Appendix A*

Participant selection and genealogical data management

This appendix provides a detailed account of how the participants were selected and their data managed in the process of reconstructing the historical population of Rio Negro using the Extended Genealogy Method (EGM). The first section describes how the genealogical population was delimited and respondents for the seed and subsequent interviews identified. The second section summarises how the primary genealogical data was recorded and transformed to obtain cross-sectional population data.

1 Defining the population of interest

Determining who counted as a member of the population was necessary for delimiting the data collection and producing a saturated family network in Rio Negro. The study aimed to reconstruct the population alive in 1981 (before the mass killings), and all of their descendants and ascendants who ever lived between 1960 and 2015. Table 1 shows the distribution of vital events in the EGM-generated data over time.

Table 1: Vital events in EGM-reconstructed population by sex and year of occurrence

	Number of reported births				Number of reported deaths			
	Female	Male	Sex unknown	Total	Female	Male	Sex unknown	Total
<1920	11	14	0	25	0	0	0	0
1920-1939	45	51	0	96	0	0	0	0
1940-1959	116	107	2	225	6	14	2	22
1960-1979	300	314	1	615	32	39	0	71
1980-1999	530	538	2	1070	241	248	0	489
2000-2015	523	536	9	1068	50	64	1	115
Date unknown	192	260	15	467	49	73	2	124
Total	1717	1820	29	3566	378	438	5	821

Data on older individuals was needed to identify kinship relations between members of the population. These records increased the ancestry depth of the genealogical data (the average number of ancestors known for every individual). The measure of ancestry depth can be used to summarise the degree to which family relations can be established in a genealogical population. An ancestry depth of two is the minimal required

*Supplementary material for the paper 'Blood is thicker than bloodshed: a genealogical approach to reconstruct populations after armed conflicts'. The appendix is fully reproducible and can be compiled from the 'Appendix_A.Rmd' file using R Markdown (Allaire et al., 2018). The source file includes detailed comments on the data and figures used for the appendix.

to identify grandparents, grandchildren, cousins, aunts, and uncles.[1] Figure 1 shows that an ancestry depth of two was available for 70% of the inhabitants of Rio Negro in the EGM-reconstructed population. This means that members of the extended family could be identified for two thirds of the population. Ancestry depth was lowest for individuals in the older birth cohorts, meaning that cousins and other parents' siblings could usually not be identified for the oldest members of the population. This was a direct result of the way in which the population was defined since information about the parents of the oldest members of the population was not recorded to limit the scope of the data collection.

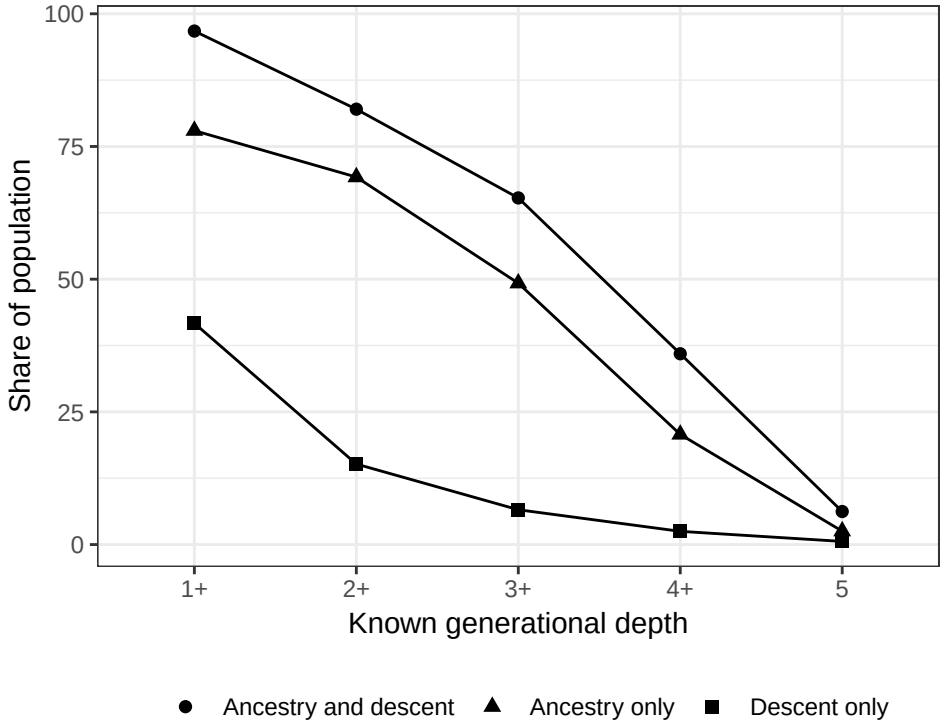

Figure 1: Generational depth in EGM-generated data for Rio Negro

The EGM interviews were conducted following a common set of guidelines. Respondents were initially asked to recall the complete marriage and birth histories of their parents (which included information about their siblings and themselves). The respondents then provided data on their own marriage and birth history. The procedure was later repeated to record the genealogical information of the respondent's siblings and children. Data on childless marriages was also recorded, along with the demographic information of present and past partners. However, no data about the partners' own families was recorded. Data on a respondent's husband was included in the questionnaire (e.g. date of birth, death, etc.) but not on the respondent's mother-in-law as this information was collected in later interviews. Data on the relatives of spouses from other

[1] An ego with known parents and grandparents has an ancestry depth of two: one step from ego to parents, another step from parents to grandparents.

communities (i.e. whose ancestry could not be traced to the 1981 Rio Negro population) was not relevant for the study. Applying these criteria consistently made it possible to delimit the population horizontally.[2] The next step was to define an appropriate sampling strategy to collect the genealogical data.

2 Seed selection and chain-referral sampling

Participants for the study were selected using a chain-referral sampling methodology (Platt, Luthra & Frere-Smith, 2015), a form of non-probabilistic network sampling in which new respondents were drawn from the pool of known records. The selection of the respondents for the initial genealogical interviews (the 'seed respondents') was a central component of this process. Two seed interviews were conducted with respondents that (a) resided in the village, (b) had been born before the 1982 killings, (c) possessed extensive genealogical knowledge, and (d) were not closely related to each other. These criteria, combined with practical considerations of access, helped narrow down the pool of potential seed respondents.

It was straightforward to determine eligibility based on the first two criteria - posterior analysis showed that 141 women and 125 men met criteria (a)-(b) at the time of the data collection. Criterion (b) reduced the pool of potential seed respondents to those aged over 34 in 2015. The chosen seed respondents were 53 and 61 years of age respectively. Previous studies have shown that older individuals tend to have more extensive kinship knowledge, but are less-well informed about recent events (e.g. vital events related to their grandchildren) (Chang et al., 2016). Younger respondents, on the other hand, are more knowledgeable about contemporary events, but are less able to recollect episodes from the more distant past. The final pool of respondents included individuals in various age groups (Table 2).

The third criterion required a definition of 'kinship knowledge'. Primary qualitative data was used for this, in the absence of other baseline data on the population. Direct observation and unstructured interviews were conducted to identify members of the population who were locally known for their extensive knowledge of the community and its history. This reduced the number of potential seed respondents and helped improve the quality of the seed interviews by making sure that the initial respondents were capable of answering the questions in the EGM questionnaire with confidence.

The fourth criterion required knowledge of the kinship relations between potential seed respondents. Establishing these links was challenging without pre-existing genealogical data. During the screening process, pairs of potential seed respondents were asked if they were related to each other in any way. The approach was useful, but kinship relations were sometimes not known or not acknowledged by participants. Two individuals were defined as 'close relatives' in this study if there were less than six degrees of separation between them in

[2] A genealogical population grows horizontally when collecting more data increases its size but not its generational depth.

the genealogical network of the village.[3] This selection criterion was introduced to ensure that interviews with seed respondents provided data on separate segments of the village's genealogy. Post-hoc analysis showed that the shortest path between the two chosen seed respondents was indeed six. The extended genealogies that grew out of the two seed interviews only converged after the thirteenth interview, when data on more than 1,000 individuals (roughly one third of all the members of the population) had already been recorded in the genealogical dataset.

Previous work on social network sampling has pointed out that the selection of seed respondents can bias the final composition of the population if seeds have a higher-than-average degree (i.e. more connections than other members of the social network) (Platt, Luthra & Frere-Smith, 2015). This is a genuine concern for networks of friendship or other types of social relations, but Table 2 shows that it was not the case in this study. Seed respondents in this study did not have considerably larger families (nuclear or extended) than the rest of the population.

Table 2: Demographic characteristics of seed respondents and other respondents in the EGM interviews

	Sex	Birth	Family size 1981		Family size 2015		Individuals reported
			Nuclear	Extended	Nuclear	Extended	
Seed 1	Female	1954	15	26.0	8	50	47.0
Seed 2	Female	1962	8	15.0	9	28	32.0
All other (median)	<NA>	1950	8	26.5	8	42	48.5

Respondents for the second wave of interviews were drawn from the genealogical data produced by the two seed interviews. All the subsequent respondents in the study fulfilled criteria (a)-(c) presented above. Participants in most interviews provided information about the current location and availability of potential future respondents. They also helped spread information about the study. Direct observation suggested that potential participants were more willing to take part in an interview if they had heard about the study in advance from a relative. Conducting the interviews in the local language was also key because most older respondents and many women were Maya Achi monolingual speakers.

The participant selection can be illustrated using EGM-generated data from Rio Negro. Figure 2 shows the ego-centric genealogy collected from one of the seed respondents. The seed respondent (id = A) had only one surviving sibling in 2015 (id = C).[4] Since the demographic information of C had already been collected during the interview with the seed respondent, the next logical step was to conduct an interview with the wife of C (id = B). The interview with B produced redundant information about her children (who had already been recorded in the seed interview with A) and new information about her parents, siblings, nephews, and

[3]The shortest path between two nodes was estimated as the minimal number of steps required to get from one member of the family network to the other. The shortest path between siblings was two; between cousins, four.

[4]The o symbol indicates that an individual had already died when the data was collected

nieces. The choice of next participant was clear in this case since all other siblings of the seed respondent had already died when the interview was conducted. In other cases, participant selection also considered practical and logistic issues related to access.

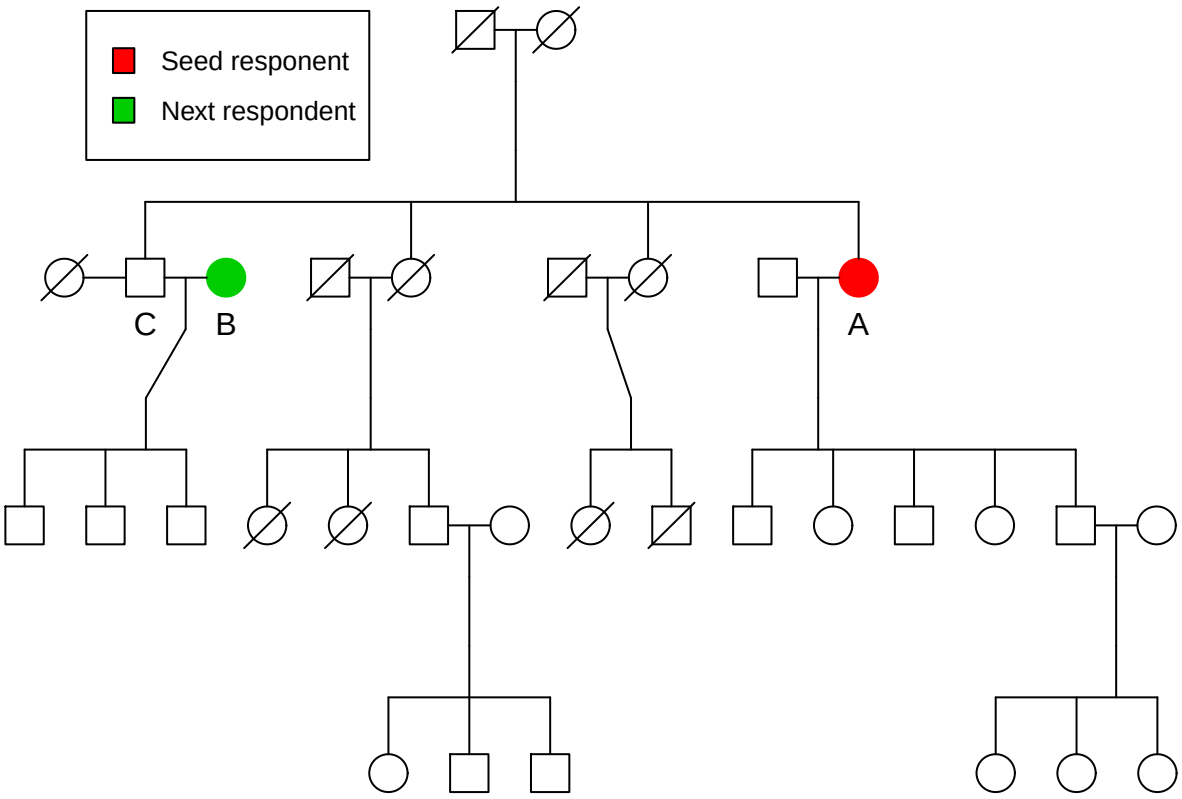

Figure 2: Genealogical diagram of individuals reported in the second seed interview

Interviewing the spouse of a previous respondent was generally discouraged. The two respondents could not always be considered 'independent' sources, especially as couples were often interviewed together. Therefore, the redundant data produced by separate interviews with spouses would have not have been useful to evaluate data consistency or data quality. It was also not advisable to conduct multiple interviews in the same household since relatives living in the household were usually present during the interviews and occasionally contributed to the genealogical interviews themselves.

Transcribing and matching the genealogical records in the field was an essential component of the data collection process. Data from the EGM questionnaires was transcribed and processed using interactive Shiny apps (Poletta, Orioli & Castilla, 2014), which were used to link the new records with those in the existing genealogical pool. Data consistency was evaluated each time a new genealogical record was added to the database. Automated R Markdown reports were produced to summarise the current state of the data

collection and highlight potential conflicts, inconsistencies, or gaps in the data. These preliminary findings were discussed in regular meetings with the research assistants of the study to make efficient choices about respondents for future interviews.

3 Data management

The EGM uses the principles of relational database design to record kinship relations. The 'Individuals Module' and the 'Marriages Module' (see Appendix B) are tables linked by ID fields that uniquely identify cases. Tables 3 and 4 show how a hypothetical genealogy would be recorded in this tabular format (the header colours clarify how the fields are linked across the tables). The 'Individuals Module Table' includes two columns that register the marriages associated with an individual. The `ParentMarriageID` field refers to the ID of the marriage formed by both parents of the individual. The field `CoupleMarriageID` refers to the ID(s) of the marriage(s) between the individual and their partner(s). A corresponding relational field is included in the 'Marriages Module Table'.

Table 3: Individuals Module Table

IndividualID	IndividualName	ParentMarriageID	CoupleMarriageID	...
1	A	2	3	...
2	B	-	3	...
3	C	-	2	...
4	D	1	2	...
5	E	-	1	...
6	F	-	1	...
7	G	1	-	...

Table 4: Marriages Module Table

MarriageID	.....	IndividualID	IndividualName	...
1	Partner 1	5	E	...
	Partner 2	6	F	...
2	Partner 1	3	C	...
	Partner 2	4	D	...
3	Partner 1	1	A	...
	Partner 2	2	B	...

Two intentional sources of redundancy were included in the EGM design to reduce human input error. The two `*MarriageID` fields in the 'Individuals Module' records the same kinship information as the `IndividualID` field in the 'Marriages Module Table'. Including the forenames of the spouses in the 'Marriages Module Table'

provided an additional way of ensuring consistency across the two tables. Simple algorithms can be used to transform the relational tables to more common genealogical or social network formats.[5]

4 Obtaining cross-sectional population data from EGM-generated genealogies

The data produced by the EGM can be used to produce 'pseudo-censuses' of the population at specific points in time. In its simplest form, cross-sectional sub-populations can be extracted from the genealogical data by filtering only the individuals that survived through a given period. The variables required for filtering the population in this way (date of birth and date of death) are available from the genealogical data. Pseudo-censuses can only be carried out after fully de-duplicating the EGM-generated records to avoid artificially inflating population size or over-representing the size of age groups. This method is also subject to error in the absence of time-variant data on the location of individuals at the time of the pseudo-census.

These 'demographic snapshots' provide valuable information on the size and composition of the population over time. Table 5, produced using the `pseudo_census` function from the EGM R package, shows this breakdown for Rio Negro in five selected years (1981, 1983, 1993, 2003, and 2013). The table gives the exact size of each demographic group at any given year, making it possible to compare the distribution of the population over time. The data, for example, shows that Rio Negro has been a young population historically. The share of adults over 45 years of age has been consistently small, whilst children under 15 constituted a clear majority before the year 2003. There were signs of a potential population ageing after this year, with the population under 15 constituting a smaller share of the total population by 2013. The table evidences a clear dip in total population size in 1982, resulting from the Rio Negro Massacres. According to the genealogical data, 38% of the pre-conflict population was killed in 1982 (366 of the 970 original inhabitants of the village), as discussed in the main text.

[5] See the R Markdown version of this document.

Table 5: Age and sex distribution of Rio Negro population: pseudo-censuses of the genealogical data for selected years

Year	1981		1983		1993		2003		2013	
Age	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
0-4	106	103	61	65	121	145	175	169	139	147
5-9	88	99	75	80	98	106	185	164	143	145
10-14	74	56	56	51	57	62	120	142	172	165
15-19	33	47	27	36	72	78	97	106	184	162
20-24	48	26	20	21	54	50	57	62	120	139
25-29	36	29	25	26	27	35	70	78	96	98
30-34	18	16	12	9	20	21	53	50	57	59
35-39	12	19	10	10	25	22	27	34	68	76
40-44	17	12	6	7	12	7	20	20	53	48
45-49	18	11	9	1	9	9	25	22	23	32
50-54	4	13	6	5	6	6	11	7	19	20
55-59	6	11	3	6	7	1	9	9	25	20
60-64	2	9	1	2	5	4	6	6	10	6
65-69	2	2	1	3	3	6	7	1	9	9
70-74	3	1	0	1	0	2	4	4	5	4
75-79	0	0	1	0	1	2	2	6	7	0
80+	1	3	0	0	0	0	1	3	5	9
Total	468	457	313	323	517	556	869	883	1135	1139

Note: Table produced with simplified filtering criteria; numbers are illustrative.

5 References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng & Winston Chang (2018) *rmarkdown: Dynamic Documents for R. R package version 1.10.*

Chang, Winston, Joe Cheng, JJ Allaire, Yihui Xie & Jonathan McPherson (2016) *shiny: Web Application Framework for R. R package version 0.14.2.*

Platt, Lucinda, Renee Luthra & Tom Frere-Smith (2015) Adapting Chain Referral Methods to Sample New Migrants: Possibilities and Limitations. *Demographic Research* 33(1): 665–700.

Poletta, Fernando a., Ieda M. Orioli & Eduardo E. Castilla (2014) Genealogical Data in Population Medical Genetics: Field Guidelines. *Genetics and Molecular Biology* 37(1): 171–185.