From secret datasets to reproducible datasets: A case study with an archaeological lithic dataset

Jonathan Roldan

Contents

1 Introduction

Archaeological collections are often first excavated and processed with great excitement. However, as time passes, analyses are published, or a new goal for the project is set, the craze for those once excavated materials starts to fade. This causes cultural collections to sit in storage rooms at or near project sites, offsite collection rooms, or at best in well catalogued/climate controlled museum basements. The raw data used for published works remains hidden away in the boxes in someone's basement, or in storage drives in the researcher's desk, thus creating problems in the accessibility of such data. As Lodwick (2019) points out, perception of data sharing and reuse is currently low among archaeologists (Lodwick, 2019).

Although traditional models of data access has been effective in advancing knowledge for archaeology, this model does not meet current norms of practice in the sciences (Marwick et al., 2017). The new norm is known as open science (Lodwick, 2019; Marwick et al., 2017), which includes data stewardship, transparency, and public involvement (Marwick et al., 2017). The strong resistance to open access has been primarily due to the "potential for harm to people and cultural heritage" if data is misused (Marwick et al., 2017, p. 10).

As Marwick et al. (2017) discusses, however, these issues could be easily addressed. But why should this matter? Afterall, archaeology has been successful at creating and advancing knowledge. Lodwick (2019) and Stodden et al. (2016) warn, however, that the lack of transparency has raised concerns on irreproducility.

In archaeology, explicit discussion on data sharing has remained mainly within zooarchaeology (with works like that of Kansa and Kansa, 2013, and Atici et al., 2013; Lodwick 2019); however, some focus has been giving to lithics data. The purpose of this project will be to contribute to the discussion on lithic data sharing. This project will use four datasets that have been stored on personal storage drives. Data access to this information is currently restricted to members of the archaeological project. Analysis and publication of this data was dated to 2014 on an interim field report. In addition, like the data, access to this report is difficult and unless a prior relationship exists between the members of the project, outside access is near impossible.

According to Marwick et al. (2017) open science practices enable archaeologists to:

- Increase transparency and reproducibility.
- Enable more ready and responsible build upon work between colleagues.
- Accelerate discovery.
- Enhance credibility.
- Promote ethical research.
- Enhance engagement between researchers and other collaborators.
- Enhance inclusiveness.

For this reason, the goal of this project will be to use datasets that have been inaccessible to the public and publish it on an open access repository. Ultimately, this project will touch on the seven key points brought up by Marwick et al. (2017), advancing data stewardship practices, open access to data, and creating an opportunity to reproduce and challenge the analysis of this project.

2 Archaeological Project Background Information

The data to be analyzed in this project was collected in 2013 by Dr. Marisol Cortes-Rincon and the Dos Hombres to Gran Cacao (DH2GC) archaeological field school crewmembers. The field season operates between the months of May and June every year. During this time, Dr. Cortes-Rincon leads a group of undergraduates with the assistance of returning students or 2 graduate students. The archaeological project is located in Belize, and its research area expands within a transect between two major Maya cities, Dos Hombres and Gran Cacao (Figure 1).

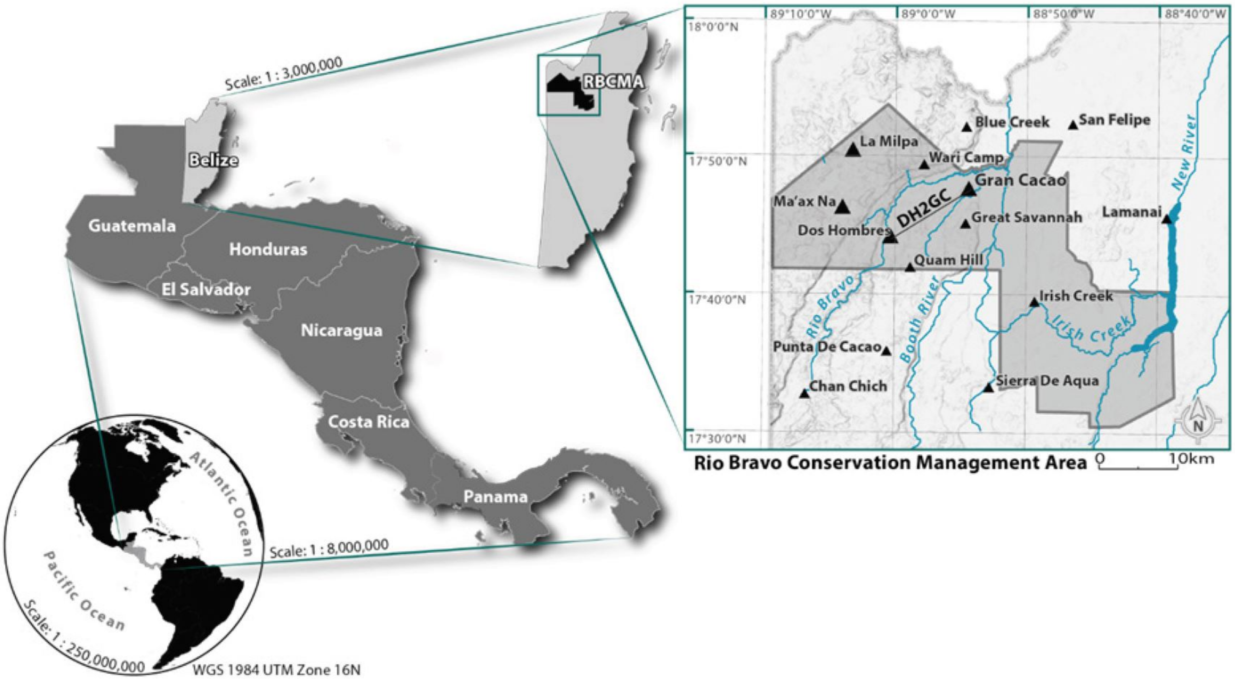


Figure 1: DH2GC Project Location (McFarland and Cortes-Rincon, 2019)

The distance between the two major cities is approximately 12km. However, the location of the 2013 field season, Figure 2, is approximately 500 (group a), 950 (group c) to 1250m (group b) - as the bird flies - from Dos Hombres. The environment near this area is composed of bajos (swamps) with some karstic hills at the far northeast of the last excavated household (b). The area is characterized by lowland vegetation, with little to no natural water sources.

The nearest natural water source is approximately 2km away (or 0.5km from Dos Hombres as the bird flies).

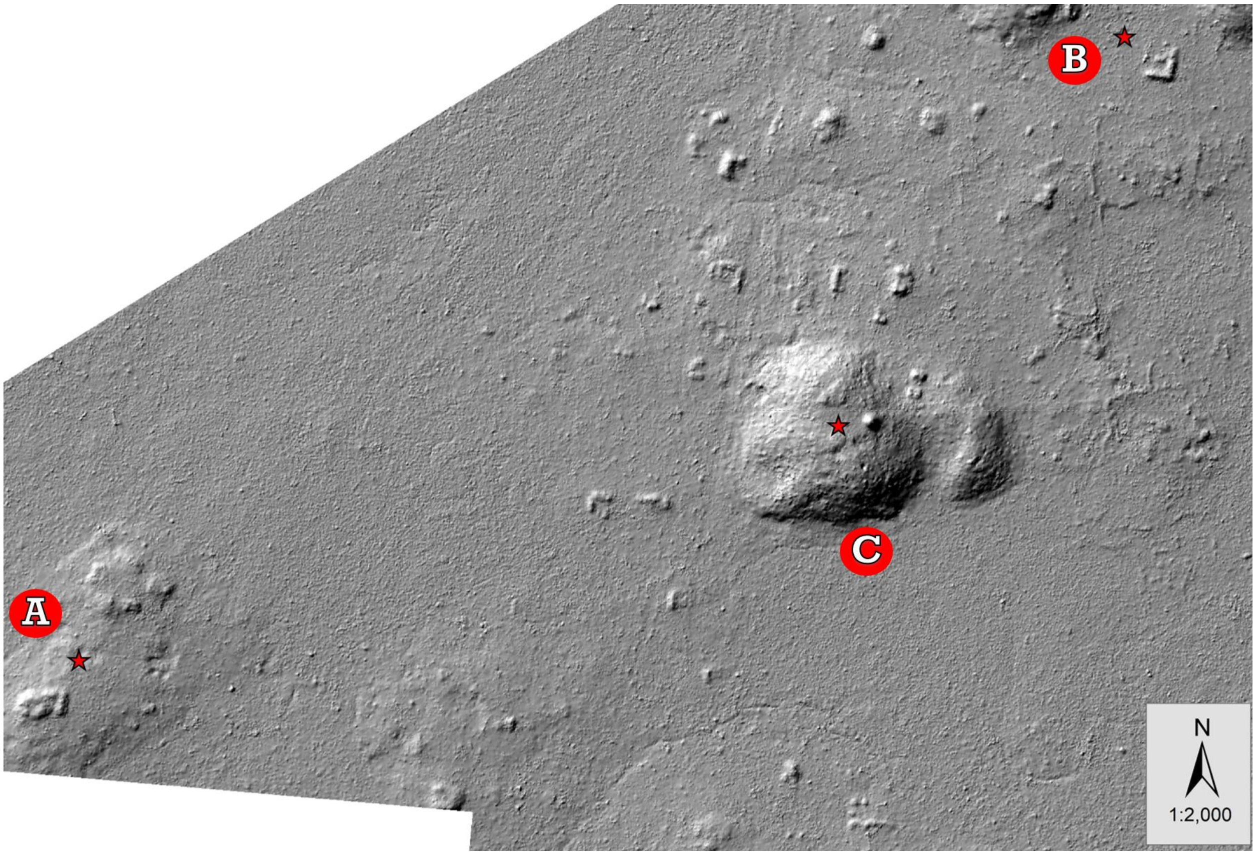


Figure 2: 2013 Excavation Area. Household 1 is at A, houseld 2 at B, and minor ritual center at C

3 Methods

I will adopt the basic workflow provided by Marwick (2017). His workflow diagram (see Marwick, 2017, fig. 1) indicated two basic key steps which enables reproductibility. First, the collection of custom R package functions, unit tests, consolidation of data in RProj, and a rendered output must all be collected in the same place. Second, research Collection (including step 1 and any manuscripts) should be stored in a GitHub repository.

In order to achieve a high degree of reproducibility (as that set by Marwick 2017), the repository files will include: a copy of this project paper, text files of raw data, script files of R, and a dockerfile that includes computational environments. GitHub will be used as the repository of the preceding files.

Table 1: Preview of Archaeological Debitage Dataset [@cortes2013]

DH2GC.Lithic.Analysis.Form..Debitage	X	X.1	X.2	X.3
Dates:				
Analyzed by:				
Artifact#	Provenience	Quantity	Weight (Grams)	Size (mm)
583	4-P-1	2	1	11.7
584	4-P-1	2	2.4	22.5
590	4-P-1	2	0.9	16.8

Table 2: Preview of Archaeological Flake Dataset [@cortes2013]

Artifact.	Provenience	Length..mm.	Width..mm.	Weight..g.
585	4-P-1	40.1	33.8	8.4
586	4-P-1	22.1	16.2	1.6
587	4-P-1	36.8	24.8	9.3
588	4-P-1	32.0	16.5	24
589	4-P-1	39.2	17.7	1.8
591	4-P-1	27.2	16.3	3.2

3.1 Data Tidying

As indicated by Wickham et al. (2014) most of the time in data analysis is spent in cleaning and preparing the data. Further, data preparation is not only the first step, but a revisited process throughout the analysis (Wickham et al., 2014). Datasets used for this project (Table **??**,**??**,**??**, and **??**) were never intended for the purposes of computational analysis. Therefore, the datasets are expected to required heavy preparation and cleaning.

These data sets will be subjected to data cleaning through the principles of tidy data. According to Wickham et al. (2014), tidy data provides a standard way to organize values

Table 3: Preview of Archaeological Tool Dataset [@cortes2013]

Artifact.Number	Provenience	Code..	Code.Name	L..mm.	W..mm.	T..mm.
602	6-B-3	NA	Poss Core	55.0	31.3	22.3
632	6-D-5	NA	Chopper	82.0	60.0	40.0
655	4-B-4	NA	Exhausted core	28.0	24.0	17.0
656	4-F-3	4	Truncated GUB	75.0	55.0	23.0
671	6-A-2	18	Scrapper	44.3	29.6	7.0
688	6-A-3	25	Test core	45.8	31.1	32.2

Table 4: Preview of Archaeological Obsidian Dataset [@cortes2013]

Prov	Artifact..	L..mm.	W..mm.	Thic..mm.	Weight..g.
4-D-1	1134	9.0	7.8	2.7	0.2
4-N-3	1135	17.2	9.3	1.6	0.4
4-N-3	1136	24.0	11.6	3.6	1.2
6-A-2	1137	20.5	8.0	3.0	0.5
6-D-6	1138	29.0	13.5	3.3	1.4
6-B-5	1139	20.9	11.3	2.7	0.9

within a dataset. In addition, this standard will facilitate initial exploration, while simpliying the development of analysis. Tidy datasets depend *on how rows, columns, and tables are matched up with observations, variables and types. In tidy data:*

1. *Each variable forms a column.*
2. *Each observation forms a row.*
3. *Each type of observational unit forms a table.* (Wickham et al., 2014, p. 4)

Lastly, in order to meet reproducibility standards as set forth by Marwick (2017), all processed datasets will be exported to a text file using the `` `write.csv()` `` function.

3.2 Archaeology

The data analyzed here was excavated during the 2013 field season. It was collected from two household groups (A and B in Figure 2) and a minor ritual center (C in Figure 2) through standard excavation techniques. Excavation untis measured 1x1m and were set up abutting the structures. Excavation units were subject to expansion if features were discovered during the excavation process (such as building corners, ritual depositions, stairs, etc).

Household A only contained one excavation unit, 4-J. This excavation was first put in place in the 2012 field season, but was expanded and finished during the 2013 season. The unit was first proposed to find what the relationship was between the household and the minor ritual center. Household B was a newly discovered group, and was the center of attention for the mapping crew and a few excavators. Units placed here were 6-A, 6-B, 6-C, and 6-D. The minor ritual center has been subjected to various years of intense excavations. The excavations are 4-B and 4-K (causeway), 4-G and 4-O (courtyard group), 4-M and 4-N (burial), 4-S (elite housing/other ritual building), 4-T (cave/xultun).

Excavated soil for each unit was screened through a 1cm mesh and stored in cloth bags for later processing. All of the excavated artifacts were processed and analyzed towards the latter end of the project by a group of students and volunteers. Artifactual recordings were analyzed through standard procedures as that set forth previously by the project. Specific lithic artifact information can be seen in the 3.3 section.

3.3 Lithic Analysis

Due to the lack of a full time lithic analyst in the DH2GC project, the principle investigator adopts an expedient method in lithic analysis. Debris is separated into two categories **debitage** and **flakes**. The only attribute to separate these categories is the presence of a flake attributes: platform, bulb of percussion, and ripples. For example, if an object does not contain a recognizable attribute the artifact is labeled as debitage, while the presence of any attribute the object is labeled as a flake.

The **debitage** dataset, (table 1), follows the Flake Aggregate Analysis methodology. This method gives attention to batches of lithic debris rather than individual artifacts (Ahler, 1989). Debris groups were organized by size-grade. However, this group sizes were not standardized. Instead, researchers grouped objects based on visual size similarities. Variables recorded for each group category were as follows:

1. Artifact Number: a unique number assigned to all artifacts found on the project
2. Provinience: a string that contains horizontal and vertical location information. Each excavation is broken down into three parts - Operation (group of excavation units within a vicinity), suboperation (a unique letter assigned to an excavation unit), and lot number (vertical level number, where 0 indicates top soil and increases if a change in strata or a feature is found during the excavation).
3. Count.
4. Mass (g).
5. Average length (mm).
6. Analyst: person(s) in charged of analyzing data.
7. Comments: any noted observation. Usually involving material and/or cortex presense.

On the other hand, the **flake** dataset, (table 2), individual objects received more attention. Variables recorded for this type of artifact were:

1. Artifact Number.
2. Provinience.
3. Dimensions: Lenght (mm), width (mm), thickness (mm), and bulb Thickness (mm).
4. Mass (g).
5. Material type and Quality.
6. Cortex Presence: 0
7. Platform: flat, cortical, abraded, or complex.

8. Termination: feathered, stepped, hindged, and overshot
9. Analyst.
10. Comments.

Lastly the **tools** and **obsidian** datasets received special attention. Data recorded for tools are:

1. Artifact Number.
2. Provinience.
3. Code and Code Name: Type of tool or finished product
4. Dimensions: Lenght (mm), thickness (mm), width (mm).
5. Mass (g).
6. Material type and Quality.
7. Comments.

Obsidian data recorded:

1. Artifact Number.
2. Provinience.
3. Dimensions: Lenght (mm), thickness (mm), width (mm).
4. Mass (g).
5. Wear: nicking, slight, dorsal, ventral, trimming, and miscellaneous.
6. Platform: Single facet, mutiple facet, abraded single facet, ground single facet.
7. Type: the type of technology.
8. Category: Lithic form.
9. Form: Flake orientation.
10. Color.
11. Comments.
12. Researcher.

Table 5: Preview of Tidy Dataset for the Debitage Database

Debitage_Id	Operation	Suboperation	Lot	Count	Mass_g	Average_Lenghtmm
583	4	P	1	2	1.0	11.7
584	4	P	1	2	2.4	22.5
590	4	P	1	2	0.9	16.8
598	6	B	3	15	7.6	16.3
599	6	B	3	10	15.5	21.0
600	6	B	3	6	1.5	10.5

4 Results

4.1 Tidy Data

All datasets were considered messy and met one or more problems according to Wickham et al. (2014) common problems of messy datasets. Recalling the debitage data (table 1), comments contained a two types of information in most of its observations, meeting the *multiple variables stored in one column* problem. Therefore, in order to bring the debitage dataset to a tidy dataset, three more variables were added. First, flake category (category) indicates the stage in the reduction sequence of the lithic assemblage; where a value of 1 are cortication flakes, and a value of 0 are cortex-free flakes

Second, a material variable was added. Lastly, the material quaility was added. Furthermore, because this data was recorded on an excel sheet, there were extra rows that did not contain any data information. This presented two problems - incorrect data type of variables and blank rows that do not provide any data information. These rows were deleted and variable names were renamed using the `` `rename()` `` function. Data types were fixed using the `` `mutate() ` `` function from tidyverse. The resulting tidy dataset was exported to a csv file named "Debitage.csv" (table 5).

The flake dataset was more tidy than that of the debitage. However, it was not free of problems. The flake dataset can still be categorized as messy because it meets the *multiple types of observational units stored in the same table.* Comments were scanned with the use of regular expressions, and with aid of if-statements, flakes that contained any wear or retouch should be considered tools and were filtered out from the dataset. The resulting tidy dataset was exported "Flake.csv" (table 6).

The tool dataset could be considered tidy. However, because the flake dataset contained observations that should have been within the tool dataset, we considered it messy. The adoption of SQL naming convention made it relatively easy to join observations (table 7) from the flake dataset into the tool dataset. Other cleaning involved the deletion of symbols (eg, question marks) within values. The resulting tidy dataset was exported to "Tools.csv" (table 8)

The obsidian dataset was the most difficult to work with. This dataset contained *multiple*

Table 6: Preview of Tidy Dataset for the Flake Databaset

Artifact_Id	Operation	Suboperation	Lot	Lenght_mm	Width_mm	Mass_g
234	1	B	5	37.9	20.9	4.9
238	1	B	5	32.4	19.3	3.7
239	1	B	5	22.7	15.9	2.2
240	1	B	5	29.0	14.1	1.5
241	1	B	5	17.6	20.5	2.2
242	1	B	5	23.8	19.8	2.8

Table 7: Preview of Tidy Dataset for the Tools Database

Artifact_Id	Operation	Suboperation	Lot	Type_collection	Lenght_mm	Width_mm
594	4	P	1	GUB	88.7	25.9
602	6	B	3	Core	55.0	31.3
625	6	B	3	Utilized Flake	27.4	22.6
629	6	B	3	Utilized Flake	18.9	15.1
632	6	D	5	GUB	82.0	60.0
641	6	B	4	Utilized Flake	25.4	17.8

types of observational units stored in the same table; consequently leading it to have *multiple variables stored in one column* and some *variables stored in both rows and columns.* This dataset is composed of both tools and flake information of one single material - obsidian. For the purposes of this project, there was not any need to keep this as a stand alone dataset. Therefore, variables were added to match both the tools and flake datasets. Values were populated with the matching observation. Then, matching observations were joined to their respective dataset.

4.2 Lithic Analysis

4.2.1 Household A

Household A show variability in lithic material with access to both chert and obsidian. In lots 4, 5, and 6 of unit 4-J we see a large quantity of finished obsidian tools, while no chert tools (Figure 3). The pattern persists when comparing flakes, with a small presence of chert flakes (Figure 4), however. On the other hand, we see lot of uncategorized debitage (Figure 5).

Household A had no lithic debri with cortex (Figure 6). Although, Lot 2 of unit 4-J shows that only 30% of lithic debri contained no cortex, this can be ommitted. There was an issue with the code which prevented an incorrect calculation.

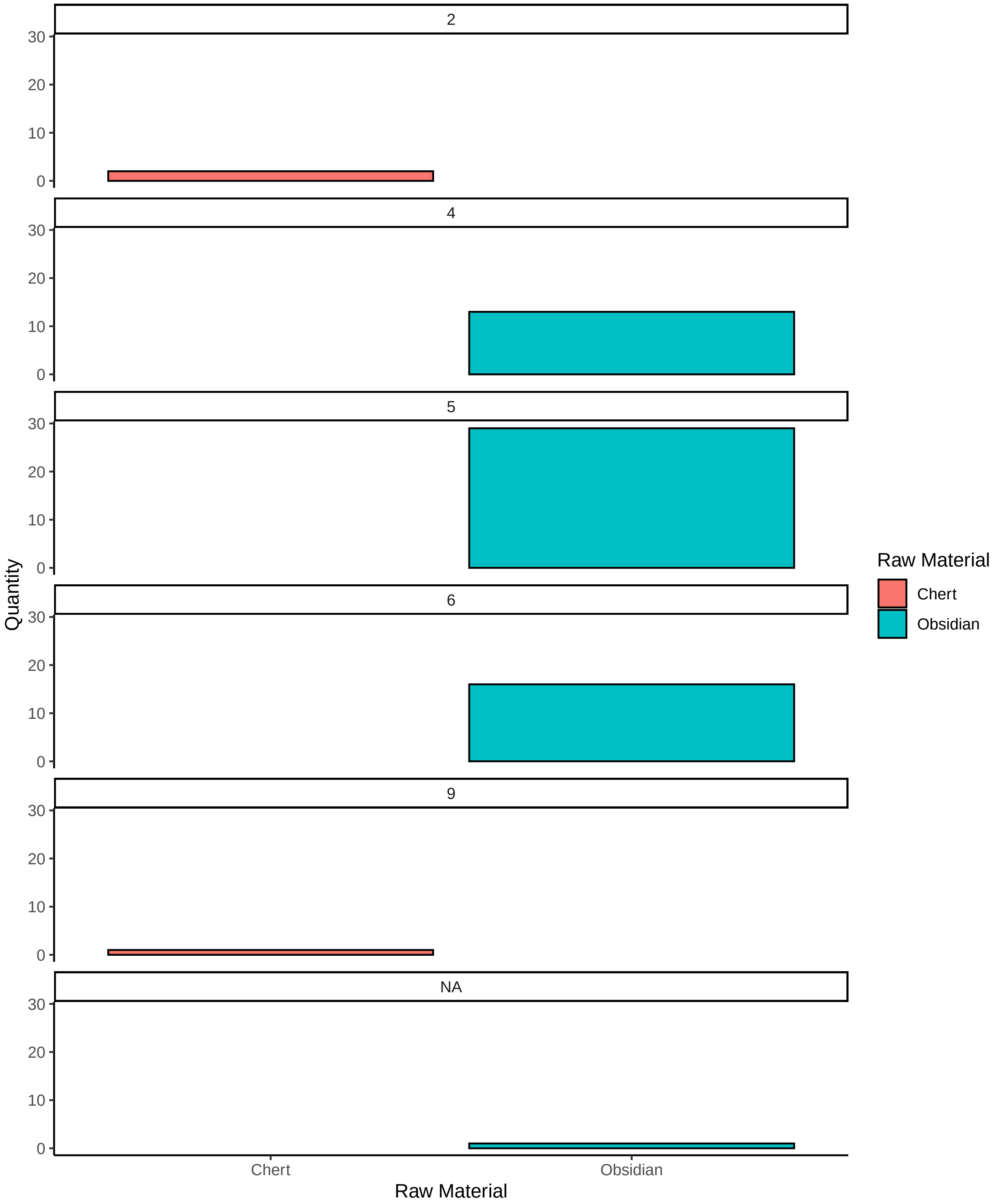


Figure 3: Tools found in Household A excavations by lot

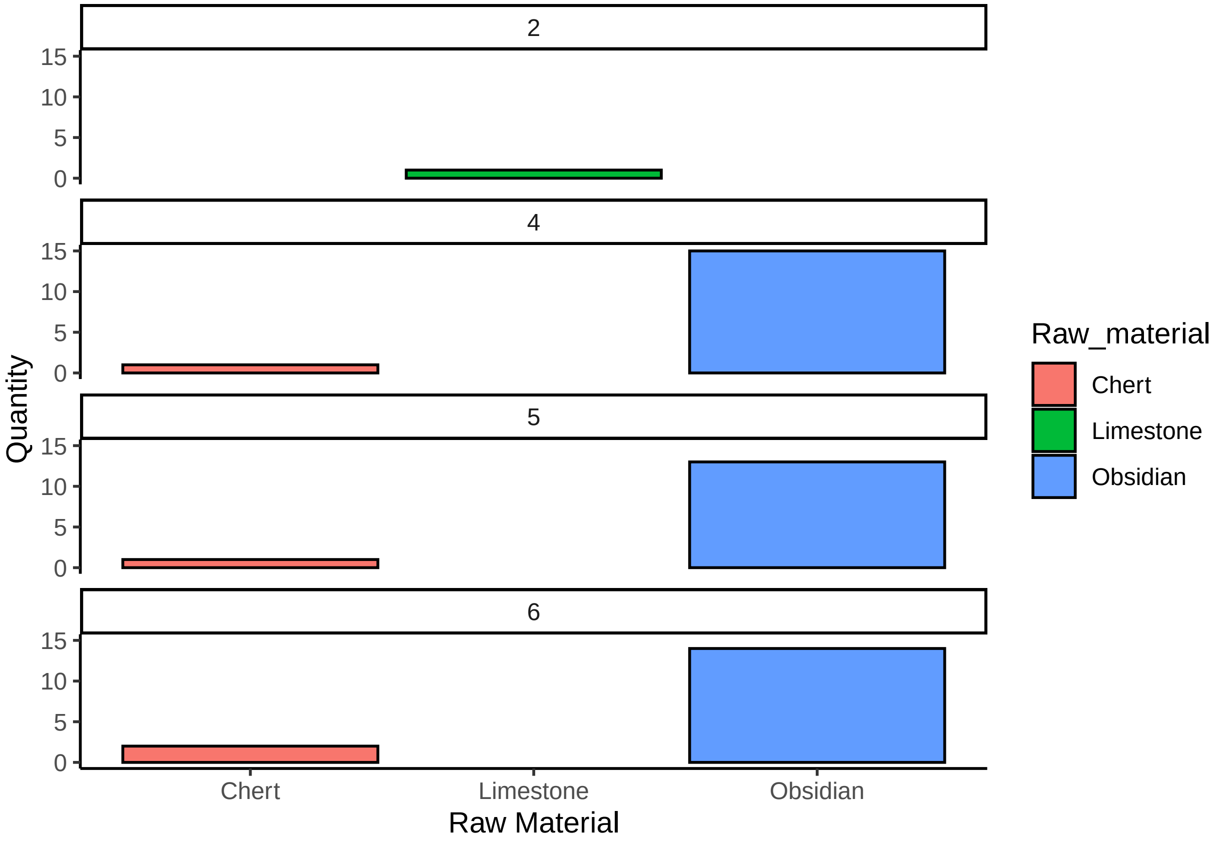


Figure 4: Flake Material found in Household A excavations by lot

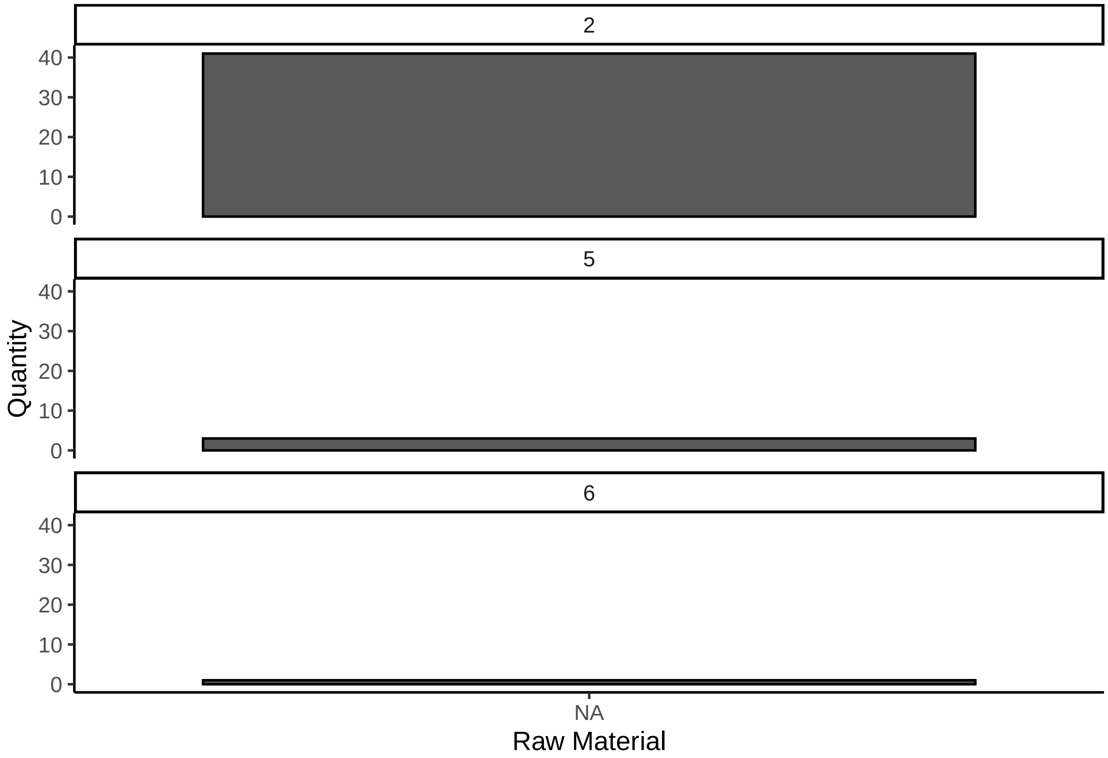


Figure 5: Debitage found in Household A excavations by lot

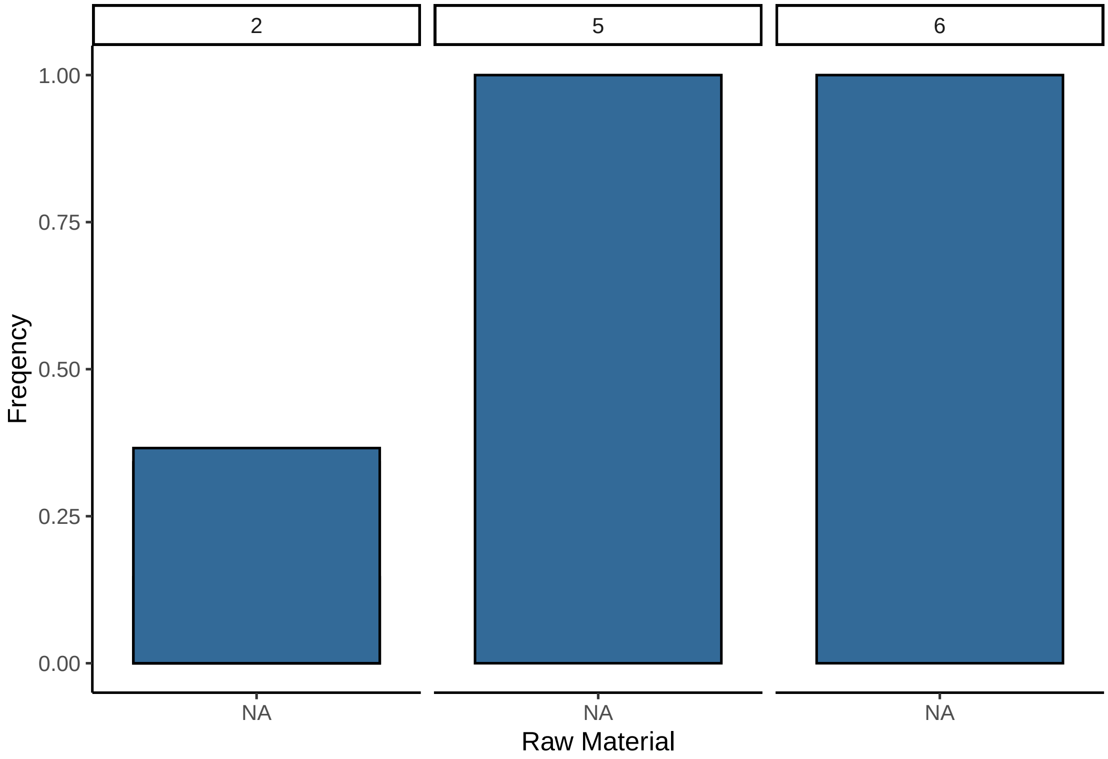
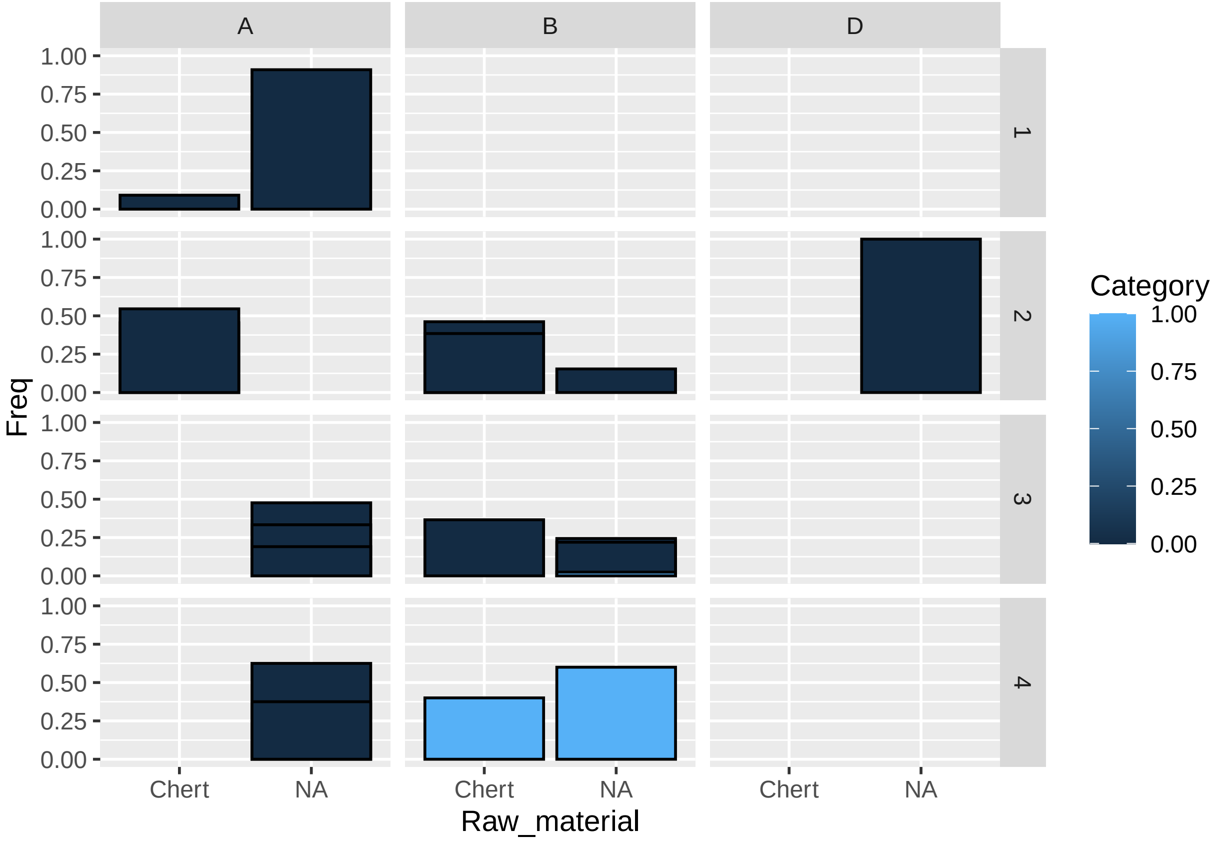


Figure 6: Frequency of Cortication and non Cortication debitage in Household A

4.2.2 Household B

This household experience less quantities of obsidian material and an increase in chert material. Although, some obsidian tools, Figure 7, can be found in 4-B lots 3 and 5, and some in 4-D lot 6, no flakes (Figure 8) or debitage (Figure 9) were present here. Instead, we see concentrations of chert debitage and flakes. This area contained, however, some limestone flakes in units 4-B lot 4 and 4-D Lot 5 (Figure 8).

For houseldhold group B, only unit 6-B at lot 4 was found to have any lithic debris with cortex (Figure **??**). All other units were found to have no cortex.



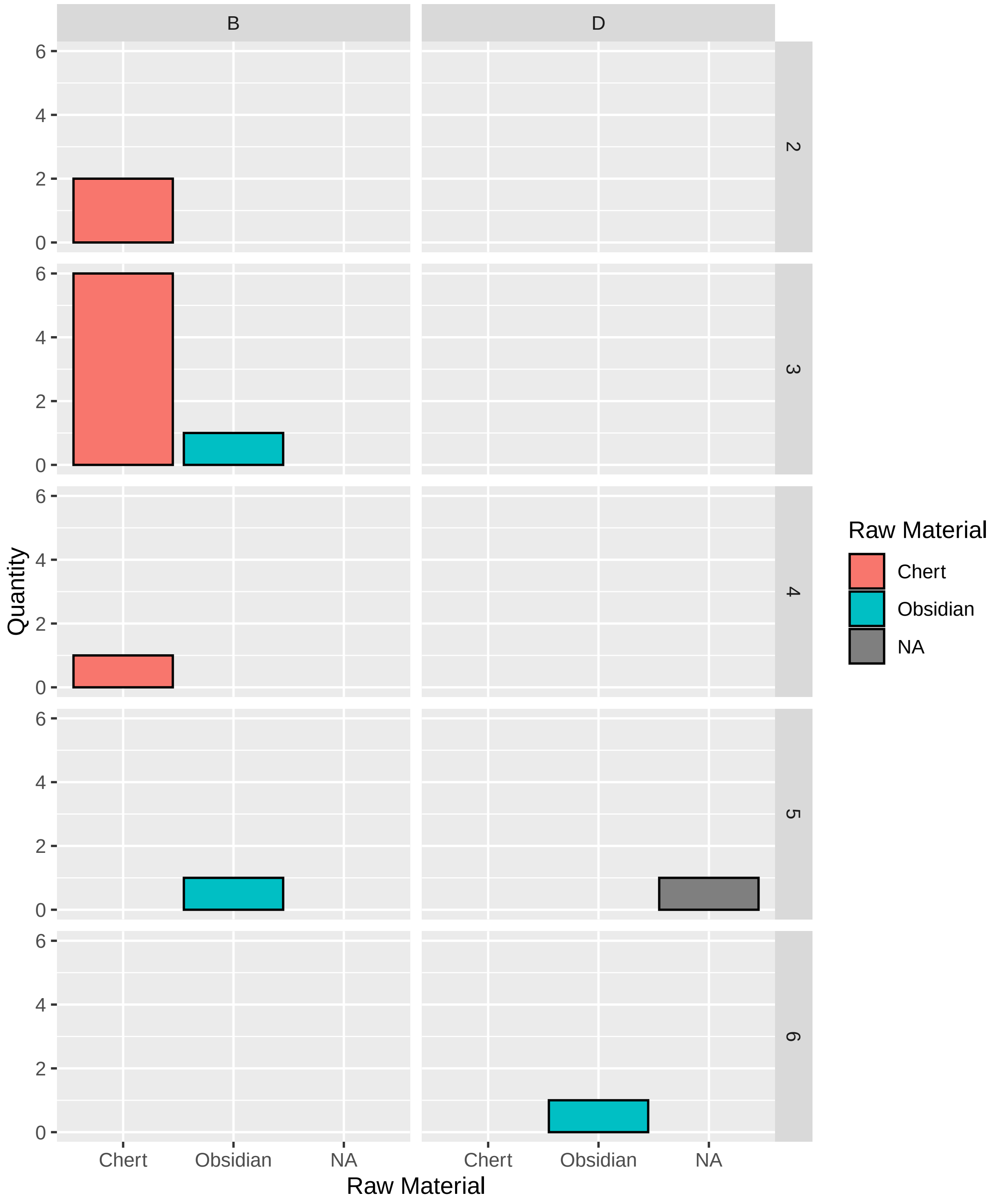


Figure 7: Tools found in Household B excavations

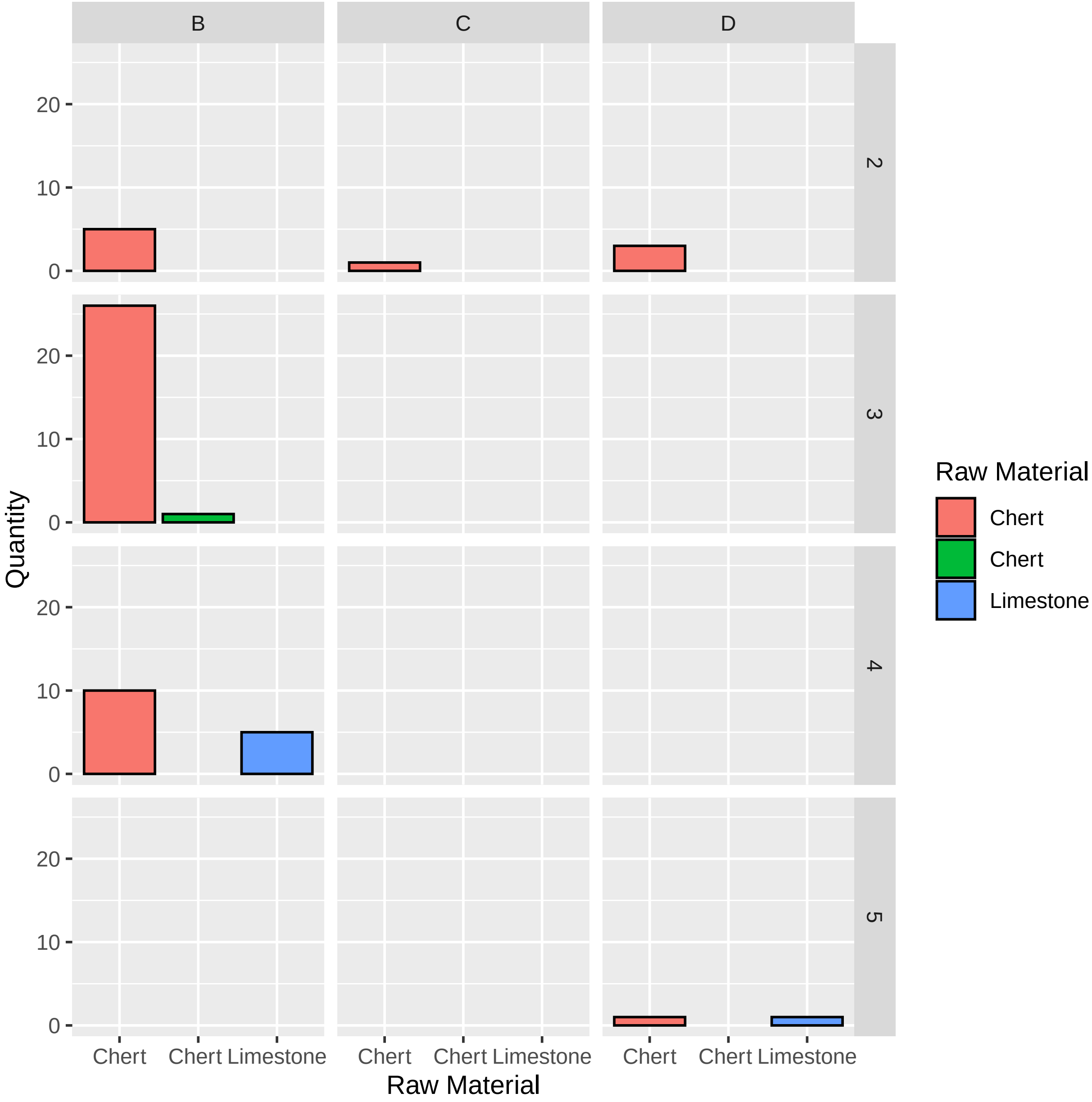


Figure 8: Flake Material found in Household B excavations

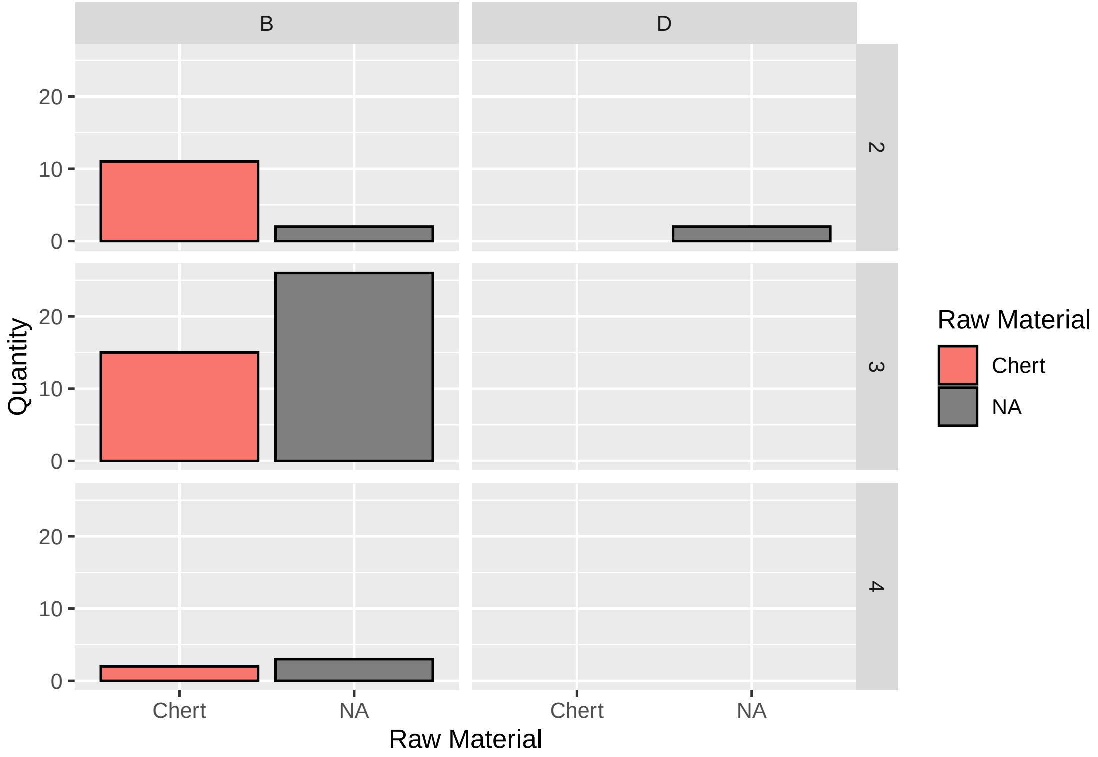


Figure 9: Debitage found in Household A excavations

4.2.3 Minor Ritual Center

It comes to no surprise that this area contains high concentrations of finished tools (Figure 10). Surprisingly, however, there are very few obsidian tools. Only Units 4-N and 4-G, Lots 3 and 4 (respectively) had obsidian tools. Lithic debris was mostly composed of chert (Figure 11 and **??**). However, some limestone lithic debris was found, specially in unit 4-G Lot 3 and unit 4-S lot 1. No obsidian lithic debris was found in this area.

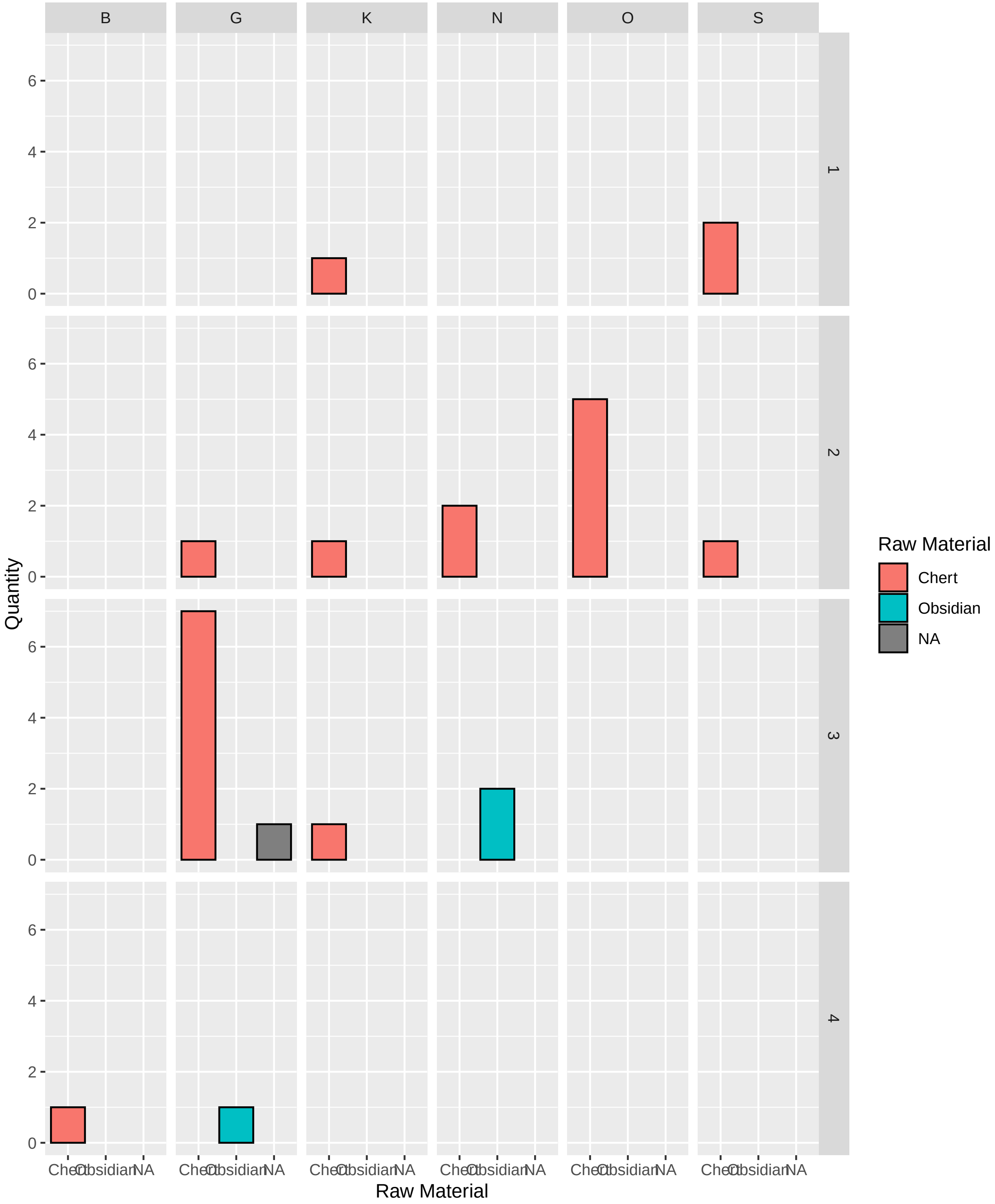


Figure 10: Tools found in the minor ritual center excavations

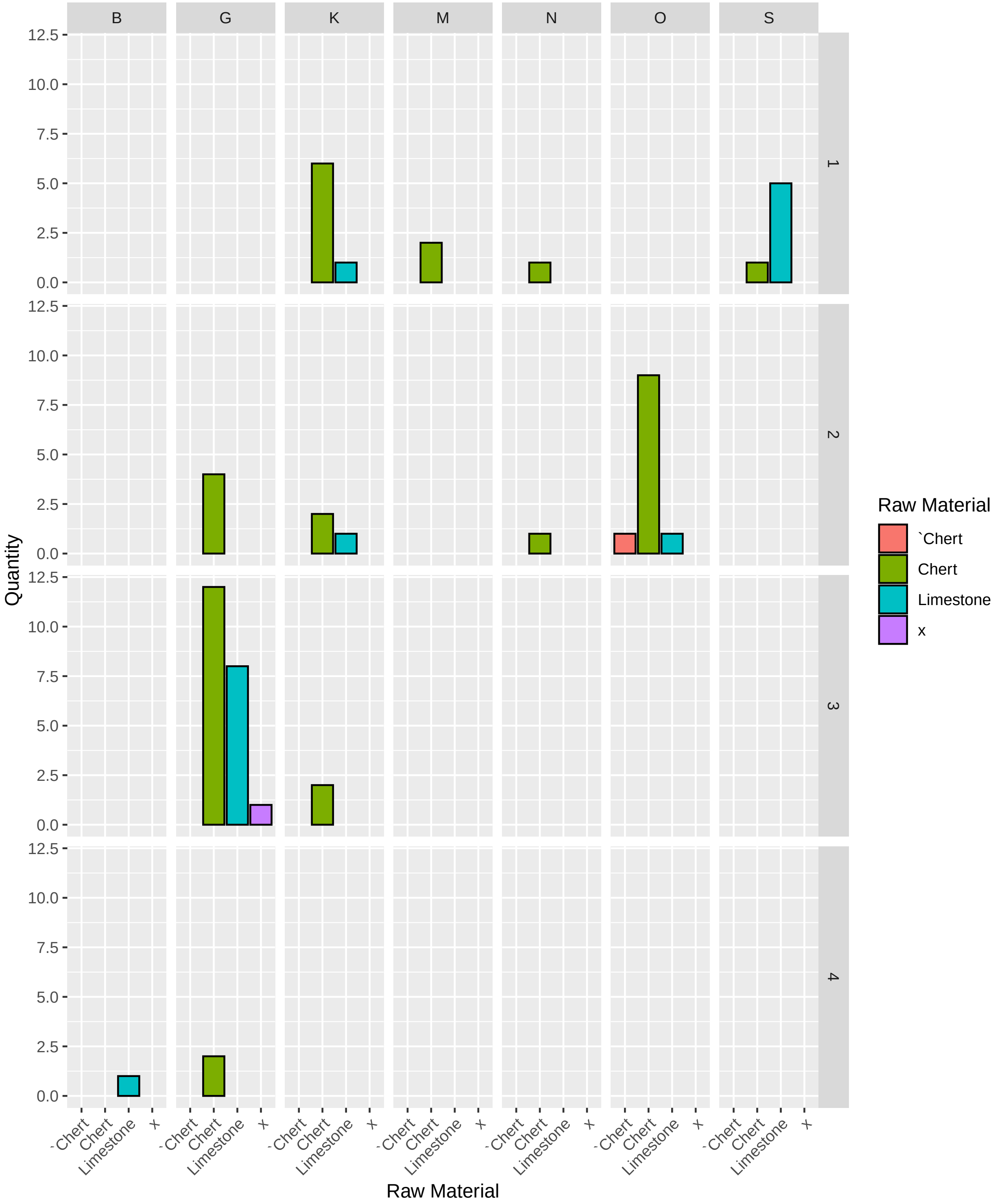
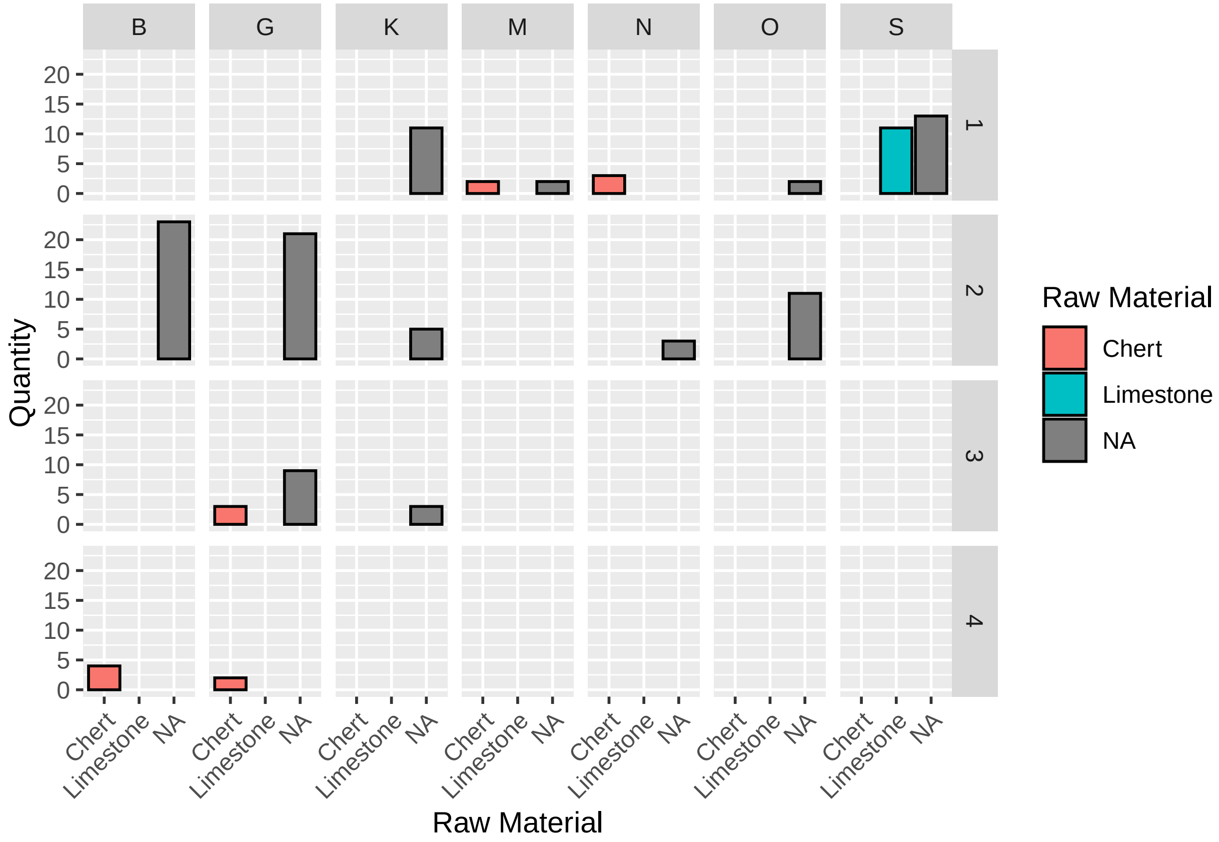


Figure 11: Flakes found in the minor ritual center excavations


B
G
K
M
N
O
S
1
2
3
4
Quantity
0
5
10
15
20
Chert
Limestone
NA
Raw Material
Raw Material
Chert
Limestone
NA

Furthermore, unsurprisingly I may add, no lithic debris was found to have any cortex (Figure 12). There was only one exception, unit 4-G at lot 3 was the only instance with approximately 25% of lithic debris with chert.

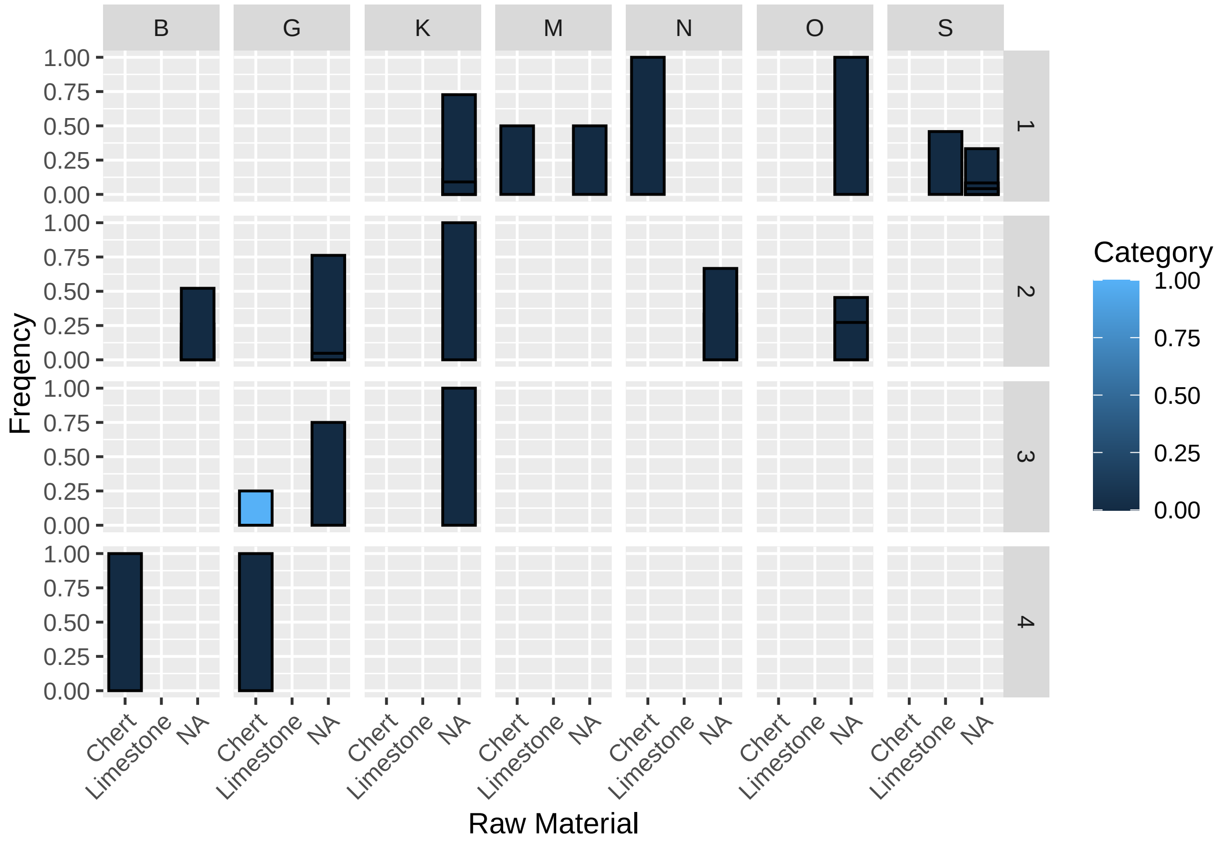


Figure 12: Frequency of Cortication and non Cortication debitage in Minor Ritual Center

5 Discussion

Household A was characterized with a majority of obsidian assamblages, with little to no chert. As oppose to Houshold A, Household B was characterized with a higher frequency of chert vs. obsidian. This can imply either specialization in raw material usage, or differntical in resources access. There is a minor complication, however. Many debitage material was not categorized or was unknown. No information was found on to the specific artifact numbers, but according to Boudreaux et al. (2014) all debitage was chert.

Furthermore, the minor ritual center exhibited surprising results. It was expected that this site was going to show a higher concentrations of exotic materials. However, the opposite occured. Althought, little to no obsidian was found, it is obvious that finished products are been handled near this site at higher frequencies. If Boudreaux et al. (2014) can be trusted, the debitage found here can be from resharpening tools rather than production.

This lack of lithic production is further exacerbated by the lack of lithic assamblages with cortex residue. This is gives us a window to the economy of the region. Although, further analysis is needed to make any conclusive remarks, we can assume that materials are coming in from outside the area in the forms of blanks (given the precenses of debris and reduction flakes). The mostlikely area this lithics originate are Colha. This site is well known for its chert and lithic productions.

6 Conclusion

Although researchers at DH2GC did a good job at recording true variables; the lack of standard procedures created messy datasets. Even under low observations a messy data can proove extremely cumbersome. As mention by Wickham et al. (2014), the majority of time during this analysis was spent in cleaning the data. However, once it was cleanned, it allowed for easy analysis and results opened new paths for research.

Saving all files pertaining to this project in GitHub (https://roldaj5.github.io/ANTH771_Project/), will allow transparency of my work. In addition, open access to this files will allow for someone to improve upon my work, research new questions, and allow for the preservation of this data. Adopting this open access proactice, will move archaeology to meet current science norms and the advancement of newer and better questions.

References

Ahler, S. A.
1989. Mass analysis of flaking debris: studying the forest rather than the tree. *Archeological papers of the American anthropological association*, 1(1):85–118.

Boudreaux, S., M. Cortes-Rincon, and M. Brennan
2014. Dos hombres to gran cacao archaeology porject (dh2gc): 2013 interim field report. *Research Reports from the Programme for Belize Archaeological Project*, Pp. 1–69.

Lodwick, L.
2019. Sowing the seeds of future research: data sharing, citation and reuse in archaeobotany. *Open Quaternary*, 5(1).

Marwick, B.
2017. Computational reproducibility in archaeological research: Basic principles and a case study of their implementation. *Journal of Archaeological Method and Theory*, 24(2):424–450.

Marwick, B., J. d'Alpoim Guedes, C. M. Barton, L. A. Bates, M. Baxter, A. Bevan, E. A. Bollwerk, R. K. Bocinsky, T. Brughmans, A. K. Carter, et al.
2017. Open science in archaeology. *SAA Archaeological Record*, 17(4):8–14.

McFarland, J. and M. Cortes-Rincon
2019. Mapping maya hinterlands. *Humboldt Journal of Social Relations*, (41):46–59.

Stodden, V., M. McNutt, D. H. Bailey, E. Deelman, Y. Gil, B. Hanson, M. A. Heroux, J. P. Ioannidis, and M. Taufer
2016. Enhancing reproducibility for computational methods. *Science*, 354(6317):1240–1241.

Wickham, H. et al.
2014. Tidy data. *Journal of Statistical Software*, 59(10):1–23.