2.2.29+ with an E-value cutoff of 1E–5. Aligned protein sequences were used for the prediction of gene regions using Exonerate v2.2.0 with default parameters [20]. A final gene set of *S. libertina* was produced with AUGUSTUS v3.2.1 using default settings [23]. Gene functions were assigned according to the best alignment attained using BLASTP against UniProt database (Last modified in January 17, 2019), NCBI nr (accecced in June 28, 2019; E-value cutoff of 1E–5), and InterProScan v5.17 [24,25].

Visualization and phylogenetic analysis

For visualization, we used R v3.6.1 and RStudio v1.2.5019 (https://cran.r-project.org/). For heatmap drawing, we used "pheatmap v1.0.10" and "heatmap3 v1.1.6" packages. From whole-mitochondrial genome and COI regions of mitochondrial DNA, the maximum likelihood tree was obtained with a Tamura-Nei model using MEGA-X v10.1.4 [26,27]. Mitochondrial DNA sequences of related species were retrieved from GenBank. Accession numbers were indicated in dendrograms.

Results

De novo assambly of *S. libertina*

A genomic DNA sample of *S. libertina* was used to construct short-insert paired end libraries. Paired end sequencing of 429 bp insert libraries generated a total of 60.99 Gb sequence data with an Illumina NovaSeq 6000 platform. Based on *k*-mer analysis, the genome size of *S. libertina* was estimated to be 3.04 Gb (3,037,193,258 bp) at a *k*-mer size of 17. The *k*-mer frequency distribution had two peaks. This is because the heterozygosity of the *S. libertina* genome is relatively high [28]. Sequence reads from paired end and mate were assembled, and gaps in scaffolds were subsequently filled with Illumina reads using GapCloser v1.12 [14]. Characteristics of the assembled genome are listed in Supplementary Table 1. The N50 size was 2,788. The total number of contigs was 748,492. Raw sequence data were deposited to NCBI SRA (PRJNA659426). Benchmarking was performed by universal single-copy orthologs software (BUSCO; v2.0) to assess the genome completeness [15]. Our assembly covered 23.0% of core genes, with 225 genes being complete genes (Supplementary Table 2).

Gene prediction and annotation

Gene prediction and structural-annotation were carried out using homology-based search. Determination of gene set was performed using transcriptome data. First, we sought to comprehensively describe ncRNA to build better coding gene models. By homology-based Blast search, a total of 935 rRNA copies were matched with 105,942 bp, accounting for 0.01% of the genome. In addition, 572 tRNA copies were estimated using tRNAscan-SEtool [18]. Using INFERNAL [19], miRNAs with 109,716 copies (9,270,754 bp) and snRNAs with 3,797 copies (426,539 bp) were found.

A total of 61,610 gene models were then predicted. The average length of genes was calculated to be 424 bp. Gene annotation databases were used to annotate gene models, find protein sequence, and search for biological functions of annotated genes. Among 61,610 gene models, 39,949 were annotated genes. A total of 10,065, 19,659, and 37,333 genes were producted hits with UniProt, NCBI nonredundant, and InterProScan databases, respectively (Table 1). Each analysis was performed under default options.

Repeat sequences

Repeat composition of the *S. libertina* genome was then investigated. We used homology and *de novo*-based approaches first. We then combined these two approaches. Using RepeatMasker, a total of 53.68% of repeats were identified in the genome. More than half of total repeat length was filled with unclassified repeats, accounting for 34.68% of the genome. DNA transposons accounted for 7.48%

Table 1. Results of gene prediction for the genome of *Semisulcospira libertina*

Parameter	Value
Total No. of gene models predicted	61,610
Annotated gene	39,949
Uniprot	10,065
NCBI nonredundant	19,659
InterProScan	37,333
Average gene length (bp)	424
Average of GC content (%)	53.68

Table 3. Summary of simple sequence repeats distribution in the genome of *Semisulcospira libertina*

Repeat type	Frequency	Frequency per million
2	512,774	364.97
3	132,734	94.47
4	86,955	61.89
5	14,883	10.59
6	1,590	1.13
7	241	0.17
8	374	0.27
9	259	0.18
10	247	0.18

of the genome. Most sequences of retrotransposons consisted of long interspersed nuclear elements (9.54%) and long terminal repeat elements (5.58%), whereas short interspersed nuclear elements (0.97%) were present at low proportions (Table 2).

We also discovered features of SSRs to provide clues for polymorphic information of other species of the genus *Semisulcospira* and molecular markers. In the genome, a total of 35,610 copies of dinucleotide repeats were detected whereas the copy number of each hexa to deca-nucoeotide repeat was <70 (Table 3). On average, a total of 512,774 dinucleotide repeats were detected and 364.97 dinucleotide repeats were detected per million basepairs. Among dinucleotide repeats, CA had the highest frequency (6,494 copies) whereas CG had the lowest frequency (656 copies). Based on these SSR data, we predicted a total of 750,057 primer sets for SSR targets that could be used for polymorphism screening across congener species of *S. libertina*.

Comparative analysis with related species

Four genomes of similar species, owl limpet, air-breathing freshwater snail, and oyster (owl limpet, *Lottia gigantea;* air-breathing freshwater snail, *Biomphalaria glabrata*; oyster, *Crassostrea gigas*) were compared. Based on PFAM database, we compared the copy number of shell formation related genes in each genome [29,30]. In the

Table 2. Number, length, and proportion of repetitive elements in the genome of *Semisulcospira libertina*

Type	No. of Elements	Length (bp)	% in genome
Retrotransposons	1,190,727	224,533,654	15.98
SINEs	102,400	13,600,944	0.97
LINEs	784,492	134,091,004	9.54
LTR elements	303,708	78,375,018	5.58
Retroposon	127	5,771	0.00
DNA transposons	596,425	105,094,630	7.48
DNA	507,157	80,734,772	5.75
RC	89,191	24,777,502	1.76
Other	77	7,150	0.00
Inserted sequence	9	437	0.00
Segmental duplication	3	134	0.00
Unclassified	3,711,311	487,237,955	34.68
Small RNA	3505	444,780	0.03
Satellites	6767	901,035	0.06
Simple repeats	847777	43,050,705	3.06
Low complexity	79335	4,196,255	0.30
Total		832,215,362	59.23

heatmap, the number of orthologous genes in each genome was depicted for 25 genes (Table 4, Fig. 1). Shell formation-related genes were retrived from previous studies [29,30]. In these four genomes, the mostly detected gene was indicated by a 'Top' row bar. A total of 25 genes used to depict heatmap and phylogenetic tree from four class (MT, metabolic transcripts; PI, protease inhibitors; SF, shell formation; SM, small matrix proteins; and TP, transmembrane proteins) were indicated by a second row bar.

We also provided a table and heatmap presenting the copy number of orthologous genes in each genome (Supplementary Table 3). Fig. 2 provides enriched PFAM domains identified as copy number. In the genome of *C. gigas*, PFAM domains were overrepresented. In the genome of *S. libertina*, domain signals from PFAM had weaker patterns than in genomes of other species. Therefore, the genome of *S. libertina* was distinctively divided into genomes of other three species.

We also provided a maximum likelihood tree for whole-mitochondrial genome (Fig. 3A) and COI regions of mitochondrial DNA (Fig. 3B). The phylogenetic tree shown in Fig. 3 reflects the relationshp of PFAM domains (Fig. 2). Dendrograms were derived from mitochondrial genome sequences obtained from GenBank database. In the phylogenetic tree of the whole-mitochondrial genome, *S. coreana* and *Turritella bacillum* were grouped (Fig. 3A). However, for COI regions, *S. libertina* and *S. coreana* were grouped as expected (Fig. 3B). As expected, *C. gigas* and *B. glabrata* were outgrouped with *S. libertina* in both analyses (Fig. 3).

https://doi.org/10.5808/gi.21039

Table 4. Shell formation related genes (ID and description were obtained from PFAM; species with the highest copy number in four genomes is indicated the in top column)

ID	Description	Highest copy number
Shell formation proteins [30]		
PF00245	Alkaline phosphatase	*S_libertina*
PF00262	Calreticulin	*B_glabrata and C_gigas*
PF03142	Chitin synthase	*C_gigas*
PF14704	Dermatopontin	*C_gigas*
PF00264	Tyrosinase	*S_libertina*
Metabolic transcripts [29]		
PF00067	Cytochrome P450	*C_gigas*
PF00151	Lipase	*L_gigantea*
PF13469	Sulfotransferase family	*C_gigas*
Protease inhibitors [29]		
PF00050	Kazal-type serine protease inhibitor domain	*C_gigas*
PF07648	Kazal-type serine protease inhibitor domain	*C_gigas*
Small matrix proteins [29]		
PF00057	Low-density lipoprotein receptor domain class A	*C_gigas*
PF00058	Low-density lipoprotein receptor repeat class B	*C_gigas*
PF00059	Lectin C-type domain	*C_gigas*
PF00090	Thrombospondin type 1 domain	*C_gigas*
PF01607	Chitin binding Peritrophin-A domain	*C_gigas*
PF02412	Thrombospondin type 3 repeat	*C_gigas*
PF03067	Chitin binding domain	*C_gigas*
PF07645	Calcium-binding EGF domain	*B_glabrata*
PF08976	EF-hand domain	*C_gigas*
PF13405	EF-hand domain	*C_gigas*
PF13499	EF-hand domain pair	*C_gigas*
PF13833	EF-hand domain pair	*C_gigas*
Transmembrane proteins [29]		
PF01146	Caveolin	*C_gigas*
PF05478	Prominin	*C_gigas*
PF14878	Death-like domain of SPT6	*B_glabrata*

Discussion

The genome of *S. libertina* could provide insights into freshwater shellfish biology such as extraction of useful components and shell body plan. Next-generation sequencing technologies have greatly reduced the cost of whole-genome sequencing. A huge amount of sequencing data have been accumulated and utilized to study substances such as venom and druggable tartgets. However, in comparison with vertebrate genome studies, freshwater snail genome study is still at its infancy. We tried to provide a source for genomics of freshwater snails. Because of its large genome size, we provided a draft genome in this study. Our draft genome has relatively lower sequencing depth ($< 20\times$). Therefore, validation steps by other methods such as PCR or targeted sequencing is needed in the future to obtain accurate genetic information. This draft genome could be used for further studies so that biological mechanisms could be elucidated.

Previous studies have shown that genomes of invertebrates have relatively high heterozygosity, and the genome of *S. libertina* might also show high heterozygosity, like genomes of *Dendronephthya gigantea* [31] and *Ruditapes philippinarum* [32]. The genome size of *C. tribblei* was 2.76 Gb [11], and the genome size *R. philippinarum* was 2.56 Gb [32]. The genome size of *S. libertina* was relatively larger than that of other species such as Gastropada class or *R. philippinarum*. *R. auricularia* has a relatively small genome size of 910 Mb [10]. Oyster, *Crassostrea gigas*, has a smaller genome size of 637 Mb [30]. Freshwater snail *B. glabrata* has a genome size of approximately 916 Mb [7]. The genome size of *S. libertina* is very large compared to other similar species, and it is similar to that of humans (3.10 Gb). Evolutionary and phylogenetic approaches to

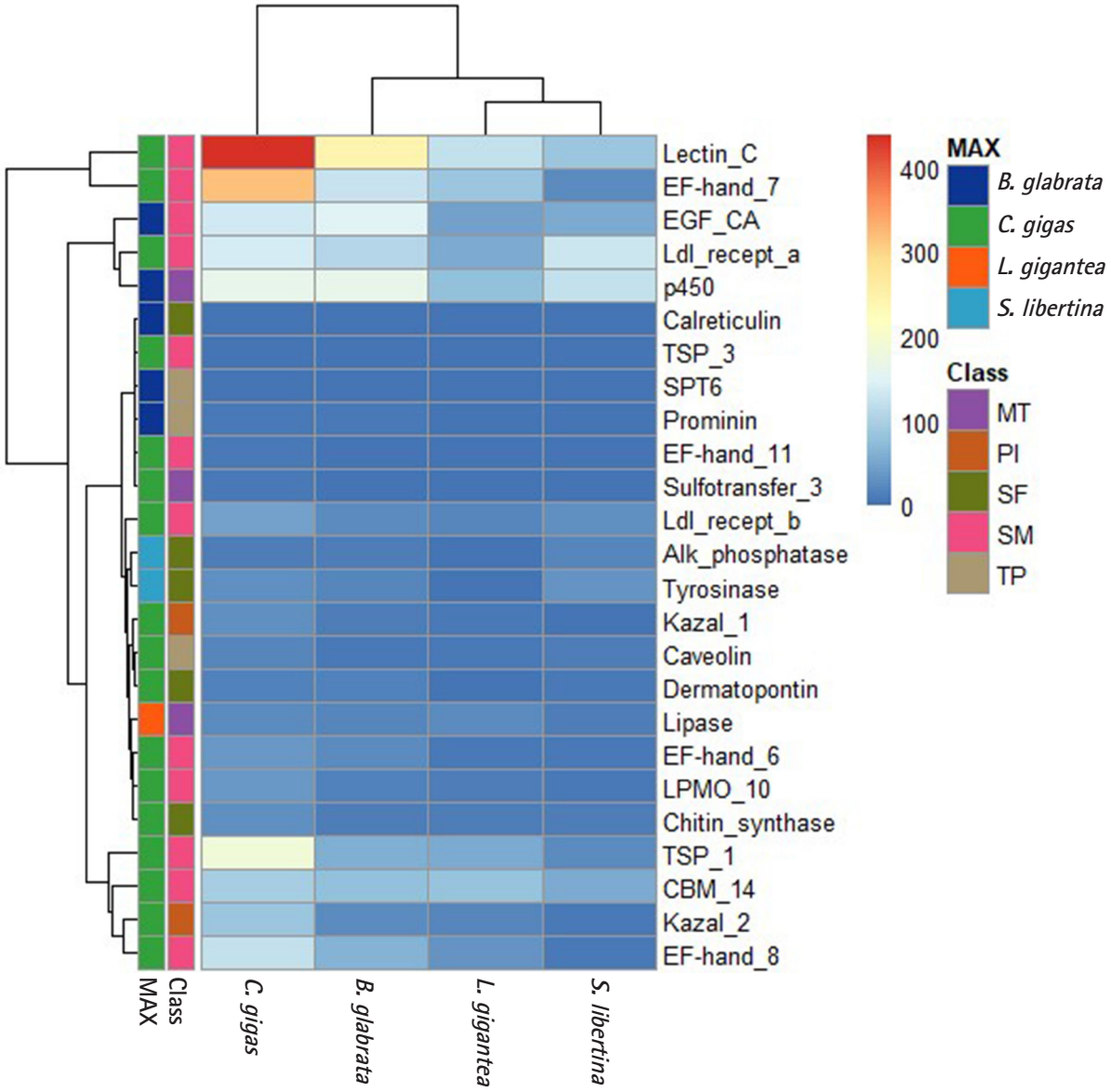

Fig. 1. Copy number of orthologous shell formation related genes calculated with PFAM in four genomes (air-breathing freshwater snail, *Biomphalaria glabrata*; oyster, *Crassostrea gigas*; owl limpet, *Lottia gigantea;* freshwater snail, *Semisulcospira libertina*). In these four genomes, genes detected with the highest frequency were indicated with 'MAX' row bar. A total of 25 genes were used to depict heatmap and construct phylogenetic tree from five class (MT, metabolic transcripts; PI, protease inhibitors; SF, shell formation; SM, small matrix proteins; TP, transmembrane proteins). They are indicated as the second row bar. Full description of each gene name is shown in Supplementary Table 3.

large genome sizes will be needed as future studies.

We calculated the copy number of orthologous genes based on PFAM dataset (Supplementary Table 3). The genome of *C. gigas* had the highest copy number for shell formation related genes. Similar copy number patterns were detected for genomes of *S. libertina* and *B. glabrata*. The genome of *B. glabrata* has different patterns of shell formation proteomes compared to the genome of *C. gigas* [7]. In the genome of *C. gigas*, PIs are highly abundant in shells. The copy number of PIs in four genomes had similar patterns in our analysis. In the genome of *C. gigas*, lectin C-type domain, EF-hand domain pair, and thrombospondin type 1 domain have dramatically higher copy numbers. Lectin C-type-containing proteins are highly expressed in the digestive gland of *C. gigas* [30]. The copy number was also highly detected in our genome. Two genes (alkaline phosphatase and tyrosinase) related to shell formation showed the highest copy number in *S. libertina* among the four species. It means that freshwater snails could have slightly different copy numbers for shellfish metabolism.

https://doi.org/10.5808/gi.21039

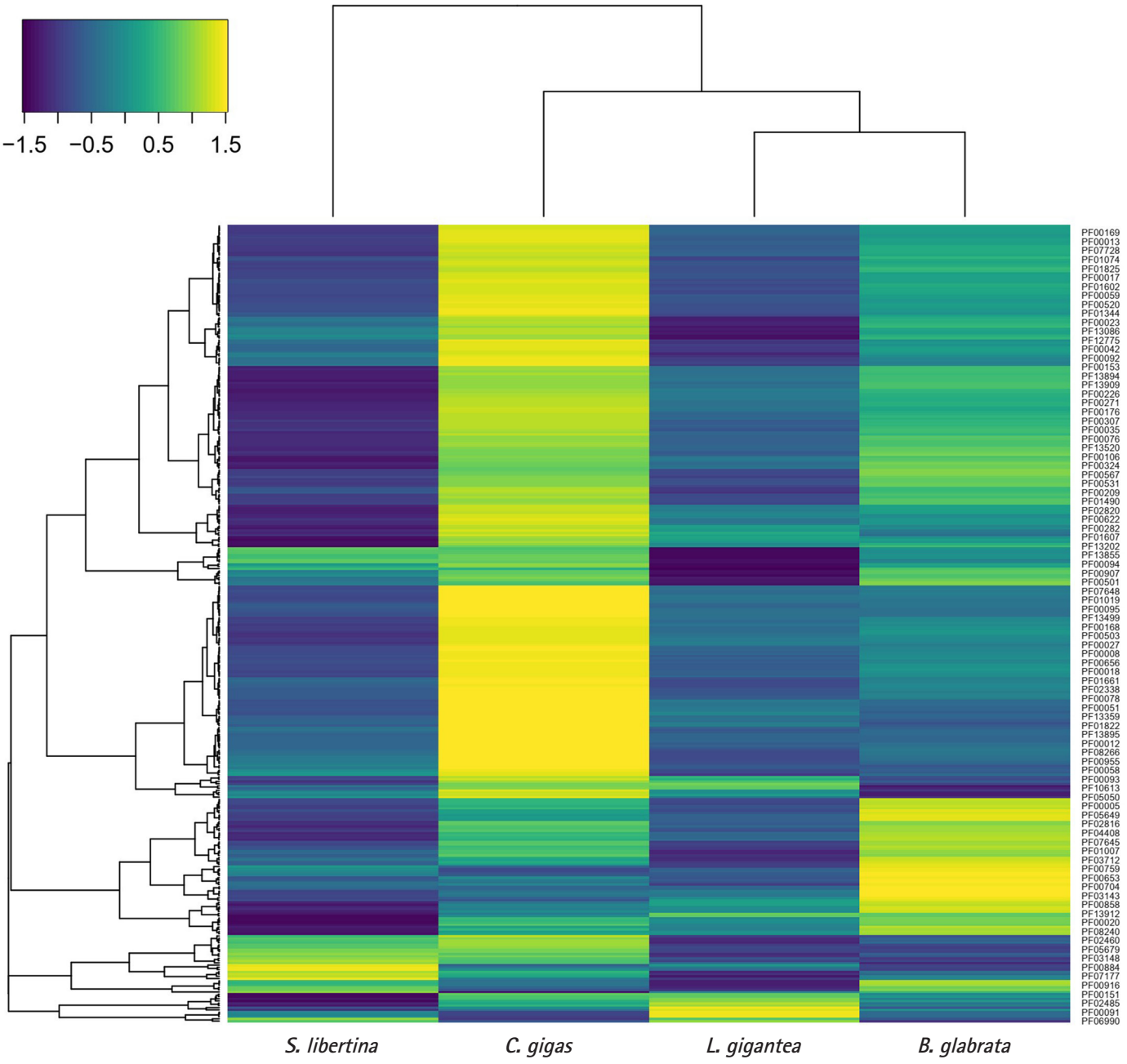

Fig. 2. Heatmap presenting copy numbers of orthologous genes in each genome. Each unit was selected if five or more copy numbers were present in the genome (air-breathing freshwater snail, *Biomphalaria glabrata;* oyster, *Crassostrea gigas*; owl limpet, *Lottia gigantea*; freshwater snail, *Semisulcospira libertina*). In the dendrogram, *B. glabrata* and *L. gigantea* were grouped whereas *S. libertina* was outgrouped.

Mitochondrial DNA sequences and COI sequences are useful for species identification. This is because each species has specific patterns in their sequences. We obtained two phylogenetic trees from whole mitochondrial and COI sequences. These trees showed slightly different patterns. COI seqeucnes tended to be more accurate evolutionarily and taxonomically than whole mitochondrial sequence. Thus, COI sequences are used to confirm species identification and geological distribution of *Semisulcospira* genus.

One of the limitations of this study was our assembly with a total length of 1.4 Gb, but the total genome size was 3 Gb. About 46.7% of the assembled sequences are available. If the complete genome is provided through additional sequencing to the sequence provided by us, it is expected to be of great help in genomics studies on Gastropoda species [33]. The sequence information in this study is in-

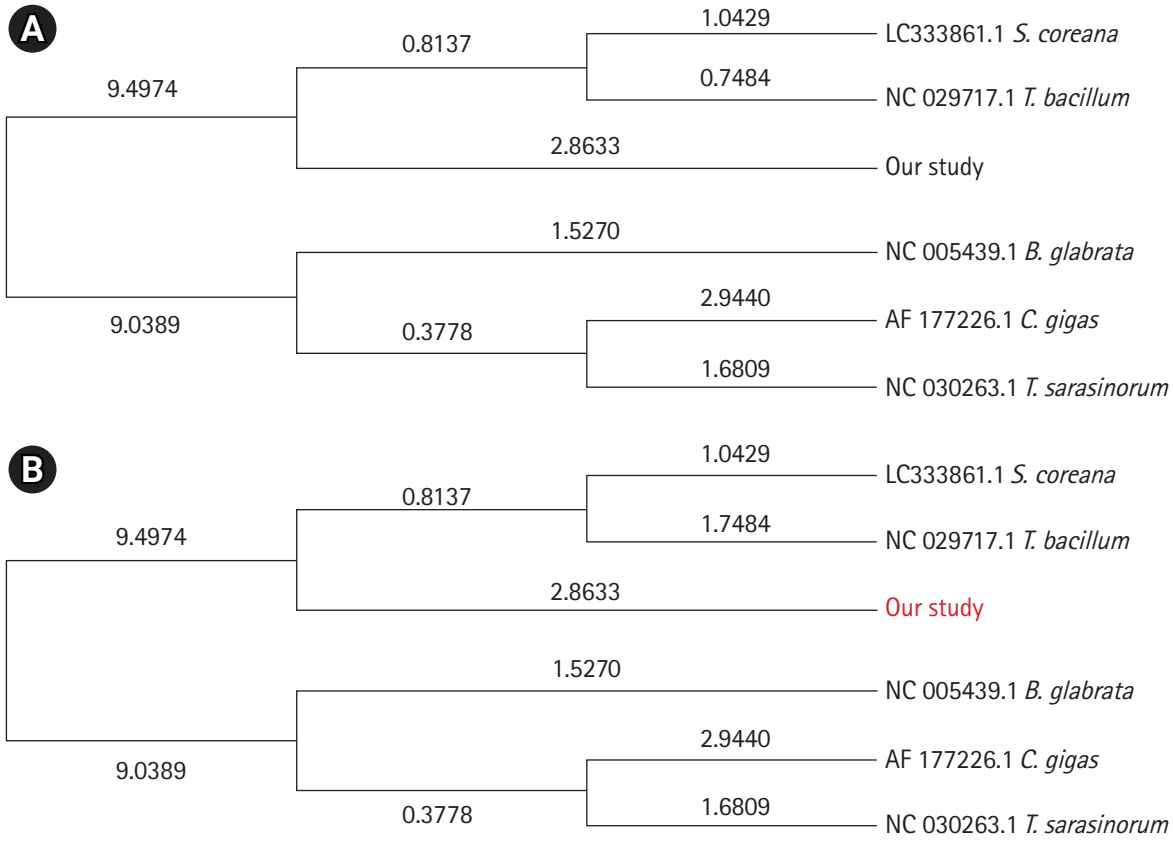

Fig. 3. Maximum likelihood tree for whole-mitochondrial genome (A) and cytochrome c oxidase subunit I regions of mitochondrial DNA (B). These dendrograms were derived from mitochondrial genome sequences identified in the GenBank database and sequences obtained from the present study. Values at nodes indicate branch lengths. Branch length is proportional to the distance between taxa.

complete, and it will not be easy to profile the genome charasteristics. Based on our study, evolutionary or phylogenetic studies in similar species could be performed by comparing gene family diversity of complete genes.

Here, we identified gene sets of *S. libertina* predicted with *de novo* genome assembly data for the first time. These results may provide clues for ecological studies of freshwater environments and immunological studies of secreted materials of *S. libertina*. Our study may also provide useful information for better understanding of the evolutionary relationship among Gastropoda species.

ORCID

Jeong-An Gim: https://orcid.org/0000-0001-7292-2520
Kyung-Wan Baek: https://orcid.org/0000-0002-8445-3773
Young-Sool Hah: https://orcid.org/0000-0002-8571-2722
Ho Jin Choo: https://orcid.org/0000-0003-1798-4360
Ji-Seok Kim: https://orcid.org/0000-0001-8829-5138

Authors' Contribution

Conceptualization: JIY, HJC, JSK. Data curation: JAG, YSH. Formal analysis: JAG, JSK, KWB. Funding acquisition: JIY, HJC. Methodology: JAG, YSH, JIY. Writing - original draft: JAG, KWB, JIY. Writing - review & editing: JAG, KWB, YSH, HJC, JSK, JIY.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This research was supported by the Ministry of SMEs and Startups' 2018 technology development project to foster regional key industries (grant number: P0002726).

https://doi.org/10.5808/gi.21039

Supplementary Materials

Supplementary data can be found with this article online at http://www.genominfo.org.

References

1. Chiu YW, Bor H, Kuo PH, Hsu KC, Tan MS, Wang WK, et al. Origins of Semisulcospira libertina (gastropoda: semisulcospiridae) in Taiwan. Mitochondrial DNA A DNA Mapp Seq Anal 2017;28:518-525.
2. Hsu KC, Bor H, Lin HD, Kuo PH, Tan MS, Chiu YW. Mitochondrial DNA phylogeography of Semisulcospira libertina (Gastropoda: Cerithioidea: Pleuroceridae): implications the history of landform changes in Taiwan. Mol Biol Rep 2014;41:3733-3743.
3. Lee T, Hong HC, Kim JJ, D OF. Phylogenetic and taxonomic incongruence involving nuclear and mitochondrial markers in Korean populations of the freshwater snail genus *Semisulcospira* (Cerithioidea: Pleuroceridae). Mol Phylogenet Evol 2007;43:386-397.
4. Zeng T, Yin W, Xia R, Fu C, Jin B. Complete mitochondrial genome of a freshwater snail, *Semisulcospira libertina* (Cerithioidea: Semisulcospiridae). Mitochondrial DNA 2015;26:897-898.
5. Kim YK, Lee SM. The complete mitochondrial genome of freshwater snail, *Semisulcospira coreana* (Pleuroceridae: Semisulcospiridae). Mitochondrial DNA B Resour 2018;3:259-260.
6. Lee SY, Lee HJ, Kim YK. Comparative analysis of complete mitochondrial genomes with Cerithioidea and molecular phylogeny of the freshwater snail, *Semisulcospira gottschei* (Caenogastropoda, Cerithioidea). Int J Biol Macromol 2019;135:1193-1201.
7. Adema CM, Hillier LW, Jones CS, Loker ES, Knight M, Minx P, et al. Whole genome analysis of a schistosomiasis-transmitting freshwater snail. Nat Commun 2017;8:15451.
8. Simakov O, Marletaz F, Cho SJ, Edsinger-Gonzales E, Havlak P, Hellsten U, et al. Insights into bilaterian evolution from three spiralian genomes. Nature 2013;493:526-531.
9. Nam BH, Kwak W, Kim YO, Kim DG, Kong HJ, Kim WJ, et al. Genome sequence of pacific abalone (*Haliotis discus hannai*): the first draft genome in family Haliotidae. Gigascience 2017;6:1-8.
10. Schell T, Feldmeyer B, Schmidt H, Greshake B, Tills O, Truebano M, et al. An annotated draft genome for *Radix auricularia* (Gastropoda, Mollusca). Genome Biol Evol 2017;9:585-592.
11. Barghi N, Concepcion GP, Olivera BM, Lluisma AO. Structural features of conopeptide genes inferred from partial sequences of the *Conus tribblei* genome. Mol Genet Genomics 2016;291:411-422.
12. Jeon T, Lee YS, Kim HJ. Hepatoprotection by *Semisulcospira libertina* against acetaminophen-induced hepatic injury in mice. Prev Nurt Food Sci 2003;8:239-244.
13. Park YJ, Lee MN, Kim EM, Park JY, Noh JK, Choi TJ, et al. Development and characterization of novel polymorphic microsatellite markers for the Korean freshwater snail *Semisulcospira coreana* and cross-species amplification using next-generation sequencing. J Ocean Limnol 2020;3:503-508.
14. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 2012;1:18.
15. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;31:3210-3212.
16. Smit AF, Hubley R, Green P. RepeatModeler Open-1.0. 2008-2015. Seattle: Institute for Systems Biology, 2015.
17. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3—new capabilities and interfaces. Nucleic Acids Res 2012;40:e115-e115.
18. Lowe TM, Chan PP. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res 2016;44:W54-W57.
19. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 2013;29:2933-2935.
20. Altschul SF. BLAST algorithm. In: Encyclopedia of Lifre Sciences Chichester: John Wiley & Sons, Ltd., 2014.
21. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15-21.
22. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 2016;32:767-769.
23. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res 2005;33:W465-W467.
24. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics 2014;30:1236-1240.
25. Mulder NJ, Apweiler R. The InterPro database and tools for protein domain analysis. Curr Protoc Bioinformatics 2008;Chapter 2:Unit 2 7.
26. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 2018;35:1547-1549.
27. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans

and chimpanzees. Mol Biol Evol 1993;10:512-526.
28. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res 2014;24:1384-1395.
29. De Wit P, Durland E, Ventura A, Langdon CJ. Gene expression correlated with delay in shell formation in larval Pacific oysters (*Crassostrea gigas*) exposed to experimental ocean acidification provides insights into shell formation mechanisms. BMC Genomics 2018;19:160.
30. Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, et al. The oyster genome reveals stress adaptation and complexity of shell formation. Nature 2012;490:49-54.
31. Jeon Y, Park SG, Lee N, Weber JA, Kim HS, Hwang SJ, et al. The draft genome of an octocoral, *Dendronephthya gigantea*. Genome Biol Evol 2019;11:949-953.
32. Mun S, Kim YJ, Markkandan K, Shin W, Oh S, Woo J, et al. The whole-genome and transcriptome of the Manila clam (*Ruditapes philippinarum*). Genome Biol Evol 2017;9:1487-1498.
33. Adachi K, Yoshizumi A, Kuramochi T, Kado R, Okumura SI. Novel insights into the evolution of genome size and AT content in mollusks. Mar Biol 2021;168:25.

https://doi.org/10.5808/gi.21039

G&I
Genomics & Informatics

Original article

eISSN 2234-0742
Genomics Inform 2021;19(3):e33
https://doi.org/10.5808/gi.21031

Received: May 21, 2021
Revised: June 29, 2021
Accepted: June 29, 2021

*Corresponding author:
E-mail: yasodha@icfre.org

Chromosome-specific polymorphic SSR markers in tropical eucalypt species using low coverage whole genome sequences: systematic characterization and validation

Maheswari Patturaj[1], Aiswarya Munusamy[1], Nithishkumar Kannan[1], Ulaganathan Kandasamy[2], Yasodha Ramasamy[1*]

[1]Institute of Forest Genetics and Tree Breeding, Coimbatore 641002, India
[2]Centre for Plant Molecular Biology, Osmania University, Hyderabad 500007, India

Eucalyptus is one of the major plantation species with wide variety of industrial uses. Polymorphic and informative simple sequence repeats (SSRs) have broad range of applications in genetic analysis. In this study, two individuals of *Eucalyptus tereticornis* (ET217 and ET86), one individual each from *E. camaldulensis* (EC17) and *E. grandis* (EG9) were subjected to whole genome resequencing. Low coverage (10×) genome sequencing was used to find polymorphic SSRs between the individuals. Average number of SSR loci identified was 95,513 and the density of SSRs per Mb was from 157.39 in EG9 to 155.08 in EC17. Among all the SSRs detected, the most abundant repeat motifs were di-nucleotide (59.6%–62.5%), followed by tri- (23.7%–27.2%), tetra- (5.2%–5.6%), penta- (5.0%–5.3%), and hexa-nucleotide (2.7%–2.9%). The predominant SSR motif units were AG/CT and AAG/TTC. Computational genome analysis predicted the SSR length variations between the individuals and identified the gene functions of SSR containing sequences. Selected subset of polymorphic markers was validated in a full-sib family of eucalypts. Additionally, genome-wide characterization of single nucleotide polymorphisms, InDels and transcriptional regulators were carried out. These variations will find their utility in genome-wide association studies as well as understanding of molecular mechanisms involved in key economic traits. The genomic resources generated in this study would provide an impetus to integrate genomics in marker-trait associations and breeding of tropical eucalypts.

Keywords: *Eucalyptus camaldulensis*, *E. grandis*, *E. tereticornis*, simple sequence repeats, whole genome resequencing

Introduction

Eucalyptus belongs to the family Myrtaceae, cultivated throughout the tropical and subtropical regions of the world. It is considered to be a major raw material for paper industry and has interesting potential in wood panel, solid wood, charcoal, biofuel and pharmaceutical sectors. Leaf extracts present a wide range of phenolic compounds having antioxidant effects [1], while its bioactive metabolites have demonstrated several ethnopharmacological properties [2]. Genetic improvement programs for limited number of eucalypt species have been implemented across many countries including India, South Africa, China, Brazil, Thailand and Australia. Owing to their great economic value, species-specific genetic and genome resources are progressively increasing. Some of the species like *E. grandis*, *E.*

camaldulensis, and *E. tereticornis* are highly significant for the tropical countries because of their unique properties in paper pulp production and abiotic stress tolerance [3,4]. These species are predominantly used in inter-specific hybridizations, where hybrid breeding strategy is always employed to combine the traits of interest and realize the genetic gains [5].

Genetic marker resources such as simple sequence repeats (SSR) and single nucleotide polymorphisms (SNPs) have been used as powerful tools for identification of individuals, analysis of population structure and genetic diversity, DNA fingerprinting, genetic mapping and localization of QTLs, marker-assisted selection and genomic selection [6]. SSRs are the popular genetic markers because of their abundance, ubiquitous distribution, high polymorphism, codominant inheritance, multiallelism and ease of assay by PCR [7]. Numerous genomic and EST-derived SSR markers have been reported in *Eucalyptus* [8-10]. SSRs can be cross-transferred between closely related species but success rate of intra-genus transferability in eucalypts varied from 40% to 96% [6]. However, identification of polymorphic SSRs between closely related individuals is often difficult because of its genome synteny and colinearity across the species [11] warranting large scale development of SSR markers having polymorphism between individuals.

Latest advances in sequencing technology and bioinformatic research have provided an unparalleled opportunity to identify high-quality, cost and time-effective polymorphic SSR markers in several plant species [12,13]. Further, continuing decrease in the cost of genome sequencing unfolded possibilities for massive identification of polymorphic SSRs as well as large scale genotyping [14]. Moreover, in the species with known genome sequence information, whole genome resequencing strategy is employed to extract polymorphic SSRs rapidly. Mapping parents and segregating populations of *Raphanus sativus* were resequenced at whole genome level and genetic map was constructed with polymorphic SNPs, SSRs, and InDels [15]. Whole genome resequencing was adopted for the development of polymorphic SSR markers between Chinese oriental melon and Korean oriental melon, and many thousands of SSRs, SNPs, and InDels were identified [16]. In *Nicotiana tabacum* whole genome resequencing was carried out in two genotypes to comparatively analyse SSR variations and identify SNPs, InDels, structural variations, and copy number variations for generation of more number of genetic markers [17]. In *Liriodendron chinense*, four genotypes were sequenced at low coverage scale and identified genome-wide SSRs, SNPs, and InDels to assist in molecular genetics, genotype identification, genetic mapping, and molecular breeding [18]. Several thousands of SSR markers were identified for *Ensete ventricosum* by analysing the genome sequence data of four landraces and *in silico* methods were adopted for the development of polymorphic markers [19]. Computational tools such as GMATA [20], PolyMorphPredict [21], SSRgenotyper [22], and MultiplexSSR [23] facilitate SSR genotype calling from resequenced data of individuals within natural populations, germplasm collections and segregating biparental mapping populations.

Accordingly, in the present study, low depth whole genome resequencing was carried out in four genotypes of three tropical species of *Eucalyptus* namely *E. grandis*, *E. camaldulensis* and *E. tereticornis*. The objectives of the study were to (1) characterize the SSR markers on different chromosomes, motif types, frequency and length distribution, (2) *in silico* identification of polymorphic SSR primers between the selected eucalypt species, functional annotation and design candidate primer pairs, (3) validate a subset of SSR primers by PCR amplification in a full-sib family of eucalypts. The distribution of SSR polymorphisms among selected individuals is discussed in relation to their application in cost-efficient genotyping of mapping populations. In eucalypt breeding programs, these markers are regarded as valuable genetic reservoir for genotype identification, genetic diversity analysis, hybrid purity testing and marker assisted selection.

Methods

Plant material and DNA isolation

Four eucalypt clonal accessions, *E. tereticornis* (ET217), *E. camaldulensis* (EC17), *E. tereticornis* (ET86), and *E. grandis* (EG9) were selected for low depth whole genome resequencing. These accessions have been frequently used as the parents for inter-specific full-sib cross generation (ET217 × EC17; ET86 × EG9). SSR polymorphism validation experiments were conducted with 80 individuals of a full-sib family, ET86 × EG9. Juvenile leaves were used for total genomic DNA isolation with DNeasy plant DNA mini kit (Qiagen Inc., Valencia, CA, USA) as per the manufacturer's instructions.

DNA quality analysis and whole genome sequencing

Quality of the DNA was checked on 0.8% Agarose (A9539, Sigma-Aldrich, St. Louis, MO, USA) gel at 120 V for approximately 60 min or until the samples reached 3/4th of the gel. Absorbance ratio at 260/280 was measured with a NanoDrop 2000 UV-Vis spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). A Qubit 2.0 Fluorometer (Q32866, Invitrogen, Carlsbad, CA, USA) was used with a Qubit dsDNA HS Assay Kit (Q32854) to confirm DNA input of 10 μg before shearing. All the DNA samples passed the QC were subjected to paired-end sequencing library preparation with NEB Ultra DNA library preparation kit (New England BioLabs, Ipswich, MA, USA). The quantity and quality check of library was carried out using Agilent TapeStation 2200 System (Agi-

https://doi.org/10.5808/gi.21031

lent Technologies, Santa Clara, CA, USA). Whole genome sequencing of the four eucalypt samples was performed by AgriGenome Labs Pvt Ltd, Hyderabad on an Illumina HiSeq 4000/X ten Genome Analyzer using 2 × 150 bp chemistry. The fastq files were pre-processed using AdapterRemoval2 (v2.2, default parameters). The raw reads were checked for presence of adapter sequences and reads that's average quality score less than 30 (< 30 phred score) in any of the paired-end reads were filtered out.

Genome annotation and analysis

Preliminary analysis was carried out to construct reference-based assembly for each of the four samples, its cleaned reads were aligned against the reference genome *Eucalyptus grandis* v2.0 (11 chromosomes) downloaded from Phytozome (http://phytozome.jgi.doe.gov/pz/portal.html). The short read sequences were assembled into 11 pseudomolecules for each individual. Only uniquely mapped reads were considered for pseudomolecule development. The reads were aligned using BWA (v0.7.17-r1188) individually to the reference genome and reference guided consensus assemblies were generated for each individual using samtools/bcftools suite (v1.9) in FASTA format for further analysis.

An online integrated genome sequence annotation pipeline GenSAS v6.0 [24] was used to annotate the pseudomolecules of four *Eucalyptus* individuals. GenSAS v6.0 was used for various analyses such as repeat masking, gene prediction, annotated gene models and mapping of predicted proteins. Repeats in the pseudomolecules were masked via RepeatMasker v.4.0.7 and RepeatModeler using *Arabidopsis thaliana* as reference. Genes were predicted using the *ab initio* tools and Augustus v.3.3.1 using models from *Arabidopsis thaliana*. Augustus was run using gene models from *Arabidopsis*, finding genes on both strands, and allowing partial models. Sequence alignments were performed using BLAST, BLAT, and PASA against NCBI plant RefSeq database. Multiple lines of evidence were integrated into a gene consensus using EVidenceModeler with default weights. Predicted proteins were compared to the NCBI plant RefSeq database and SwissProt using BLASTP. Protein families were classified using the InterPro database and InterProScan v. 5.8-49.0. An estimate of the completeness of the predicted proteins was calculated using the program BUSCO v. 3.0.2. The GO-Slim and Enzyme-code annotation were performed using Blast2GO for the predicted proteins. Pathway annotation was conducted by mapping the sequences obtained from Blast2GO to the contents of the Kyoto Encyclopedia of Gene and Genomes Automatic Annotation Server (KAAS; http://www.genome.jp/kegg/kaas/ (1 April 2020, date last accessed). The Venn diagrams were generated using jvenn to differentiate common genes across individuals [25]. Transcription factors, transcriptional regulators and protein kinases were identified and classified into different families using the iTAK pipeline v1.7 [26].

Genome-wide SNP and InDel detection

Genome-wide SNPs and InDels were analysed in the four *Eucalyptus* genomes using reference-based assembly of *E. grandis* to document the genetic variants. The reference genome was indexed and the mapping was done using Bowtie2 Aligner [27]. SAMtools was used to convert the generated SAM file to BAM format. The BAM file was sorted and indexed. The reference was also indexed using faidx command of SAM tools [28]. The sorted BAM file was used to generate BCF file using mpileup command of the same package. The variant calling was conducted using bcftools by converting the BCF file to VCF with parameters such as low quality filter > 20 and DP > 100.

Identification of SSRs and detection of polymorphism

FASTA formatted pseudomolecules of *Eucalyptus* were analyzed for frequency and density of SSRs using the Perl script MIcroSAtelitte (MISA; http://pgrc.ipk-gatersleben.de/misa/). Initially SSRs of 2–6 nucleotides motifs were identified with the minimum repeat unit defined as 10 for mono-nucleotides, 7 for di-nucleotides, 5 for tri-and tetra-nucleotides, and four each for penta- and hexa-nucleotides. Compound SSRs were defined as ≥ 2 SSRs interrupted by ≤ 100 bases. SSR length was classified into three categories in accordance with repeat lengths as less than 20 bp (< 20), 20–40 bp, and above 40 bp (> 40). Microsatellites located on the 11 pseudomolecules were used to amplify the genomic sequences of ET86 × EG9 and ET217 × EC17 employing the ePCR module of GMATA software [20]. The primer nucleotide mismatch allowed was no more than one nucleotide and other parameters were set as default. The polymorphic primers were selected based on difference in number of repeat units present in between the genomes of ET86 × EG9 and ET217 × EC17 and polymorphic information content value greater than 0.3 to ensure the SSR polymorphism.

Genotyping of a full-sib family

A subset of 58 primer pairs which were polymorphic *in silico* in the cross ET86 × EG9 was randomly selected for validating the SSR loci amplification in the 80 full-sib progenies. PCR amplification was performed following the protocol of [29] and the products were separated on 7% polyacrylamide gel electrophoresis. The gel was run at 220 V constant power for 3 h and bands visualised by standard silver staining methods. Allele size variations were measured with Alpha Ease FC 5 software (Alpha Innotech, San Leandro, CA, USA).

Results and Discussion

Annotation of genes and repetitive elements

In the present study, four individuals belonging to *E. grandis*, *E. camaldulensis*, and *E. tereticornis* were subjected to short read sequencing with approximate of genome coverage of 10× for each genotype. The final assemblies had a total length of 611.8 to 612.2 Mb (Table 1), and each assembly was arranged into 11 pseudomolecules and deposited in NCBI (Biosample SAMN14826404, SAMN14826405, SAMN14826406, and SAMN14826407). Size of the individual pseudomolecule varied between 37.7 and 83.9 Mb, with an average of 55.6 Mb. The percentage of genes identified in the *Eucalyptus* genome sequence showed that nearly 82%-85% of the genome was represented (Table 2). The results were in accordance with *E. pauciflora*, where the BUSCO genes varied from 70.5%–91.3% in different assemblies [21].

Repeats in the genome of four *Eucalyptus* individuals were identified and masked which comprised to maximum 53.65 % (ET86) and minimum 35.82% (ET217) of the assemblies (Supplementary Table 1). Unclassified repeats occupied the maximum amount of genome repeats totalling to 64.57, 51.04, 48.16, and 43.56% of the ET86, EC17, EG9 and ET217, respectively. The LTR elements including Gypsy and Copia repeats had occupied next highest type of repeat classes across the individuals analysed. Gene prediction with NCBI RefSeq resulted in maximum of 57,075 (EG9) to minimum of 49,515 (ET86) protein-coding sequences (Table 3). Out of 2,005 gene families analysed 1,807 common genes were identified, accounting for 90.0% of the total protein-predicted genes highlighting the close relationship among the species (Fig. 1). Very limited number of genes was found to be species-specific, some of the unique genes such as disease resistance protein RPS4 and chitin elicitor receptor kinase 1 involved in defense activation were identified in ET86 and EC17, respectively. Gene ontology classification revealed higher proportion of genes related to molecular function followed by biological process, and cellular components (Fig. 2). Analysis of transcription factors (TFs), transcriptional regulators, and protein kinases (PKs) identified an average of 1,807 (from 69 families), 393 (from 24 families), and 2,137 (from 120 families) genes respectively in the four genomes analysed (Supplementary Table 2). Further, the eucalypt genome encoded majority of PKs

Table 1. Description of sequence data generated for four *Eucalyptus* individuals

Sample ID	No. of raw reads	No. of bases (Mb)	GC percent	% Squences with Q30	Genome assembly size (Mb)
E.camaldulensis (EC17)	50,810,080	7,621.5	41.2	91.1	611.9
E.tereticornis (ET217)	41,507,796	6,226.2	38.8	89.5	611.9
E.tereticornis (ET86)	43,514,058	6,527.1	40.0	90.3	611.9
E.grandis (EG9)	46,840,164	7,026.0	39.6	90.3	612.3
Average	45,668,025	6,850.2	40.0	90.0	612.0

Table 2. Summary of BUSCO analysis results for the four *Eucalyptus* assemblies

BUSCO assessment	*E. camaldulensis* (EC17)	*E. tereticornis* (ET217)	*E. tereticornis* (ET86)	*E. grandis* (EG9)
Complete BUSCOs (C)	1,206 (83.8)	1,211 (84.1)	1,181 (82.1)	1,219 (84.7)
Complete and single-copy BUSCOs (S)	1,141 (79.2)	1,162 (80.7)	1,124 (78.1)	1,153 (80.1)
Complete and duplicated BUSCOs (D)	65 (4.5)	49 (3.4)	57 (4.0)	66 (4.6)
Fragmented BUSCOs (F)	114 (7.9)	99 (6.9)	115 (8.0)	99 (6.9)
Missing BUSCOs (M)	120 (8.3)	130 (9.0)	144 (9.9)	122 (8.4)
Total BUSCO groups searched	1,440 (100)	1,440 (100)	1,440 (100)	1,440 (100)

Values are presented as number (%).

Table 3. Descriptive details on genome annotation of four *Eucalyptus* individuals

Sample ID	NCBI RefSeq proteins					No. of predicted proteins with Swiss-Prot database	No. of proteins annotated for functionality
	Total	Less than 100 amino acids	Percent similarity to Eucalyptus	Uncharacterized protein	Maximum and minimum amino acid length		
E. camaldulensis (EC17)	54,656	7,231	90.9	17,991	5,164:34	30,892	10,514
E. tereticornis (ET217)	54,852	7,415	91.1	21,251	5,422:34	28,067	10,665
E. tereticornis (ET86)	49,515	6,492	94.1	16,789	5,372:34	27,148	10,424
E. grandis (EG9)	57,075	7,333	90.6	22,985	5,904:34	31,807	10,753

https://doi.org/10.5808/gi.21031

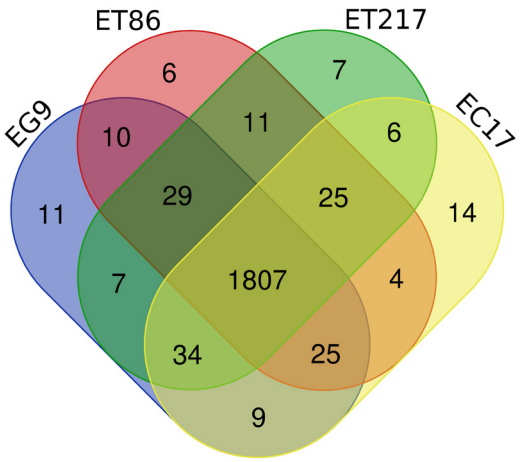

Fig. 1. Venn diagram shows the number of shared and unique gene families among the four *Eucalyptus* individual analyzed. Each color represents one individual (*E. camaldulensis* [EC17], *E. tereticornis* [ET86 and ET217], and *E. grandis* [EG9]).

belong to the receptor-like kinase family.

The reference assemblies generated in this study were primarily in accordance with the *Eucalyptus* annotation release 101 (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Eucalyptus_grandis/101/). The released assembly of *E. grandis* had totally 55,643 genes belonging to various classes like protein coding, non-coding, pseudogenes and genes with variants [30]. Recent studies in eucalypts highlighted the role of TFs and PKs associated with secondary cell wall development [31], biotic resistance [32] and abiotic tolerance [33,34]. Accordingly, results of this study offer a comprehensive view of regulatory sequences associated with almost all essential cellular functions and provides a foundation for further characterization.

Identification of SNPs and InDels

Sequences mapped to the assembled chromosomes were analysed to predict the putative SNPs and InDels (Supplementary Table 3).

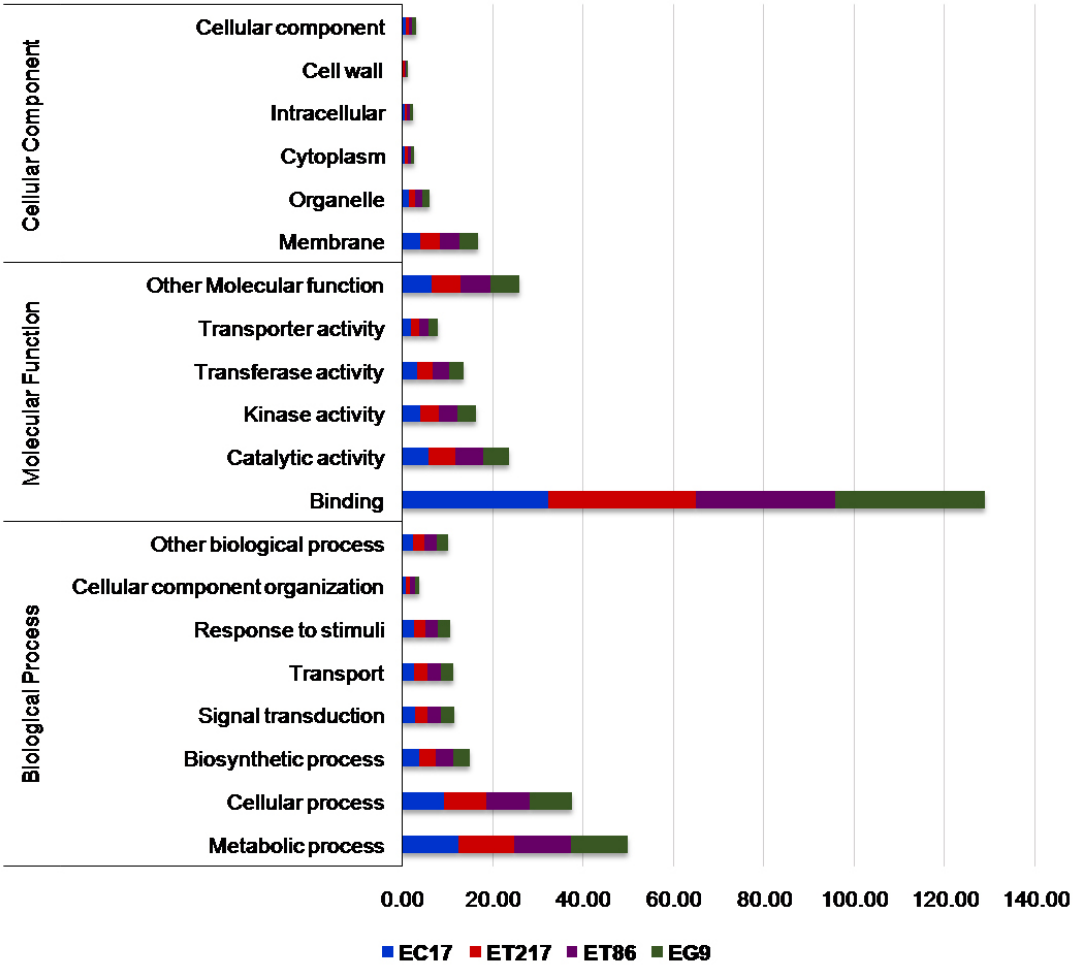

Fig. 2. Gene ontology classification of four *Eucalyptus* genome analyzed (*E. camaldulensis* [EC17], *E. tereticornis* [ET86 and ET217], and *E. grandis* [EG9]) for biological process, cellular component, and molecular function.

The number of SNPs recorded across the four eucalypt genomes analysed was 727,996 (EG09), 1,225,836 (ET86), 1,207,912 (ET217), and 1,170,967 (EC17). Similarly, the InDels observed were 104,542 (EG09), 141,591 (ET86), 142,384 (ET217), and 134,986 (EC17). Maximum number of SNPs and InDels were recorded in longer pseudomolecules such as 03, 08, and 05. In the recent past, genome-wide association and genomic selection approaches were implemented using SNPs and InDels in eucalypt species [35-37].

SSR distribution and polymorphism

SSRs were detected using MISA in assembled pseudomolecules and SSR prediction statistics are presented in Table 4 and 5. Average number of SSR loci identified was 95,513 and their distribution across species were 94,889 (EC17), 95,373 (ET217), 95,425 (ET86), and 96,365 (EG9), respectively. The number of SSRs was found to be correlated with the chromosome length. Longer chromosomes like 03, 05 and 08 had more than 11,000 SSRs whereas shorter ones like 04, 09 and 10 had lower number of SSRs (Table 6). The highest frequency of the grouped SSR motif units was dimer AG/CT (44.2%–46.9%) and among the tri, tetra, penta and hexamers AAG/TTC, ACAT/ATGT, AAAAT/ATTTT and AAAAAG/CTTTTT was most common, respectively (Supplementary Table 4). Tri-nucleotide motif type AAG/TTC was reported to be the most common in eucalypts and in many other dicot plants such as *Arachis*, cucumber, soybean, *Arabidopsis* and grape [38,39]. Among the 17 different tri-nucleotides ACA/TGT was the least commonly present motif type. The predominant repeat motif types are in accordance with earlier reports on eucalypts [38,40]. SSR class with the length of < 20 bp was more abundant followed by 20–40 bp and > 40 bp in all the four individuals analysed (Supplementary Table 5).

Eucalyptus species are closely related with overlapping geographical locations having high amount of gene flow among species [41], thus pose difficulty in choosing polymorphic SSRs. In this study, SSR polymorphism in perfect repeat motifs was determined by *in silico* characterization of SSR length variation between the parents of the cross, ET217 × EC17 and ET86 × EG9. The cross, ET217 × EC17, had an average of 95,131 SSRs of which 13.4% (12,725) markers from pseudomolecule 1, 8 and 10 were showing polymorphism. In the case of ET86 × EG9, all the chromosomes harboured polymorphic SSRs except 3, 6, 9, and 11. Although the cross had an average of 95,895 SSRs, only 25.7% (24,688) of the SSRs could be converted into usable markers. The number of genic SSRs which were polymorphic among the clonal accessions analysed are shown in Fig. 3. In both the crosses, among the five repeat motif types, the di-nucleotide showed maximum polymorphism (76.0 %) followed by tri-nucleotides (18.0%). The *in silico* polymorphic markers can be utilized not only for high density genetic linkage map generation but also for variety of purposes including genome-wide marker-trait associations and population genetic studies.

Validation of polymorphic SSR markers

Among the 58 primer pairs, 46 (81%) successfully amplified in 80 full-sib progenies of the cross ET86 × EG9 and 12 (19%) failed to generate the PCR products. Out of 46 primer pairs, 35 (62%) gen-

Table 4. Distribution of SSRs in *Eucalyptus* genome

Parameter	*E. camaldulensis* (EC17)	*E. tereticornis* (ET217)	*E. tereticornis* (ET86)	*E. grandis* (EG9)
Total No. of SSR loci	94,889	95,373	95,425	96,365
Loci distance (kb)	6.45	6.42	6.41	6.35
Density (SSRs/Mb)	155.08	155.87	155.96	157.39
SSR length < 20 bp	86,824	87,628	87,585	88,719
SSR length 20–40 bp	7,649	7,326	7,421	7,222
SSR length > 40 bp	416	419	419	424

SSR, simple sequence repeat.

Table 5. Different types of SSR loci identified in four *Eucalyptus* individuals

SSR type (%)/Sample ID	*E. camaldulensis* (EC17)	*E. tereticornis* (ET217)	*E. tereticornis* (ET86)	*E. grandis* (EG9)
Di-nucleotide	62.5	59.9	59.9	59.6
Tri-nucleotide	23.7	27.0	27.0	27.2
Tetra-nucleotide	5.6	5.3	5.2	5.3
Penta-nucleotide	5.3	5.0	5.0	5.2
Hexa-nucleotide	2.9	2.7	2.8	2.8

SSR, simple sequence repeat.

https://doi.org/10.5808/gi.21031

Table 6. Various SSR types and their distribution among the 11 pseudomolecules of the four Eucalyptus clonal accessions

SSR type/Chromosome	1	2	3	4	5	7	8	9	10	11
E. camaldulensis (EC17)										
Di-nucleotide	4,389	5,543	6,839	3,862	6,588	5,275	6,946	3,606	3,888	4,495
Tri-nucleotide	1,916	2,518	3,063	1,656	2,796	2,219	3,169	1,679	1,840	2,029
Tetra-nucleotide	396	409	624	351	647	430	673	329	350	385
Penta-nucleotide	374	432	639	340	534	446	582	326	322	361
Hexa-nucleotide	192	256	325	159	272	222	347	182	192	217
E. tereticornis (ET217)										
Di-nucleotide	4,440	5,600	6,878	3,880	6,614	5,278	6,870	3,615	3,952	4,499
Tri-nucleotide	1,940	2,565	3,085	1,704	2,855	2,260	3,229	1,708	1,898	2,031
Tetra-nucleotide	390	430	665	346	646	435	659	332	342	380
Penta-nucleotide	380	427	630	339	528	439	580	317	319	352
Hexa-nucleotide	187	262	295	160	291	226	335	193	179	197
E. tereticornis (ET86)										
Di-nucleotide	4,442	5,590	6,919	3,882	6,577	5,274	6,922	3,653	3,937	4,471
Tri-nucleotide	1,954	2,552	3,086	1,687	2,869	2,255	3,206	1,705	1,888	2,021
Tetra-nucleotide	371	420	648	351	668	458	677	330	348	365
Penta-nucleotide	377	432	647	345	508	438	579	330	326	371
Hexa-nucleotide	181	265	302	156	294	244	313	195	176	228
E. grandis (EG9)										
Di-nucleotide	4,429	5,545	6,877	3,950	6,653	5,253	6,925	3,665	4,003	4,521
Tri-nucleotide	1,984	2,567	3,202	1,759	2,893	2,317	3,265	1,723	1,879	2,005
Tetra-nucleotide	401	411	665	367	644	453	700	327	333	356
Penta-nucleotide	406	442	652	348	548	440	582	344	340	384
Hexa-nucleotide	182	285	310	173	297	225	339	197	178	222

SSR, simple sequence repeat.

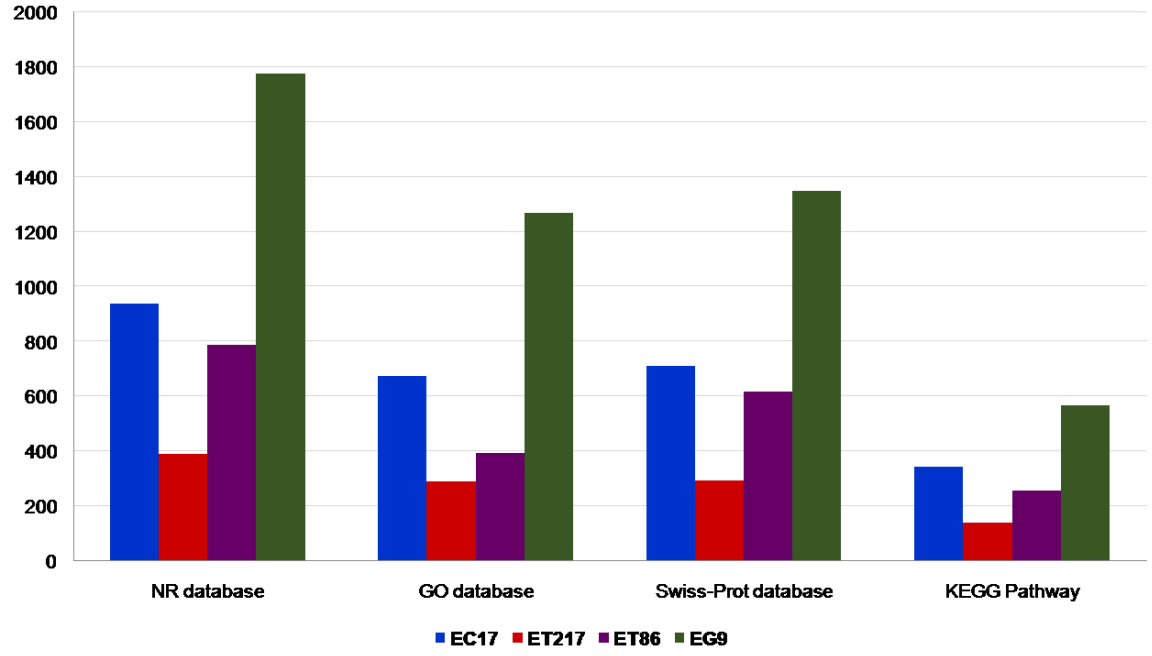

Fig. 3. Annotation of polymorphic genic simple sequence repeats associated to the RefSeq non-redundant (NR), Gene Ontology (GO), SwissProt, and Kyoto Encyclopedia of Gene and Genomes (KEGG) database (*E. camaldulensis* [EC17], *E. tereticornis* [ET86 and ET217], and *E. grandis* [EG9]).

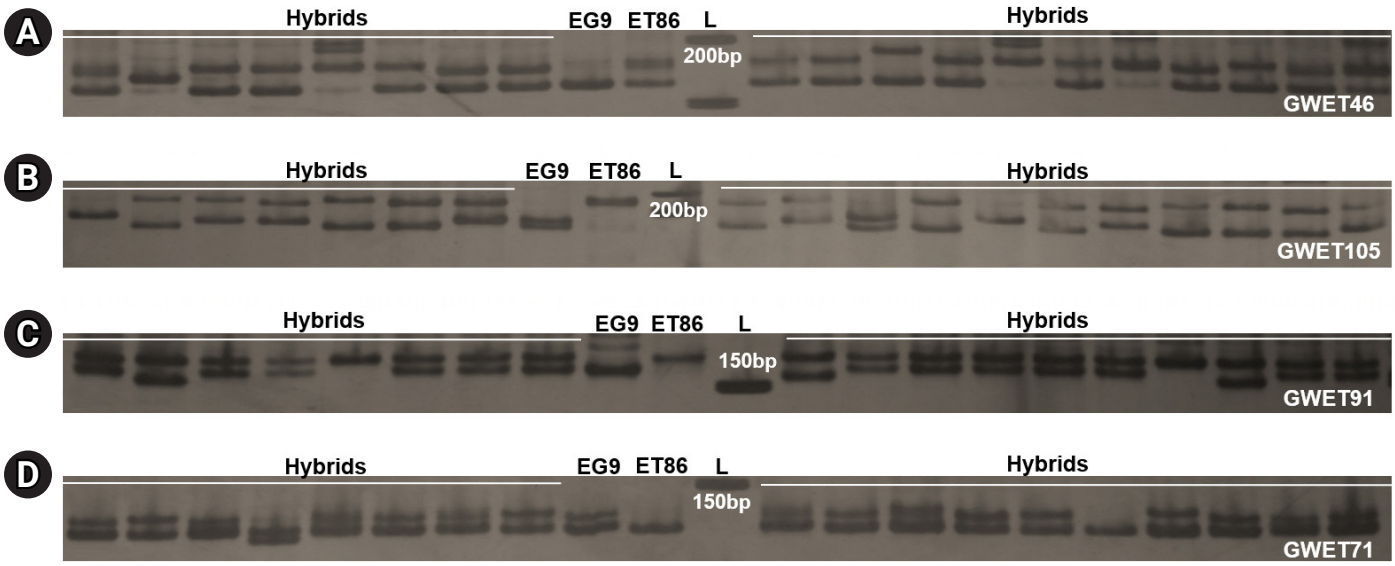

Fig. 4. Gel images of simple sequence repeat markers with primers GWET 46 (A), GWET 105 (B), GWET 91 (C), and GWET 71 (D) for the cross ET86 × EG9. Hybrids, full-sib progenies of ET86 × EG9; ET86, *E. tereticornis*; EG9, *E. grandis*; L, 50 bp DNA ladder.

erated one or two polymorphic alleles and remaining 11 showed monomorphic products (Fig. 4, Supplementary Table 6). A similar study in two closely related species *Capsicum chinense* and *C. annuum* employed *in silico* produced SSRs to ascertain their effectiveness and 71.2% polymorphism was recorded [42]. In the full-sib family of *Trachinotus ovatus*, a computational pipeline Multiplex SSR was employed to predict polymorphism and 85% success in PCR amplification was recorded [23]. Further, these results confirmed that predicted SSRs polymorphism can be utilized for accurate genotyping in capillary based systems economically and efficiently.

Conclusion

Experiments on the whole genome resequencing are becoming increasingly frequent, and eucalypts are no exception. The majority of resequencing experiments were able to detect significant genetic variations between the sequenced accessions and the reference genome. In the genetic advancement of eucalypts, genome-enabled approaches have become indispensable. Some of the examples include integration of DNA markers in commercial breeding of eucalypts by paper industries for quality control in clonal forestry and hybrid purity. International research consortia are using genomics to identify chromosomal locations governing commercially important traits. Abundant SSRs have been discovered in the released genomes of *E. grandis*, *E. camaldulensis*, and *E. pauciflora*. However, the inherent drawbacks associated with identification of polymorphic SSRs for mapping population genotyping, varietal fingerprinting and population genomics continue to exist. Affordable whole genome resequencing technology along with appropriate bioinformatics tools makes the prediction of SSR variations across highly similar genomes possible. The findings of this study improved our knowledge on genetic variations between eucalypt individuals and developed pre-screened SSR markers to genotype the mapping populations. Further, a large set of chromosome anchored markers and TFs were discovered. The SNPs generated would find its application in high density genetic linkage map generation. It forms a valuable genomic resource with promising applications in QTL based selection and breeding, genomic selection, conservation of genetic resources and improvement of eucalypt germplasm.

ORCID

Yasodha Ramasamy: https://orcid.org/0000-0001-9992-9992

Authors' Contribution

Conceptualization: YR. Formal analysis: MP, AM, NK, UK. Methodology: AM, MP, NK. Writing - original draft: MP. Writing -review & editing: YR.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

The authors acknowledge the funding support from the Depart-

https://doi.org/10.5808/gi.21031

ment of Biotechnology, Government of India.

Supplementary Materials

Supplementary data can be found with this article online at http://www.genominfo.org.

References

1. Gullon B, Muniz-Mouro A, Lu-Chau TA, Moreira MT, Lema JM, Eibes G. Green approaches for the extraction of antioxidants from eucalyptus leaves. Ind Crops Prod 2019;138:111473.
2. Salehi B, Sharifi-Rad J, Quispe C, Llaique H, Villalobos M, Smeriglio A, et al. Insights into *Eucalyptus* genus chemical constituents, biological activities and health-promoting effects. Trends Food Sci Technol 2019;91:609-624.
3. Vena PF, Garcia-Aparicio MP, Brienzo M, Gorgens JF, Rypstra T. Effect of alkaline hemicellulose extraction on kraft pulp fibers from *Eucalyptus grandis*. J Wood Chem Technol 2013;33:157-173.
4. Booth TH. Assessing the thermal adaptability of tree provenances: an example using *Eucalyptus tereticornis*. Aust For 2019;82:176-180.
5. Chen S, Weng Q, Li F, Li M, Zhou C, Gan S. Genetic parameters for growth and wood chemical properties in *Eucalyptus urophylla* × *E. tereticornis* hybrids. Ann For Sci 2018;75:16.
6. Grattapaglia D, Vaillancourt RE, Shepherd M, Thumma BR, Foley W, Kulheim C, et al. Progress in Myrtaceae genetics and genomics: *Eucalyptus* as the pivotal genus. Tree Genet Genomes 2012;8:463-508.
7. Vieira ML, Santini L, Diniz AL, Munhoz Cde F. Microsatellite markers: what they mean and why they are so useful. Genet Mol Biol 2016;39:312-328.
8. Brondani RP, Brondani C, Grattapaglia D. Towards a genus-wide reference linkage map for *Eucalyptus* based exclusively on highly informative microsatellite markers. Mol Genet Genomics 2002;267:338-347.
9. Grattapaglia D, Mamani EM, Silva-Junior OB, Faria DA. A novel genome-wide microsatellite resource for species of *Eucalyptus* with linkage-to-physical correspondence on the reference genome sequence. Mol Ecol Resour 2015;15:437-448.
10. Zhou C, He X, Li F, Weng Q, Yu X, Wang Y, et al. Development of 240 novel EST-SSRs in *Eucalyptus* L'Hérit. Mol Breed 2014;33:221-225.
11. Gion JM, Hudson CJ, Lesur I, Vaillancourt RE, Potts BM, Freeman JS. Genome-wide variation in recombination rate in *Eucalyptus*. BMC Genomics 2016;17:590.
12. Srivastava S, Kushwaha B, Prakash J, Kumar R, Nagpure NS, Agarwal S, et al. Development and characterization of genic SSR markers from low depth genome sequence of *Clarias batrachus* (magur). J Genet 2016;95:603-609.
13. Taheri S, Lee Abdullah T, Yusop MR, Hanafi MM, Sahebi M, Azizi P, et al. Mining and development of novel SSR markers using next generation sequencing (NGS) data in plants. Molecules 2018;23:399.
14. Yang J, Zhang J, Han R, Zhang F, Mao A, Luo J, et al. Target SSR-Seq: a novel SSR genotyping technology associate with perfect SSRs in genetic analysis of cucumber varieties. Front Plant Sci 2019;10:531.
15. Mun JH, Chung H, Chung WH, Oh M, Jeong YM, Kim N, et al. Construction of a reference genetic map of *Raphanus sativus* based on genotyping by whole-genome resequencing. Theor Appl Genet 2015;128:259-272.
16. Song WH, Chung SM. Whole genome re-sequencing and development of SSR markers in oriental melon. J Plant Biotechnol 2019;46:71-78.
17. Yang H, Geng X, Zhao S, Shi H. Genomic diversity analysis and identification of novel SSR markers in four tobacco varieties by high-throughput resequencing. Plant Physiol Biochem 2020;150:80-89.
18. Li B, Lin F, Huang P, Guo W, Zheng Y. Development of nuclear SSR and chloroplast genome markers in diverse *Liriodendron chinense* germplasm based on low-coverage whole genome sequencing. Biol Res 2020;53:21.
19. Biswas MK, Darbar JN, Borrell JS, Bagchi M, Biswas D, Nuraga GW, et al. The landscape of microsatellites in the enset (*Ensete ventricosum*) genome and web-based marker resource development. Sci Rep 2020;10:15312.
20. Wang X, Wang L. GMATA: an integrated software package for genome-scale SSR mining, marker development and viewing. Front Plant Sci 2016;7:1350.
21. Wang W, Das A, Kainer D, Schalamun M, Morales-Suarez A, Schwessinger B, et al. The draft nuclear genome assembly of *Eucalyptus pauciflora*: a pipeline for comparing de novo assemblies. Gigascience 2020;9:giz160.
22. Lewis DH, Jarvis DE, Maughan PJ. SSRgenotyper: a simple sequence repeat genotyping application for whole-genome resequencing and reduced representational sequencing projects. Appl Plant Sci 2020;8:e11402.
23. Guo L, Yang Q, Yang JW, Zhang N, Liu BS, Zhu KC, et al. MultiplexSSR: a pipeline for developing multiplex SSR-PCR assays from resequencing data. Ecol Evol 2020;10:3055-3067.
24. Humann JL, Lee T, Ficklin S, Main D. Structural and functional annotation of eukaryotic genomes with GenSAS. Methods Mol

Biol 2019;1962:29-51.
25. Bardou P, Mariette J, Escudie F, Djemiel C, Klopp C. jvenn: an interactive Venn diagram viewer. BMC Bioinformatics 2014;15:293.
26. Zheng Y, Jiao C, Sun H, Rosli HG, Pombo MA, Zhang P, et al. iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. Mol Plant 2016;9:1667-1670.
27. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357-359.
28. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078-2079.
29. Sumathi M, Bachpai VKW, Mayavel A, Dasgupta MG, Nagarajan B, Rajasugunasekar D, et al. Genetic linkage map and QTL identification for adventitious rooting traits in red gum eucalypts. 3 Biotech 2018;8:242.
30. Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, et al. The genome of *Eucalyptus grandis*. Nature 2014;510:356-362.
31. Hussey SG, Grima-Pettenati J, Myburg AA, Mizrachi E, Brady SM, Yoshikuni Y, et al. A standardized synthetic *Eucalyptus* transcription factor and promoter panel for re-engineering secondary cell wall regulation in biomass and bioenergy crops. ACS Synth Biol 2019;8:463-465.
32. Santos SA, Vidigal PM, Guimaraes LM, Mafia RG, Templeton MD, Alfenas AC. Transcriptome analysis of *Eucalyptus grandis* genotypes reveals constitutive overexpression of genes related to rust (*Austropuccinia psidii*) resistance. Plant Mol Biol 2020;104:339-357.
33. Sasaki K, Ida Y, Kitajima S, Kawazu T, Hibino T, Hanba YT. Overexpressing the HD-Zip class II transcription factor EcHB1 from *Eucalyptus camaldulensis* increased the leaf photosynthesis and drought tolerance of *Eucalyptus*. Sci Rep 2019;9:14121.
34. Zhang J, Wu J, Guo M, Aslam M, Wang Q, Ma H, et al. Genome-wide characterization and expression profiling of *Eucalyptus grandis* HD-Zip gene family in response to salt and temperature stress. BMC Plant Biol 2020;20:451.
35. Supple MA, Bragg JG, Broadhurst LM, Nicotra AB, Byrne M, Andrew RL, et al. Landscape genomic prediction for restoration of a *Eucalyptus* foundation species under climate change. Elife 2018;7:e31835.
36. Kainer D, Padovan A, Degenhardt J, Krause S, Mondal P, Foley WJ, et al. High marker density GWAS provides novel insights into the genomic architecture of terpene oil yield in *Eucalyptus*. New Phytol 2019;223:1489-1504.
37. Thavamanikumar S, Arnold RJ, Luo J, Thumma BR. Genomic studies reveal substantial dominant effects and improved genomic predictions in an open-pollinated breeding population of *Eucalyptus pellita*. G3 (Bethesda) 2020;10:3751-3763.
38. Sumathi M, Yasodha R. Microsatellite resources of *Eucalyptus*: current status and future perspectives. Bot Stud 2014;55:73.
39. Cavagnaro PF, Senalik DA, Yang L, Simon PW, Harkins TT, Kodira CD, et al. Genome-wide characterization of simple sequence repeats in cucumber (*Cucumis sativus* L.). BMC Genomics 2010;11:569.
40. Liu G, Xie Y, Zhang D, Chen H. Analysis of SSR loci and development of SSR primers in *Eucalyptus*. J For Res 2018;29:273-282.
41. Arumugasundaram S, Ghosh M, Veerasamy S, Ramasamy Y. Species discrimination, population structure and linkage disequilibrium in *Eucalyptus camaldulensis* and E*ucalyptus tereticornis* using SSR markers. PLoS One 2011;6:e28252.
42. Uncu AT. Genome-wide identification of simple sequence repeat (SSR) markers in *Capsicum chinense* Jacq. with high potential for use in pepper introgression breeding. Biologia 2019;74:119-126.

https://doi.org/10.5808/gi.21031

G&I
Genomics & Informatics

Application note

eISSN 2234-0742
Genomics Inform 2021;19(3):e34
https://doi.org/10.5808/gi.21035

Received: July 6, 2021
Accepted: August 11, 2021

*Corresponding author:
E-mail: kyudong.han@gmail.com

High-accuracy quantitative principle of a new compact digital PCR equipment: Lab On An Array

Haeun Lee[1], Cherl-Joon Lee[1], Dong Hee Kim[2], Chun-Sung Cho[3], Wonseok Shin[4], Kyudong Han[1,5,6*]

[1]Department of Bioconvergence Engineering, Dankook University, Yongin 16890, Korea
[2]Department of Anesthesiology and Pain Management, Dankook University Hospital, Cheonan 31116, Korea
[3]Department of Neurosurgery, Dankook University College of Medicine, Cheonan 31116, Korea
[4]NGS Clinical Laboratory, Dankook University Hospital, Cheonan 31116, Korea
[5]Center for Bio Medical Engineering Core Facility, Dankook University, Cheonan 31116, Korea
[6]Department of Microbiology, College of Science and Technology, Dankook University, Cheonan 31116, Korea

Digital PCR (dPCR) is the third-generation PCR that enables real-time absolute quantification without reference materials. Recently, global diagnosis companies have developed new dPCR equipment. In line with the development, the Lab On An Array (LOAA) dPCR analyzer (Optolane) was launched last year. The LOAA dPCR is a semiconductor chip-based separation PCR type equipment. The LOAA dPCR includes Micro Electro Mechanical System that can be injected by partitioning the target gene into 56 to 20,000 wells. The amount of target gene per wells is digitized to 0 or 1 as the number of well gradually increases to 20,000 wells because its principle follows Poisson distribution, which allows the LOAA dPCR to perform precise absolute quantification. LOAA determined region of interest first prior to dPCR operation. To exclude invalid wells for the quantification, the LOAA dPCR has applied various filtering methods using brightness, slope, baseline, and noise filters. As the coronavirus disease 2019 has now spread around the world, needs for diagnostic equipment of point of care testing (POCT) are increasing. The LOAA dPCR is expected to be suitable for POCT diagnosis due to its compact size and high accuracy. Here, we describe the quantitative principle of the LOAA dPCR and suggest that it can be applied to various fields.

Keywords: digital PCR, LOAA dPCR, Micro Electro Mechanical System, point of care testing, Poisson distribution
Availability: The overall principle of Lab On An Array described in this article is available at http://optolane.com/technology/.

Introduction

Digital PCR (dPCR) assay, a third-generation PCR method, is the latest PCR method capable of real-time absolute quantification of target genes without reference materials [1]. Commercial companies in the molecular diagnostic field have developed various dPCR systems because they have advantages over the existing quantitative PCR (qPCR) in several ways [2]. Following the evolving trend of this technique, a new dPCR equipment (Lab On An Array [LOAA] digital real-time PCR analyzer; LOAA dPCR, Optolane, Seongnam, Korea) was launched in South Korea last year. The LOAA dPCR is a separation type dPCR using semiconductor chip-based Micro Electro Mechanical System (MEMS) technology

[3]. This LOAA dPCR equipment has several technical advantages. The amplification of the target gene and the fluorescence analysis of each well can be sequentially performed in only one device. For this reason, the false-negative probability is low because the effect of carry-over contamination is negligible.

As the coronavirus disease of 2019 (COVID-19) has now spread worldwide, diagnostic equipment capable of point of care testing (POCT) in various diagnostic spaces is required [4,5]. Currently, COVID-19 diagnosis is carried out through qPCR equipment, which is less mobile because it consists of a thermal control device and a light source scanning device. Compared to other PCR systems, the semiconductor cartridge used in the LOAA dPCR is combined with the thermal control unit within the fluorescent sensor part, making it possible to dramatically reduce the equipment size (Fig. 1). These suggest that LOAA dPCR has potential in POCT with advantages of compact and highly accurate absolute quantification.

In 2019, the LOAA dPCR was approved as a class II medical device in South Korea. In addition, it was approved as an emergency use product that Viral Load 20K kit (Optolane) diagnose COVID-19. Here, we describe the principle of the LOAA dPCR, which can be quantified with high accuracy, and the POCT potential of this equipment with various advantages.

Highly Accurate Quantification Principle of the LOAA dPCR

The LOAA dPCR follows the semiconductor plate method using MEMS technology with excellent performance such as small size, light weight, rapid response, and precise measurement [3]. The MEMS structure can be injected by dividing the same amount of target molecules into as few as 56 wells and as much as about 20,000 wells. As the number of wells increases, the size of well is gradually partitioned into smaller volumes (Fig. 2). To verify that LOAA dPCR follows Poisson distribution statistics, we confirmed the four types of Poisson distribution when injecting 2,000 copy genes in 56 wells, 400 wells, 5,000 wells, and 20,000 wells [1]. As a result, we could identify that when the number of wells increases, 99.5% of wells are digitized to 0 or 1 (Fig. 3). In other words, the LOAA dPCR allows absolute quantification because it follows the

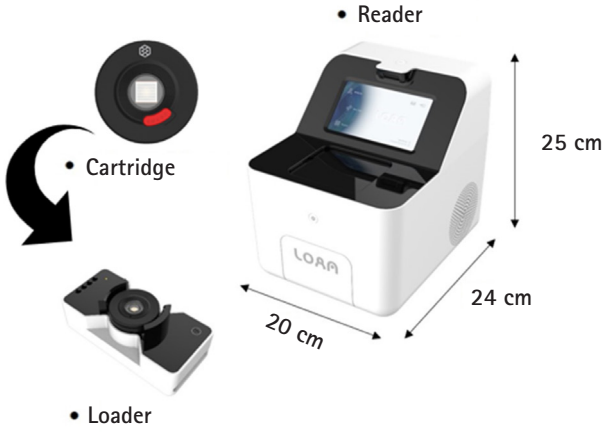

Fig. 1. The components and size of the Lab On An Array (LOAA) digital PCR (dPCR). The LOAA dPCR has a compact size with length 24 cm, width 20 cm, and height 25 cm.

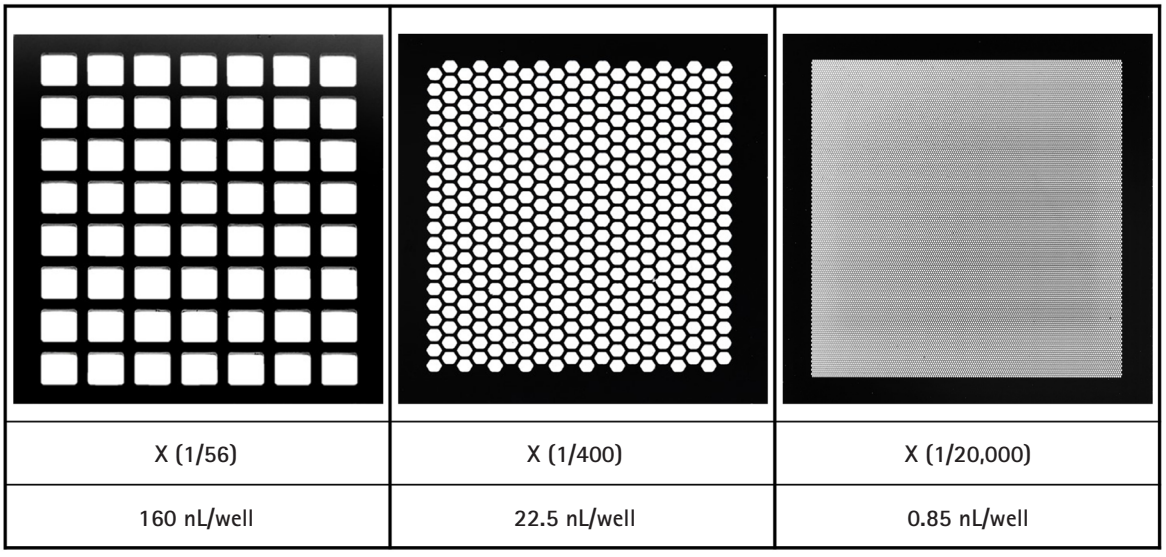

Fig. 2. Micro Electro Mechanical System structures with 56 wells, 400 wells, and 20,000 wells. As the number of the well increases, the size of well gets partitioned into smaller volumes.

https://doi.org/10.5808/gi.21035

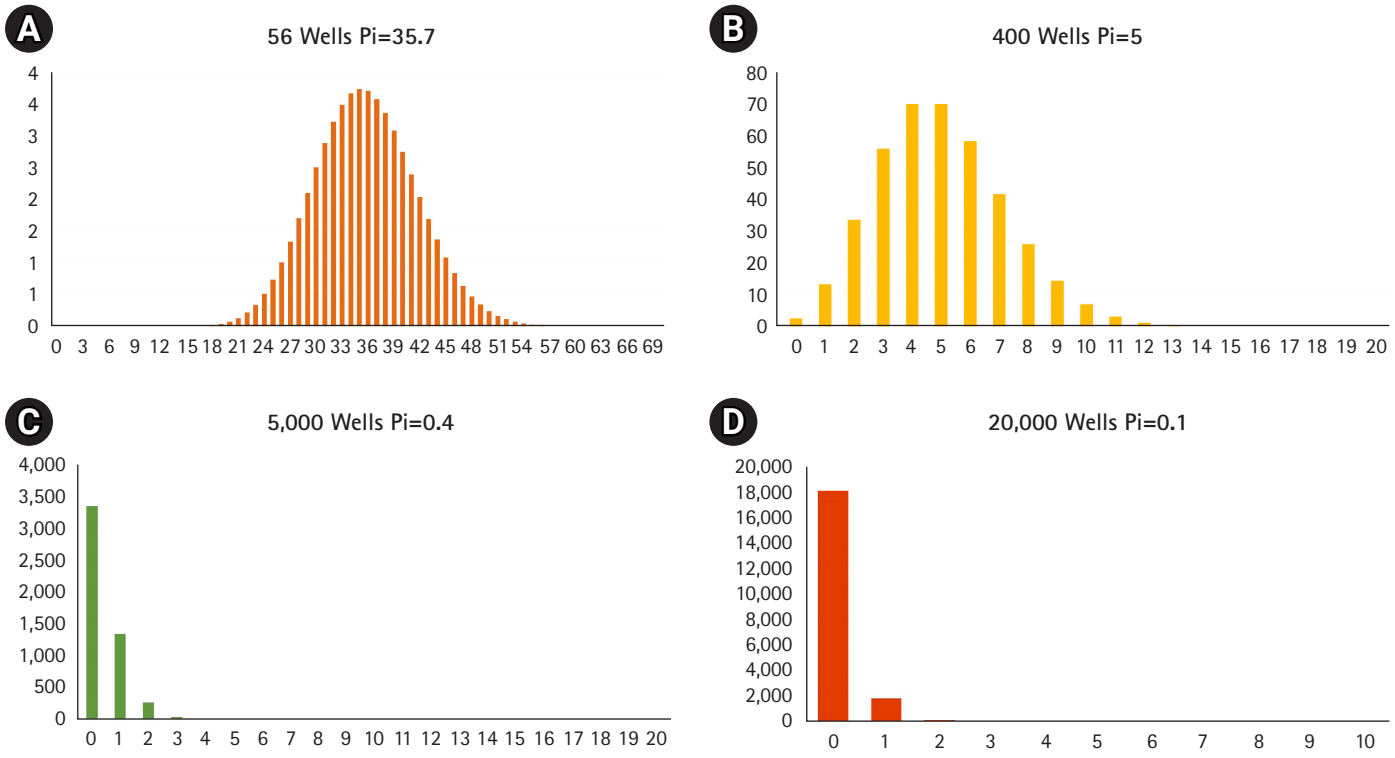

Fig. 3. Poisson probability distribution when injecting 2,000 copy genes in 56 wells (A), 400 wells (B), 5,000 wells (C), and 20,000 wells (D). As the number of well increases from 56 wells to 20,000 wells, 99.5% of wells are digitized into 0 or 1.

Poisson distribution containing either 0 or 1 molecules in the wells divided into the partitions.

Micro-statistics, based on Monte Carlo Simulation [6], which randomly places target molecules with temporal and spatial equivalences, are represented by Bose-Einstein Statistics. Then it becomes equal to a random photon in a sensor array without physical effect. The statistical standard deviation of the target gene based on the mean value, called photon noise, can be represented as Eq. 1.

$$\sigma_{SHOT}(P_I)^2=P_I\ \frac{e_{kT}^{E}}{e_{kT}^{E}-1} \tag{1}$$

P_I means the average number of photons. E is the energy value of the photon. K stands for Boltzmann's constant (1.38×10^{-23} J/K), and T refers to the absolute temperature. When $E \gg kT$, it is expressed as Eq. 2 which is the shot noise characteristic.

$$\sigma_{SHOT}(P_I)=\sqrt{P_I} \tag{2}$$

The shot noise distribution above can be described by the classical Poisson distribution shown below, where p_i represents the probability that there are i target genes per well.

$$p_i=\frac{P_I^i}{i!}\,e^{-PI} \tag{3}$$

The graph would appear in Gaussian distribution form if pi increases as the number of wells decreases (Fig. 3). This means as the pi increases, the standard deviation increases.

$$p_0=\frac{\textit{the number of negative PCR well}}{\textit{the total number of available PCR well}} \tag{4}$$

On the other hand, in the dPCR, p_0 can be calculated using the Poisson probability distribution represented in Eq. 5.

$$p_0=e^{-PI} \tag{5}$$

P_I represents the average number of target molecules injected per well and can be specifically expressed as Eq. 6, according to the experiment definition.

$$p_1=\frac{\textit{the number of target genes in well }(x)}{\textit{the total number of available PCR well }(AW)} \tag{6}$$

Applying Eqs. 4, 5, and 6, p_0, the probability that no target molecule is in the well, can be expressed as Eq. 7.

$$\frac{\textit{the number of negative PCR well}}{\textit{the total number of available PCR well}}=e^{-\frac{x}{AW}} \tag{7}$$

If we calculate x, the first number of target genes in a sample, it can be expressed as Eq. 8.

$$x = -(\text{the number of availabel pcr well}) \times log_e = \left(\frac{\text{the number of target genes in well}}{\text{the total number of available PCR well}}\right) \quad (8)$$

Finally, we can confirm the accuracy of the LOAA dPCR by analyzing the Poisson distribution of Eqs. 2 and 6.

$$\sigma_{SHOT}(P_I) = \sqrt{P_I} = \sqrt{\frac{\text{the number of PCR}}{\text{the total number of available PCR}}} \quad (9)$$

$$\text{the number of target genes in well} \gg \text{the number of available PCR well} \quad (10)$$

When the condition Eq. 10 is met, the standard deviation becomes small. In other words, the accuracy of the LOAA PCR increases when the amount of target gene decreases.

Region of Interest Selection

The first and important step in sample detection prior to PCR is to locate the region of interest within the cartridge divided into 20,000 wells. A way to differentiate between valid wells and invalid wells is to examine Otsu Filter to investigate pixel differences [7]. When there is one channel (1 channel: FAM), well data is formed based on the average value of 12 pixels. On the other hand, if there are two channels (2 channels: FAM/FRET-Cy5), the data is generated using the average of each six samples (Fig. 4).

Invalid Well Filtering Methods

When plotting a histogram using acquired well data after the run ends, the data is generally distributed from 3 σ below to 4 σ above of the mean. For brightness filter, data exceeding two multiples (-6 σ to 8 σ) in 1 to 22 cycles is judged to be invalid (Fig. 5A).

In the entire PCR cycle, the slope of available well data is generally between 20 and 50. Considering the noise data that occurs at the beginning of the run, the data is excluded if the slope exceeds 10 multiples (−200 to 500) (Fig. 5B). In addition, for the baseline filter, invalid wells can be determined by negative wells which have none of the target genes. Negative well cycle data below −7 σ and above 7 σ is treated invalid. Because when the threshold range is narrowed, the noise well and a number of normal wells are considered non-functional, resulting in well loss (Fig. 5C).

Detecting invalid well also involves filtering noise based on the algorithm written in Eq. 11. If the curve is not continuous, it is considered unavailable (Fig. 5D)

$$AccNoise = \frac{\sum cycle((Noise\text{-}Normal)^2)}{total\ cycle} \quad (11)$$

Discussion

Since the COVID-19 pandemic has occurred, many countries have come to require equipment capable of POCT diagnosis. POCT equipment requires two conditions: size and sensitivity. Compared with existing commercialized qPCR and dPCR devices in the diagnostic market, the LOAA dPCR has a relatively compact size. In addition, the LOAA dPCR has twice the detection sensitivity of DNA and three times the sensitivity of RNA compared to Bio-rad equipment, called gold-standard equipment [8]. Therefore, there is a growing interest in compact and accurate diagnostic equipment such as LOAA dPCR.

Accuracy and sensitivity of the LOAA dPCR may contribute to the advancement of GMO tests, drug resistance research, and personalized cancer treatment [9]. LOAA dPCR is a prominent potential candidate for a comparative data method as a reference data in the clinical area by detecting infectious diseases caused by viruses and bacteria [10].

The LOAA dPCR has a compact size due to its structural specificity and has high accuracy based on the Poisson statistics. Based on these advantages, it is expected to be utilized in more various fields.

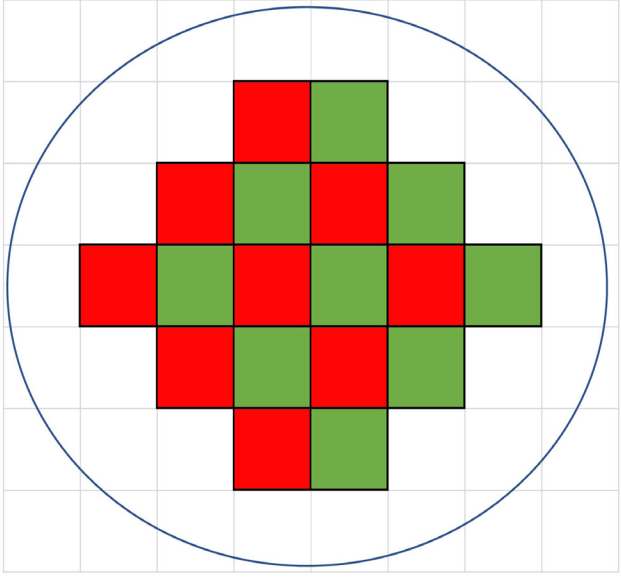

Fig. 4. Well region of interest with two channels. When there are two channels (FAM/FRET-Cy-5), data is generated using an average of six samples.

https://doi.org/10.5808/gi.21035

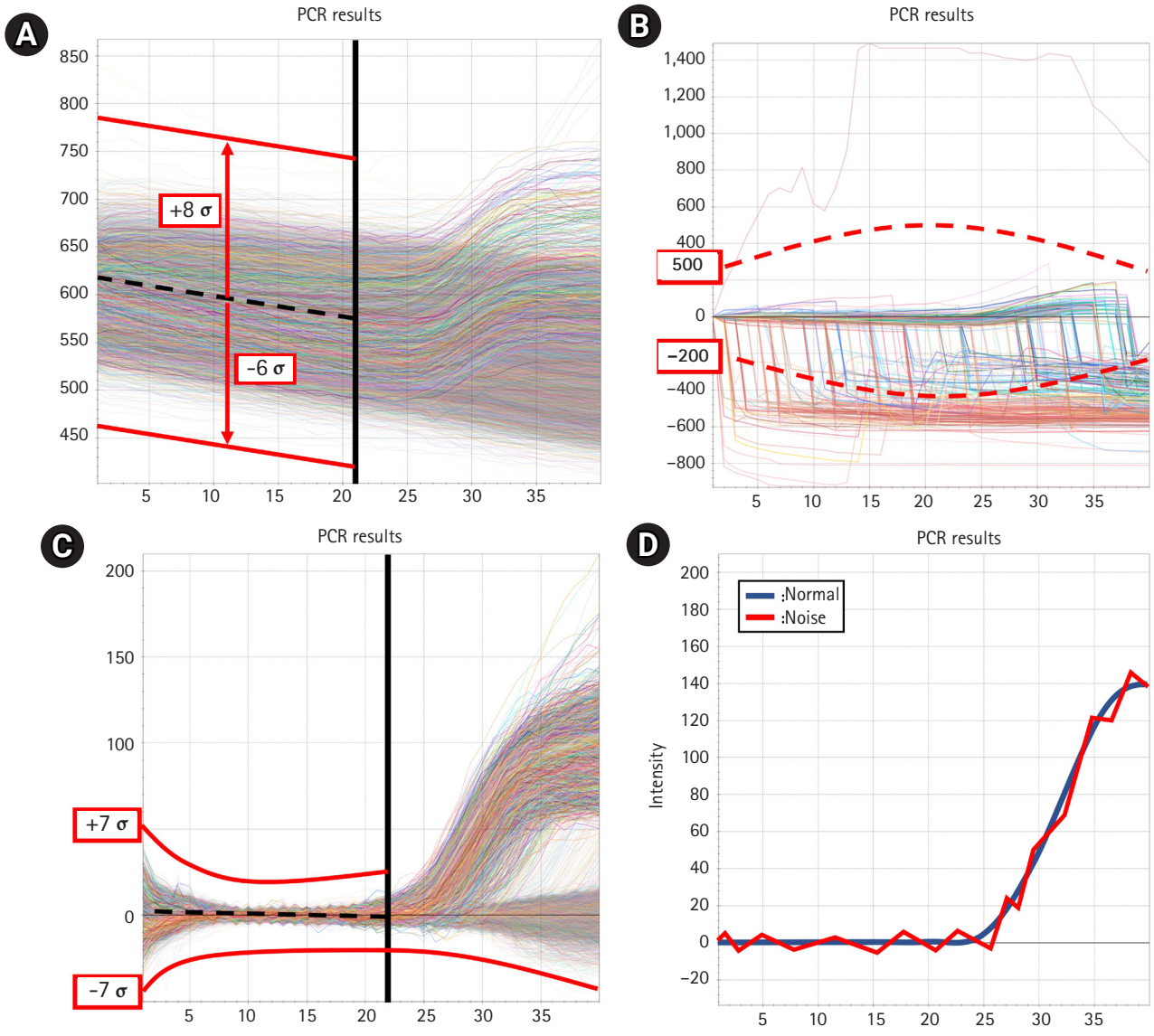

Fig. 5. Filtering methods. (A) In the brightness filter threshold, data exceeding two multiples (−6 σ to 8 σ) within 22 cycles is considered invalid. (B) For the slope filter threshold, data exceeding 10 multiples (−200 to 500) is judged invalid. (C) The baseline filter threshold is based on the negative cycle without the target gene. If the data is less than -7 σ or exceeds 7 σ, it is considered invalid. (D) Normal and noise graph. Depending on the algorithm, when the curve is not continuous, it is judged to be noise.

ORCID

Haeun Lee: https://orcid.org/0000-0002-9415-0557
Cherl-Joon Lee: https://orcid.org/0000-0001-8282-2216
Dong Hee Kim: https://orcid.org/0000-0002-5056-7406
Chun-Sung Cho: https://orcid.org/0000-0001-6077-652X
Wonseok Shin: https://orcid.org/0000-0001-5964-1425
Kyudong Han: https://orcid.org/0000-0001-6791-2408

Authors' Contribution

Conceptualization: KH. Data curation: KH. Formal analysis: HL, CJL, WS, KH. Funding acquisition: KH. Methodology: HL, CJL, WS, KH. Writing - original draft: HL, CJL, WS, KH. Writing - review & editing: HL, CJL, DHK, CSC, WS, KH.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

The Department of Microbiology was supported through the Research-Focused Department Promotion Project as a part of the University Innovation Support Program for Dankook University in 2021. The authors gratefully acknowledge the Center for Bio-Medical Engineering Core Facility at Dankook University for providing valuable reagents and research space. The authors have appreciatively acknowledged the Optolane for providing valuable data.

References

1. Morley AA. Digital PCR: a brief history. Biomol Detect Quantif 2014;1:1-2.
2. Baker M. Digital PCR hits its stride. Nat Methods 2012;9:541-544.
3. Khan MS, Tariq MO, Nawaz M, Ahmed J. MEMS sensors for diagnostics and treatment in the fight against COVID-19 and other pandemics. IEEE Access 2021;9:61123-61149.
4. Gupta N, Augustine S, Narayan T, O'Riordan A, Das A, Kumar D, et al. Point-of-care PCR assays for COVID-19 detection. Biosensors (Basel) 2021;11:141.
5. Keni R, Alexander A, Nayak PG, Mudgal J, Nandakumar K. COVID-19: emergence, spread, possible treatments, and global burden. Front Public Health 2020;8:216.
6. Bonate PL. A brief introduction to Monte Carlo simulation. Clin Pharmacokinet 2001;40:15-22.
7. Talab AM, Huang Z, Xi F, HaiMing L. Detection crack in image using Otsu method and multiple filtering in image processing techniques. Optik 2016;127:1030-1033.
8. Lee SS, Park JH, Bae YK. Comparison of two digital PCR methods for EGFR DNA and SARS-CoV-2 RNA quantification. Clin Chim Acta 2021;521:9-18.
9. Tong Y, Shen S, Jiang H, Chen Z. Application of digital PCR in detecting human diseases associated gene mutation. Cell Physiol Biochem 2017;43:1718-1730.
10. Pavsic J, Devonshire AS, Parkes H, Schimmel H, Foy CA, Karczmarczyk M, et al. Standardization of nucleic acid tests for clinical measurements of bacteria and viruses. J Clin Microbiol 2015;53:2008-2014.

https://doi.org/10.5808/gi.21035

Instructions for authors

Enacted January 2003
Recently revised January 9, 2019

Genomics & Informatics (Genomics Inform) is owned and published by the Korea Genome Organization (KOGO). It is published four times per year (Mar, Jun, Sep, and Dec) in an online version. *Genomics & Informatics* welcomes high-quality research papers presenting novel data on the topics of gene discovery, comparative genome analyses, molecular and human evolution, informatics, genome structure and function, technological innovations and applications, statistical and mathematical methods, cuttingedge genetic and physical mapping, and DNA sequencing and other reports that present data where sequence information is used to address biological concerns. The journal publishes papers based on original research that are judged after editorial review to make a substantial contribution to the understanding of any area of genomics or informatics. Only manuscripts written in English under the *Genomics & Informatics* author guidelines are accepted. *Genomics & Informatics* follows the open access journal policy. All of the content of *Genomics & Informatics* is freely available online. Digital files can be read, downloaded, and printed without charge.

Manuscripts for submission to *Genomics & Informatics* should be prepared according to the following instructions. *Genomics & Informatics* follows the Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals (http://www.icmje.org) from ICMJE and Principles of Transparency and Best Practice in Scholarly Publishing (joint statement by COPE, DOAJ, WAME, and OASPA; (http://doaj.org/bestpractice) if otherwise not described below.

Research and publication ethics

For the policies on research and publication ethics that are not stated in these instructions, the Good Publication Practice Guidelines for Medical Journals (http://kamje.or.kr/intro.php?body=publishing_ethics) and the Guidelines on Good Publication(http://publicationethics.org/resources/guidelines) can be applied. The Editor-in-Chief reserves the right to reject manuscripts that do not comply with the below requirements. The author will be held responsible for false statements or failure to fulfill the below requirements.

Statement of Informed Consent

Copies of written informed consent and Institutional Review Board (IRB) approval for clinical research should be kept. If necessary, the editor or reviewers may request copies of these documents to resolve questions about IRB approval or study conduct.

Statement of Human and Animal Rights

All human investigations must be conducted according to the principles expressed in the Declaration of Helsinki. All studies involving animals must state that the guidelines for the use and care of laboratory animals of the authors' institution, or of any national law, were followed. Registration of clinical trial research: Any research that deals with a clinical trial should be registered with the primary national clinical trial registry site, such as the Korea Clinical Research Information Service (CRiS, http://cris.nih.go.kr), other primary national registry sites accredited by the World Health Organization (http://www.who.int/ictrp/network/primary/en/), or ClinicalTrials.gov (http://clinicaltrials.gov/), a service of the United States National Institutes of Health.

Authorship

Authorship credit should be based on 1) substantial contributions to conception and design, acquisition of data, and/or analysis and interpretation of data; 2) drafting the article or revising it critically for important intellectual content; 3) final approval of the version to be published; and 4) agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Every author should meet all of these four conditions. After the initial submission of a manuscript, any changes whatsoever in authorship (adding author(s), deleting author(s), or re-arranging the order of authors) must be explained by a letter to the editor from the authors concerned. This letter must be signed by all authors of the paper. Copyright assignment must also be completed by every author.

Corresponding author and first author

It does allow multiple corresponding authors for one article. Only one author should correspond with the editorial office. It does accept notice of equal contribution for the first author when the study was clearly performed by co-first authors.

Correction of authorship after publication

It does not correct authorship after publication unless a mistake

has been made by the editorial staff. Authorship may be changed before publication but after submission when an authorship correction is requested by all of the authors involved with the manuscript.

Conflict of Interest Statement

The corresponding author must inform the editor of any potential conflicts of interest that could influence the authors' interpretation of the data. Examples of potential conflicts of interest are financial support from or connections to pharmaceutical companies, political pressure from interest groups, and academically related issues. In particular, all sources of funding applicable to the study should be explicitly stated. As a guideline, any affiliation associated with a payment or financial benefit exceeding $10,000 per annum or 5% ownership of a company or research funding by a company with related interests would constitute a conflict that must be declared. This policy applies to all submitted research manuscripts and review material.

Originality and Duplicate Publication

No part of the accepted manuscript should be duplicated in any other scientific journal without the permission of the Editorial Board. If duplicate publication or plagiarism related to the papers of this journal is detected, the authors will be announced in the journal, their institutes will be informed, and the authors will be penalized. All submitted manuscripts are screened by CrossCheck (Similarity Check), a plagiarism detection program provided by iThenticate. The authors assure that no substantial part of the work has been published or is being considered for publication elsewhere. When any of the results is to appear in another journal, details must be submitted to the Editor-in-Chief, together with a copy of the other paper(s) and the expected date(s) of publication.

Secondary Publication

It is possible to republish manuscripts if the manuscripts satisfy the condition of secondary publication of the Uniform Requirements for Manuscripts Submitted to Biomedical Journals by the International Committee of Medical Journal Editors (ICMJE), available from http://www.icmje.org/. These are:

- The authors have received approval from the editors of both journals (the editor concerned with the secondary publication must have access to the primary version).
- The priority for the primary publication is respected by a publication interval negotiated by editors of both journals and the authors.
- The paper for secondary publication is intended for a different group of readers; an abbreviated version could be sufficient.
- The secondary version faithfully reflects the data and interpretations of the primary version.
- The secondary version informs readers, peers, and documenting agencies that the paper has been published in whole or in part elsewhere—for example, with a note that might read, "This article is based on a study first reported in the [journal title, with full reference]"—and the secondary version cites the primary reference.
- The title of the secondary publication should indicate that it is a secondary publication (complete or abridged republication or translation) of a primary publication. Of note, the United States National Library of Medicine (NLM) does not consider translations to be "republications" and does not cite or index them when the original article was published in a journal that is indexed in MEDLINE.

Process to manage research and publication misconduct: When the Journal faces suspected cases of research and publication misconduct, such as a redundant (duplicate) publication, plagiarism, fabricated data, changes in authorship, undisclosed conflicts of interest, an ethical problem discovered with the submitted manuscript, a reviewer who has appropriated an author's idea or data, complaints against editors, and other issues, the resolving process will follow a flowchart provided by the Committee on Publication Ethics (http://publicationethics.org/resources/flowcharts). The discussion and decision on suspected cases are done by the Editorial Board of *Genomics & Informatics*.

Preparation of manuscripts

General requirement

Authors are recommended to keep the length of papers below 10 printed pages (30 typed pages of manuscript, including figures and tables) for original articles, four printed pages for research communications, and two printed pages (approximately 1,400 words or 1,000 words plus one figure) for application notes. All sections of the typescript should be double-spaced on one side of A4 paper (210 × 297 mm), and all pages must be numbered in order.

Manuscript type

Original articles

Original research articles are full scientific reports of original research. The manuscript should be organized as follows: Title Page, Abstract & Keywords, Introduction, Methods, Results, Discussion, Acknowledgments, References, Tables, and Figure Legends. The Results and Discussion can be combined.

Application notes

Application notes are short communications about novel software, new algorithm implementations, databases, and network services (web servers and interfaces). The manuscripts include the following: Title Page, Abstract & Keywords, Availability, Introduction, Main Text, References, and Supplementary Information.

Clinical genomics

Clinical genomics is for a short report of all kinds of genome analysis data from clinical fields, such as cancer, diverse complex diseases, and genetic diseases. Especially, *Genomics & Informatics* would encourage submitting cancer panel analysis data for a single cancer patient or a group of patients. *Genomics & Informatics* also would encourage depositing genome data into the *Genomics & Informatics* database. The manuscript should be organized as follows: Title Page, Abstract & Keywords, Introduction, Methods, Results, Discussion, Acknowledgments, References, Tables, and Figure Legends. The Introduction, Methods, Results, and Discussion can be combined.

Genome archives

Genome Archives is for a short manuscript announcing the genetic information of recently sequenced prokaryotic and eukaryotic genomes. *Genomics & Informatics* would encourage depositing the genome data into the *Genomics & Informatics* database. These genome archive data can make the rationale for sequencing a specific organism. The manuscripts include the following: Title Page, Abstract & Keywords, Introduction, Main Text, References, Tables, and Figure Legends.

Letters to the editor

Critical comments are welcomed for correcting errors of published facts and for providing alternative interpretations of published data. The sequence for a Letter to the Editor is the following: Title Page, Text, References, and Names and Affiliations of Authors. If needed, tables and figures can be included. A Letter to the Editor should not be longer than a printed page.

Review articles

Review Articles are usually solicited by the Editor-in-Chief. Authors wishing to prepare a review article should contact the Editor-in-Chief to discuss the suitability of the subject for the journal. There is no specific requirement for subsections of the body text of the paper.

Opinions / Commentaries

An opinion or commentary piece is a short article that conveys the author's viewpoint on a research publication, including interpretation of data, value of methods used, and strengths/weaknesses, regarding any topic relevant to the field of research. Opinion (or commentary) articles provide insight, interpretation, and evaluation of specific issues, within the scope of the journal. Opinions should explain the implications of the article and describe the most important conclusions of the paper they are commenting on, highlight controversial issues, mention the strengths and weaknesses of the paper, highlight the presenter's omission of key facts, and mention supporting arguments that would create a stronger presentation. Opinions are relatively short articles, around 1000 words, allowing maximum freedom of authors' viewpoints, and are peer-reviewed. The articles are copyedited, citable, published in both PDF and HTML formats, and submitted for indexing in digital archives (e.g., PubMed Central). Authors are not required to pay a fee to publish an opinion (or commentary) article. Commentaries have no set format beyond the basic building blocks of a regular article, i.e., title, manuscript text, subheadings as needed, references, and author information.

Minireviews

Minireview articles are similar to review articles, except for their word limit and references. Minireviews focus on clearly defined topics of current interest, and recent developments in specific fields. Therefore, they offer a fast and easy means to keep abreast of exciting new developments and/or concepts. The word limit for minireview articles is 1000 words (or 2 double-spaced pages), with no more than 30 references. Minireview articles are peer-reviewed, copyedited, citable, published in both PDF and HTML formats, and submitted for indexing in digital archives, such as PubMed Central. Authors are required to pay a fee to publish a minireview.

Research communications

Research communication (RC) intends to deliver significant scientific discovery with broad interest in a short format. RCs may contain unstructured main text that includes introduction, results and discussion. RCs typically have no more than 2 display items (figures and tables) and the main text (not including abstract, references, tables and figure legends) is limited to 1,500 words. RCs may have online supplementary section.

Manuscript Format

Title

The title page should include (1) the full names of all authors with their Open Researchers and Contributors ID (ORCID), and the name(s) and address(es) of the institution(s) at which the work was carried out; (2) the telephone and fax numbers, and the

E-mail address of the corresponding author; and (3) a running title of no more than 50 characters, including spaces. Place an asterisk (*) after the corresponding author.

Abstract

The abstract should be unstructured and a single paragraph of fewer than 250 words. References should not be cited in the abstract. Six or fewer keywords should be appended to the abstract in alphabetical order. When possible, the keywords should be those found in the Medical Subject Headings of Index Medicus.

Main text:

All papers should be divided into the following sections and appear in this order:

(1) **Introduction:** The paper begins with an introduction without subheadings that reviews the literature and states and justifies the purpose of the research.

(2) **Methods:** This section should contain sufficient detail so that all procedures can be repeated, in conjunction with the cited references. The manufacturer and model number should be stated in this section—for example, as Sigma Chemical Co. (St. Louis, MO, USA).

(3) **Results:** This section should describe the results of the experiments. Extensive interpretation should be reserved for the Discussion section. The results should be presented as concisely as possible. Footnotes should not be used and will be transferred to the text. Gene symbols should be italicized; protein products are not italicized.

(4) **Discussion:** This section should provide an interpretation of the results in relation to previously published work and to the experimental system at hand. The Results and Discussion may be combined.

(5) **Acknowledgments:** Information concerning the sources of financial support should be included in the acknowledgments.

Authors' contribution

If the number of authors is equal to or greater than two, the authors' roles should be described according to their specific role. *Genomics & Informatics* participates in the CRediT standard for author contributions. The contributions of all authors must be described using the CRediT Taxonomy of author roles. For each of the categories below, please enter the initials of the authors who contributed in that category. If listing more than one author in a category, separate each set of initials with a space. If no one contributed in a category, you may leave that box blank. The corresponding author is responsible for completing this information at submission, and it is expected that all authors will have reviewed, discussed, and agreed to their individual contributions ahead of this time.

- Conceptualization: AB
- Data curation: EFG
- Formal analysis: AB
- Funding acquisition: CD
- Methodology: AB, CD, EFG
- Writing – original draft: AB, EFG
- Writing – review & editing: AB, CD, EFG

Reference

The references should include only articles that are published or in press. Unpublished data, submitted manuscripts, abstracts, and personal communications should be cited within the text only. References are to be numbered in the order of citation within the article in brackets. References with up to six authors must list all names; for more than six authors, the first six names should be listed, followed by "et al." Journal name titles should be abbreviated in accordance with the NLM Catalog, available from: https://www.ncbi.nlm.nih.gov/nlmcatalog/journals, or the ISO 4 standard, available from: http://www.issn.org/services/online-services/access-to-the-ltwa/?letter=a.

Examples of references are given below:

Journal article

- Park J, Lappe M, Teichmann SA. Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. J Mol Biol 2001;307:929-938.
- Cho SM, Jung SH, Chung YJ. A variant in RUNX3 is associated with the risk of ankylosing spondylitis in Koreans. Genomics Inform 2017;15:65-68.
- Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 2003;13:2129-2141.

Books

- Cowan WM, Jessell TM, Zipursky SL. Molecular and Cellular Approaches to Neural Development. New York: Oxford University Press, 1997.

Book sections

- Sorenson PW, Caprio JC. Chemoreception. In: The Physiology of Fishes (Evans DH, ed.). Boca Raton: CRC Press, 1998. pp. 375-405.

Online document

- Puniyani AR, Lukose RM. Growing random networks under

constraints. Ithaca: Cornell University Library, 2001. Accessed 2011 Oct 3. Available from: http://xxx.lanl.gov/abs/cond-mat/0107391.

Conference paper

- Han H. Nonnegative principle component analysis for mass spectral serum profiles and biomarker discovery. In: The 8th Asia-Pacific Bioinformatics Conference (Parida L, Myers G, eds.), 2010 Jan 18-21, Bangalore.

Dissertation/Thesis

- Hwang KB. Hierarchical probabilistic graphical models for large-scale data analysis. Ph.D. Dissertation. Seoul: Seoul National University, 2005.

Tables and figures

Figure legends and tables should be included in the submitted manuscript as separate sections and should be formatted following the style of the journal. Each figure legend should have a brief, separate title that describes the entire figure without citing specific panels. The manuscript should be submitted with a set of figures of sufficient quality for reviewers to judge the data. All figures may be provided in color for the electronic version of the journal, even if the print version is in black and white. Figures will be printed in color only when in the reviewers' opinions the color is essential.

Photographs and illustrations should be of professional quality. Images should be provided as TIFF files. JPEG is also acceptable when the original format is JPEG. Each figure must be of 300 dpi or higher resolution with good contrast and sharpness. If a figure is to be reduced, all elements, including labels, should be able to withstand reduction and remain legible. Electron and light microscopic figures must be original or scanned copies from the original. The magnification should be indicated on each micrograph with a scale bar.

Tables are to be organized in portrait view and may run, if necessary, to subsequent pages in the vertical direction only. Tables should be designed for printing within two (17.5 cm) columns of width in no less than 10-point font and should not exceed more than the width of a journal page. If a table does not fit into this format, consider shortening row or column labels, using more than one table to display the data, eliminating unnecessary data, or converting table data into a figure or transferring part of the table data to the supplement.

Scientific names

The full formal Latin name for a taxon (e.g., Homo sapiens) should be provided the first time that the taxon is mentioned and should be italicized. In subsequent sentences, the scientific name of all taxa in the same genus should be abbreviated to the first initial of the generic name and the species name (e.g., H. sapiens), except where this usage creates confusion or ambiguity. When common names are used, the scientific name should be provided the first time the taxon is mentioned in the abstract and again the first time that taxon is mentioned in the main manuscript [e.g., "red pine (Pinus densiflora)..."]. Other taxonomic designations (e.g., family names) should not be italicized, and common names should not be capitalized.

Units and equations

Standard metric units should be used for describing length, height, weight, and volume. The unit of temperature is given in degrees Celsius (°C). All others are in terms of the International System of Units (SI). All unit symbols must be preceded by one space except percentage (%) and temperature (°C). All equations should be numbered in Arabic numerals.

Abbreviations

Abbreviations must be used as an aid to the reader, rather than as a convenience of the author, and therefore, their use should be limited. Generally, avoid abbreviations that are used less than 3 times in the text, including the tables and figure legends. In addition to abbreviations for SI units, common molecular, chemical, immunological, and hematological terms can be used without definition in the title, abstract, text, tables, and figure legends—e.g., bp, kb, kDa, DNA, cDNA, RNA, mRNA, and PCR. Other common abbreviations are as follows (the same abbreviations are used for plural forms): h (hour; use 0-24:00 h for time), s (second), min (minute), day (not abbreviated), week (not abbreviated), month (not abbreviated), year (not abbreviated), L (liter), mL (milliliter), μL (microliter), g (gram), kg (kilogram), mg (milligram), μg (microgram), ng (nanogram), pg (picogram), g (gravity; not ×g), n (sample size), SD (standard deviation of the mean), and SE (standard error of the mean).

Supplementary materials

Supplementary materials can be provided to support and enhance scientific information. Supplementary files offer additional possibilities for publishing supporting applications, sequence alignment, background datasets, microarray hybridization experiments, high-resolution images, movies, sound clips, and more. Supplementary files will be published alongside the online version of the article on the *Genomics & Informatics* web site. This material will not be edited or formatted; thus, the authors are responsible for the accuracy and presentation of all such material. Accepted file formats for supplementary materials:

- Quick Time files (.mov)

- Graphical image files (.gif)
- HTML files (.html)
- MPEG movie files (.mpg)
- JPEG image files (.jpg)
- Sound files (.wav)
- Plain ASCII text (.txt)
- Acrobat files (.pdf)
- MS Word documents (.doc)
- Postscript files (.ps)
- MS Excel spreadsheet documents (.xls)
- PowerPoint (.ppt)
- TeX and LaTeX

File sizes must be as small as possible, for quick downloading. Recommended specifics are:

- Videos
  - File size: <150 MB
  - Frame rate: 30 frames per second
  - Field order: none (progressive, not interlaced)
  - Aspect ratio: widescreen 16:9
  - Video codec: H.264
  - Video bitrate: 2 Mbps
  - Audio codec: AAC
  - Audio bitrate: 128 kbps
- Images
  - Frame size: 300 dpi in resolution
  - Frame rate: 300 dpi in resolution and 10-15cm in width

Please seek advice from the editorial office before sending files larger than our recommended size to avoid delays in publication.

Accession numbers

Please provide accession numbers for any new data (SNPs, gene sequences, protein sequences, CNVs, microarray data, or structures), which must be deposited in the appropriate genome- or locusspecific database, in a separate section entitled "Accession Numbers," following the Web Resources section (or the Acknowledgments section if no online resources or appendices have been used), directly above the reference list. Please use the following format to list accession numbers: "The accession number(s) for the _____ sequence(s) reported in this paper is/are [database]: [accession number]."

Gender equity (Described according to ICMJE recommendation available from

http://www.icmje.org/recommendations/browse/manuscript-preparation/preparing-for-submission.html)

Selection and Description of Participants

Clearly describe the selection of observational or experimental participants (healthy individuals or patients, including controls), including eligibility and exclusion criteria and a description of the source population. Because the relevance of such variables as age, sex, or ethnicity is not always known at the time of study design, researchers should aim for inclusion of representative populations into all study types and at a minimum provide descriptive data for these and other relevant demographic variables. Ensure correct use of the terms sex (when reporting biological factors) and gender (identity, psychosocial or cultural factors), and, unless inappropriate, report the sex and/or gender of study participants, the sex of animals or cells, and describe the methods used to determine sex and gender. If the study was done involving an exclusive population, for example in only one sex, authors should justify why, except in obvious cases, (e.g., prostate cancer)." Authors should define how they determined race or ethnicity and justify their relevance.

Submission of Manuscript

The manuscript should be submitted in MS Word file format. The recommended font is Times New Roman with a 11-point font size. All manuscripts must be submitted online through the *Genomics & Informatics* e-submission system at http://submit.genominfo.org. Any questions concerning manuscript submission should be directed to: Editor, *Genomics & Informatics*, Korea Genome Organization, The Korean Federation of Science and Technology Societies, Room No. 806, 193 Mallijae-ro, Jung-gu, Seoul 04501, Korea (http://www.kogo.or.kr, Tel: +82-2-558-9394, Fax: +82-2-558-9434, E-mail: kogo@kogo.or.kr).

Peer review and revision of manuscripts

Peer review

A manuscript is generally reviewed by at least two peer reviewers qualified to evaluate the manuscript. It is a single blind peer review. An initial decision will normally be made within one month of receipt of a manuscript. A manuscript that has been published or of which a substantial portion has been published elsewhere will not be accepted. The Editor-in-Chief is responsible for final decisions regarding the acceptance of a peer-reviewed paper.

Manuscript revision

When a manuscript is returned to the corresponding author for revision, the reviewed manuscript must be re-submitted within one month, unless the authors request an extension. A galley proof

and reprint order form will be sent to the corresponding author. The corresponding author is responsible for communicating with the other authors about revisions and final approval of the proofs. The first proofreading is the author's responsibility, and the proof should be returned within three days from the date of receipt.

Copyrights, open access policy and open data policy

Copyright

The regulations for acceptance of a manuscript for publication automatically include the consent of the author(s) to transfer the copyright or license to KOGO. Authors should complete a Copyright Agreement Form (CAF) at the time of proofreading. The corresponding author can sign on behalf of any co-authors. The CAF can be obtained from the editorial office. Acceptance of the agreement will ensure full copyright protection and help to disseminate the article to the widest possible readership in print and electronic formats. The authors are responsible for obtaining permission to reproduce copyrighted material from other sources

Open access policy

Genomics & Informatics is an open access journal. Articles are distributed under the terms of the Creative Commons Attribution license (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited for non-commercial purposes. To use the tables or figures of *Genomics & Informatics* in other periodicals, books, or media for scholarly, educational, or even commercial purposes, the process of permission request to the Publisher is not necessary. This is in accordance with the Budapest Open Access Initiative definition of open access. It also follows the open access policy of PubMed Central at the United States National Library of Medicine (http://www.ncbi.nlm.nih.gov/pmc/). All of the content of the journal is available immediately upon publication without an embargo period.

Archiving policy

It is accessible without barrier from Korea Citation Index (https://kci.go.kr), National Library of Korea (http://nl.go.kr), or PubMed Central (https://www.ncbi.nlm.nih.gov/pmc/journals/1928/) in the event a journal is no longer published.

Deposit policy (Self-archiving policy) according to Sherpa/Romeo

(http://www.sherpa.ac.uk/): Author can not archive pre-print (i.e., pre-refereeing). Author can archive post-print (i.e., final draft post-refereeing). Author can archive publisher's version/PDF.

Open data policy

Data sharing is recommended. If the data are already public, the URL site or sources should be disclosed. If data can not be publicized, it can be negotiated with the editor. If there are any inquiries on depositing data, author s should contact the editorial office.

Clinical data sharing policy

This journal follows the data sharing policy described in "Data Sharing Statements for Clinical Trials: A Requirement of the International Committee of Medical Journal Editors" (https://doi.org/10.3346/jkms.2017.32.7.1051). As of July 1, 2018, manuscripts submitted to ICMJE journals that report the results of clinical trials must contain a data sharing statement as described below. Clinical trials that begin enrolling participants on or after January 1, 2019 must include a data sharing plan in the trial's registration. The ICMJE's policy regarding trial registration is explained at www.icmje.org/recommendations/browse/publishingand-editorial-issues/clinical-trial-registration.html. If the data sharing plan changes after registration, this should be reflected in the statement submitted and published with the manuscript and updated in the registry record. Data sharing statements must indicate the following: whether individual deidentified participant data (including data dictionaries) will be shared; what data in particular will be shared; whether additional, related documents will be available (e.g., study protocol, statistical analysis plan, etc.); and when the data will become available and for how long; by what access criteria data will be shared (including with whom, for what types of analyses, and by what mechanism). Illustrative examples of data sharing statements that would meet these requirements are in Table 1.

Detailed Description of Use of Articles of *Genomics & Informatics*

Reader benefit

Publisher applies the Creative Commons Attribution Non-Commercial license to works it publishes and allows free immediate access to, and unrestricted reuse of, original works of all types.

Reuse benefit

Publisher applies the Creative Commons Attribution Non-Commercial license to works it publishes and allows free immediate access to, and unrestricted reuse of, original works of all types.

Copyrights

Publisher applies the Creative Commons Attribution Non-

Table 1. Examples of data sharing statements that fulfill ICMJE requirements[a]

Element	Example 1	Example 2	Example 3	Example 4
Will individual participant data be available (including data dictionaries)?	Yes	Yes	Yes	No
What data in particular will be shared?	All of the individual participant data collected during the trial, after deidentification.	Individual participant data that underlie the results reported in this article, after deidentification (text, tables, figures, and appendices).	Individual participant data that underlie the results reported in this article, after deidentification (text, tables, figures, and appendices).	Not available
What other documents will be available?	Study protocol, statistical analysis plan, informed consent form, clinical study report, analytic code	Study protocol, statistical analysis plan, analytic code	Study protocol	Not available
When will data be available (start and end dates)?	Immediately following publication. No end date.	Beginning 3 months and ending 5 years following article publication.	Beginning 9 months and ending 36 months following article publication.	Not applicable
With whom?	Anyone who wishes to access the data.	Researchers who provide a methodologically sound proposal.	Investigators whose proposed use of the data has been approved by an independent review committee ("learned intermediary") identified for this purpose.	Not applicable
For what types of analyses?	Any purpose	To achieve aims in the approved proposal.	For individual participant data meta-analysis.	Not applicable
By what mechanism will data be made available?	Data are available indefinitely at (link to be included).	Proposals should be directed to xxx@yyy. To gain access, data requestors will need to sign a data access agreement. Data are available for 5 years at a third-party website (link to be included).	Proposals may be submitted up to 36 months following article publication. After 36 months, the data will be available in our University's data warehouse but without investigator support other than deposited metadata. Information regarding submitting proposals and accessing data may be found at (link to be provided).	Not applicable

ICMJE, International Committee of Medical Journal Editors.
[a]These examples are meant to illustrate a range of, but not all, data sharing options.

Article processing charge

Neither page charge, article processing fee nor submission fee will be applied since 2019. It is the platinum open access journal

Contact address

Editorial office

Room No. 806, 193 Mallijae-ro, Jung-gu, Seoul 04501, Korea
Tel: +82-2-558-9394
Fax: +82-2-558-9434
E-mail: kogo3@kogo.or.kr

Copyright transfer agreement

The copyright to this article is transferred to Genomics & Informatics effective if and when the article is accepted for publication. The author warrants that his/her contribution is original and that he/she has full power to make this grant. The author signs for and accepts responsibility for releasing this material on behalf of any and all co-authors. The copyright transfer covers the exclusive right to reproduce and distribute the article, including reprints, translations, photographic reproductions, microform, electronic form (offline, online) or any other reproductions of similar nature.

According to the deposit policy (self-archiving policy) of Sherpa/Romeo (http://www.sherpa.ac.uk), authors cannot archive pre-print (i.e. pre-refereeing), but they can archive post-print (i.e. final draft post-refereeing). Authors can archive publisher's version/PDF.

Title of article	
Author(s)	
Author's signature	
Date	

Taesung Park
Editor in Chief
Genomics & Informatics
Korea Genome Organization (KOGO)

Publication ethics

For the policies on research and publication ethics that are not stated in these instructions, the Good Publication Practice Guidelines for Medical Journals (http://kamje.or.kr/intro.php?body=publishing_ethics) and the Guidelines on Good Publication(http://publicationethics.org/resources/guidelines) can be applied. The Editor-in-Chief reserves the right to reject manuscripts that do not comply with the below requirements. The author will be held responsible for false statements or failure to fulfill the below requirements.

Statement of Informed Consent

Copies of written informed consent and Institutional Review Board (IRB) approval for clinical research should be kept. If necessary, the editor or reviewers may request copies of these documents to resolve questions about IRB approval or study conduct.

Statement of Human and Animal Rights

All human investigations must be conducted according to the principles expressed in the Declaration of Helsinki. All studies involving animals must state that the guidelines for the use and care of laboratory animals of the authors' institution, or of any national law, were followed. Registration of clinical trial research: Any research that deals with a clinical trial should be registered with the primary national clinical trial registry site, such as the Korea Clinical Research Information Service (CRiS, http://cris.nih.go.kr), other primary national registry sites accredited by the World Health Organization (http://www.who.int/ictrp/network/primary/en/), or ClinicalTrials.gov (http://clinicaltrials.gov/), a service of the United States National Institutes of Health.

Authorship

Authorship credit should be based on 1) substantial contributions to conception and design, acquisition of data, and/or analysis and interpretation of data; 2) drafting the article or revising it critically for important intellectual content; 3) final approval of the version to be published; and 4) agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Every author should meet all of these four conditions.

After the initial submission of a manuscript, any changes whatsoever in authorship (adding author(s), deleting author(s), or re-arranging the order of authors) must be explained by a letter to the editor from the authors concerned. This letter must be signed by all authors of the paper. Copyright assignment must also be completed by every author.

Corresponding author and first author

It does allow multiple corresponding authors for one article. Only one author should correspond with the editorial office. It does accept notice of equal contribution for the first author when the study was clearly performed by co-first authors.

Correction of authorship after publication

It does not correct authorship after publication unless a mistake has been made by the editorial staff. Authorship may be changed before publication but after submission when an authorship correction is requested by all of the authors involved with the manuscript.

Conflict of Interest Statement

The corresponding author must inform the editor of any potential conflicts of interest that could influence the authors' interpretation of the data. Examples of potential conflicts of interest are financial support from or connections to pharmaceutical companies, political pressure from interest groups, and academically related issues. In particular, all sources of funding applicable to the study should be explicitly stated. As a guideline, any affiliation associated with a payment or financial benefit exceeding $10,000 per annum or 5% ownership of a company or research funding by a company with related interests would constitute a conflict that must be declared. This policy applies to all submitted research manuscripts and review material.

Originality and Duplicate Publication

No part of the accepted manuscript should be duplicated in any other scientific journal without the permission of the Editorial Board. If duplicate publication or plagiarism related to the papers of this journal is detected, the authors will be announced in the journal, their institutes will be informed, and the authors will be penalized. All submitted manuscripts are screened by CrossCheck

(Similarity Check), a plagiarism detection program provided by iThenticate. The authors assure that no substantial part of the work has been published or is being considered for publication elsewhere. When any of the results is to appear in another journal, details must be submitted to the Editor-in-Chief, together with a copy of the other paper(s) and the expected date(s) of publication.

Secondary Publication

It is possible to republish manuscripts if the manuscripts satisfy the condition of secondary publication of the Uniform Requirements for Manuscripts Submitted to Biomedical Journals by the International Committee of Medical Journal Editors (ICMJE), available from http://www.icmje.org/. These are:

- The authors have received approval from the editors of both journals (the editor concerned with the secondary publication must have access to the primary version).
- The priority for the primary publication is respected by a publication interval negotiated by editors of both journals and the authors.
- The paper for secondary publication is intended for a different group of readers; an abbreviated version could be sufficient.
- The secondary version faithfully reflects the data and interpretations of the primary version.
- The secondary version informs readers, peers, and documenting agencies that the paper has been published in whole or in part elsewhere—for example, with a note that might read, "This article is based on a study first reported in the [journal title, with full reference]"—and the secondary version cites the primary reference.
- The title of the secondary publication should indicate that it is a secondary publication (complete or abridged republication or translation) of a primary publication. Of note, the United States National Library of Medicine (NLM) does not consider translations to be "republications" and does not cite or index them when the original article was published in a journal that is indexed in MEDLINE.

Process to manage research and publication misconduct: When the Journal faces suspected cases of research and publication misconduct, such as a redundant (duplicate) publication, plagiarism, fabricated data, changes in authorship, undisclosed conflicts of interest, an ethical problem discovered with the submitted manuscript, a reviewer who has appropriated an author's idea or data, complaints against editors, and other issues, the resolving process will follow a flowchart provided by the Committee on Publication Ethics (http://publicationethics.org/resources/flowcharts). The discussion and decision on suspected cases are done by the Editorial Board of Genomics & Informatics.

Author's checklist

☐ 1. Typed double-spaced with 12-point font in Times New Roman font on A4 sized paper and prepared with an MS-word file.

☐ 2. Title page: (1) complete title, (2) manuscript type, (3) authors' name, (4) affiliations, (5) telephone, facsimle and E-mail address of corresponding author, (6) running title (no more than 50 characters).

☐ 3. Abstract in unstructured format within 250 words.

☐ 4. Six or fewer keywords, preferably MeSH terms.

☐ 5. Manuscript is structured as follows:
Original Article: Abstract, Keywords, Introduction, Materials and Methods, Results, Discussion, References, Table and Figure.
Research Communication: Abstract, Keywords, Main Text, and Conclusion (if applicable), References, Table and Figure.
Application Note: Abstract, Keywords, Availability, Introduction, Main Text, and Supplementary Information, References, Table and Figure.

☐ 6. Reference in proper format. Check that all references listed in the references section are cited in the text and vice versa.

☐ 7. All figures and tables referenced in the text and numbered in order of its appearance in the text.

☐ 8. Figures as a separate files, in TIFF or JPG format, minimum 300 dpi.

☐ 9. Each necessary permission statement signed by the appropriate source.

☐ 10. Elucidation of research or project support/funding.