Cell Genomics, Volume *4*

Supplemental information

The Mutographs biorepository:

A unique genomic resource

to study cancer around the world

Sandra Perdomo, Behnoush Abedi-Ardekani, Ana Carolina de Carvalho, Aida Ferreiro-Iglesias, Valérie Gaborieau, Thomas Cattiaux, Hélène Renard, Priscilia Chopard, Christine Carreira, Andreea Spanu, Arash Nikmanesh, Ricardo Cortez Cardoso Penha, Samuel O. Antwi, Patricia Ashton-Prolla, Cristina Canova, Taned Chitapanarux, Riley Cox, Maria Paula Curado, José Carlos de Oliveira, Charles Dzamalala, Elenora Fabianova, Lorenzo Ferri, Rebecca Fitzgerald, Lenka Foretova, Steven Gallinger, Alisa M. Goldstein, Ivana Holcatova, Antonio Huertas, Vladimir Janout, Sonata Jarmalaite, Radka Kaneva, Luiz Paulo Kowalski, Tomislav Kulis, Pagona Lagiou, Jolanta Lissowska, Reza Malekzadeh, Dana Mates, Valerie McCorrmack, Diana Menya, Sharayu Mhatre, Blandina Theophil Mmbaga, André de Moricz, Péter Nyirády, Miodrag Ognjanovic, Kyriaki Papadopoulou, Jerry Polesel, Mark P. Purdue, Stefan Rascu, Lidia Maria Rebolho Batista, Rui Manuel Reis, Luis Felipe Ribeiro Pinto, Paula A. Rodríguez-Urrego, Surasak Sangkhathat, Suleeporn Sangrajrang, Tatsuhiro Shibata, Eduard Stakhovsky, Beata Świątkowska, Carlos Vaccaro, Jose Roberto Vasconcelos de Podesta, Naveen S. Vasudev, Marta Vilensky, Jonathan Yeung, David Zaridze, Kazem Zendehdel, Ghislaine Scelo, Estelle Chanudet, Jingwei Wang, Stephen Fitzgerald, Calli Latimer, Sarah Moody, Laura Humphreys, Ludmil B. Alexandrov, Michael R. Stratton, and Paul Brennan

The Mutographs biorepository: A unique genomic resource to study cancer around the world

Sandra Perdomo[1‡§], Behnoush Abedi-Ardekani[1‡], Ana Carolina de Carvalho[1‡], Aida Ferreiro-Iglesias[1‡], Valérie Gaborieau[1], Thomas Cattiaux[1], Hélène Renard[1], Priscilia Chopard[1], Christine Carreira[2], Andreea Spanu[1], Arash Nikmanesh[1], Ricardo Cortez Cardoso Penha[1], Samuel O. Antwi[3,4], Patricia Ashton-Prolla[5,6], Cristina Canova[7], Taned Chitapanarux[8], Riley Cox[9], Maria Paula Curado[10], José Carlos de Oliveira[11], Charles Dzamalala[12], Elenora Fabianova[13], Lorenzo Ferri[14], Rebecca Fitzgerald[15], Lenka Foretova[16], Steven Gallinger[17], Alisa M. Goldstein[18], Ivana Holcatova[19,20], Antonio Huertas[21], Vladimir Janout[22], Sonata Jarmalaite[23,24], Radka Kaneva[25], Luiz Paulo Kowalski[10,26], Tomislav Kulis[27,28], Pagona Lagiou[29], Jolanta Lissowska[30], Reza Malekzadeh[31], Dana Mates[32], Valerie McCorrmack[33], Diana Menya[34], Sharayu Mhatre[35], Blandina Theophil Mmbaga[36], André de Moricz[37], Péter Nyirády[38], Miodrag Ognjanovic[39], Kyriaki Papadopoulou[40], Jerry Polesel[41], Mark P. Purdue[42], Stefan Rascu[43], Lidia Maria Rebolho Batista[44], Rui Manuel Reis[44,45], Luis Felipe Ribeiro Pinto[46], Paula A. Rodríguez-Urrego[47], Surasak Sangkhathat[48], Suleeporn Sangrajrang[49], Tatsuhiro Shibata[50,51], Eduard Stakhovsky[52], Beata Świątkowska[53], Carlos Vaccaro[54], Jose Roberto Vasconcelos de Podesta[55], Naveen S. Vasudev[56], Marta Vilensky[57], Jonathan Yeung[58], David Zaridze[59], Kazem Zendehdel[60], Ghislaine Scelo[61], Estelle Chanudet[62], Jingwei Wang[63], Stephen Fitzgerald[63], Calli Latimer[63], Sarah Moody[63], Laura Humphreys[63], Ludmil B. Alexandrov[64,65,66], Michael R. Stratton[63], Paul Brennan[1*]. On behalf of the Mutographs Study

Summary

Initial submission: Received : 7/17/2023

Scientific editor: Laura Zahn

First round of review: Number of reviewers: 2
Revision invited : 8/23/2023
Revision received : 10/24/2023

Second round of review: Number of reviewers: 2
Accepted : 1/10/2023

Data freely available: Yes

Code freely available: Yes

This transparent peer review record is not systematically proofread, type-set, or edited. Special characters, formatting, and equations may fail to render properly. Standard procedural text within the editor's letters has been deleted for the sake of brevity, but all official correspondence specific to the manuscript has been preserved.

Referees' reports, first round of review

Reviewer #1: The article focuses on a description of the Mutographs Grand Challenge, funded by Cancer Research UK. The Mutographs project aims at collecting thousands of cancer whole genomes, as well as patient information, in 30 countries across the world. This is an important project and first one of its kind, which will with no doubt improve our understanding of the mutagenic and environmental causes of cancer. Other cancer sequencing efforts are often limited to one or very few countries involved, and the patient information tends to be incomplete.

Mutographs tries to overcome these limitations by coordinating sequencing and efforts and patient data collection and harmonisation across several countries.

While the article reads well and does a good job to motivate the Mutographs effort, most of the article, especially from line 138, feels to me like a long materials and methods section. I believe that the article would benefit from reducing the description of how the samples and the data are collected, and from adding more information pertinent to the Perspective style. Some examples could be more description and details of past and current efforts and their findings and limitations, or looking ahead perhaps more at the potential impacts or at need and the obstacles to bring efforts like the Mutographs to even more countries, even those that lag behind in technology and investments.

In conclusion, I believe the article requires some work to make it a more informative and enjoyable read.

Reviewer #2: This project presents an unprecedented data source for cancer research, offering a comprehensive collection of phenotypic information across various research centers. However, there is room for improvement especially in genetic data generation and processing. Here are my comments.
1. On page 3, regarding the questions asked by the journal, I believe the rationales of both questions are "Yes". It is crucial to provide the analysis code for Whole Genome Sequencing (WGS) results, including, at a minimum, information about the software version and parameters employed. It's also noteworthy that this project aims to generate a substantial, large-scale dataset.
2. How does this dataset correlate with existing datasets in ICGC or TCGA? Is there any overlap, particularly since a significant portion of the data were from aggregating data from pre-existing sources like data pools and biorepositories? Could this imply an overlap with previous patient cases? Furthermore, could you clarify the concept of pooling data from existing sources and biorepositories? Does this signify that cases were sourced from established databases or biobanks? A verification of genetic relatedness would be essential.
3. From lines 217 to 222, there are some abbreviations in the text that do not appear in Table 1, and vice versa. For enhanced clarity, aligning the information consistently would be beneficial.
4. In the same section (lines 217-222), why are only these three cancer sites mentioned? Does this imply that PCAWG lacks exposure information for other cancer types? Is there a specific rationale behind highlighting the place of residence?
5. Regarding lines 255-256, it's imperative to describe the sequencing protocol employed. If the protocol aligns with reference 38, it's important to acknowledge that the sequencing depth is notably lower than that of PCAWG (https://www.nature.com/articles/s41586-020-1969-6#Sec14). Additionally, differences in sequence read length need clarification. Providing a concise overview of the data analysis pipeline, including software and version details, is advisable.
6. Within Table 1, the presence of samples with unknown gender raises questions. Given that gender can be determined from genetic data, could you elucidate the reason behind this discrepancy?

Authors' response to the first round of review

We thank the reviewers and the editor for your valuable time and for the enriching feedback that undoubtedly will improve the quality of this article titled "The Mutographs biorepository: A unique genomic resource to study cancer around the world". In response to the reviewers' and Editor's comments, we have made the following changes to the manuscript:

Reviewers' Comments:
Reviewer #1:
While the article reads well and does a good job to motivate the Mutographs effort, most of the article, especially from line 138, feels to me like a long materials and methods section. I believe that the article would benefit from reducing the description of how the samples and the data are collected, and from adding more information pertinent to the

Perspective style. Some examples could be more description and details of past and current efforts and their findings and limitations, or looking ahead perhaps more at the potential impacts or at need and the obstacles to bring efforts like the Mutographs to even more countries, even those that lag behind in technology and investments.
In conclusion, I believe the article requires some work to make it a more informative and enjoyable read.

We agree with the reviewer's suggestion, and we restructured the manuscript to include more information that could emphasize the uniqueness of the Mutographs study collection, the opportunities to extend and/or develop similar initiatives in other countries and the lessons learned during the development of this biorepository.

-We have included a new section explaining the overall rationale of the Mutographs study and more details explaining why this is an example of how genomic epidemiology studies in comparison to classical epidemiological studies constitute better approaches to identify new causes of cancer globally (lines 34-53).
-We included additional examples (line 332) of future uses of the samples and analyses of the data generated by the Mutographs study.
-A final section (lines 381-456) highlights some of the fundamental aspects that contributed to the creation of this large-scale cancer biorepository and that could be used in future similar initiatives

Reviewer #2:
1. On page 3, regarding the questions asked by the journal, I believe the rationales of both questions are "Yes". It is crucial to provide the analysis code for Whole Genome Sequencing (WGS) results, including, at a minimum, information about the software version and parameters employed. It's also noteworthy that this project aims to generate a substantial, large-scale dataset.

We have included additional information as part of the Data repository and sharing section (lines 307-327). We incorporated the study reference for the data submitted to EGA, references to the sequencing pipelines already published and the links to bioinformatic repositories which include the algorithms, software versions and codes used in the analyses.

2. How does this dataset correlate with existing datasets in ICGC or TCGA? Is there any overlap, particularly since a significant portion of the data were from aggregating data from pre-existing sources like data pools and biorepositories? Could this imply an overlap with previous patient cases? Furthermore, could you clarify the concept of pooling data from existing sources and biorepositories? Does this signify that cases were sourced from established databases or biobanks? A verification of genetic relatedness would be essential.

We have clarified this point in the description of methods (lines 159-162). The retrospective studies and biorepository collections contributing to Mutographs have been selected from among those that have not been previously analysed or included in other large international genomic projects, mainly ICGC and TCGA. Therefore, the sequencing data and metadata contributing to Mutographs are not integrated into any other publicly available dataset. Similarly, samples selected in Mutographs from existing biorepositories have never been selected for sequencing in previous genomics initiatives.

In the description of data collection, we refer to harmonising rather to pooling data (line 226). For instance, data on smoking\alcohol history from existing biorepositories was harmonised to evaluate history of exposure, quantity, and frequency using the same definition for each variable. This has been described in the published data from Mutographs (references added in the corresponding section).

We have modified the sentence in line 124. We removed the term "pooling data" to avoid confusion.

3. From lines 217 to 222, there are some abbreviations in the text that do not appear in Table 1, and vice versa. For enhanced clarity, aligning the information consistently would be beneficial.

We have included the abbreviations in table 1 and in the table legend to keep consistency with the text.

4. In the same section (lines 217-222), why are only these three cancer sites mentioned? Does this imply that PCAWG lacks exposure information for other cancer types? Is there a specific rationale behind highlighting the place of residence?

We compared the available exposure information among patients from cancer sites included in both PCAWG and Mutographs to estimate the extent of the Mutographs project metadata. These three cancer sites: Esophageal, Head and Neck and Pancreas were the only overlapping cancer sites between the two studies with exposure information. PCAWG only included partial information on history of alcohol and tobacco consumption as highlighted in the text (lines 209-216) and in Table 1. The updated version of Table 1 includes the countries from where cases were selected in both studies to emphasise a higher geographical diversity in our study.

Information regarding residential history was important to highlight specific lifestyle and/or environmental and risk exposures for different regions, i.e consumption of opium for specific regions in the north of Iran and mate drinking in the south of Brazil. To understand exposure to Aristolochic acid in the Balkan region and boundary countries, we intended to use residential history to track a possible environmental source of exposure to this carcinogen beyond the use of herbal remedies.

5. Regarding lines 255-256, it's imperative to describe the sequencing protocol employed. If the protocol aligns with reference 38, it's important to acknowledge that the sequencing depth is notably lower than that of PCAWG (https://www.nature.com/articles/s41586-020-1969-6#Sec14). Additionally, differences in sequence read length need clarification. Providing a concise overview of the data analysis pipeline, including software and version details, is advisable.

We have added specific details on the depth of coverage for both tumor (40X) and normal tissues (20X)(lines 282-287) and minimal depth considered for further analyses. As mentioned by the reviewer, in PCAWG, the mean read coverage was 39X (higher than in our study) for normal samples, whereas tumours had a bimodal coverage distribution with modes at 38X and 60X (within the range for tumors sequenced in Mutographs).
We also added the references to the sequencing pipelines from published articles and the links to bioinformatic repositories which include all algorithms, software versions and bioinformatic codes used in the analyses as clarified above in point 1.

6. Within Table 1, the presence of samples with unknown gender raises questions. Given that gender can be determined from genetic data, could you elucidate the reason behind this discrepancy?

In the new version of Table 1. we have completed the missing information on sex from cases in Mutographs. However, metadata publicly available and published for the PDAC cases in the PCAWG collection lacked information on sex for 2 cases. https://doi.org/10.1038/s41586-020-1969-6. Supplementary Table 1.

Referees' reports, second round of review

Reviewer #1: I would like to thank the authors for working to address my previous comments. I am satisfied with the changes, and I believe that the article is now much improved. The authors provide a clear motivation for the Mutographs projects, contextualise it and provide a discussion of their workflow and how they addressed the challenges of the project. Finally, they illustrate how their initiative can serve as an example and pave the way for other projects that can build on it.

My only remaining concern is that the project is still ongoing and, while one article and its data have been published in 2021 (Moody et al, about 552 esophageal cancers), most of the data and

the results from the Mutographs project as it has been described here (4,400 successfully processed samples) are not yet available/published. This limits to some extent the discussion of the impact and effects that this project has had so far. At the same time, I think that it might be up to the editor to decide whether we should wait or not for more results to be published, so that a summary of the results and their impact can be included. The presented article certainly has already the potential to be a very influential piece, illustrating how efforts of cancer sequencing involving many countries globally with highly harmonised metadata are possible and still very much needed.

Reviewer #2: 1. The line numbers in authors' responses do not align with the resubmitted revision. For example, the author indicated the depth of coverage contents were in line 282-278, but it actually are in line 248-250. Please make sure line numbers are updated to correspond with the latest revision of the manuscript.

2. If the line numbers are correct, there appears to be a discrepancy between lines 124-125 and the response to question 2 from reviewer 2 concerning the term "pooling data." If this phrase does not accurately represent your intended meaning, please amend the text in lines 124-125 to ensure consistency across the manuscript.

3. The author did not modify the color of Figure 1 accordingly.

Authors' response to the second round of review

We thank the reviewers and the editor for their final comments and remarks. In response to those comments, we have made the following changes to the manuscript:
Reviewer #1:
I would like to thank the authors for working to address my previous comments. I am satisfied with the changes, and I believe that the article is now much improved. The authors provide a clear motivation for the Mutographs projects, contextualise it and provide a discussion of their workflow and how they addressed the challenges of the project. Finally, they illustrate how their initiative can serve as an example and pave the way for other projects that can build on it.

The comments from the reviewer were extremely useful to reshape the focus of the manuscript. We emphasized the uniqueness of the Mutographs study collection, the opportunities to extend and/or develop similar initiatives in other countries and the lessons learned during the development of this biorepository.

My only remaining concern is that the project is still ongoing and, while one article and its data have been published in 2021 (Moody et al, about 552 esophageal cancers), most of the data and the results from the Mutographs project as it has been described here (4,400 successfully processed samples) are not yet available/published. This limits to some extent the discussion of the impact and effects that this project has had so far. At the same time, I think that it might be up to the editor to decide whether we should wait or not for more results to be published, so that a summary of the results and their impact can be included. The presented article certainly has already the potential to be a very influential piece, illustrating how efforts of cancer sequencing involving many countries globally with highly harmonised metadata are possible and still very much needed.

We acknowledge that the analyses included in the Mutographs

project are still ongoing. The end of the project was extended until January 2025. However, a great progress has been accomplished until now. For instance, the kidney cancer analysis (in biorchives REF43) is now under review after resubmission to Nature. New publications will follow in 2024, two manuscripts are currently in preparation, the Head and Neck cancer analysis to be submited in mid January and the Colorectal cancer manuscript in Spring 2024. We envision that the publication of the Mutographs Biorepository will enhance visibility to the new upcoming publications and vice versa.
As suggested by the editor we also included a paragraph on Limitations of the Study in the discussion section and added new supporting references.

Response to Reviewers
Reviewer #2:
3. The author did not modify the color of Figure 1 accordingly.
The Map in Figure 1B (Now labeled Figure 2) has been updated accordingly. For clarity, we decided to color in blue all the countries included in Mutographs and point to the cities included in the patient collection. The complete list of cities per country is included in the corresponding figure legend.

2. If the line numbers are correct, there appears to be a discrepancy between lines 124-125 and the response to question 2 from reviewer 2 concerning the term "pooling data." If this phrase does not accurately represent your intended meaning, please amend the text in lines 124-125 to ensure consistency across the manuscript.

We have modified the sentence in line 124. We removed the term "pooling data" to avoid confusion.

1. The line numbers in authors' responses do not align with the resubmitted revision. For example, the author indicated the depth of coverage contents were in line 282-278, but it actually are in line 248-250. Please make sure line numbers are updated to correspond with the latest revision of the manuscript.

We have added specific details on the depth of coverage for both tumor (40X) and normal tissues (20X)(lines 246-250) and minimal depth considered for further analyses. As mentioned by the reviewer, in PCAWG, the mean read coverage was 39X (higher than in our study) for normal samples, whereas tumours had a bimodal coverage distribution with modes at 38X and 60X (within the range for tumors sequenced in Mutographs).