*Master of Science in Bioinformatics*

# Assessing dyadic relatedness in rhesus macaques using pedigree data

by

Hendrikje Westphal

Prof. Dr. Peter Stadler

Institute of Informatics
Faculty of Mathematics and Informatics
Leipzig University

Prof. Dr. Anja Widdig

Institute of Biology
Faculty of Life Science
Leipzig University

Leipzig, September 2023

# Abstract

text for abstract...

# Contents

# 1 Introduction

**kin selection**

Especially in behavioural ecology, dyadic relatedness remains one of the most studied key factors for the explanation of social and affiliative behaviour between individuals with a varying level of relatedness e.g., parent-infant interaction, protection, social grooming, coalition forming, cooperation, food sharing/foraging, or spatial proximity (Perry et al. 2008, Wellens et al. 2022, Gompper et al. 1997, Foroughirad et al. 2019).

The underlying idea goes back to the kin selection theory (Smith 1964, Hamilton 1964) which is based on Darwin's natural selection theory (Darwin 1956) and the individual's aim to maximize its inclusive fitness. Inclusive fitness comprises direct (reproductive success of the individual itself) as well as indirect fitness effects (reproductive success of close kin), like supporting a close relative to raise its offspring. The differentiation between both fitness types was originally made public by Fisher in 1930, but the innovation of Hamilton's extension was, that it supplies an explanation for the evolution of altruistic behaviour. Per definition, altruism refers to behaviour, for which the individuals does not gain any benefit for itself, like warning other individuals by alarm calls while exposing themselves to the predator (Silk 2002). Depending on the level of relatedness, Hamilton predicts the likeliness of expressing altruistic behaviour towards another individual (Hamilton's Rule: $rb - c > 0$, with c as the donor's fitness costs, b as the recipient's fitness benefits and r as the genetic relatedness).

Despite ongoing disputes regarding the encompassing validity and specific applicability of Hamilton's Rule and its derivates (Gardner et al. 2011, Birch and Okasha 2015, Nowak et al 2010), the concept of relatedness affecting the disposition to affiliative behaviour is recurrently basis for multiple studies, especially if it is taken into account that kin-biased assortativity can also be caused by reciprocal altru-

ism, mutualism or selfish behaviour to gain (in)direct fitness benefits by cooperative behaviour (Silk 2006, Chapais 2001). Besides, depending on population structure, dispersal pattern and society, several co-factors may also influence individual preferences e.g. age, gender, or social status/dominance rank (Carter et al. 2013, Perry et al. 2008, Pinter-Wollman et al. 2014). But even though - or especially because of it - it might be difficult to exactly distinguish between altruism, mutualism and selfish social affiliative behaviour as well as other cofactors, the inevitable basis to study kin preference in general is always an as exact as possible relatedness estimation (Foroughirad et al., 2019).

**Measuring relatedness and pedigree construction**

The relatedness coefficient $r$ is generally defined as the probability of two alleles, each from another individual, being identity-by-descent (IBD). IBD refers to a gen locus of identical state in both individuals which is caused by relatedness/inheritance from a common ancestor instead of chance. Since an offspring inherits at average 50% of the genes from the mother and 50% of the father, the statistically expected relatedness coefficient - without any mutation or recombination and referring to an ideal population with unrelated founders - can be easily derived from pedigree. On average, a parent-offspring dyad share 50% of their genes ($r = 0.5$) while half-siblings or grandparent-offspring dyads share 25% (r = 0.25).

Even though there are multiple strategies to determine kinship and measure relatedness, the most common approach especially in behavioural ecology is to obtain it from a pedigree which is usually reconstructed by observation and genetic analysis based on microsatellite markers (discovered by Tautz in 1989) to determine unknown parents (, , ). Microsatellites or STR marker (short tandem repeats) are DNA sequences with repeated subunits of typically di-, tri-, or tetra-nucleotide sequences which represent co-dominant polymorphic alleles due to their varying number of repeats (Jones et al., 2010). For an effective parentage analysis, a specific STR marker set needs to be established for each population which includes marker who are highly polymorphic on a specific gene locus and are characterized by high amplification success (Zane et al., 2002). If either mother or father is known, the missing parent can be determined by exclusion, because STRs are inherited from both parents based on the rules of Mendelian inheritance (Jones et al., 2010). Therefore, theoretically at least one of the offspring's alleles at a certain locus has to be shared with the mother,

and the other one with the father which means that contradictory parent candidates can be excluded if the offspring's alleles do not match with the alleles from the candidate. But due to possible mutation events (Eckert and Hile, 2009), programs like FindSire (Krawczak, 1999) recommend multiple degrees of quality: strict/relaxed rule and unsolved paternity) which addresses potential allelic mismatches even with correctly assigned parent candidates.

As a result of proceeding technical innovation and rapid development in the field of next generation sequencing, whole genome sequencing became more accessible and affordable while offering accurate high-throughput sequencing data, which provides the possibility of SNP genotyping (Park and Kim, 2016). SNPs or single nucleotide polymorphisms can be used to estimate the dyadic relatedness coefficient too by measuring the amount of shared DNA throughout the genome based on characteristics of individual IBD segments ([xxx]?). Due to recombination during meiosis and relatedness values based on shared IBD segments instead of statistical average, realized relatedness and pedigree-derived relatedness can greatly differ; especially the more distantly kins are related.

Technically, STR markers can also be used to estimate the dyadic relatedness [xxx] but SNPs are lot more reliable due to a considerably higher abundance within the genome as well as an even lower mutation rate (Schneider 2012, Rengmark et al. 2006). Additionally, STR based genetic analysis often are vulnerable to homoplasy, the occurence of null-alleles as well as genotyping errors (Glover et al. 2010, Jones et al. 2010). Besides, pedigree-derived relatedness highly depends on the accuracy (correctly assigned mother/father), available genealogical depth and completeness of the pedigree itself but also on the choice of an appropriate analysis program. Especially in behavioural or conservational studies of wild populations, pedigree analysis programs have to deal with a more or less pronounced number of missing values (unknown parents/generations). Also, the greater the generational time of a species, the longer it takes to gather breeding records over multiple generations. Hence more distant kinclasses which would only be found in extended pedigrees, often could not be surveyed and included in the analysis if the data set covers only recent breeding records (as done in [xxx], [xxx]).

In the literature, multiple genealogy or pedigree analysis tools can be found, but

the applicability in regards to the Cayo Santiago data set needs improvement. Often pedigree programs are either limitated by the maximum number of individuals/time efficiency due to a big data set [ribd package?]; or the considered genealogical depth in general (LK, [DESCENT]). Other programs calculate various coefficients (like kinship or coancestry coefficient) but not the desired relatedness coefficient [GeneticsPed, PEDIGREE VIEWER]; rely solely on genetic data [KINSHIP, WHODAD]; or depend on other programs to create suitable input formats [KING, PMx]. Ultimately, some more programs seemed promising but could not be set to work properly with the available data set [PedKin, PEDIGRAPH]. Therefor, the main aim of this thesis comprises the development of a time efficient pedigree analysis tool which is able to deal with big data sets, missing values as well as the calculation of the relatedness coefficient and further paths characteristics without being restricted to a certain generational depth.

**Rhesus macaques**

Also, to emphasize the importance of data completeness and pedigree depth our data set was collect over 80 years from a population of rhesus macaques in Cayo Santiago (n = 12049 individuals). Generally, rhesus macaques mate promiscuously which implies that kinship among females can usually be observed by mother-infant interactions while paternity data is lacking consistently and has to be assessed by genetic analysis. That is caused by females, mating with multiple males to confuse paternity to ultimately gather care and protection for their offspring from many males to maximize her offspring's fitness and survival rate as well as to prevent infanticide (as surveyed in multiple promiscuous nonhuman primate species: Van Schaik et al. 2004, , ).

Males on the other hand try to ensure the paternity by monopolization, but due to limited information about the fertile phase (), reproductive synchrony (), female choice (Manson, 1992) and alternative male reproductive strategies, mate-guarding by high ranking males was observed to be successful only in 30-40% fertilizations (Dubuc et al., 2012). Sneaky copulations (Berard et al., 1994) are for instance another strategy to increase male mating success which occur both within or outside the social group due to overlapping home ranges. In fact, Ruiz-Lambides et al. (2017) were able to detect an average of 16% extra-group paternities, ranging from 0 to 64.7% in the Cayo Santiago colony.

Nonetheless, male reproductive success of rhesus macaques is highly skewed in relation to the male dominance rank (Smith, 1994). A study from 2004 byWiddig et al. states that 74% of all infants have at least one paternal half-siblings in their birth cohort, because the majority of males do not have any reproductive success while high-ranking males sired the most. In combination with a relatively short male tenure (**?**) and the maternal effort to diversify the genpool over all of her offspring (), most of the siblings in the population are half-siblings which leads to quite complex genetic linkages between the individuals.

Generally, rhesus macaques live with multiple males and females in female-philopatric groups with approximately 8 to 98 individuals as reported in Lindburg (1971) but under semi-natural conditions group sizes up to 300 individuals were observed too (). While females usually stay within their natal group, forming stable matrilines, males tend to migrate multiple times during their life, starting at the age of 3-4 years when they leave their natal group as subadults in search for better mating and breeding opportunities (Boelkins and Wilson 1972). In view of the whole population, the dispersal additionally serves as a strategy to prevent inbreeding (). ...

Even though a consistently genetic analysis is established in the population of Cayo Santiago for over 40 years, half of all sires remain unsolved. Therefor, the second aim of the thesis is to examine whether it is possible to find the most likely parental candidates in case of missing parental information by simulated annealing. To achieve this, the discrepancy between realized relatedness values and the pedigree-derived relatedness is used as function which has to be minimized. In the following it shall be tested upon a randomly filled subset whether this approach can be used to reconstruct an incomplete pedigree based on whole genome sequencing data.

# 2 Methods

## 2.1 Data set and study site

### 2.1.1 Cayo Santiago

To test the functionality of the developed pedigree programme, a long-term demographic and genetic data set was analyzed. The data was gathered over time from a colony of Rhesus macaques at Cayo Santiago - a 15.2ha sized island, located in the Caribbean sea, approximately 1km off the coast of Puerto Rico, nearby the city Punta Santiago.

The primordial population was introduced to the island in 1938 on behalf of the School of Tropical Medicine of the University of Puerto Rico. 409 wild-captured Rhesus macaques from India were released as founder individuals to form a disease-free breeding and free-ranging colony at Cayo Santiago (Kessler and Rawlins, 2016). In 1956, they started to consistently collect demographic data, for instance day of birth/death, social group, matriline, migration status and the observed mother for each individual. Additionally, since 1992, an obligatory blood sample is taken of each individual (during the annual trapping period to register all yearlings) especially to confirm the observed mothers and to assess paternity by genetic analysis.

Under semi-natural conditions, continuous observation, and close monitoring the colony has grown to a colony of approximately 2050 individuals (in 2019, for more detailed information see Fig. 2.1) and is one of the most well studied population of Rhesus macaques throughout time since foundation (Ruiz-Lambides et al. 2017, Widdig et al. 2016, Kazem and Widdig 2013, Manson 1992, Missakian 1972, Altmann 1962). The magnitude of well-documented available data in regards of genetics, social behavioural, and phenotype data makes it exceptionally insightfull

for behavioural and genetic research.

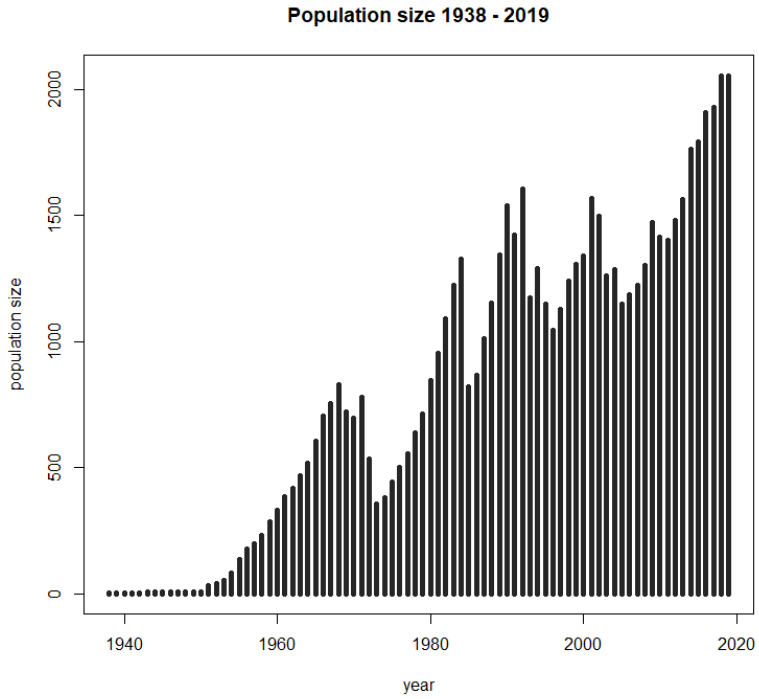



**Figure 2.1: Population size from 1938 to 2019**
Total number of registered individuals in Cayo Santiago for each year from 1938 to 2019. Early values should be interpreted carefully because the comprehensive gathering of demographic data started only in 1956; hence numbers before that date might be unreliable/underestimated.

### 2.1.2 Pedigree data and missing parental information

Today, the population and documentation is managed by the Caribbean Primate Research Center (CPRC) who kindly provide us access to their demographic data which is available for 12049 individuals as well as the genetic data for 5539 individuals (1564 females, 1542 males, 7 individuals of unknown sex). In total, the whole data set comprises 5826 females, 6026 males and 197 individuals with unknown sex (especially not yet registered infants).

Generally, mothers were determined by observation but if ever possible, were confirmed by genetic analysis too. 5138 of 5149 (99.7%) could be confirmed (at least one time in independent analyses run by the CPRC and/or Leipzig University), but for 11 individuals the observed and genetically determined mothers do not match. In this case (if a sample swap could not be excluded) or if no further maternal genetic data was available, we use the observed mother in reference to Widdig et al.

(2017) who pointed out that 98.7% of observed mothers were also the genetically confirmed mothers (see also the supplement of Widdig et al. 2017 for a detailed description of the genetic analysis and combination of both data sets, cumulated by CPRC only - 3113 samples, UL only - 87 samples, or both - 2339 samples). Sires, on the other hand, were determined by genetic analysis by the amount of compliant STR markers. Even if it is not possible to determine a definite sire (due to allelic mismatches in all potential sires who were analyzed), we are able to gain list of explicitly exclude sires which reduces the pool of the remaining potential parents distinctly.

In Figure 2.2, the distribution of missing parental data is plotted in connexion to the birth rate per year (stepped line in the background). The dark bars show for how

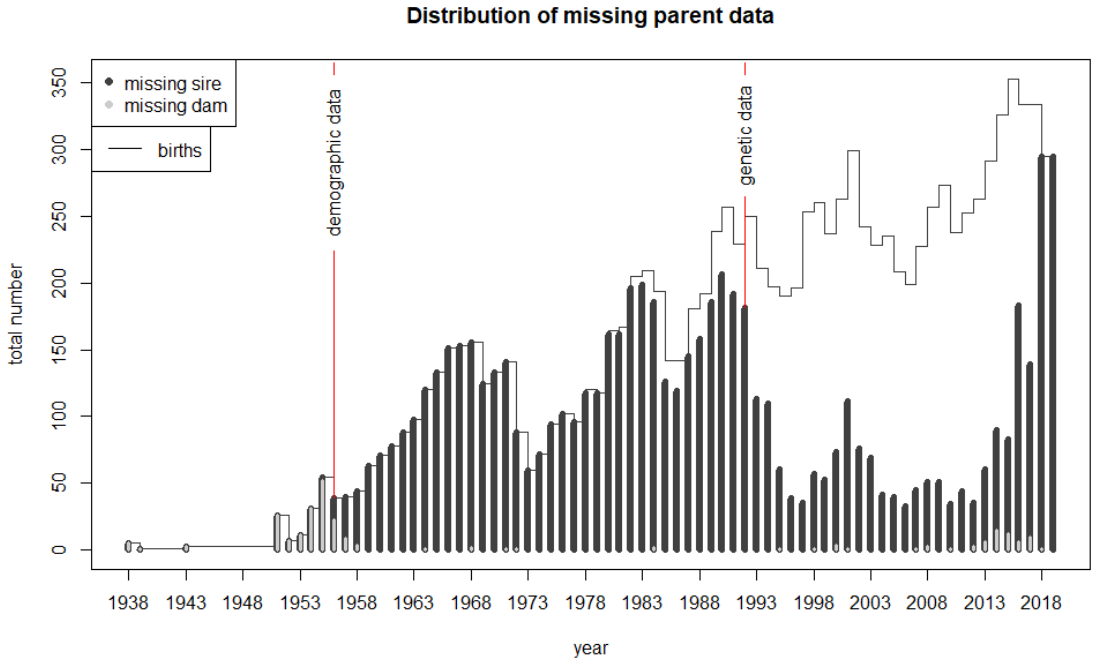



**Figure 2.2: Distribution of missing parent data**
Number if missing sire (dark) and dams (light grey) per birth cohort from 1938 to 2019. Sires can only be genetically determined in case of a known mother, therefore the number of missing sires is restricted to be equal or greater than the number of missing dams per birth cohort.

many of the born offspring the sire hood is uncertain while the lighter ones represens the number of missing dams. It is nicely shown that after starting to consistently collect demographic data, the missing dams decline to a low percentage. The same effect is visible for parental gaps after 1992 when the consistently genetic sampling

started, except for the recent years 2018/2019 for which the genetic data is not yet analyzed respectively included in the data set. In the end 247 mothers (2.02%) and over 7063 sires (58.62%) were unknown.

## 2.2 Implementation

### 2.2.1 Initial programme requirements

Due to the magnitude of the pedigree itself (12049 individuals) and the high amount of gaps within (7310 gaps), the programme itself needs to be able to handle missing parental data as well as to achieve an efficient run time even with a total number of 72 million dyads. Secondly, instead of using partial or truncated pedigrees, it is desirable to use an approach which is not limited by a specific number of generations to get the most exact relatedness coefficients as possible. Therefore, the programme uses a graph theoretical approach, developed in C++, as described in the following.

### 2.2.2 Graph theoretical background

Innately, pedigrees are often visualized in form of a graph because it is an intuitive way to represent genealogy and kinship structures. For our purpose, the genealogy G is delineated as a directed, acyclic graph with two distinct classes of vertices, $V_1$ (males) and $V_2$ (females), with each node referring to an unique individual. Edges are defined as unidirectional parent/offspring relationships which means that each child has one edge coming from its mother, and one coming from its father. To ensure that each individual has a maternal and paternal edge, the graph comprises of two additionally, imaginary nodes $\rho_1 \epsilon V_1$ and $\rho_2 \epsilon V_2$ which are the compensatory substitute in case of an unknown dam or sire. Each node is furthermore characterized with the raw attributes ID-name, sex, birthseason, day of birth (DOB), respectively day of death (DOD), nonsire and nondam (due to genetical analysis excluded parental candidates).

### 2.2.3 Relatedness coefficient calculation

To determine whether two individuals are related or not, the programme searches for kinship paths. Generally, paths are defined as an alternating sequence between incident edges and adjacent nodes, without repeated nodes or edges. Therefore, individuals connected by a path along directed edges from an ancestor to the focal

are related, for instance grandparent/offspring. But due to the hereditary nature of a pedigree, the most common kinship structures consists of two single paths, linked together by the lowest common ancestor (like the uncle/niece pair in Fig. 2.3).

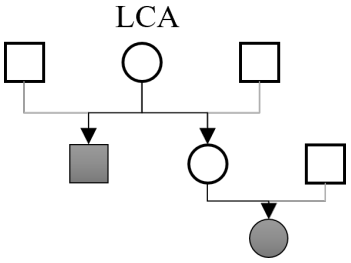



**Figure 2.3: Paths in a pedigree to determine relatedness**
The dark colored individuals are related to each other due to the existence of a common ancestor who is related to both, independently. Since both paths need to be completely independent (without any homolog nodes), only the lowest common anchestor (LCA) is of interest for defining relatedness.

Imaginary individuals were considered as unrelated to each other as well as unrelated to any other individual $x \in V$: $f(\rho_1, \rho_2) = f(x, \rho_1) = f(x, \rho_2) = 0$, while the relatedness of an individual $x \in V$ to itself is set as $f(x, x) = 1$. Otherwise, the relatedness coefficient $f(x, y)$ of a dyad, consisting of the two vertices $x,\ y \in V$, is given by the recursive formula

$$f(x, y) = \frac{1}{4}\left[f(x_1, y_1) + f(x_1, y_2) + f(x_2, y_1) + f(x_2, y_2)\right]$$

with $x_1,\ x_2$ as parental vertices of $x$ and $y_1,\ y_2$ as parents of $y$ ($x_1,\ y_1 \in V_1$ and $x_2,\ y_2 \in V_2$). Furthermore, the relatedness coefficient calculation between an individual $x$ and its ancestor $x_i$ is expressed as

$$f(x, x_i) = \frac{1}{2}\left[f(x_1, x_i) + f(x_2, x_i)\right]$$

with $x,\ x_i \in V$; $x_1 \in V_1$ and $x_2 \in V_2$ as parents of $x$; and more specific in case $x_i \equiv x_1 \vee x_2$, the relatedness between parent and offspring is calculated by

$$f(x, x_1) = \frac{1}{2}\left[1 + f(x_1, x_2)\right]$$

Based on these recursive functions, the programme computes the relatedness be-

tween a dyad stepwise - comparable with a breadth-first-search - until either their lowest common ancestor is found, or it terminates due to a trivial solution.

To enhance the efficiency of the potentially exponential relatedness calculation, a feature was implemented to reduce the node space beforehand by using the intersection of all ancestors of both focals. But since in an additionally second step all descendants of these selected individuals needs to be added to the subset too, the memory expenditures increases significantly. As a result, the efficiency of prefiltering depends highly on the pedigree structure, especially regarding the completeness. In fact, for our data set, comprising a pedigree with over 7000 gaps, it does not improve the execution time at all.

### 2.2.4 Validation by simulated pedigree

In order to validate the functionality and accuracy of the relatedness coefficient calculation, an artificial pedigree was simulated. Parameters to set for simulation are: start year, simulation duration, number of start individuals, gestation length, maturation age of females and males and the maximal age. For testing purpose, a pedigree of approximately 1300 individuals distributed over five generations was created, while further parameters were taken from a previous study (Widdig et al., 2017) about inbreeding assessment in rhesus macaques (gestation length = 200 days - a liberal adjustment due to a reported mean gestation length of 166.5 days by Silk et al. (1993), female maturation age = 1095 days and male maturation age = 1250 days, see also Bercovitch et al. 2003). Additionally, a random sex and birth date (day and month) was simulated for each individual. Deduced from the original rhesus macaque pedigree, the population growth is also characterized by an increasing number of birth and deaths every consecutive year. In case of start individuals (representing founder individuals, born in start year), the mother as well as sire is set as unknown. The actual pedigree simulation starts in the year after all start individuals reached maturity. From this point on, descendants are simulated yearwise. Therefor, the function filters for a subset of potential mothers and sire, selects randomly a parent pair for each offspring from the subset, and removes each selected mother afterwards from the pool of potential mothers to avoid twins as well as from the pool of potentially deceasing individuals, which is also determined randomly.

During simulation, an extra class SIMPATH is used to track the relatedness for all simulated dyads. Therefor a map with an entry for each individuals and its belonging simpaths is used. A simpath consists of the path name (e.g. A_B_C), the associated relatedness coefficient and the dyad name (comprising the first and last individual from path name, e.g. A_C). When ever an individuals was chosen as parent, all paths associated to the parents are examined whether they needs to be expanded and updated. In the end, all relatedness paths are listed and successfully compared with the calculated relatedness coefficients from above.

### 2.2.5 Additional path information

As main output, the programme provides the relatedness coefficients of either preselected or all possible dyads. But if required, the programme exports all relevant paths between the dyads in the pedigree too as well as further path information like *lca*, *depth*, *kinlabel*, *pathline*, *kinline*, *fullhalf* and the minimal dyadic genealogical depth (*min_DGD*) per path (see explanation and example in Tbl. 2.1 and Fig. 2.4).

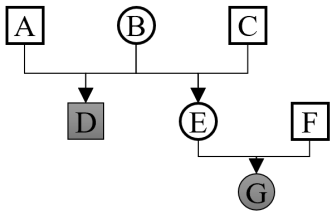



**Figure 2.4: Pedigree example (Tbl. 2.1)**
Simple pedigree example to illustrate the implications of the calculated path characteristics. Squares depict male individuals; circles females. The focal individuals $D$ and $G$ are related only by maternal ancestors (kinline = mat), whereby the lowest common ancestor $B$ is one edge apart from $D$ and two from $G$ (depth = 1/2) which codes for the kinlabel uncle/niece. For each focal all parents are resolved but both' their grandparent generation is incomplete, therefore the min_DGD is 2.

The original path information is collected during the calculation of the relatedness coefficient, i.e. in each recursive substep, whenever a relatedness coefficient greater than zero could be detected, the relevant individual(s) are appended to each path of the current dyad, either on one side or both ends - depending on the kind of kinship.

**Table 2.1: Overview and explanation of all path characteristics obtained from the pedigree programme**

| name | explanation | example |
|---|---|---|
| path | consecutive list of nodes along the relatedness path (edge directions are left unregarded) | D@B@E@G |
| lca | lowest common ancestor within path | B |
| pathline | sequence of sexes (f/m/u) along the path | mfff |
| kinline | whether the path consists solely of maternal or paternal ancestors; "mixed" if both sexes occur | mat |
| depth | path length from LCA to each focal | 1/2 |
| kinlabel | kinclass label based on the table of consanguinity (Fig. 4.1) | uncle-niece |
| fullhalf | whether two identical paths exist with different common ancestors, e.g. differentiation between full- and half-siblings | half |
| min_DGD | minimal dyadic genealogical depth states the pedigree completeness for the dyad; i.e. the minimal amount of fully resolved generations starting from both focals | 2 |

## 2.3 Gap filling by simulated annealing

### 2.3.1 General idea of simulated annealing

To achieve the second aim of the thesis, a simulated annealing algorithm was implemented to try to ascertain the missing parents based on the realized relatedness. The concurrent fundamental idea of simulated annealing goes back to Kirkpatrick et al. (1983) and Černỳ (1985). Originally, it was developed to solve optimization problems whenever the underlying algorithm cannot be solved in polynomic time or is in general either unknown or too complex to apply (e.g, minimizing the route in a traveling salesman problem).

Generally, simulated annealing mirrors - metaphorically speaking - a gradually cooling process derived from thermodynamics which allows molecules to order themselves stage-by-stage in a state of minimal energy instead of cooling them rapidly down which often results in the end in somewhat less than robust or energetically optimal structures (Brooks and Morgan, 1995). For optimization problems, the step-

wise declining procedure is especially beneficial to find the global minimum instead of getting stuck within a local minimum, since the algorithm allows also slightly worse solution along the way to the global optimal. Whereby the probability of favouring a worse solution, respectively leaving a (potential) local minimum, highly depends on the current temperature parameters (see Fig. 2.5).

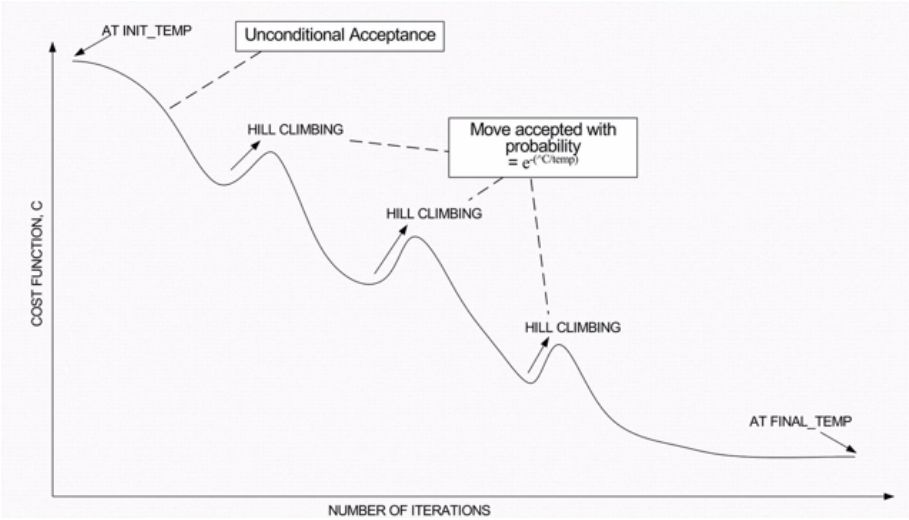



*Figure from Aydemir and Karagül (2020)*

**Figure 2.5: Overview simulated annealing**
The algorithm starts with a chosen start temperature, slowly declining throughout, and stops as soon as the current temperature falls below the final temperature. Beginning with a random start solution, in each iteration a neighbour solution from the solution before is analyzed in regards to the fitness function and if the new solution (with fitness cost of $F_n$) performs better than the current one ($F_c$), the algorithm accepts it unconditionally as the new current solution. Otherwise, it is a matter of accepting a worse solution which is usually controlled by an acceptance criterion like the Metropolis acceptance criterion $e^{\frac{F_n - F_c}{T}}$. This implies that, the lower the temperature $T$, the more restrictive the criterion will be, whereas the probability depends also on the total difference of $F_n$ and $F_c$ (Bertsimas and Tsitsiklis, 1993).

### 2.3.2 Adapted algorithm

Adapted, to fit our gap problem, the algorithm can be simplified as in the following outline:

1. Get all gaps to fill in the incomplete pedigree
2. Create start solution by randomly determining missing parents from the pool of fitting parents for each gap

3. Relatedness coefficient calculation for each relevant dyad (dyads with IBD values to compare to)
4. Compare old versus new relatedness values for each relevant dyad and get the total difference between start solution and incomplete pedigree which ultimately serves as fitness function to be minimized:

$$F = \Sigma \, |f(x, y) - g(x, y)| \rightarrow min$$

(with $f(x, y)$ as the pedigree-based dyadic relatedness and $g(x, y)$ as the dyadic realized relatedness)
5. Save current total difference as (currently) best difference and start solution as (currently) best pedigree
6. While current temperature > stop temperature:
   (a) Create neighbor solution (exchange one potential parent with another fitting candidate)
   (b) Calculate new relatedness values for each dyad in associated to one of the three altered individuals (the individual with the missing parent, the newly chosen parent candidate and the previously chosen parent candidate)
   (c) Compare old vs. new relatedness values for each relevant dyad and get the total difference between current pedigree (starting point for neighbor solution) and neighbor solution
   (d) if the new solution is worse, use the metropolis acceptance criterion to determine whether the new solution will be rejected or not:

$$e^{\frac{F_n - F_c}{T}} > X \rightarrow [0, 1]$$

   (with $F_n$ as fitness function of the new solution and $F_c$ of the current solution; T as temperature and X as random number in the range between 0 and 1)
   (e) if true, set new solution as new current solution (starting point for next solution), else reject new solution and previous current solution endures
   (f) if necessary update best difference/pedigree
7. save last pedigree solution in file

### 2.3.3 Data subset

For testing purpose, whether or not the algorithm is able to find the best solution in an appropriate time, the original pedigree was truncated to 8744 individuals (72.6%), respectively 38 224 396 dyads. In the pedigree subset, all individuals are included who were born in or before 1986 which equals three times the Rhesus macaques' generation time of 11 years (Xue et al., 2016), so that it remains to be a multi-generational pedigree, but it mainly encompass the time (since 1992), when they have started to run systematic genetic analysis for all individuals. Therefor, the number of individuals is reduced to three quarter while the number of gaps has actually more than halved (42.8%). Additionally, all known parents who did not occur in the subset so far, were added but with *unknown* labeled parents for themselves. In case of Baby-IDs (infants who were not yet cataloged, usually either because they are from the current birth cohort or they died in the first year) with missing DOD dates (day of death), an artificial DOD date was set to exclude these individuals as potential parents.

Generally, the pool of potential parental candidates comprises all individuals of the respective sex, who were alive and mature at the time of conception (males) or time of birth (females) to be a reasonable parent candidate. As established in Widdig et al. (2017), the female maturation age is set to 1095 days and for males to 1250 days. The length of gestation was again estimated with 200 days. Additionally, males excluded by STR marker during paternity assessment, were excluded as potential sire too. Furthermore, only females, who did not have an assigned offspring in the respective birth cohort were considered as a potential dam because the probability to bear twins in Rhesus macaques is very low ([xxx]%, as mentioned in [xxx]). In a few occasions no sire could be assigned because of an empty parent pool due to the truncation. In this case, one of the previously by STR markers excluded sires was chosen as the *real* sire to prevent the pedigree to expand further.

Otherwise, as the *real* underlying pedigree without gaps, a random solution was used, generated similarly to the start solution described previously. Despite using a truncated pedigree, a total of 3129 gaps with at average 250 potential parent

candidates for each gap remains which generates a landscape with $250^{3129} = 1426 \times 10^{7500}$ potential solutions. So to test the general functionality, the algorithm iterate over approximately 10 million solutions to test whether or not it is possible to achieve a suitable solution in reasonable time (mean calculation time for each new solution: 0.15-0.208s; temperature decay = $\sqrt[10000000]{\frac{t_{stop}}{t_{start}}} = \sqrt[10000000]{\frac{1}{1191.61}} = 0.9999992916945$ with a start temperature of 1191.61 and a stop temperature of 1).

### 2.3.4 Simulation of WGS data

To simulate whole genome sequencing data, the possibility of recombination have to be taken into account which ultimately results into varying degrees of relatedness in similar kinclasses (see Tbl. 2.2). These simulated IBD values were used to adapt the relatedness values from the real pedigree by choosing a random value from within the simulated range in Tbl. 2.2.

**Table 2.2: Simulated IBD values (Freudiger unpublished)**

| kinlabel | mean | maximum | minimum |
|---|---|---|---|
| full siblings | 0.501 | 0.624 | 0.333 |
| parent-offspring | 0.5 | 0.5 | 0.5 |
| half siblings | 0.254 | 0.34 | 0.173 |
| full aunt/uncle-niece/nephew (1st) | 0.252 | 0.325 | 0.147 |
| grandparent-offspring | 0.248 | 0.335 | 0.152 |
| half aunt/uncle-niece/nephew (1st) | 0.127 | 0.213 | 0.059 |
| full aunt/uncle-niece/nephew (2nd) | 0.127 | 0.196 | 0.074 |
| full cousins (1st) | 0.124 | 0.194 | 0.06 |
| greatgrandparent-offspring | 0.123 | 0.194 | 0.06 |
| half cousins (1st) | 0.066 | 0.126 | 0.019 |
| full aunt/uncle-niece/nephew (3rd) | 0.063 | 0.116 | 0.024 |
| half aunt/uncle-niece/nephew (2nd) | 0.062 | 0.113 | 0.01 |
| full cousins (1st) once rem. | 0.061 | 0.129 | 0.029 |
| greatgreatgrandparent-offspring | 0.061 | 0.121 | 0.016 |
| half cousins (1st) once rem. | 0.032 | 0.073 | 0.002 |
| full cousins (1st) twice rem. | 0.031 | 0.073 | 0.006 |
| half aunt/uncle-niece/nephew (3rd) | 0.031 | 0.077 | 0.007 |
| full cousins (2nd) | 0.03 | 0.071 | 0.005 |
| full cousins (2nd) once rem. | 0.016 | 0.047 | 0.002 |
| half cousins (1st) twice rem. | 0.016 | 0.049 | 0 |
| half cousins (2nd) | 0.016 | 0.059 | 0 |
| full cousins (3rd) | 0.01 | 0.028 | 0 |
| half cousins (2nd) once rem. | 0.009 | 0.044 | 0 |
| half cousins (3rd) | 0.006 | 0.032 | 0 |

# 3 Results

## 3.1 Pedigree analysis

### 3.1.1 Path distribution in Cayo Santiago

To gain a greater insight into the population structure of Cayo Santiago, all 72 583 176 dyads were analyzed by the developed programme. All in all, it lasted 2 days and 7 hours without parallelization. With multithreading based on 60 cores, the run time could be reduced to [xxx].

Althogether, approximately 32 million dyads (44.15%) were related by over 143 million paths in total. Thereby, the n umber of paths a dyad is related by, ranges from 1 to 120, with a mean of 4.5 paths per dyad, see Fig 3.1.

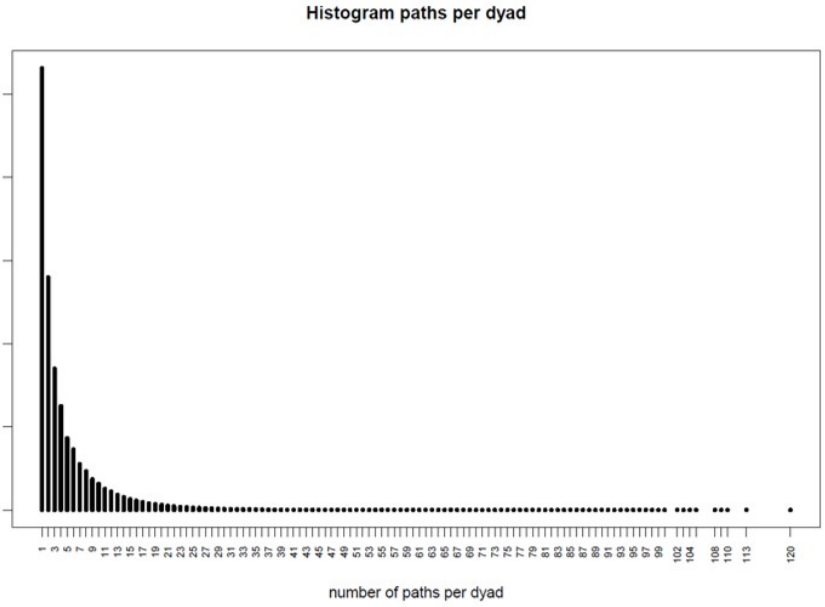



**Figure 3.1: Paths per dyad**
Histogram which shows the number of paths a dyad is related by, n = 32 042 071 related dyads.

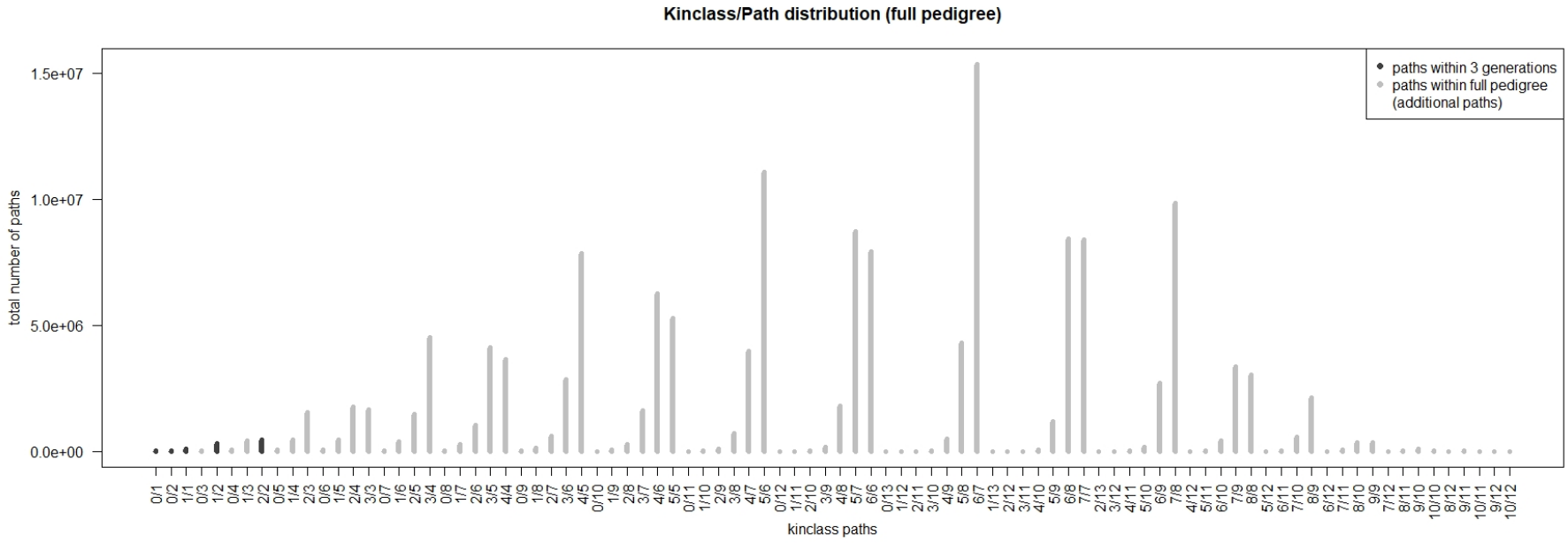



**Figure 3.2: Path distribution**

The plot shows the path distribution of all 32 042 071 kins in a population of 12 049 individuals. The x axis lists all detected path depths - respectively the different kinclasses (e.g. 0/1 = parent/offspring; 1/1 = siblings) while the y axis represents the frequency. The left number from the depth label codes for the number of edges or generational steps between the first focal and the lowest common ancestor, and similarly the right value from the second focal to the lowest common ancestor. The darker colored bars marks paths which would be covered by a programme which only calculate the relatedness up to three generations from the focals (grandparent generation) while all other ancestors are considered unrelated. Contrary to that, the lighter bars originate from a unrestricted programme.

The majority of paths are characterized with a mixed kinline (95.9%). For the remaining paths, the ratio between maternal paths and the number of paths consisting only of paternal ancestors is highly skewed with 95:5 (5 656 001 vs. 311 042) which is unexpected since almost all mothers are known (98%) while every second sire is missing.

Per average, a single individual is related to 44.7% of all of the other individuals in the pedigree (min: 0%, max: 86.3%). But the most dyads represent rather distant kins, as it is shown in Fig. 3.2. For example, some paths have actually a total path length of 22 which is equivalent to a relatedness value of 0.00000024. And the most prominent peak with 15 million instances is at 6/7 which are *fifth cousins once removed* with a relatedness value of 0.0001.

Ultimately, the mean relatedness coefficient in the population is about 0.0055. But generally it is important to note that this is the minimum value because the data set consists of multiple gaps and the programme is only able to offer the reliably verifiable kinship paths while considering any unknown parent as unrelated to every other individual.

### 3.1.2 Generational restriction - a comparison

Originally, the idea to develop a new pedigree analysis tool was inspired because most studies were conducted with either small pedigrees or a pedigree analysis programme which is restricted in regards of the considered generations, which both ultimately results in rather categorial relatedness values instead of continuous ones. To investigate the amount of possibly lost data, the following section focus on a comparison between the new programme and a previously used programme (Widdig et al., 2017) which considers for a dyadic relatedness coefficient calculation only pedigree data up to the level of the focals' grandparents.

Both programmes were compared based on a subset of 385 000 dyads while the original pedigree with 12049 individuals remains unaltered. All in all, the previous programme differentiate in ten different relatedness levels, like 0 for nonkin and 0.5 for parent-offspring pairs or full-siblings; see x-axis of Fig. 3.3. The plot shows further how the newly calculated relatedness coefficients are distributed over each of these relatedness levels. As the mean in Fig. 3.3 indicates, the relatedness

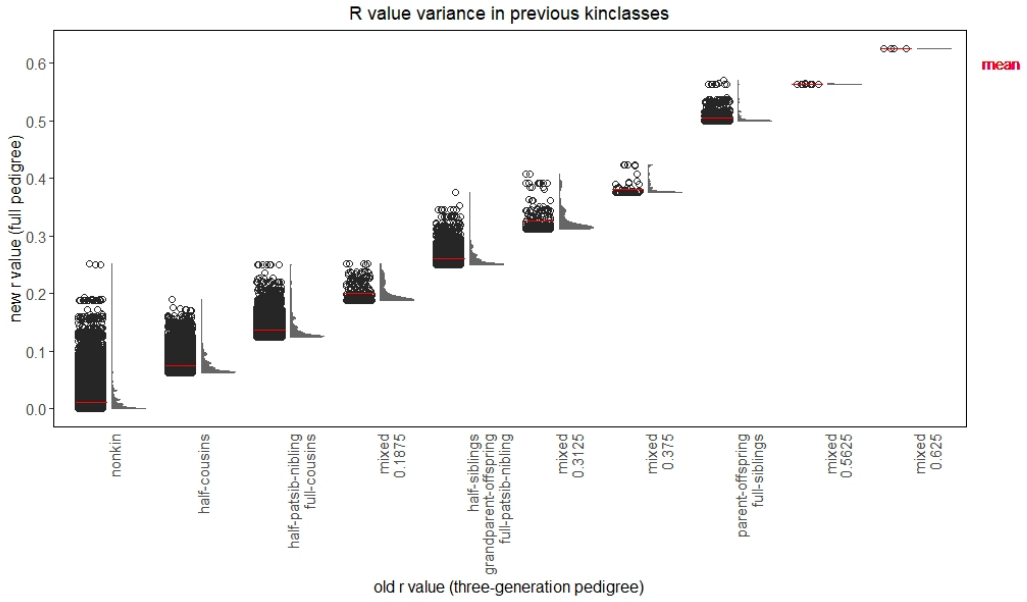



**Figure 3.3: Relatedness variance in basic kinlevels**
The plot compares the outcome of the relatedness coefficient calculation of two pedigree analysis programmes - one unrestricted while the other programme calculation was restricted to take only two generations of ancestors from the focals into account. Essentially, both axes show relatedness values but while the y-axis constitute a rather continuous scale, the x-axis is divided into categorial kinclass levels whose respective relatedness coefficients can be visually derived from the lowest "line" of data points, e.g. half-cousins = 0.0625, full-cousins = 0.125, or half-siblings = 0.25. Each kinlabel is splitted into the data points itself on the left side and a violin plot to unveil the distribution of the data points too. The red line represents the average relatedness coefficient from the new programme for each relatedness level.

coefficient increases distinctly, especially in these levels that were rather low to begin with. Half-cousins are for example at average 20.7% more related than the previous programme would have implied (range in the other kinlevels: 9 to 0.07%).

Moreover, the proportion of related individuals to nonkin increases from 8 to 89% and the mean relatedness shifts from approximately 0.01 to 0.02. All in all, the new programme found over 1.5 million paths between all dyads (compare to approximately 41 000 paths). Also 83.33% of all kins share multiple common ancestors while the previous programme could detect multiple paths only in 20.28% of all related dyads.

### 3.1.3 Community detection

To examine potentially subgroups/communities in the rhesus macaques colony of Cayo Santiago, the python modules PYTHON-LOUVAIN by Aynaud (2020) and NET-

WORKX (Hagberg et al., 2008) was used. The Louvain Algorithm, originally developed by Blondel et al. (2008), starts with each community consisting of a single node and tries to merge these communities together in order to maximize the modularity of a graph. Basically, modularity measures the amount of connectivity within a subgroup in comparison to nodes outside of the group. Optimally, nodes within a community are highly connected by heavy weighted edges while the edges to members of other communities are a lot less dense and exhibit a lower edge weight. Adapted to our pedigree, it tries to answer the question whether there are subpopulations in the colony which can be detected by pedigree/relatedness. After the pedigree assessment and dyadic relatedness calculation, 148 individuals were detected who are not related to any other individual (mainly individuals with unknown mother and sire). But for the remaining 11 901 individuals a community clustering was performed based on the Louvain algorithm. All in all, the unsupervised learning algorithm detected [48] communities, but only [5] of them are of a notable size ( [100]) which are plottet in Fig. 3.4.

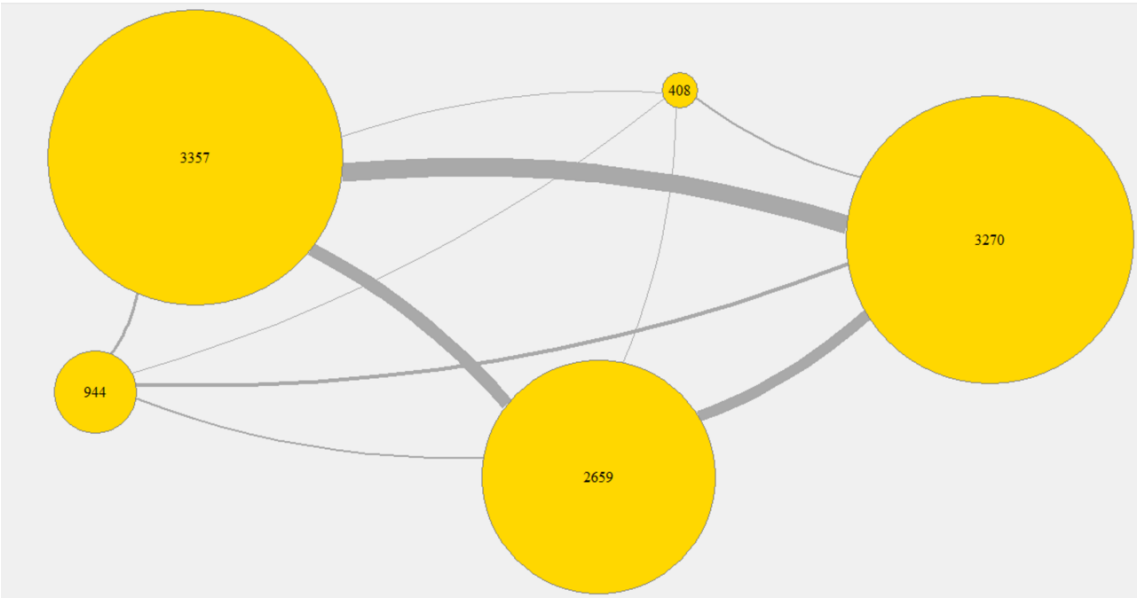



**Figure 3.4: Community detection**
A simplified network plot in order to illustrate the relatedness between the found subgroups (communities) within the population of Cayo Santiago based on pedigree data, analyzed by the Louvain algorithm (Blondel et al., 2008). The diameter as well as the numbers within the community circles state the size of each community; while the edge width represents the relative amount of relatedness between the detected community (total relatedness divided by the number of individuals from both communities).

### 3.1.4 Half-siblings

The analysis of half-sibling pairs suggests a notable difference beween maternal and paternal half-siblings as shown in Fig 3.5. The figure might suggest that paternal seem to be slightly more related than maternal half-siblings but the median difference remains low (paternal: 0.2559, maternal: 0.25). On the other hand, maternal half-siblings are distributed over multiple cohorts (median = 4), but the median of paternal ones is distincly lower (2).

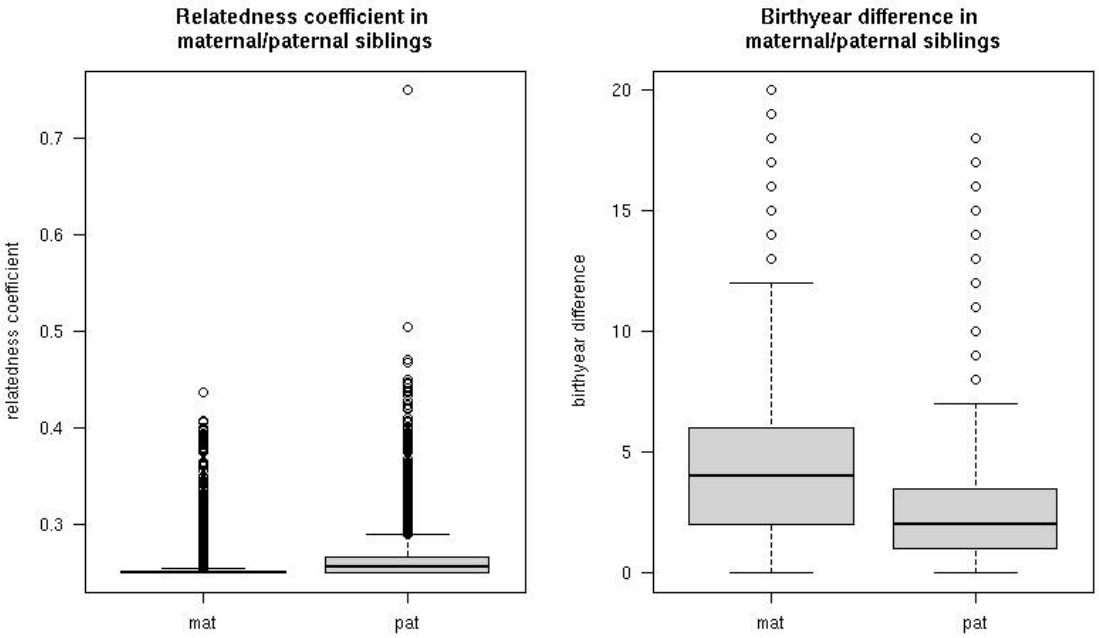



**Figure 3.5: Half-siblings**
Boxplots of maternal versus paternal half-siblings in regards to relatedness and birth year difference between the focals of the respective dyad. Full-siblings were excluded.

### 3.1.5 Inbreeding assessment

Additionally to the relatedness coefficient and the path characteristics, the programme provides the minimal inbreeding value as well as the number of fully known generations for each individual. The inbreeding value $F$, is defined as the product of 0.5 and the parental dyadic relatedness (Wright, 1922). But since the pedigree contains multiple gaps, this measurement respresents only the minimal inbreeding which was traceable within the current data set. Therefor, multiple subsets, filtered by the number of full generations, were distinguished to examine the percentual

shift of inbreeding depending on the available pedigree data (see Fig. 3.6). A generation counts as fully known if all associated ancestors within the respective depth from the focal are known. For instance, an individual with a known mother and a missing sire, would fit in the category of one full generation (the generation of the individual itself). But once mother and sire are known, two generations were fully known because all possible ancestors within a pedigree depth of one are known.

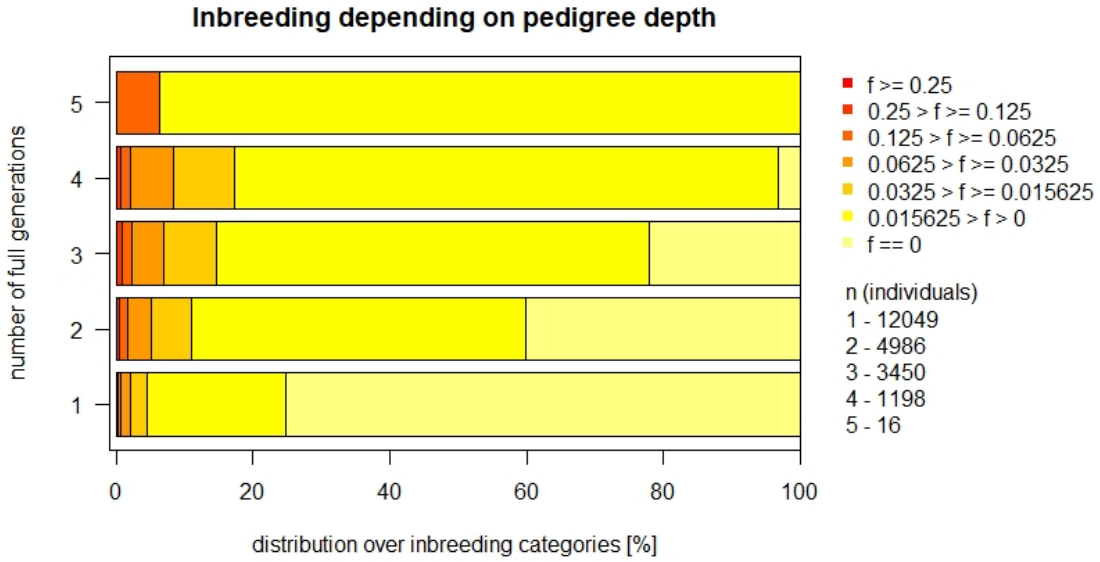



**Figure 3.6: Inbreeding in correlation to pedigree depth**
The x axis shows the percentage distribution of the color-coded inbreeding categories based on the total number of individuals who fit into the subsets, labeled on the y axis which codes for the number of completely known generations. An inbreeding value of 0.25 would implicate that parents of an individual would be full-siblings or parent and offspring; while an individual with $F$ between 0.125 and 0.0625 would have parents between the level of half-siblings and for instance full cousins.

The more generations are known, the more pronounced is the shift towards slight inbreeding. The bottommost bar, comprising all individuals regardless how much ancestors are known, is dominated by an inbreeding value of zero. But for example when at least four generations are fully known, which covers all individuals with a complete pedigree at least up to the level of its great-grandparents, almost all of their parents are related at some level, even though they are mostly very distant kins.

Furthermore, Fig. 3.7 suggests that the relatedness over time does increase. In the plot, all dyads were sorted by their birthyear combination (e.g. 2010_2015)

to plot the average dyadic relatedness coefficient as a heatmap for each specific cohort and birthyear combination. But eventually, it might be biased by the depth of available pedigree information (the number of generations ahead as well as the amount of gaps in close distance).

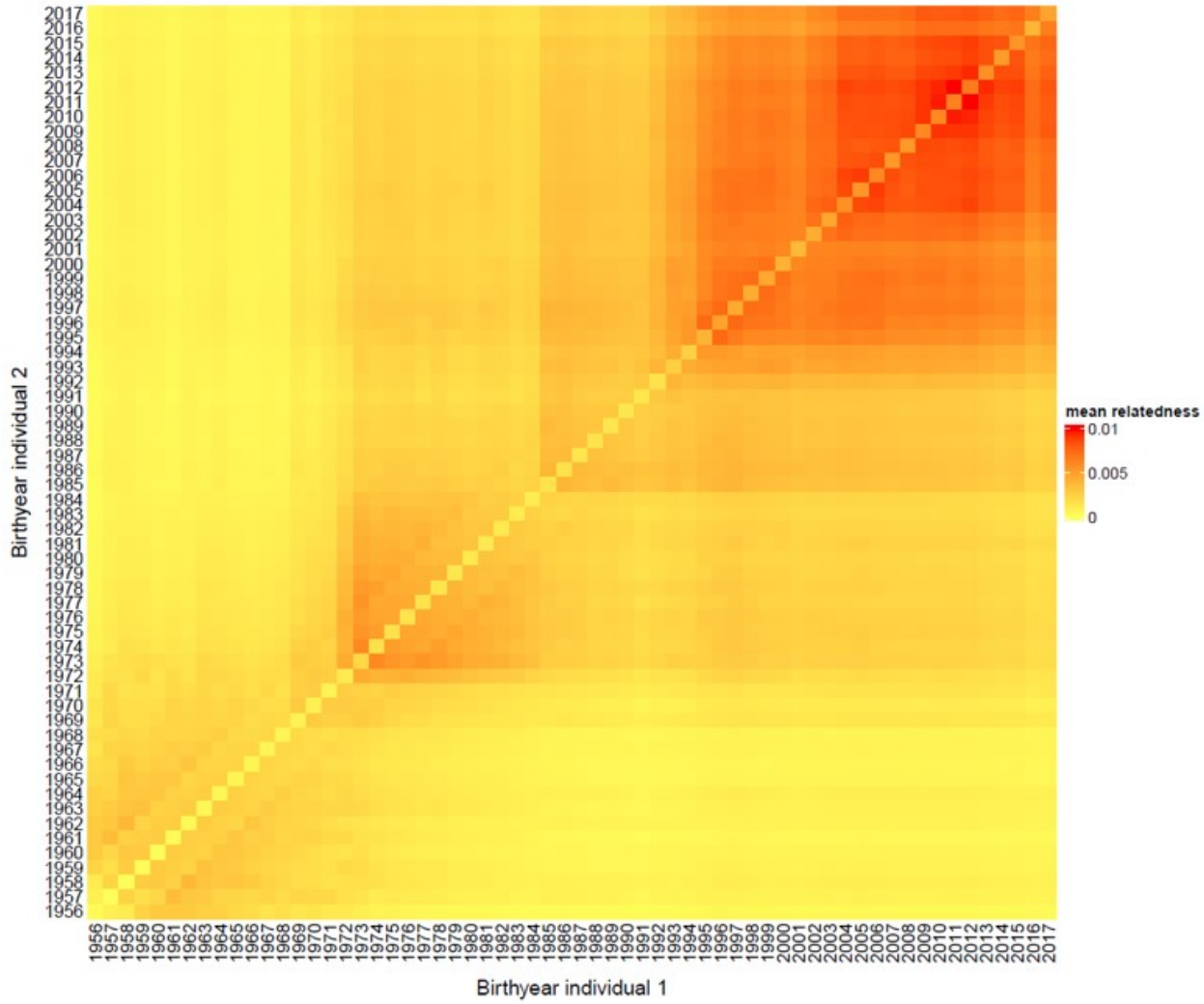


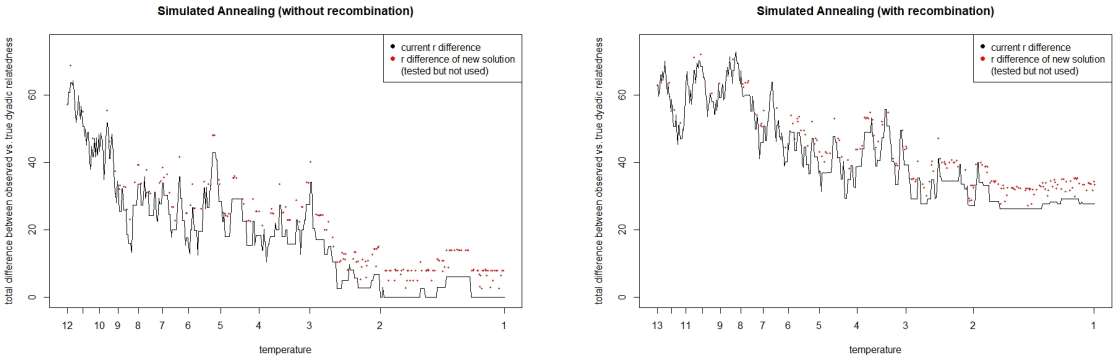

**Figure 3.7: Relatedness over time**

...

## 3.2 Simulated Annealing

... At first, to demonstrate the general functionality of the implemented simulated annealing function, the algorithm was tested in a small isimulated pedigree example with 37 individuals. In Fig. 3.8a and 3.8b, the impact of realized relatedness (obtained from whole genome sequencing) contrary to pedigree-derived values in regards of pedigree reconstruction is shown. In the end, in both plots the discrepancy

between the relatedness values of the current and the real solution declines distinctly over the course of the annealing algorithm which means that the current solution converges more and more towards the real pedigree. All in all, 496 respectively 512 solutions were tested which resulted in a correct pedigree reconstruction (Fig. 3.8a) and one false assigned sire in Fig. 3.8b.




**(a) without recombination** **(b) with added recombination values**

**Figure 3.8: Simulated annealing example**

...

# 4 Discussion

During the course of this master thesis, it was possible to develop a pedigree programme which provides validated dyadic relatedness coefficients, inbreeding values, as well as some further path attributes in order to characterize the type of relatedness based on available pedigree data. At the same time, it could be proven that the functionality is not constrained, regardless the genealogical depth (up to 12 (incomplete) generations), the high amount of unknown sire ( 50%xxx), or the big data set in general (72 million dyads due to a total of 12049 individuals which were still analysed within a decent time of 2.3 days). However, the pedigree programme can provide exact minimal values (regarding relatedness and inbreeding), but the overall accuracy highly depends on the completeness of the pedigree itself. To evaluate values regarding their reliability, the programme supplies the minimal genealogical depth for each dyad which I highly suggest to consider in further interpretation of the dyadic relatedness/inbreeding values. The same applies to the following discussion of the results.

As expected, the results show that the depth-independent approach offers a lot more detailed relatedness values than a limited version. The more historical pedigree data is available to analyze, the more continuous the values will be instead of "categorical" (0.5, 0.25, 0.125). That is because the probability to be related by multiple paths as well as the combinatorial possibilities increase with pedigree depth. In the data set from Cayo Santiago multiple dyads could be identified who are related via over 100 paths at once. Nonetheless, these respective dyads do not feature a particularly high relatedness or inbreeding value (for instance dyad [xxx_xxx], related by 120 paths, has a relatedness coefficient of [xxx] and inbreeding values of xxx (individual xxx) and xxx (individual xxx)). And even though the average number of paths between dyads is around 4-5, inbreeding values do not

suggest a highly inbred population which would coincide with the findings of a study by (Widdig et al., 2017) from 2017 who investigated the same population (except the two most recent cohorts) but with a pedigree programme which restricts the depth to three generations. The researchers suggest that 7% of all individuals seemed to be inbred while due . Such similar results... use only individuals with at least 3 full generations (all parents and grandparents have to be known) which corresponds to approximately 29% of the population.

Dyadic relatedness values have a broad application, e.g. population structure, ... This thesis aims to demonstrate the amount of information could be obtained from pedigree as well as to emphasize the importance of deep, long-time, complete pedigree - especially in hindsight that much of past (but also recent) studies relied on results based on less solid pedigree data, e.g. consideration of only 1 or 2 generations above.

Colony (Jones and Wang, 2010)! kanthaswamy2017 genetic

priorly reduce parent pool by Mini-SA

# References


Altmann, S. A. (1962). A field study of the sociology of rhesus monkeys, macaca mulatta. *Annals of the New York Academy of Sciences*.

Aydemir, E. and Karagül, K. (2020). Solving a periodic capacitated vehicle routing problem using simulated annealing algorithm for a manufacturing company. *Brazilian Journal of Operations  Production Management*, 17.

Aynaud, T. (2020). python-louvain x.y: Louvain algorithm for community detection. https://github.com/taynaud/python-louvain.

Berard, J. D., Nurnberg, P., Epplen, J. T., and Schmidtke, J. (1994). Alternative reproductive tactics and reproductive success in male rhesus macaques. *Behaviour*, 129(3-4):177–201.

Bercovitch, F. B., Widdig, A., Trefilov, A., Kessler, M. J., Berard, J. D., Schmidtke, J., Nürnberg, P., and Krawczak, M. (2003). A longitudinal study of age-specific reproductive output and body condition among male rhesus macaques, macaca mulatta. *Naturwissenschaften*, 90:309–312.

Bertsimas, D. and Tsitsiklis, J. (1993). Simulated annealing. *Statistical science*, 8(1):10–15.

Birch, J. and Okasha, S. (2015). Kin selection and its critics. *BioScience*, 65(1):22–32.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Boelkins, R. C. and Wilson, A. P. (1972). Intergroup social dynamics of the cayo santiago rhesus (macaca mulatta) with special reference to changes in group membership by males. *Primates*, 13:125–139.

Brooks, S. P. and Morgan, B. J. (1995). Optimization using simulated annealing. *Journal of the Royal Statistical Society Series D: The Statistician*, 44(2):241–257.

Carter, K. D., Brand, R., Carter, J. K., Shorrocks, B., and Goldizen, A. W. (2013). Social networks, long-term associations and age-related sociability of wild giraffes. *Animal Behaviour*, 86(5):901–910.

Černỳ, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of optimization theory and applications*, 45:41–51.

Chapais, B. (2001). Primate nepotism: what is the explanatory value of kin selection? *International Journal of Primatology*, 22:203–229.

Darwin, C. (1956). The origin of species: By means of natural selection or the preservation of favoured races in the struggle for life. Technical report, Oxford University Press,.

Dubuc, C., Muniz, L., Heistermann, M., Widdig, A., and Engelhardt, A. (2012). Do males time their mate-guarding effort with the fertile phase in order to secure fertilisation in cayo santiago rhesus macaques? *Hormones and Behavior*, 61(5):696–705.

Eckert, K. A. and Hile, S. E. (2009). Every microsatellite is different: Intrinsic dna features dictate mutagenesis of common microsatellites present in the human genome. *Molecular Carcinogenesis: Published in cooperation with the University of Texas MD Anderson Cancer Center*, 48(4):379–388.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

Foroughirad, V., Levengood, A. L., Mann, J., and Frère, C. H. (2019). Quality and quantity of genetic relatedness data affect the analysis of social structure. *Molecular ecology resources*, 19(5):1181–1194.

Gardner, A., West, S. A., and Wild, G. (2011). The genetical theory of kin selection. *Journal of evolutionary biology*, 24(5):1020–1043.

Glover, K. A., Hansen, M. M., Lien, S., Als, T. D., Høyheim, B., and Skaala, Ø. (2010). A comparison of snp and str loci for delineating population structure and performing individual genetic assignment. *BMC genetics*, 11:1–12.

Gompper, M. E., Gittleman, J. L., and Wayne, R. K. (1997). Genetic relatedness, coalitions and social behaviour of white-nosed coatis, nasua narica. *Animal Behaviour*, 53(4):781–797.

Hagberg, A., Swart, P., and S Chult, D. (2008). Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Hamilton, W. D. (1964). The genetical evolution of social behaviour. ii. *Journal of theoretical biology*, 7(1):17–52.

Jones, A. G., Small, C. M., Paczolt, K. A., and Ratterman, N. L. (2010). A practical guide to methods of parentage analysis. *Molecular ecology resources*, 10(1):6–30.

Jones, O. R. and Wang, J. (2010). Colony: a program for parentage and sibship inference from multilocus genotype data. *Molecular ecology resources*, 10(3):551–555.

Kazem, A. J. and Widdig, A. (2013). Visual phenotype matching: cues to paternity are present in rhesus macaque faces. *PLoS One*, 8(2):e55846.

Kessler, M. J. and Rawlins, R. G. (2016). A 75-year pictorial history of the cayo santiago rhesus monkey colony. *American Journal of Primatology*, 78(1):6–43.

Kirkpatrick, S., Gelatt Jr, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *science*, 220(4598):671–680.

Krawczak, M. (1999). Documentation of find sire. *https://www.uni-kiel.de/medinfo/mitarbeiter/krawczak/download/findsire.txt* Accessed: 2023-10-10.

Lindburg, D. G. (1971). The rhesus monkey in north india: an ecological and behavioral study. *Primate behavior: developments in field and laboratory research*, 2:1–106.

Manson, J. H. (1992). Measuring female mate choice in cayo santiago rhesus macaques. *Animal Behaviour*, 44:405–416.

Missakian, E. A. (1972). Genealogical and cross-genealogical dominance relations in a group of free-ranging rhesus monkeys (macaca mulatta) on cayo santiago. *Primates*, 13:169–180.

Park, S. T. and Kim, J. (2016). Trends in next-generation sequencing and a new era for whole genome sequencing. *International neurourology journal*, 20(Suppl 2):S76.

Perry, S., Manson, J. H., Muniz, L., Gros-Louis, J., and Vigilant, L. (2008). Kin-biased social behaviour in wild adult female white-faced capuchins, cebus capucinus. *Animal Behaviour*, 76(1):187–199.

Pinter-Wollman, N., Hobson, E. A., Smith, J. E., Edelman, A. J., Shizuka, D., De Silva, S., Waters, J. S., Prager, S. D., Sasaki, T., Wittemyer, G., et al. (2014). The dynamics of animal social networks: analytical, conceptual, and theoretical advances. *Behavioral Ecology*, 25(2):242–255.

Rengmark, A. H., Slettan, A., Skaala, Ø., Lie, Ø., and Lingaas, F. (2006). Genetic variability in wild and farmed atlantic salmon (salmo salar) strains estimated by snp and microsatellites. *Aquaculture*, 253(1-4):229–237.

Ruiz-Lambides, A. V., Weiß, B. M., Kulik, L., Stephens, C., Mundry, R., and Widdig, A. (2017). Long-term analysis on the variance of extra-group paternities in rhesus macaques. *Behavioral Ecology and Sociobiology*, 71:1–11.

Schneider, P. M. (2012). Beyond strs: the role of diallelic markers in forensic genetics. *Transfusion Medicine and Hemotherapy*, 39(3):176–180.

Silk, J., Short, J., Roberts, J., and Kusnitz, J. (1993). Gestation length in rhesus macaques (macaca mulatta). *International Journal of Primatology*, 14:95–104.

Silk, J. B. (2002). Kin selection in primate groups. *International Journal of Primatology*, 23:849–875.

Silk, J. B. (2006). Practicing hamilton's rule: kin selection in primate groups. In *Cooperation in primates and humans: Mechanisms and evolution*, pages 25–46. Springer.

Smith, D. G. (1994). Male dominance and reproductive success in a captive group of rhesus macaques (macaca mulatta). *Behaviour*, 129(3-4):225–242.

Smith, J. M. (1964). Group selection and kin selection. *Nature*, 201(4924):1145–1147.

Tautz, D. (1989). Hypervariability of simple sequences as a general source for polymorphic dna markers. *Nucleic acids research*, 17(16):6463–6471.

Van Schaik, C. P., Pradhan, G. R., and van Noordwijk, M. A. (2004). Mating conflict in primates: infanticide, sexual harassment and female sexuality. *Sexual selection in primates: new and comparative perspectives*, pages 131–150.

Wellens, K. R., Lee, S. M., Winans, J. C., Pusey, A. E., and Murray, C. M. (2022). Female chimpanzee associations with male kin: trade-offs between inbreeding avoidance and infanticide protection. *Animal Behaviour*, 190:115–123.

Widdig, A., Bercovitch, F. B., Jürgen Streich, W., Sauermann, U., Nürnberg, P., and Krawczak, M. (2004). A longitudinal analysis of reproductive skew in male rhesus macaques. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(1541):819–826.

Widdig, A., Kessler, M. J., Bercovitch, F. B., Berard, J. D., Duggleby, C., Nürnberg, P., Rawlins, R. G., Sauermann, U., Wang, Q., Krawczak, M., et al. (2016). Genetic studies on the cayo santiago rhesus macaques: a review of 40 years of research. *American Journal of Primatology*, 78(1):44–62.

Widdig, A., Muniz, L., Minkner, M., Barth, Y., Bley, S., Ruiz-Lambides, A., Junge, O., Mundry, R., and Kulik, L. (2017). Low incidence of inbreeding in a long-lived primate population isolated for 75 years. *Behavioral ecology and sociobiology*, 71:1–15.

Wright, S. (1922). Coefficients of inbreeding and relationship. *The American Naturalist*, 56(645):330–338.

Xue, C., Raveendran, M., Harris, R. A., Fawcett, G. L., Liu, X., White, S., Dahdouli, M., Deiros, D. R., Below, J. E., Salerno, W., et al. (2016). The population genomics of rhesus macaques (macaca mulatta) based on whole-genome sequences. *Genome research*, 26(12):1651–1662.

Zane, L., Bargelloni, L., and Patarnello, T. (2002). Strategies for microsatellite isolation: a review. *Molecular ecology*, 11(1):1–16.

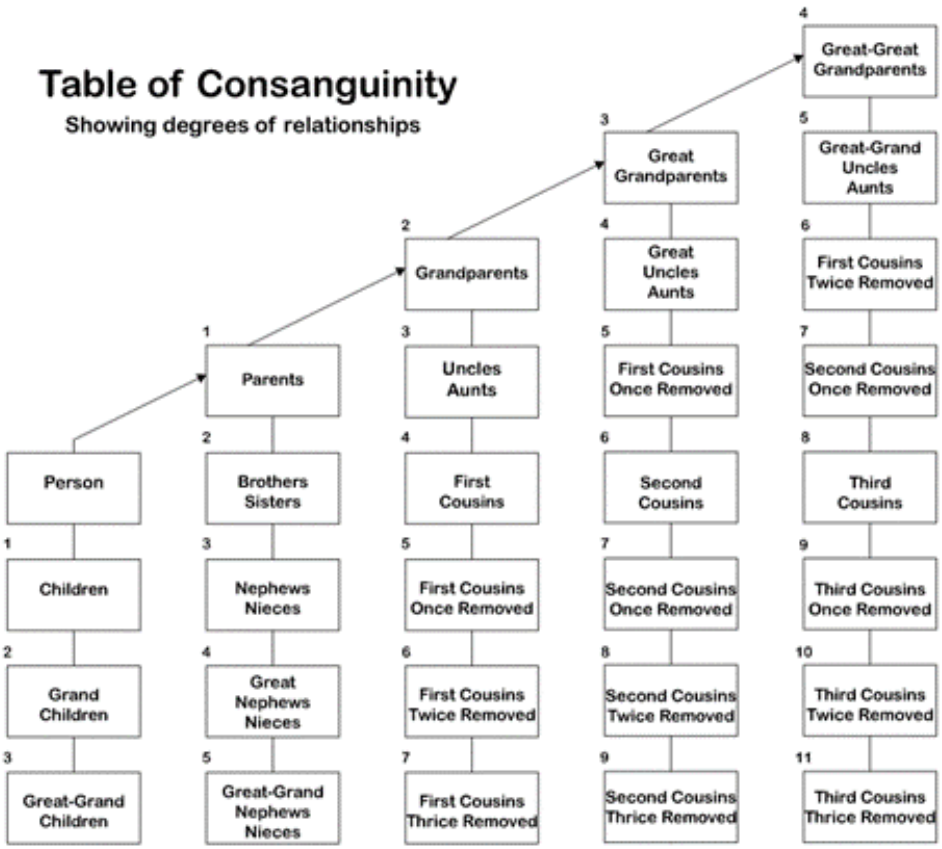



**Figure 4.1: Table of consanguinity**

# List of Figures

# List of Tables

# Acknowledgements

text for acknowledgement...

# Erklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die im Literaturverzeichnis angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder noch nicht veröffentlichten Quellen entnommen sind, sind als solche kenntlich gemacht. Die Zeichnungen oder Abbildungen in dieser Arbeit sind von mir selbst erstellt worden oder mit einem entsprechenden Quellennachweis versehen. Diese Arbeit ist in gleicher oder ähnlicher Form noch bei keiner anderen Prüfungsbehörde eingereicht worden.

Leipzig, den 31.09.2023