

GenoProof®



Theory Manual

Qualitype GmbH

Moritzburger Weg 67 | 01109 Dresden

Phone +49 (0)351 8838 2800 | Fax +49 (0)351 8838 2809

info@qualitype.de

Editorial Work

Dr. Frank Götz, Viktoria Decker, Cordula John, Holger Schönborn

2. Edition, Dresden, 2012

Version 2013-02-19

Copyright © 2006 – 2012 Qualitype GmbH Dresden. All rights reserved.

Production and Sales

Qualitype GmbH

Moritzburger Weg 67

01109 Dresden, Germany

Phone +49 (0)351-8838 2800

Legal Information and Trademarks

This document is property of Qualitype GmbH and protected by copyright.

The content of this document must not be reproduced in whole or in part, translated, transmitted, stored or disclosed to third parties without prior written approval by Qualitype GmbH.

Qualitype GmbH continually develops and enhances its products. Hence the information in this manual may be changed without prior notice due to product improvements, standardizations or technical reasons.

Some products referred to in the documentation are trademarks or registered trademarks of Qualitype GmbH. Other mentioned product names or company names may be trademarks, trade names or registered trademarks of their respective owners.

For further information please read the Terms and Conditions and the Software License Terms of Qualitype GmbH.

Table of Contents

1	Introduction	5
1.1	Area of Application.....	5
1.2	Theory Manual.....	5
2	Basics.....	6
2.1	Terms and Definitions	6
2.2	Theory of Probability	6
2.3	Population Genetics	8
3	Kinship Examinations	9
3.1	Forming of Hypotheses.....	9
3.2	Calculation of Pedigree Likelihoods.....	9
3.3	Paternity and Maternity Tests.....	11
3.3.1	Statistical Parameters	11
3.3.2	Trio- and Duo Cases.....	16
3.3.3	Paternity and Maternity Tests with Grandparents	22
3.3.4	Paternity and Maternity Tests in Cases of Incest	22
3.4	Sibling Analyses.....	23
3.5	Complex Kinship Examinations.....	26
3.5.1	Combinatorial Deficiency Analysis	26
3.5.2	Kinship Algorithm	28
3.6	Further Kinship Calculations	34
3.6.1	Avuncular Index	34
3.6.2	Probability of Monozygosity in Twins.....	36
3.6.3	X- and Y-chromosomal Comparisons	39
3.7	Parameter of calculations	40
3.7.1	Population Data.....	40
3.7.2	Linkage Groups.....	41
3.7.3	Mutation Models	42
3.7.4	Silent Alleles	45
3.7.5	Subpopulations	46
4	Population Studies.....	47

4.1	Allele Frequencies.....	47
4.2	Genotype Frequencies.....	48
4.3	Polymorphic Information Content (PIC)	48
4.4	Homozygoty (h) und Heterozygoty (HET)	48
4.5	Power of Exclusion (PE).....	49
4.6	Paternity Index (PI)	50
4.7	Power of Discrimination (PD)	50
4.8	Mean Exclusion Chance (MEC)	51
4.9	Gene Diversity (GD)	52
4.10	Hardy-Weinberg Equilibrium (HWG).....	53
5	Chimerism Analyses.....	56
6	Forensic Examinations	58
7	Index of Tables and Figures	60
8	References.....	61

1 Introduction

1.1 Area of Application

GenoProof® is an all-in-one software solution for the evaluation of STR analyses and the conduction of biostatistical calculations for kinship investigations, population studies and chimerism analyses.


With GenoProof® you can conduct a complete raw data analysis including peak detection, size calling and allele calling. The program supports various import formats, such as the file format of the ABI sequencers (.fsa). Hence, the raw data analysis can be performed without need for any additional software. All results of the raw data analysis can be viewed and edited easily by means of user-friendly representation tools that have been developed especially for this purpose. There are several quality assurance functionalities included in the program, e.g. artifact detection and automatic contamination checks.

One area of application of GenoProof® is the kinship investigation. GenoProof® determines values for standard trio and duo paternity and maternity cases but also for incest and deficiency cases. In addition, you can perform sibling analyses and calculate avuncular indices and monozygoty probabilities. All calculations can be conducted with all sorts of STR test kits. One can also consider events like mutations, silent alleles and subpopulation membership. Batch analyses allow the conduction of more than 100 tests in one examination.

With GenoProof® you can conduct and evaluate population studies. The determined frequencies can then be used for calculations. Furthermore, you can use GenoProof® to investigate the success of bone marrow transplantations.

All required reference data (markers, test kits, size standards) are already included in the software. The data can be edited and extended by own datasets, when needed. In addition, the program offers a population database with more than 50 populations maintained by hand. Qualitytype offers regular updates for all reference data.

1.2 Theory Manual

The document at hand is a theory manual. It explains the theoretical background of the calculations integrated in GenoProof®. It does not contain detailed instructions. Explanations of all functions of the program as well as the complete content of this document can be found in the **program's internal help**. The help can be opened by pressing the  **Help** button in the toolbar of in the main window. If you use the software for the first time, we also want to recommend you the **short guide**.

2 Basics

Parentage analyses are based on statistic tests and calculations. Thereto, biological and mathematical basics on kinship examinations are introduced in the following sections.

2.1 Terms and Definitions

Alleles: An allele is a form or variation of a gene. The number of alleles at a genetic locus is variable and locus-dependent.

Short tandem repeats (STR's): Short tandem repeats are repetitive sequences on the DNA consisting of 2 to 7 base pairs. The repetitions at this genetic locus are variable and highly diverse. Depending on the number of repetitions different alleles are classified.

Allele frequencies: The relative frequency of an allele is depending on the population. That means that the frequencies of alleles vary between different demographic groups.

DNA marker: DNA markers are genes or DNA sequences at a known location encoding easily identifiable characteristics and are used for instance for identification of characteristics, persons or species. Different forms of a marker are called alleles. The more alleles a marker has the lower are his allelic frequencies and the more certain are the results of a kinship analysis.

2.2 Theory of Probability

The analysis of parentage cases is based on probability calculation.

To begin with, the **probability P(A)** of a random event is the ratio of the number of favorable outcomes for an event and the total number of possible outcomes:

$$P(A) = \frac{\text{number of favorable outcomes}}{\text{total number of possible outcomes}} \quad \text{mit } 0 \leq P(A) \leq 1 \quad (1)$$

Passing alleles from the father on to the child is not depending on passing genes from the mother on to the child or the other way round. In this way the events are stochastically independent. Hence, the probability of the co-occurrence of both events is calculated by multiplication:

Independent events

Two events A and B are stochastically independent if:

- The occurrence of A is not influencing the occurrence of B.
- The occurrence of B is not influencing the occurrence of A.

Multiplication theorem

The probability of two statistically independent events A and B is calculated by multiplication:

$$P(A \cap B) = P(A) \cdot P(B) \quad (2)$$

Example: Let us assume a simple paternity test for a marker with the alleles a, b, c and d. Involved persons are the father (a-c), the mother (b-d) and the child (a-b). With it, the probability that the child inherited the genotype a-b is:

$$P(A \cap B) = P(A) \cdot P(B) = 0,5 \cdot 0,5 = 0,25$$

Since parents can pass only one of their two alleles at a time on to their child, the events are incompatible. Thus, the probability of having one of two different genotypes is the summation of the single genotype probabilities.

Incompatibility of events

Two events A and B are incompatible if they are mutually exclusive.

$$P(A \cap B) = \emptyset \quad (3)$$

Addition theorem

The probability of two incompatible events is calculated by addition:

$$P(A \cup B) = P(A) + P(B) \quad (4)$$

Example: Let us assume a simple paternity test for a marker with the alleles a, b, c and d. Involved persons are the father (a-c), the mother (b-d) and the child (a-b). The probability that the child inherited the genotype a-b or the genotype c-d is:

$$P((A \cap B) \cup (C \cap D)) = P(A) \cdot P(B) + P(C) \cdot P(D) = 0,5 \cdot 0,5 + 0,5 \cdot 0,5 = 0,5$$

Combining the addition and multiplication theorem you can deduce the Hardy-Weinberg law (chapter 2.3):

In case of a homozygous genotype the same allele a is situated on both chromatids. Both are inherited independently. Therefore, you have to square the allele frequency to calculate the genotype probability:

$$p(a - a) = p(a) \cdot p(a) = p(a)^2 \quad (5)$$

On the other hand the different alleles a and b can be located on chromatids 1 and 2 or on chromatids 2 and 1. Both events are incompatible. Hence, it is:

$$p(a - b) = p(a) \cdot p(b) + p(b) \cdot p(a) = 2 \cdot p(a) \cdot p(b) \quad (6)$$

2.3 Population Genetics

To enable kinship calculations it is necessary to suppose an **idealized population**. The assumptions are:

- The population is very big (infinite).
- The number of males and females is equal.
- Individuals are diploid and can mate unrestrictedly.
- Generations do not overlap.
- There is no selection.
- There is no migration.
- There are no mutations.

This model is static so that no evolution is possible. This situation is considered as a current equilibrium and is sufficient for parentage analyses and population studies. Based on these assumptions it is possible to calculate in concordance to the Mendelian laws.

If genotype frequencies and allele frequencies in a sample are in the Hardy-Weinberg equilibrium, the frequency of homozygosity and heterozygosity for this marker will remain constant for generations.

As a consequence, the assumption of an idealized population means that genotype frequencies can be calculated from allele frequencies and that both frequencies do not change (supposing a Mendelian heritage). With it, the probability of a genotype with the alleles a and b is calculated by:

$$p(a - b) = \begin{cases} p(a) \cdot p(a) & , \text{ if } b=a \\ 2 \cdot p(a) \cdot p(b) & , \text{ if } b \neq a \end{cases} \quad (7)$$

Obviously, it is crucial to check the Hardy-Weinberg equilibrium when determining the allele frequencies. If the equilibrium is pertained for one marker, all following calculations of allele frequencies are not reasonable.

3 Kinship Examinations

The purpose of a kinship investigation is to determine or exclude an assumed family relation. The calculations in GenoProof® are based on the genotypes of the examined persons.

3.1 Forming of Hypotheses

Analysis of parentage examinations take place similarly to statistical tests by forming of hypotheses.

Thereby, the null hypothesis $H(X)$ usually describes the assumed kinship constellation whereas the alternative hypothesis assumes the tested person to be unrelated. Null hypothesis and alternative hypothesis are mutually exclusive and involve the same number of persons.

Based on these assumptions the pedigree likelihoods are calculated. To this, probands are tested for at least 12 DNA markers and the genotypes are determined. Finally, the pedigree likelihoods are used for calculation of statistical parameters and the probabilities of hypotheses can be obtained. The more DNA markers you test and the more alleles are covered by a marker, the more certain is the result.

To give an example Figure 1 shows the hypotheses of a simple paternity test.

- **H(X)** The aligned father is the biological father of the child.
- **H(Y)** An unknown person is the biological father of the child.

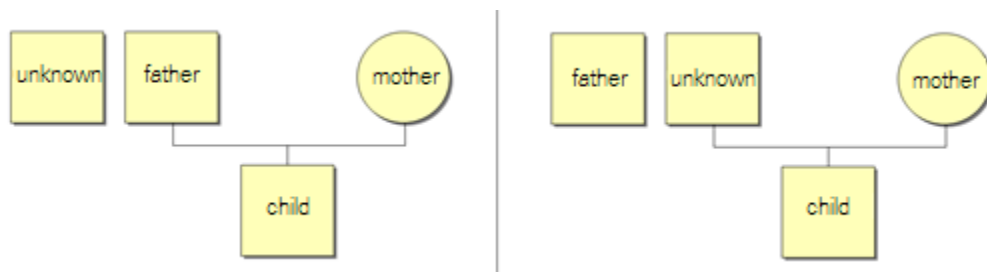


Figure 1: Null hypothesis and alternative hypothesis of a paternity test

3.2 Calculation of Pedigree Likelihoods

The calculation of pedigree likelihoods is based on the genotype frequencies of the involved persons. To this end, the calculation of the independent genotype probability of parents and the calculation of the conditional probability of children has to be distinguished.

Genotype probability of parents

The genotype probability of parents and unattached probands in a pedigree for a **fully typed** person is calculated according to the Hardy-Weinberg equilibrium (formula (7)):

$$p(a-b) = \begin{cases} p(a) \cdot p(a) & , \text{ if } b=a \\ 2 \cdot p(a) \cdot p(b) & , \text{ if } b \neq a \end{cases}$$

where

p(a-b) is the probability for the occurrence of a genotype with alleles a and b,

p(a) is the probability for the occurrence of allele a in the population (allele frequency for allele a) and

p(b) is the probability for the occurrence of allele b in the population (allele frequency for allele b).

Genotype probability of children

Since the genotype of a child is dependent on the genotype of its parents a conditional probability has to be calculated:

$$p(a-b) = p(\text{father is passing allele a}) \cdot p(\text{mother is passing allele b})$$

where

$$p(\text{parent is passing an allele}) = \begin{cases} 1 & , \text{ if the parent is homozygous} \\ 0,5 & , \text{ if the parent is heterozygous} \\ 1 & , \text{ if the parent is partially typed} \end{cases} \quad (8)$$

where

p(a-b) is the probability for the occurrence of a genotype with alleles a and b,

p(a) allele frequency for allele a and

p(b) allele frequency for allele b.

The calculation of pedigree likelihoods is enabled with help of the calculated genotype frequencies of the involved persons.

Genotypical pedigree likelihood

In a genotypical pedigree of a null hypothesis H(X) with n persons with the genotype frequencies p_i the probability X of a genotypical pedigree is calculated according to the multiplication theorem (formula (2)):

$$X = p_1 \cdot p_1 \cdots p_n = \prod_{i=1}^n p_i \quad (9)$$

where

X is the pedigree likelihood of the hypothesis H(X),

n is the number of persons in the pedigree,

p_i is the respective genotype probability of the n involved persons of a pedigree
 i attains values from 1 to n .

Thereby, the probability of the null hypothesis $H(X)$ is X and the probability of the alternative hypothesis $H(Y)$ is Y .

3.3 Paternity and Maternity Tests

Paternity tests aim to include or exclude a man (alleged father - AF) as biological father of a child (C). Analogously, one can include or exclude a woman (alleged mother - AM) from biological maternity by a maternity test. Based on the determined probability $W(X)$ (chapter 3.3.1.2) the program will provide an attribute based on HUMMEL (1997) as in Table 1.

3.3.1 Statistical Parameters

These values are calculated within paternity and maternity tests for each marker:

- Paternity index (PI)
- Essen-Möller value (EM)
- Individual power of exclusion (PE or A)

These values of paternity and maternity tests are determined for all investigated markers:

- Combined paternity index (CPI)
- Combined Essen-Möller value (CEM)
- Combined power of exclusion (CPE)
- Probability of paternity or maternity W after ESSEN-MÖLLER

Values for the single markers cannot be shown, if the constellation of genotypes for the referring marker is an exclusion constellation (mismatch). In this case, one can consider special events such as mutations (chapter 3.7.3) or silent alleles (chapter 3.7.4). Too many incompatible constellations lead to an automatic exclusion of paternity or maternity. The number of incompatible constellations allowed can be set by the user himself. By default, the program allows up to 2 exclusion constellations.

The program is also capable to regard if the examined persons are members of subpopulations (chapter 3.7.5).

3.3.1.1 Paternity Index (PI), Combined Paternity Index (CPI) und Likelihood Ratio (LR)

The **paternity index PI** indicates how much more likely is the null hypothesis than the alternative hypothesis. It is calculated for one examined marker according to BUTLER (2005):

$$PI = \frac{X}{Y} \quad (10)$$

where

X is the probability of the null hypothesis (formula (9)) and

Y is the probability of the alternative hypothesis (formula (9)).

X and Y are mutually exclusive and assume non-relatedness (chapter 3.1). The calculation of the pedigree likelihoods X and Y are detailed described in chapter 3.2.

To strengthen the null hypothesis usually at least 12 different markers are examined. The paternity indexes are multiplied resulting in the **combined paternity index CPI**:

$$CPI = PI_1 \cdot PI_2 \cdot \dots \cdot PI_n \quad (11)$$

where

n is the number of examined markers and

PI_{1...n} is the paternity index (formula (10)) of the markers 1 to n.

The higher the PI or CPI the more likely is the null hypothesis.

As a reciprocal of the paternity index the likelihood ratio LR indicates how much more likely is the alternative hypothesis than the null hypothesis. The likelihood ratio for one examined marker is calculated according to BUTLER (2005):

$$LR = \frac{Y}{X} \quad (12)$$

where

X is the probability of the null hypothesis (formula (9)),

Y is the probability of the alternative hypothesis (formula (9)),

n is the number of examined markers and

LR_{1...n} is the likelihood ratio of the markers 1 to n.

The **combined likelihood ratio CLR** is calculated as the product of the LRs of the single markers:

$$CLR = LR_1 \cdot LR_2 \cdot \dots \cdot LR_n \quad (13)$$

where

n is the number of examined markers and

LR_{1...n} is the likelihood ratio (formula (12)) of the markers 1 to n.

3.3.1.2 Values by Essen-Möller (W, EM und CEM)

The predicates for the assessment of the probability of the kinship relation investigated (Table 1) are assigned on basis of the probability **W(X)** by ESSEN-MÖLLER. The probability for the assumption of the hypothesis H(X) is:

$$W(X) = \frac{CPI \cdot P_R}{(CPI \cdot P_R + (1 - P_R))} \quad (14)$$

where

CPI is the combined paternity index (formula (11)) and

PR is the a priori probability.

The **a priori probability** represents the evidence beyond the genetic examination. By default, one assumes an a priori probability of 0.5 meaning the circumstance known so far (without genetic data) would favor both hypotheses equally. Using an a priori probability of 0.5 allows cancelling the W(X) value formula down to:

$$W(X) = \frac{CPI}{CPI + 1} = \frac{1}{1 + CLR} \quad (15)$$

where

CPI is the combined paternity index (formula (11)) and

CLR is the combined likelihood ratio (formula (13)).

Table 1: Predicates of paternity testing according to Hummel (1997)

Probability of Paternity*	Predicate
> 99,73 %	Paternity* practically proven
> 99 – 99,73 %	Paternity* highly likely

> 95 – 99 %	Paternity* very likely
> 90 – 95 %	Paternity* likely
> 80 – 90 %	certain indication of biological paternity*
> 70 – 80 %	formal indication of biological paternity*
> 30- 70 %	(without predicate)
> 20 – 30 %	formal indication of potential non-paternity*
> 10 – 20 %	certain indication of non-paternity*
> 5 – 10 %	Paternity* unlikely
> 1 – 5 %	Paternity* very unlikely
0,27 – 1 %	Paternity* highly unlikely
< 0,27 %	Paternity* practically excluded

* The predicates for maternity are composed analogously.

Besides the W value there is the so-called Essen-Möller value EM. It is determined for each single marker:

$$EM = \log \frac{Y}{X} + 10 \quad (16)$$

where

X is the probability of the null hypothesis X (formula (9)) and

Y is the probability of the alternative hypothesis Y (formula (9)).

The exact formulas for X and Y can be found in Buckleton (2005).

Based on the Essen-Möller value, there is the **combined Essen-Möller value CEM** as the sum of the EM of all examined markers minus a correction:

$$CEM = \sum_{i=1}^n EM_i - (n-1) \cdot 10 \quad (17)$$

where

n is the number of all examined markers and

EM_i is the Essen-Möller value (formula (16)) of the *i*th i-ten Markers and

i can attain all values from 1 to n.

The lower CEM the more likely is the null hypothesis (the alleged parent is the true biological parent of the child).

3.3.1.3 Individual Power of Exclusion (PE)

Not every non-paternity or non-maternity is recognized correctly as such. The **individual chance of exclusion from paternity PE** (power of exclusion) indicates the probability of a non-father or a non-mother to be excluded correctly (true negatives). At this, one can exclude all men who do not carry the paternal allele of the child (if it is possible to determine the paternal allele) or who do not carry any of two child alleles (if it is not possible to excluded the paternal allele). The power of exclusion for women works analogously.

Note: The individual chance of exclusion from paternity or maternity is independent from the genotype of the alleged parent.

If it is possible to determine which of the child's alleles is the paternal one, the PE is calculated as follows (BAUR, 1993):

$$PE=(1-p(a))^2 \quad (18)$$

where

p(a) is the frequency of the allele originating from the alleged parent in the population of the alleged parent.

If both of the child's alleles might originate from the alleged parent, the PE is calculated as following:

$$PE=(1-p(a)-p(b))^2 \quad (19)$$

where

p(a) is the frequency of the first child allele in the population of the alleged parent and

p(b) is the frequency of the second child allele in the population of the alleged parent.

Besides the PE, there is the **combined power of exclusion CPE**, which summaries the PE values of all examined markers.

$$CPE=(1-(1-PE_1) \cdot (1-PE_2) \cdot \dots \cdot (1-PE_n)) \quad (20)$$

where

n is the number of all examined markers and

PE_{1...n} is the power of exclusion (formula (19)) of the markers 1 to n.

3.3.2 Trio- and Duo Cases

Testing paternity or maternity, one compares different genetic markers (loci) of the child with the markers of the examined parents. In an ideal situation, one can conclude which of child's allele must originate from mother and which comes from the father. For instance, a child carrying the alleles 12 and 13 with its mother carrying the alleles 13 and 14 must have a biological father carrying allele 12 at least once.

In order to make a statement about a putative paternity and maternity, one needs to compare the alleles of several genetic markers. The significance of the statement increases with the number of markers examined. At the present day, one usually compares at least 12 markers. These must be unlinked, that is they must be inherited independently. Furthermore, they have to be located on 10 different chromosomes at least (HOPPE, 2002).

There are two possible scenarios for paternity and maternity tests:

1. **Trio case:**

Genetic information is available about all three persons (child (C), mother (M) and alleged father (AF), respectively, child (C), father (F) and alleged mother (M)).

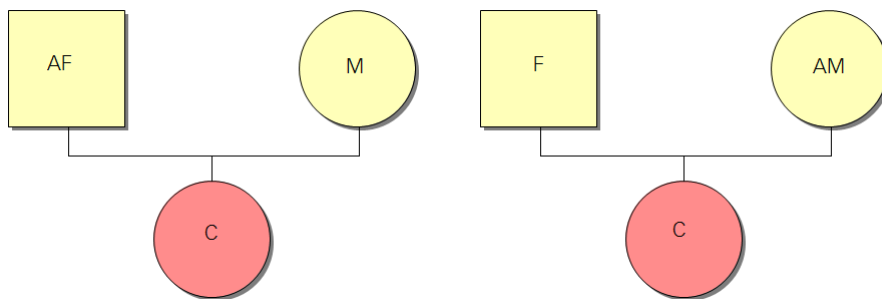


Figure 2: Trio cases for paternity tests (on the left) and maternity tests (on the right)

2. Duo case:

Genetic information is only available for the child (C) and the alleged father (AF) or alleged mother (AM). The information about the second parent (mother (M) or father (F)) is missing.

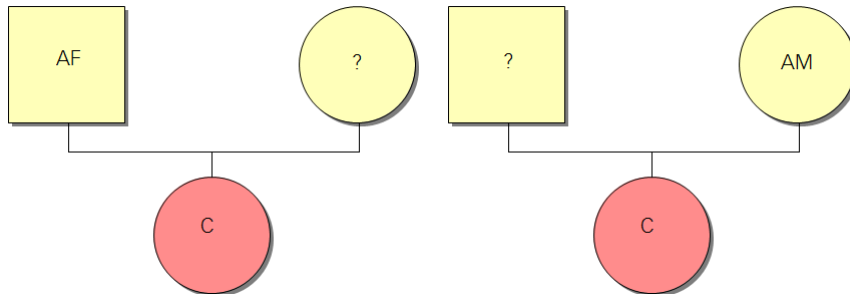


Figure 3: Duo cases for paternity tests (on the left) and maternity tests (on the right)

The investigation of trio cases presumes that: 1.) the mother is not related to the alleged father or other possible fathers, and that 2.) different alleged fathers are not related to each other (maternity tests analogously). In addition, you need to know the ethnical background of the examined persons (Hummel, 1997).

Note: Trio maternity tests presume that the tested father is the biological father of the child. Exclusion of maternity can then be caused by two reasons: 1.) the examined woman is not the mother of the child or 2.) the father set is not the biological father of the child. Exclusion of maternity due to a wrong father can be avoided by performing a duo test (with child and alleged mother only). However, duo tests are, in case of inclusion, less significant than trio tests.

3.3.2.1 Calculation of a Trio Case

As an example the calculation of the pedigree likelihood of a classical trio case are described following.

The hypothesis $H(X)$ involves the following constellation and genotypes of persons (Figure 4): alleged father (a-b), mother (c-d), child (b-c) and a unknown person (untyped).

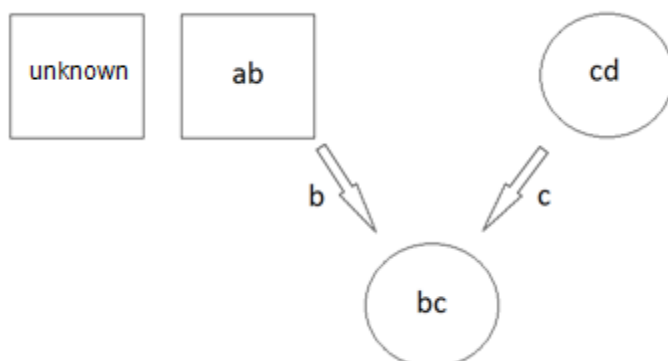


Figure 4: Null hypothesis H(X) in a paternity test

According to the Hardy-Weinberg equilibrium the genotype frequencies of the parents can be calculated via formula (7) since they can be assumed as two random individuals of a population:

$$p(a-b) = 2 \cdot p(a) \cdot p(b)$$

$$p(c-d) = 2 \cdot p(c) \cdot p(d)$$

The unattached untyped unknown could have any genotype in the null hypothesis. Therefore, a probability of 1 is assumed.

$$p(x-x) = 1$$

Since the child's genotype is a result of inheriting one allele from the father and one from the mother its condition probability is calculated according to formula (8).

$$p(b-c) = 0,5 \cdot 0,5 = 0,25$$

For the calculation of the pedigree likelihood X of the null hypothesis H(X) the single genotype probabilities have to be multiplied according to formula (9):

$$X = 2 \cdot p(a) \cdot p(b) \cdot 2 \cdot p(c) \cdot p(d) \cdot 0,25 \cdot 1$$

The calculation of the probability of the **alternative hypothesis** is performed analogously. The hypothesis H(Y) involves the following constellation and genotypes of persons (Figure 5): alleged father (a-b), mother (c-d), child (b-c) and a unknown person (untyped).

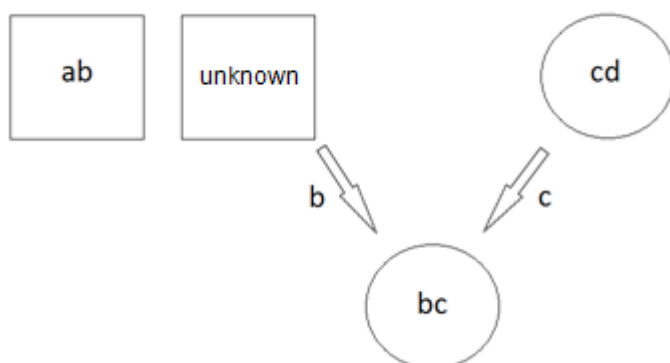


Figure 5: Alternative hypothesis H(Y) in a paternity test

According to the Hardy-Weinberg equilibrium the genotype frequencies of the mother and the unrelated alleged father can be calculated via formula (7) since they can be assumed as two random individuals of a population:

$$p(a-b)=2 \cdot p(a) \cdot p(b)$$

$$p(c-d)=2 \cdot p(c) \cdot p(d)$$

The untyped unknown is the father in the alternative hypothesis. In this case, he has to have the allele he passed to the child. Therefore, a probability of the allele frequency b is assumed.

$$p(b-x)=p(b)$$

Since the child's genotype is a result of inheriting one allele from the unknown and one from the mother its condition probability is calculated according to formula (8).

$$p(b - c)=1 \cdot 0,5=0,5$$

For the calculation of the pedigree likelihood Y of the alternative hypothesis $H(Y)$ the single genotype probabilities have to be multiplied according to formula (9):

$$Y=2 \cdot p(a) \cdot p(b) \cdot 2 \cdot p(c) \cdot p(d) \cdot 0,5 \cdot p(b)$$

Finally, you can calculate the paternity index as a ratio of both probabilities of the hypotheses according to formula (10):

$$PI = \frac{X}{Y} = \frac{2 \cdot p(a) \cdot p(b) \cdot 2 \cdot p(c) \cdot p(d) \cdot 0,25 \cdot 1}{2 \cdot p(a) \cdot p(b) \cdot 2 \cdot p(c) \cdot p(d) \cdot 0,5 \cdot p(b)} = \frac{0.5}{p(b)}$$

It becomes clear that the calculation of the PI is only depending on the inherited allele of the father. If you go through all possible genotype constellations you will get the following table showing the calculation of paternity indexes for the particular constellation.

Table 2: Formula for trio cases

Mother	Child	Father	Paternity Index
aa	aa	ab	$\frac{0,5}{p(b)}$
aa	ab	ab	$\frac{0,5}{p(b)}$
aa	ab	bc	$\frac{0,5}{p(a)}$
ab	aa	ab	$\frac{0,5}{p(a)}$
ab	aa	ac	$\frac{0,5}{p(a)}$
bc	ab	ab	$\frac{0,5}{p(a)}$
bc	ab	ac	$\frac{0,5}{p(a)}$
bd	ab	ac	$\frac{0,5}{p(a)}$
aa	aa	aa	$\frac{1}{p(a)}$
ab	aa	aa	$\frac{1}{p(a)}$

bb	ab	aa	$\frac{1}{p(a)}$
bc	ab	aa	$\frac{1}{p(a)}$
ab	ab	ac	$\frac{0,5}{p(a)+p(b)}$
ab	ab	aa	$\frac{1}{p(a)+p(b)}$
ab	ab	ab	$\frac{1}{p(a)+p(b)}$

3.3.2.2 Calculation of a Duo Case

The calculation of duo cases is analogous to trio cases. Due to the missing parent more possible combinations of genotypes are feasible. The table for calculation of paternity indexes for all possible duo constellations is the following:

Table 3: Formula for duo cases

Alleged parent	Child	Paternity Index
ac	ab	$\frac{0,25}{p(a)}$
ab	ab	$\frac{p(a)+p(b)}{4 \cdot p(a) \cdot p(b)}$
aa	ab	$\frac{0,5}{p(a)}$
ac	aa	$\frac{0,5}{p(a)}$
aa	aa	$\frac{1}{p(a)}$

3.3.3 Paternity and Maternity Tests with Grandparents

If there is no genetic information available for a parent or alleged parent because the person is deceased or cannot be found, one can conduct a test using the information of the corresponding grandparents instead.

The program supports all possible scenarios: Mother, father or both can be replaced by the referring grandparents. The test can be conducted one-sided or both-sided, that means considering only the paternal or maternal side or considering both sides.

Please regard that a test like this is evaluating the grandparent-child relation. It is not suitable to safely exclude paternity or maternity. For example: An alleged father is replaced by the paternal grandfather and the result of the tests excludes grandfather hood. It is not possible to determine the true cause for this result. It might be that 1.) the missing alleged father is not the father of the child or that 2.) the examined grandfather is not the biological father or the alleged father (HUMMEL, 1997). We recommend performing further tests in this case, e.g. sibling analyses (chapter 3.4) with single grandparents or X- or Y-chromosomal comparisons (chapter 3.6.3).

In a principle, one tries to combine all possible genotypes the missing parents might have (assuming the known genotypes). With, for instance, a paternal grandfather with the alleles A and B, and a paternal grandmother with the alleles A and C, the missing father must have either the genotype AA or AC or BC. If one then has the child's mother with a genotype DD and the child with AD, there are only two compatible genotypes left for the missing father: AA and AC. In this case, one would determine PI for both case and then average them (FUNG, 2003).

If there is no compatible genotype combination for a locus one has to assume a mismatch or mutation. Both cases are treated like in normal paternity or maternity cases.

For paternity and maternity tests with grandparents, the program determines all parameters that are investigated in normal trio and duo cases, too.

3.3.4 Paternity and Maternity Tests in Cases of Incest

If the (alleged) parents of a child are related to each other, the probability of paternity might be affected by this.

In the case of mother and alleged father (or father and alleged mother) being cousins, uncle and niece or aunt and nephew, there is only less influence. The probabilities of paternity or maternity determined when regarding the relationship hardly deviate from the results obtained without consideration of such. Thus, trio cases of this type are evaluated like normal trio cases (HUMMEL, 1997).

However, if mother and alleged father (or father and alleged mother) are parent and child or siblings, one needs to use other formulas for the calculation of the PI (chapter 3.3.1.1) because the genotypes of

the parents are not independent from each other. The formulas used in GenoProof® are based on the three-allele system formulas by MINATAKA (1996). The formulas have been adapted for multi-allele systems. They can be found in Table 4.

Incest cases can not be considered in duo cases.

Table 4: Formula for calculation of the paternity index for paternity test in case of incest

Child	Father	Mother	formula for	
			Father-daughter incest /Mother-son incest	Brother-sister incest
AA	AA	AA	$1 / (0.5 a + 0.5)$	$1 / (0.5 a + 0.5)$
AA	AA	AB	$1 / (a + 0.5)$	$1 / (a + 0.5)$
AA	AB	AA	$1 / (a + 1)$	$1 / (a + 1)$
AA	AB	AB	$1 / (a + 0.5)$	$1 / (a + 0.5)$
AA	AC	AB	$0.5 / a$	$0.5 / a$
AB	AA	AB	1	1
AB	AA	BB	exclusion	$1 / (0.5 a)$
AB	AA	BC	exclusion	$1 / a$
AB	AB	AA	$1 / b$	$1 / b$
AB	AB	AB	1	1
AB	AB	BC	$0.5 / a$	$0.5 / a$
AB	AC	AB	$0.5 / (a + b)$	$0.5 / (a + b)$
AB	AC	BC	$0.5 / a$	$0.5 / a$
AB	AC	BD	exclusion	$0.5 / a$
AB	BC	AA	exclusion	$0.5 / b$

A, B, C and D are different alleles of a multi-locus system. Their allele frequencies are a, b, c and d.

The probability of paternity W (formula (14)) determined considering the incest is often lower than the one determined without consideration (HUMMEL, 1997). In addition, there is a higher number of incompatible constellations (at least for parent-child incest). For instance, if there is a mother with the genotype aa, it is not possible that the alleged has genotype bb, at least not if he was the true biological father of the mother.

3.4 Sibling Analyses

The sibling analysis is used to calculate the probability of two persons having a certain (or no) degree of relationship.

For this example, we assume to persons P_K and P_L with the genotypes K and L. The algorithm first determines the probability of the genotype combination K-L under the condition of a certain degree of kinship for P_K and P_L . WEIR (1996) distinguishes 5 different degrees of relationship:

- 1) Full siblings
- 2) Parent and child
- 3) Half siblings, grandparents and child or aunt/uncle and nephew/niece
- 4) First cousins
- 5) Unrelated persons.

The degree of relationship of two persons influences the probability of them exhibiting the same genotype. For this reason, one introduces so-called IBD coefficients (Φ_2 , Φ_1 and Φ_0) into the formula. Φ_2 represents the probability that two persons of a certain degree of relationship have 2 alleles of a marker in common and that these alleles originate from the same ancestor. These are so-called IBD alleles (IBD = Identical by Descent). Φ_1 represents the probability that two persons of a certain degree of relationship share exactly 1 IBD allele and Φ_0 is the probability that they do not share any IBD allele. All IBD coefficients are values between 0 and 1. Their total sum is 1.

The values of the IBD coefficients depend on the marker type (autosomal, X-gonosomal and Y-gonosomal), too. For gonosomal markers they also depend on the sex. The derivation of IBD coefficient values is explained in chapter 3.6.2 using the example of autosomal markers in dizygote twins (full siblings).

Table 5: IBD coefficients for autosomal marker

Degree of Relationship	Φ_2	Φ_1	Φ_0
Full siblings	1/4	1/2	1/4
Parent-child	0	1	0
Half siblings or grandparents-grandchild or Aunt/uncle-niece/nephew	0	1/2	1/2
First cousins (all sex combinations)	0	1/4	3/4
Unrelated persons	0	0	1

Now, one calculates the probability **P** of the genotype combination K-L under the condition of a given degree of relationship. The probability P is calculated for each marker:

$$P = (p_2 \cdot \Phi_2) + (p_1 \cdot \Phi_1) + (p_0 \cdot \Phi_0) \quad (21)$$

where

p₂ is the probability of two random persons of the population having two alleles in common (formula (37)),

p₁ is the probability of two random persons of the population having one allele in common,

- p_0 is the probability of two random persons of the population not having any allele in common,
- Φ_2 is the probability of two persons sharing 2 IBD alleles (under the condition of the examined degree of relationship),
- Φ_1 is the probability of two persons sharing 1 IBD allele (under the condition of the examined degree of relationship) and
- Φ_2 is the probability of two persons sharing no IBD allele (under the condition of the examined degree of relationship).

The P value of each hypothesis is now divided by the P value of the fifth hypothesis (unrelated) in order to proportion the results. The value derived this way is the **joint probability** w:

$$w_i = P_i / P_5 \quad (22)$$

where

- i is the hypothesis examined and
- P_i is the p value (formula (21)) of the hypothesis i and
- P_5 is the p value (formula (21)) of the hypothesis 5 (unrelated).

In case of autosomal and female X-chromosomal markers one, now multiplies the values of all markers:

$$CW = w_1 \cdot w_1 \cdot \dots \cdot w_n \quad (23)$$

where

- n is the number of all examined markers and
- $w_{1...n}$ is the joint probability w (formula (22)) of the markers 1 to n. If w is 0, the 0 is replaced by a 1.

Linked markers (e.g. Y-chromosomal markers and X-chromosomal markers of the same linkage group) can not be combined this way since they are not inherited independently from each other. In this case, one may only use one marker per linkage group.

Finally, the probability of every hypothesis is calculated as percentage:

$$W_i = \frac{CW_i}{CW_1 + CW_2 + CW_3 + CW_4 + CW_5} \quad (24)$$

where

- i is the hypothesis examined and
- CW_i is the CW value (Formel (23)) of hypothesis i and

CW_{1...5} is the CW value (Formel (23)) of the hypotheses 1 to 5.

It might be that the results of different hypotheses are quite similar to each other (e.g. hypothesis parent-child and hypothesis full siblings). In such cases one can only conclude a close blood relationship. For a safe statement one requires a large CW value and a great distance to the CW values of the other hypotheses.

3.5 Complex Kinship Examinations

In practice a large part of parentage analyses are paternity and maternity tests that can be calculated by the given formula (chapter 3.3). Even so, there are more complex questions such as deficiency cases in which probands are untyped.

Figure 6 shows an example for a complex kinship case. Crossed out probands are untyped. Based on their typed relatives it shall be investigated if the two children in the last generation are really brother and sister or if the boy is not related with the female child at all.

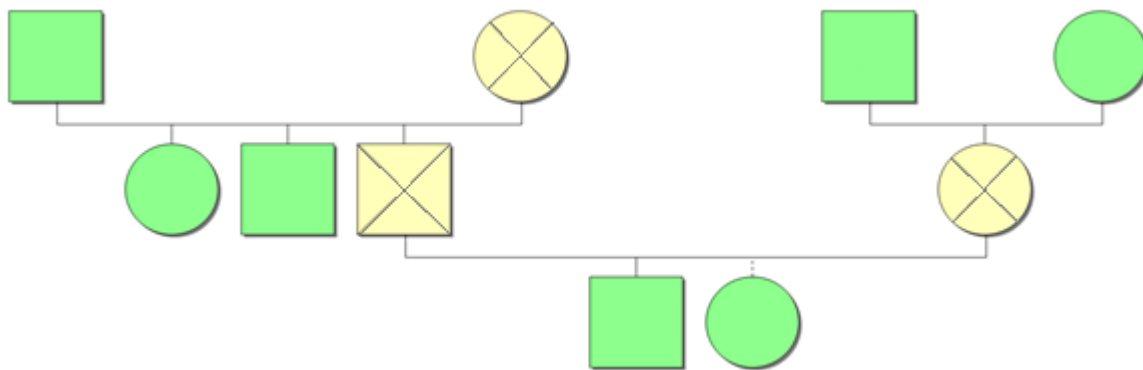


Figure 6: Complex parentage case

To solve such complex cases special solutions have been created to enable the analysis of any kinship constellation. For it, GenoProof® is operating using the combinatorial deficiency analysis for the determination of all possible pedigree constellations and the kinship algorithm for calculation of statistical parameters. Both methods are introduced in the following sections.

3.5.1 Combinatorial Deficiency Analysis

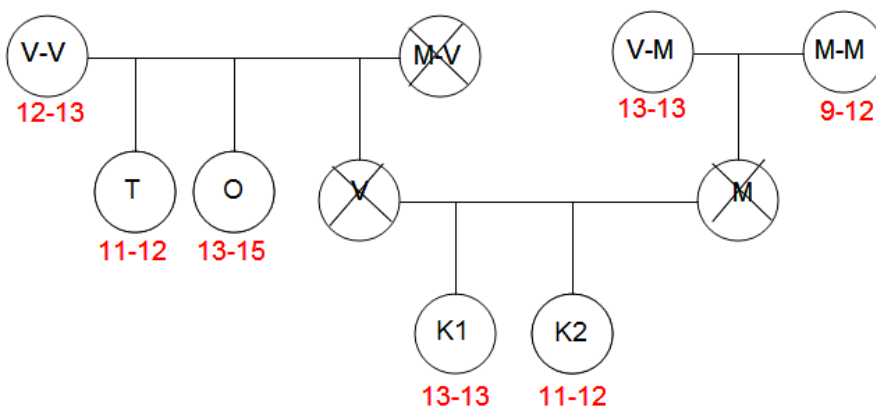
The combinatorial deficiency analysis (CONRADT, 1983) is solving the deficiency problem and examines possible solutions for a genotypical pedigree. With it, we can assume the alternative hypothesis to be true, if the genotypes in the null hypothesis are incompatible. In case that also the genotypes in the alternative hypotheses are incompatible, no proposition about the kinship relations can be made for this marker. The latter can take place if at least one other person's genotype does not match or other effects that can not be considered (such as mutations) are involved.

The method is operating by finding solutions for the combinatorial problem step by step. Doing so, genotype constellations that are compatible with the given pedigrees are found. For that to happen, a genotypical pedigree is examined from bottom to top going through every possible genotype combination of the probands and taking into account the compatibility of all persons involved. Thereby, compatibility tests are done to check the compatibility of a proband's genotype with those of his children. If all compatibility tests are positive a possible pedigree has been found.

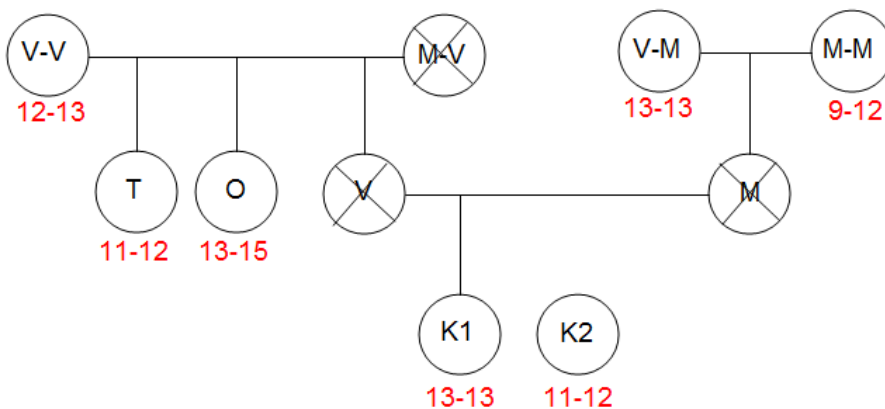
Example

As an example we assume the following null hypothesis and alternative hypothesis of a complex sister-sister-case for the marker D4S2366 with the alleles 11,12,13,14 and 15.

- **H(X)** K2 is the sister of K1.

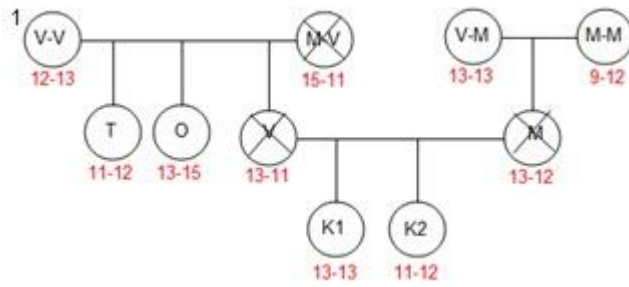


- **H(Y)** K2 is not related to K1.

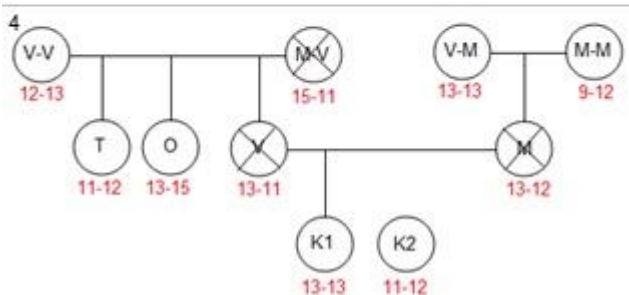
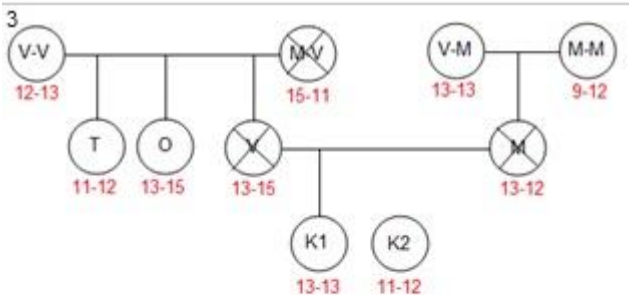
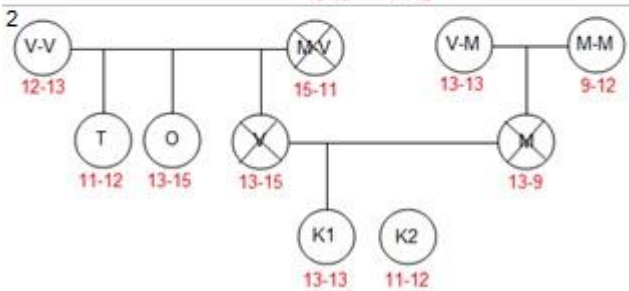
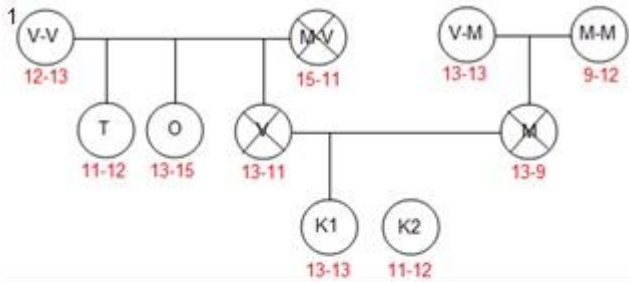


With help of the combinatorial deficiency analysis we get one possible pedigree for the null hypothesis H(X) and 4 possible pedigrees for the alternative hypothesis H(Y):

- **H(X)** K2 is the sister of K1.



- **H(Y)** K2 is not related to K1.



3.5.2 Kinship Algorithm

The kinship algorithm determines the genotypical pedigree likelihoods according to HUMMEL (1997). Thereby, the calculation of pedigree likelihoods is based on the genotype frequencies of the involved persons.

Genotype probability of terminal parents

The genotype probability of terminal parents (persons without parents in the pedigree) and unattached probands in a pedigree for a **fully typed** person is calculated according to the Hardy-Weinberg equilibrium (formula (7)):

$$p(a-b) = \begin{cases} p(a) \cdot p(a) & , \text{ if } b=a \\ 2 \cdot p(a) \cdot p(b) & , \text{ if } b \neq a \end{cases}$$

where

p(a-b) is the probability for the occurrence of a genotype with alleles a and b,

p(a) is the probability for the occurrence of allele a in the population (allele frequency for allele a) and

p(b) is the probability for the occurrence of allele b in the population (allele frequency for allele b).

If an untyped person attains an allele in the course of the combinatorial deficiency procedure the parent becomes **partially typed**. This means, that only one allele is known and the second allele could be every other allele. The genotype probability is depending on the number of his children.

$$p(a-x) = p(a) \cdot (p(a) + 2^{1-k} \cdot (1-p(a))) \quad (25)$$

where

p(a-x) is the probability for the occurrence of a genotype with the allele a and another unknown allele x,

p(a) is the allele frequency for allele a and

k is the number of the person's children.

Unattached **untyped** persons are often part of a pedigree. Since the allele frequencies are unknown, untyped persons could have any genotype in the hypothesis. Therefore, a probability of 1 is assumed.

$$p(x-x) = 1 \quad (26)$$

where

p(x-x) is the probability of the occurrence of a genotype of a untyped person that is 1.

Genotype probability of children

Since the genotype of a child is dependent on the genotype of its parents a conditional probability has to be calculated (formula (8)):

$$p(a-b)=p(\text{father is passing allele } a) \cdot p(\text{mother is passing allele } b)$$

where

$$p(\text{parent is passing an allele}) = \begin{cases} 1 & , \text{ if the parent is homozygous} \\ 0,5 & , \text{ if the parent is heterozygous} \\ 1 & , \text{ if the parent is partially typed} \end{cases}$$

where

p(a-b) is the probability for the occurrence of a genotype with alleles a and b,

p(a) is the allele frequency for allele a and

p(b) is the allele frequency for allele b.

Following, the calculation of pedigree likelihoods is enabled with help of the calculated genotype frequencies of the involved persons.

Genotypical pedigree likelihood

In a genotypical pedigree of a null hypothesis H(X) with n persons with the genotype frequencies p_i the probability X of a genotypical pedigree (formula (9)) is calculated according to the multiplication theorem (formula (2)):

$$X = p_1 \cdot p_1 \cdots p_n = \prod_{i=1}^n p_i$$

where

X is the pedigree likelihood of the hypothesis H(X),

n is the number of persons in the pedigree,

p_i is the respective genotype probability of the n involved persons of a pedigree and

i attains values from 1 to n.

Thereby, the probability of the null hypothesis H(X) is X and the probability of the alternative hypothesis H(Y) is Y.

With the given formula all possible genotype cases in a genotypical pedigree can be determined. What is still remaining is the way of calculation in cases of one pedigree is consisting of several partial pedigrees and in cases of more than one pedigree as a solution of a combinatorial deficiency analyses.

Probability of several genotypical partial pedigrees

A pedigree as a solution of the combinatorial deficiency analyses can consist of several partial pedigrees that are stochastically independent from each other. Such a partial pedigree can even consist of a single person only. Due to the independence of the pedigrees the multiplication theorem (formula (2)) is used for calculation of the probability of the partial pedigree:

$$X_{\text{total}} = X_{i1} \cdot X_{i2} \cdots X_{im} = \prod_{j=1}^m X_{ij} \quad (27)$$

where

X_{total} is the probability of a genotypical pedigree of a hypothesis $H(X)$,

X_{ij} is the particular probability of the m^{th} independent genotypical partial pedigree of a hypothesis $H(X)$,

m is the number of independent genotypical partial pedigrees of a hypothesis $H(X)$,

i attains values from 1 to n and

j attains values from 1 to m .

Undefined genotypical pedigree likelihood

If more than one pedigree is possible as a solution of a combinatorial deficiency analysis, these pedigrees are incompatible. The occurrence of at least one of these pedigrees is calculated according to the addition theorem (formula (4)):

$$X_{\text{total}} = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^n X_i \quad (28)$$

where

X_{total} is the undefined genotypical pedigree likelihood for the occurrence of at least one of the possible pedigrees,

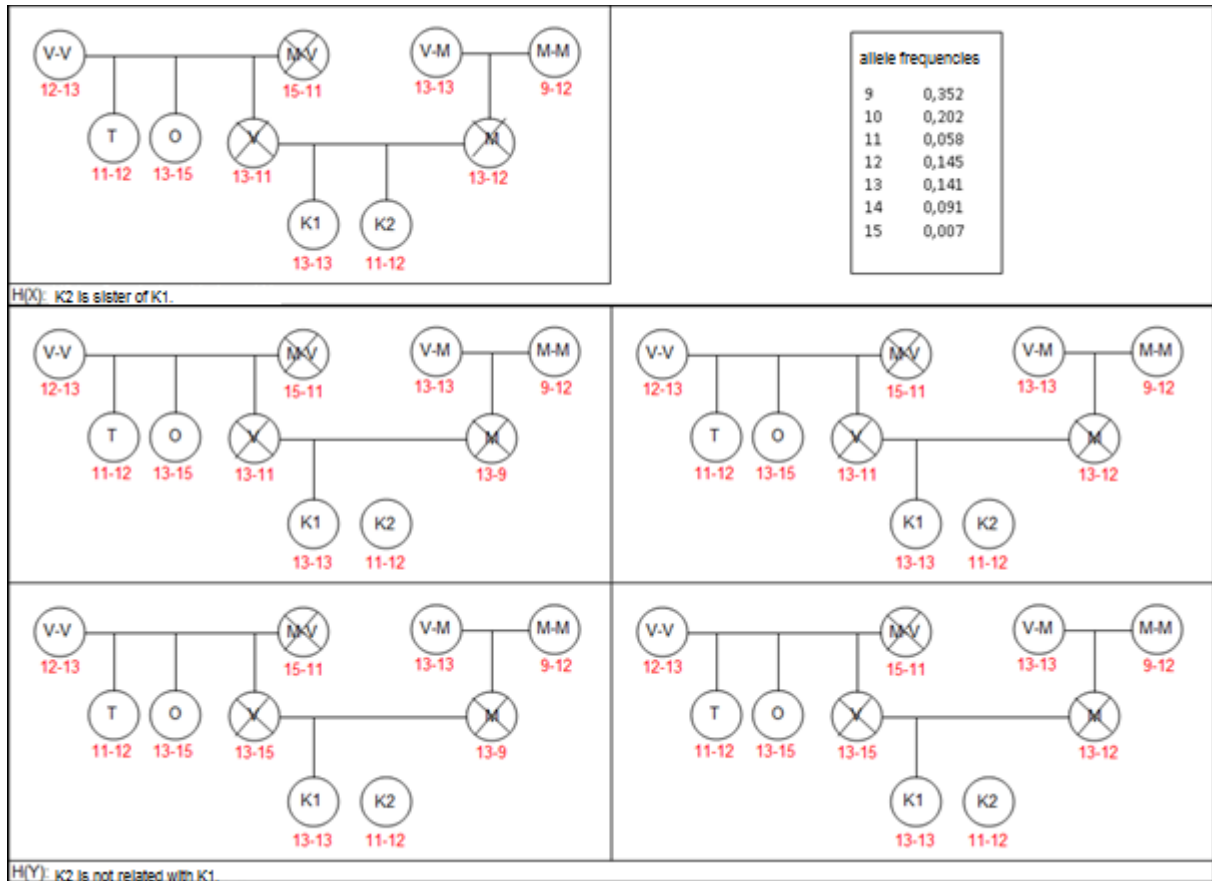
n is the number of the incompatible genotypical pedigrees of a hypothesis $H(X)$,

X_i is the particular pedigree likelihood of the n^{th} incompatible genotypical pedigree of a hypothesis $H(X)$ and

i attains all values from 1 to n .

Example

To give an example we use the solutions of the of the combinatorial deficiency analysis we have got in chapter 3.5.1 displayed in the following figure.



In order to calculate the probability of the null hypothesis H(X) in relation to the alternative hypothesis H(Y) the probabilities are calculated with help of the kinship algorithm:

Null hypothesis H(X):

- Genotype probability of all (terminal) parents:

$$p(V-V)=2 \cdot 0.145 \cdot 0.141=0.04089$$

$$p(M-V)=2 \cdot 0.007 \cdot 0.058=8.12 \cdot 10^{-4}$$

$$p(V-M)=0.141 \cdot 0.141=0.019881$$

$$p(M-M)=2 \cdot 0.352 \cdot 0.145=0.10208$$

- Genotype probability of all children:

$$p(K1|V,M)=0.5 \cdot 0.5=0.25$$

$$p(K2|V,M)=0.5 \cdot 0.5=0.25$$

$$p(T|V-V,M-V)=0.5 \cdot 0.5=0.25$$

$$p(O|V-V,M-V)=0.5 \cdot 0.5=0.25$$

$$p(V|V-V,M-V)=0.5 \cdot 0.5=0.25$$

$$p(M|V-M,M-M)=1 \cdot 0.5=0.5$$

- Genotypical pedigree likelihood:

$$\begin{aligned} X &= 0.04089 \cdot 8.12 \cdot 10^{-4} \cdot 0.019881 \cdot 0.10208 \cdot 0.25 \cdot 0.25 \cdot 0.25 \cdot 0.25 \cdot 0.25 \cdot 0.5 \\ &= 3.2901983 \cdot 10^{-11} \end{aligned}$$

Alternative hypothesis H(Y), pedigree 1:

- Partial pedigree 1: genotype probabilities of all terminal parents:

$$p(V-V)=2 \cdot 0.145 \cdot 0.141=0.04089$$

$$p(M-V)=2 \cdot 0.007 \cdot 0.058=8.12 \cdot 10^{-4}$$

$$p(V-M)=0.141 \cdot 0.141=0.019881$$

$$p(M-M)=2 \cdot 0.352 \cdot 0.145=0.10208$$

- Partial pedigree 1: genotype probabilities of all children:

$$p(K1|V,M)=0.5 \cdot 0.5=0.25$$

$$p(T|V-V,M-V)=0.5 \cdot 0.5=0.25$$

$$p(O|V-V,M-V)=0.5 \cdot 0.5=0.25$$

$$p(V|V-V,M-V)=0.5 \cdot 0.5=0.25$$

$$p(M|V-M,M-M)=1 \cdot 0.5=0.5$$

- Partial pedigree 2: genotype probabilities of the unattached proband:

$$p(K2)=2 \cdot 0.058 \cdot 0.145=0.01682$$

- Genotypical pedigree likelihood:

$$Y_1 = (0.04089 \cdot 8.12 \cdot 10^{-4} \cdot 0.019881 \cdot 0.10208 \cdot 0.25 \cdot 0.25 \cdot 0.25 \cdot 0.25 \cdot 0.5) \cdot 0.01682$$

$$= 2.21364542 \cdot 10^{-12}$$

The other 3 possible pedigrees have to be calculated analogously (the calculation is not presented here). Due to the particular genotype constellation the genotypical pedigree likelihoods are the same as for the first possible pedigree. Therewith, the probability of the alternative hypothesis is the summation of the pedigree likelihoods of these 4 possible pedigrees:

$$Y = 4 \cdot 2.21364542 \cdot 10^{-12} = 8.85458168 \cdot 10^{-12}$$

Knowing the pedigree likelihoods of the null hypothesis and alternative hypothesis you can calculate the relevant statistical parameter as described in chapter 3.3.1:

$$PI = \frac{3.2901983 \cdot 10^{-11}}{8.85458168 \cdot 10^{-12}} \approx 3.7158$$

$$W(X) = \frac{3.7158}{3.7158 + 1} \approx 0.7879 = 78.79\%$$

The results imply that the null hypothesis is 4 times more probable than the alternative hypothesis. Supposing a further examination of 14 other markers would come to similar results the probability of the two children being sisters in proportion to being not related would be 99,99999972 %.

3.6 Further Kinship Calculations

3.6.1 Avuncular Index

The **avuncular index** AI determines the probability that the tested man is the brother of the biological father (that is the uncle) of the child or the father of the biological father (grandfather of the child). There is no analogous value for maternity tests.

The avuncular index is calculated for single markers:

$$AI = \frac{H_1}{H_2} \quad (29)$$

where

H1 is the probability of the tested man being the brother (or father) of the biological father of the child and

H₂ is the probability that another random man of similar ethnical background is the true biological father of the child. H_2 thus corresponds to the value Y applied for the calculation of the PI (formula (10)).

Thereby, **H₁** is calculated as follows:

$$H_1 = \frac{(X+Y)}{2} = 0,5 \cdot X + 0,5 \cdot Y \quad (30)$$

where

X is the probability of the null hypothesis $H(X)$ (formula (9)) and

Y is the probability of the alternative hypothesis $H(Y)$ (formula (9)).

The exact formulas for X and Y can be found in BUCKLETON (2005). With it, the formula for the calculation of the avuncular index can be rearranged as follows:

$$AI = \frac{H_1}{H_2} = \frac{0,5 \cdot X + 0,5 \cdot Y}{Y} = \frac{PI+1}{2} \quad (31)$$

where

H₁ is the probability of the tested man being the brother (or father) of the biological father of the child,

H₂ is the probability that another random man of similar ethnical background is the true biological father of the child,

X is the probability of the null hypothesis $H(X)$ (formula (9)),

Y is the probability of the alternative hypothesis $H(Y)$ (formula (9)) and

PI the paternity index (formula (10)).

Based on the avuncular index there is the **combined avuncular index** CAI. The CAI is the product of the AI of all examined markers:

$$CAI = AI_1 \cdot AI_2 \cdot \dots \cdot AI_n \quad (32)$$

where

n is the number of all examined markers and

AI_{1...n} is the avuncular index (formula (29)) of the markers 1 to n .

The larger AI and CAI the more likely is hypothesis H_1 (the tested man is the brother or father of the true biological father of the child).

In addition, one can use the AI to calculate how much more likely it is that the tested man is not the father of child but its uncle or grandfather:

$$\frac{H_0}{H_1} = \frac{CPI}{CAI} \quad (33)$$

where

H₀ is the probability of the null hypothesis H(X) (formula (9)),

H₁ is the probability of the tested man being the brother or father of the true biological father of the child (formula (30)),

CPI is the combined paternity index (formula (11)) and

CAI is the combined avuncular index (formula (32)).

3.6.2 Probability of Monozygosity in Twins

Knowing the zygosity (mono- or dizygosity) of twins can be of important for medical, scientific and personal reasons. One unambiguous way to solve the questions of zygosity is the use of genetic data. The following calculations (NYHOLT, 2006) expect the twins to exhibit the same genotype for all examined markers. The calculations determine the probability that two twins with a random equal genotype for the examined markers are monozygote twins.

The results of the tests apply to whole population. They solely depend on the frequency distribution of the markers. The genotype of examined twins does not affect the results.

According to the Mendelian laws, the offspring of parents with the genotypes AB and CD (where the A, B, C and D might also describe the same alleles) can have four different genotypes (AC, AD, BC and BD). Of these sixteen combinations, there are 4 cases in which the twins will not share any IBD allele (chapter 3.4), 8 cases in which they will share 1 IBD allele and 4 cases in which they will share two IBD alleles.

One now calculates the **probability of dizygosity** DZ. DZ is the probability that two dizygote twins exhibit the same genotype for one STR marker:

$$DZ = \frac{4}{16}Z_0 + \frac{8}{16}Z_1 + \frac{4}{16}Z_2 = \frac{1}{4}Z_0 + \frac{1}{2}Z_1 + \frac{1}{4}Z_2 \quad (34)$$

where

Z₀ is the probability of two random persons of the population not having any allele in common and

Z₁ is the probability of two random persons of the population having one allele in common and

Z₂ is the probability of two random persons of the population having two alleles in common.

If twins do not have any allele in common they can be handled like unrelated persons. If a population is in Hardy-Weinberg equilibrium (chapter 2.3), the probability M_h of two random persons having the same homozygous genotype is:

$$M_h = \sum_{i=1}^n (p_i^2 \cdot p_i^2) = \sum_{i=1}^n p_i^4 \quad (35)$$

where

n is the number of different alleles for the examined marker and

p_i is the frequency of the i^{th} allele in the given population and

i can attain values from 1 to n .

The probability M_H of two random persons having the same heterozygous genotype is (for a population is in Hardy-Weinberg equilibrium):

$$M_H = \sum_{i=1}^n \sum_{j=i+1}^n (2 \cdot p_i \cdot p_j)^2 \quad (36)$$

where

n is the number of different alleles of the examined marker,

p_i is the frequency of the i^{th} allele in the given population,

p_j is the frequency of the j^{th} allele in the given population,

i can attain values from 1 to n and

j can attain values from 1 to n .

Thus, for a population in Hardy-Weinberg equilibrium, the probability M_0 (or RPM, short for: Random Match Probability) that two random person of a population have the same genotype, is:

$$M_0 = Z_0 = \sum_{i=1}^n p_i^4 + \sum_{i=1}^n \sum_{j=i+1}^n (2 \cdot p_i \cdot p_j)^2 \quad (37)$$

where

n is the number of different alleles of the examined marker,

p_i is the frequency of the i^{th} allele in the given population,

p_j is the frequency of the j^{th} allele in the given population,

i can attain values from 1 to n and

j can attain values from 1 to n .

The probability **a₂** that two dizygote twins share one random allele is the same as the probability of a parent and a child sharing one random allele:

$$a_{2=Z_1} = \sum_{i=1}^n p_i^2 \quad (38)$$

where

- n** is the number of different alleles of the examined marker,
- p_i** is the frequency of the *i*th allele in the given population and
- i** can attain values from 1 to n.

If dizygote twins have both alleles in common, the probability is the same as the probability of monozygote twins having both alleles in common: 1; (as described above one expects a pair of twins having the same genotype).

Inserting these values into formula (34) delivers:

$$\begin{aligned} DZ &= \frac{1}{4}Z_0 + \frac{1}{2}Z_1 + \frac{1}{4}Z_2 \\ &= \frac{1}{4} \left(\sum_{i=1}^n p_i^4 + \sum_{i=1}^n \sum_{j=i+1}^n (2 \cdot p_i \cdot p_j)^2 \right) + \frac{1}{4} \sum_{i=1}^n p_i^2 + \frac{1}{4} \end{aligned} \quad (39)$$

where

- z₀** is the probability of two random persons of the population not having any allele in common,
- z₁** is the probability of two random persons of the population having one allele in common,
- z₂** is the probability of two random persons of the population having two alleles in common,
- p_i** is the frequency of the *i*th allele in the given population,
- p_j** is the frequency of the *i*th allele in the given population
- n** is the number of different alleles of the examined marker,
- i** can attain values from 1 to n and
- j** can attain values from 1 to n.

The calculations expect the persons of the pairs being twins. Hence, the probability of twins is 1. If the twins are not dizygote they have to be monozygote. The probability **MZ** of the twins being monozygote twins is thus:

$$MZ = 1 - DZ \quad (40)$$

where

DZ is the probability of dizygosity (formula (34)).

Besides, there is the **combined probability of dizygosity CDZ** and the **combined random match probability CM₀**. The values are the product of the probabilities of dizygosity of all examined markers, respectively, the product of the M₀ of all examined markers.

The **combined probability of monozygosity CMZ** is the probability of monozygosity when regarding all examined markers:

$$CMZ = (1 - (DZ_1 \cdot DZ_2 \cdot \dots \cdot DZ_n)) \cdot 100 \quad (41)$$

where

n is the number of different alleles of the examined marker,

DZ_{1...n} is the probability of dizygosity (formula (34)) of the markers 1 to n.

The CMZ is presented as percentage. The larger the percentage the more likely it is that two random twins having the same alleles for all examined markers are monozygote twins. Please regard that the values describe the whole population, not a distinct twin pair.

3.6.3 X- and Y-chromosomal Comparisons

X-chromosomal comparison is useful if there is no genetic data about parents or alleged parents but data about blood relatives such as grandparents or siblings available. X-chromosomal comparison do not calculate values or probabilities. It just examines whether the examined X-marker linkage groups might have the same haplotype or not. In case of females, the program will combine and test all possible haplotypes.

X-chromosomal comparisons are based on the sex-specific inheritance of X-chromosomal markers. The X-chromosome of men does always originate from mother, while women inherit one X-chromosome from each parent. Since men possess only one X-chromosome, they inherit this chromosome to all of their female offspring.

This fact can be used when investigating, for example, a potential sister relation. Sisters and half sister of the same father have to share at least one X-chromosome. If the data of the maternal grandparents is available, too, one can determine whether they are half or full siblings. In case of full siblings, the second X-chromosome must also be found in the maternal grandparents.

X-chromosomal comparison can be conduct for every constellation of persons. The interpretation of the results is the responsibility of the reviewer.

Y-chromosomal comparisons are useful if there is no genetic data available about parents or alleged parents but data about male blood relatives on the paternal side. Y-chromosomal comparisons do not

calculate values or probabilities. They just examine whether the examined Y-marker combination is the same or not.

Y-chromosomal comparisons are based on the sex-specific inheritance of Y-chromosomal markers. Only men carry Y-chromosomes. The Y-chromosomes is always inherited from the father.

This can be used, for example, to clarify an alleged paternity in case of male children if there are no samples available for the alleged father, but, for instance, for his brother or father. In case of paternity, the uncle and grandfather must carry the same Y-chromosome as the male child.

All Y-chromosomal markers are inherited as linkage group. As a result, one cannot apply the product rule since it requires independence of the involved partial events. Hence, one cannot determine any combined values. Instead, one can use haplotype frequencies. Haplotype frequencies for Y-chromosomal markers can be in the YHRD (Y Chromosome Haplotype Reference Database) at www.yhrd.org.

3.7 Parameter of calculations

For evaluation of parentage expertises different factors plays a role. These elements can be considered in GenoProof® and are described below.

3.7.1 Population Data

Biostatistical evaluations of blood relationships require knowledge about the distribution of alleles (or haplotypes) in the relevant ethnical group of the population. When selecting frequency data (population data), one must consider several issues:

- Which alleles represent the given marker?
- To which ethnic groups or populations do the examined persons belong?
- Is the selected frequency data reliable?

There is no definite answer to the question, by which alleles a locus is represented. Every system might comprise very rare alleles. For this reason, we have used the selection of commercial test kit manufactures for GenoProof®. The program is capable to identify all alleles investigate be test kits of the ABI, Promega, Serac, Biotype and Qiagen. However, it is also possible to investigate other markers or alleles. You just have to add them to the reference data.

The quality of population data depends on several aspects. One important factor is the sampling number n , that is the number of individuals (or chromosomes) examined to determine the frequencies. Most studies rely on 100 to 200 persons. Both CHAKRABORTY (1992) and EVETT und GILL (1991) concluded that for a determination of proper allele frequency and the calculation of robust likelihood quotients an inquiry

of 100 to 150 person would suffice. However, larger studies improve the accuracy of the results (FOREMAN und EVETT, 2001).

Alleles that are rare in the relevant population and whose frequency is thus not known occur frequently. In order to calculate parameters, one needs apply a frequency to these unknown alleles. An allele frequency for unknown alleles can either be a conservative allele frequency. The German LKA's (State Investigation Bureaus) for instance use an allele frequency of 0.001. Another approach is recommended by the National Research Council (NRCII, 1996). The NRC states that reliable frequencies should occur at least five times. Hence, the minimal allele frequency (and allele frequency for unknown alleles) should be $5/2n$, where n is the number of individuals examined in the study. GenoProof® offers both methods. The desired approach (and the frequency, if applicable) can be chosen in the Preferences.

Other quality characteristics are discussed in chapter 4.

GenoProof® contains a hand-maintained population database with more than 1000 frequency tables for more than 50 populations around the world. The data has been compiled from different sources such as journal publications, online population databases and documentations of test kit manufacturers. The sources are not listed here due to their great number. The reference of each table can be found in the reference data in the population detail dialog.

It is also possible to use other population data. These have to must be added to the reference data first. You can add them yourself, manually, via import (if is it export data from Qualitytype software), or by transfer from own population studies. Please read the internal help for more information. Alternatively, you can ask our support (info@genoproof.de) to add the population data for you.

3.7.2 Linkage Groups

Results of markers belonging to the same linkage group can not be just combined, because all formulas for the calculation of combined values require independent inheritance, which is not the case for linked markers.

One option is to use only one marker of the linkage group. The second option is replacing the referring allele frequencies by one haplotype frequency. Haplotype frequencies can not be calculated. They have to be determined empirically, just as allele frequencies.

GenoProof® offers both approaches. The linkage group a marker belongs to can be defined in the marker reference data. Frequency tables for haplotype frequencies are deposited in the population database. The linkage group of each marker (if there is any) is then be shown during the execution a kinship test in a column of the same name on the input page of the test wizard. You can unselect markers now. If you did not unselect a marker, the program will automatically employ haplotype frequencies instead of allele

frequencies. If there are no haplotype frequencies for the marker combination the program will employ the frequency for unknown alleles which is defined in the preferences.

3.7.3 Mutation Models

There are two possible reasons for incompatible genotype constellations:

- 1.) the constellation is a true mismatch or
- 2.) a mutation has occurred.

One usually excludes paternity or maternity in case of paternity or maternity testing if there are mismatch constellations on three or more chromosomes. However, if there are less than three incompatible constellations, one should consider the possibility of a mutation and also acknowledge this biostatistically.

The ISFG guide lines recommend three different approaches for the calculation of PI concerning the occurrence of a lower number of mismatches (mutations): The RFLP model, the AABB model and the step model. GenoProof® supports all three methods. Mutations can only be regarded by the program if you investigate a direct lineage (paternity, maternity or grandparenthood).

In order to include the possibility of mutation, one needs mutation rates. Paternal mutation rates are often larger than the maternal ones. However, one often refrains from a distinction and uses a total mutation rate for the whole marker system instead. Mutation rates have to be determined empirically. The determined rates are most probably lower than the real rates occurring in nature, since not all mutational events are recognized or measurable by default methods. The total rate rather serves the statistical acknowledgment (HENKE, 1993).

The mutation rates already included in GenoProof® have been obtained from the National Institute of Standards and Technology of the USA (NIST, 2005) and from KAYSER & SAJANTILA (2001). They can be viewed and edited in the marker reference data.

The calculations in the program use maternal and paternal mutation rates: Paternal rates are used in paternity tests, maternal rates in maternity tests. If there is no rate defined for this marker, the program will use a default mutation rate instead. The default rates are 0.001 for autosomal markers, 0.003 for X-chromosomal markers and 0.0028 for Y-chromosomal markers.

In case of the step model (chapter 3.7.3.3) it is possible to calculate a maternal mutation rate from the paternal one (and vice versa) by multiplying the maternal rate with a coefficient (by default: 3.5, respectively dividing the paternal rate by it). Both coefficient and default mutation rates are defined in the preferences.

Hereafter, all mutations rates will be designated by the character μ .

3.7.3.1 The RFLP Model

The model equates the PI with the mutation rate μ . It is the earliest method to consider mutational events. The name of the model comes from its development in connection with restriction fragment length polymorphisms (RFLP).

3.7.3.2 The AABB Model

According to the Annual Report Summary for Testing in 2003 by the American Association of Blood Banks (AABB) 2/3 of the interviewed laboratories are using following formula (GARBER & MORRIS):

$$PI = \frac{\mu}{\bar{A}} \quad (42)$$

where

μ is the mutation rate and

\bar{A} is the average power of exclusion (formula (43)) for the given population.

The average power of exclusion \bar{A} for a population is calculated as follows:

$$\bar{A} = \sum_{i=1}^n p_i \cdot (1-p_i)^2 + \sum_{i < j}^n (p_i p_j)^2 \cdot (3p_i + 3p_j - 4) \quad (43)$$

where

p_i is the frequency of the i^{th} allele in the given population,

p_j is the frequency of the j^{th} allele in the given population,

n is the number of alleles of this marker,

i attains values from 1 to n and

j attains values from 1 to n .

3.7.3.3 Step Model

The step model has been suggested by BRENNER in 2006. It is based on two assumptions: 1.) An increase of STR length is as likely as a reduction of length. 2.) Length changes by one repeat unit are much more likely than changes of several repeat units length. How much more likely a one-unit-change is (compare to a two-unit-change) is not stated uniquely. One often assumes a ratio of 10 while an evaluation of published studies suggests a ratio of 7. The ratio used in GenoProof® is defined in preferences (by default: 7).

The following example shall explain how the ratio is applied: When using a ratio of 10 (for the sake of simplicity), the probability of

- an increase by 1 repeat unit was $0.5 * \mu$
- an decrease by 1 repeat unit was $0.5 * \mu$
- an increase by 2 repeat units was $(0.5 * \mu) / 10$
- an decrease by 2 repeat units was $(0.5 * \mu) / 10$
- an increase by 3 repeat units was $(0.5 * \mu) / (10 * 10)$
- an decrease by 3 repeat units was $(0.5 * \mu) / (10 * 10)$
- etc.

The 0.5 in the numerator is derived from the first assumption (one half of all mutational events result in length increase, the other half results in length decrease).

Non-integer changes (micro-variants) can not be regarded with the step model. According to BRENNER events like these were so unlikely that practically they would never occur.

After BRENNER one does now determine all mutational events that might have resulted in the examined genotype constellation. In paternity tests one assumes that the mutation has happened in the alleged father, in maternity one assumes it has happened in the mother. For example: Mother and child both carry the alleles 12 and 14. The alleged has the alleles 15 and 16. There are four possible events that might have caused this constellation:

- 1.) allele 15 converts to allele 14 (one-step-reduction)
- 2.) allele 15 converts to allele 12 (three-step-reduction)
- 3.) allele 16 converts to allele 14 (two-step-reduction)
- 4.) allele 16 converts to allele 12 (four-step-reduction)

Now, one uses the case with the least steps (case 1 in the example) to calculate the probability X of the observed genotype constellation assuming the alleged father was the true biological father of the child:

$$X = t \cdot 0,5 \cdot \mu \cdot \left(\frac{1}{F}\right)^{s-1} \quad (44)$$

where

- X** is the probability of the observed genotype constellation (including mutation) given the alleged parent is the true biological parent of the child and
- t** is the probability of the examined alleles originating from the alleles parent ($t = 0.5$ if the person carries the allele exactly once) and
- μ** is the mutation rate of the examined marker and

F is the ratio defining how much more likely a one-step-change is compared to a two-step-change and

s is the number of repeat units that have been added or removed by this mutation.

The PI is now calculated by formula (10):

$$PI = \frac{X}{Y}$$

Where

X is the probability of the null hypothesis (the probability of the observed genotype constellation if one assume the alleged parent to be the true biological parent of the child) (formula (9)) and

Y is the probability of the alternative hypothesis (the probability that another random is the true biological parent of the child) (formula (9)).

3.7.4 Silent Alleles

Silent alleles or null alleles are alleles that could not be detected although there are present in an examined person. Silent alleles can be caused by mutations at the primer bindings sites used by a STR test kit (if the site has changed so much that the primer is not able to bind anymore or hardly able to bind, so that there are no or hardly PCR products for this fragment). In electropherograms, one will then find only one peak for the marker. As a result, one might (mistakenly) conclude the genotype for this marker was homozygous.

Silent alleles are extremely rare. Still the guide lines of ISFG demand them to be acknowledged statistically. The first recommendation for a potential silent allele is to examine the marker again using another test kit. However, it is also possible to employ special formulas.

GenoProof® offers a special algorithm by BUCKLETON (2005) to this end. This algorithm calculates the PI (chapter 3.3.1.1) using a frequency for silent alleles. The referring frequency is defined in the preferences. The default value set is 0.0037.

In order to consider potential silent alleles you have to

- 1.) activate the corresponding function on the input page of the test execution wizard and
- 2.) mark the marker you expect to exhibit an silent allele in the child and one parent or grandparents (if there is only one potential silent allele the program will use a mutation model instead).

A marker is marked by removing its homozygosity label in the electropherogram representation or in the genotype of the corresponding person. A typical sign of silent alleles is that the second peak detected

for the marker is only about as high as the peak of heterozygous allele (peak of markers that are truly homozygous are usually larger).

With GenoProof®, silent alleles can be considered when testing maternity or paternity with or without grandparents and without incest.

3.7.5 Subpopulations

Allele frequencies are usually not homogeneous with a population. There are regional differences that are however mostly that small that they do not affect results significantly. Still, in some cases the deviations might be larger so that one can define subpopulations within a population. According to the ISFG guide lines there are two possibilities to regard subpopulations:

- 1.) one can either use special frequency data for the subpopulation or
- 2.) use an algorithm that considers subpopulation when calculating a PI (chapter 3.3.1.1).

GenoProof® contains one of these algorithms (by BUCKLETON, 2005). If persons examined in a paternity or maternity belong to a subpopulation one can calculate the PI using alternative formulas containing a correction coefficient θ . The formulas depend on the genotype constellation. It is also important whether the examined person belong to the same subpopulation or to different ones.

The closer the relationship of the people within a subpopulation the higher is the correction coefficient θ . In smaller subpopulations, in which one can find also mating of distant relatives, one uses a larger θ than in bigger subpopulation whose member are hardly and not related at all (e.g. Afro-Americans).

With GenoProof®, subpopulations can be considered when testing maternity or paternity with or without grandparents and without incest. You have to choose whether to use formulas for greater or for smaller subpopulations. The correction coefficient θ is 0.01 (by default) in both cases. The default values can be changed in the preferences.

4 Population Studies

With GenoProof® you can conduct and evaluate population studies. At this, the program determines both allele frequencies and biostatistical parameters for the evaluation of the results. The results of the study are calculated on basis of the examined peoples' genotypes. The genotypes can be imported, entered manually or identified by raw data analysis.

If a study complies with your standards you can transfer its results to the reference data and use it for calculations.

These are the parameters determined by GenoProof®:

- allele frequencies
- genotype frequencies
- polymorphic information content PIC
- homozygoty h and heterozygoty HET
- power of exclusion PE (for each marker and for the whole study)
- paternity index PI (for each marker and for the whole study)
- power of discrimination PD (for each marker and for the whole study)
- mean chance of exclusion MEC (after Krüger, Desmarais, Chakraborty and Kishida)
- gene diversity (only for Y-chromosomal markers)
- Hardy-Weinberg equilibrium (asymptotic and as exact test)

The next chapters will explain more about these parameters.

4.1 Allele Frequencies

The formula for the calculation of an allele frequency depends on the marker type:

Autosomal markers:

$$f_i = \frac{n_i}{2N} \quad (45)$$

X-chromosomal markers:

$$f_i = \frac{f_m + 2 \cdot f_F}{3} \quad (46)$$

Y-chromosomal markers:

$$f_i = \frac{n_i}{N} \quad (47)$$

where

- f_i is the frequency of allele i and
- n_i is the count of allele i occurrences in the study and
- i becomes all values from 1 to n and
- f_M is the frequency of allele i in men (calculated using the formula for Y-chromosomal markers) and
- f_F is the frequency of allele i in women (calculated using the formula for autosomal markers) and
- N is the number of persons examined for this marker.

4.2 Genotype Frequencies

The **genotype frequency** f of a genotype G consisting of the allele i and j (where i and j might be the same allele) is calculated as follows:

$$f = \frac{n}{N} \quad (48)$$

where

- n is the number of examined individuals having genotype G in the study and
- N is the number of persons examined for this marker.

Genotype frequency and allele frequency (formula (46) und (47)) are the same for a marker, if the marker has only one copy per cell (e.g. Y-chromosomal markers and X-chromosomal markers in men).

4.3 Polymorphic Information Content (PIC)

The **polymorphic information content** **PIC** is a measure for distinctiveness of a marker. It is calculated as follows:

$$PIC = 1 - \sum_{i=1}^n p_i^2 - \left(\sum_{i=1}^n p_i^2 \right)^2 + \sum_{i=1}^n p_i^4 \quad (49)$$

where

- p_i is the frequency of the i^{th} allele in the given population,
- n is the number of persons examined for this marker and
- i attains values from 1 to n .

4.4 Homozygoty (h) und Heterozygoty (HET)

Homozygosity h and **heterozygosity HET** represent the expected fraction of homozygous and heterozygous genotypes in a population. Hence it is:

$$h + HET = 1 \quad (50)$$

The **homozygosity** is determined by summing up the frequencies of all homozygous genotypes. This requires that the population is in Hardy-Weinberg equilibrium (chapter 2.3)

$$h = \sum_{i=1}^n p_i^2 \quad (51)$$

where

p_i is the frequency of the *i*th allele in the given population,

n is the number of persons examined for this marker and

i attains values from 1 to n.

The expected **heterozygosity** is determined by subtracting the homozygosity from 1:

$$HET = 1 - h = 1 - \sum_{i=1}^n p_i^2 \quad (52)$$

where

p_i is the frequency of the *i*th allele in the given population,

n is the number of persons examined for this marker and

i attains values from 1 to n.

The higher the heterozygosity the higher is the allele diversity and the less likely is it that two random persons of the population exhibit the same genotype.

4.5 Power of Exclusion (PE)

The **power of exclusion PE** has first been described by FISHER (1951). It is calculated as follows:

$$PE = HET^2 \cdot (1 - (1 - HET) \cdot HET)^2 \quad (53)$$

where

HET is the heterozygosity of the given marker (formula (52)).

Besides the PE, there is the **combined power of exclusion CPE** summarizing the PE values of all examined markers:

$$CPE = PE_1 \cdot PE_2 \cdots PE_n \quad (54)$$

where

n is the number of examined markers and

PE_{1...n} is the power of exclusion PE (formula (53)) of the markers 1 to n.

4.6 Paternity Index (PI)

The **paternity index PI** of a population study indicates how much more likely it is that a child's genotype will support the null hypothesis X (the alleged parents is the true biological parent of the child) compared to the chance that it would support the alternative hypothesis Y (another random person unrelated with the alleged parent is the true biological parent of the child).

$$PI = \frac{HET+h}{2h} = \frac{(1-h)+h}{2h} = \frac{1}{2h} = \left(2 \cdot \sum_{i=1}^n p_i^2 \right)^{-1} \quad (55)$$

where

HET is the heterozygoty of the given marker (formula (52)),

h is the homozygoty of the given marker (formula (51))

p_i is the frequency of the *i*th allele in the given population,

n is the number of persons examined for this marker and

i attains values from 1 to n.

Based on the paternity index, there is the **combined paternity index CPI**. The CPI is product of the PI of all examined markers (formula (11)):

$$CPI = PI_1 \cdot PI_2 \cdots PI_n$$

where

n is the number of examined markers and

PI_{1...n} is the paternity index PI (formula (55)) of the markers 1 to n.

4.7 Power of Discrimination (PD)

The **power of discrimination PD** is the probability that two different individuals of a population have a different genotype for a marker:

$$PD = 1 - 2 \cdot \left(\sum_{i=1}^n p_i^2 \right)^2 - \sum_{i=1}^n p_i^4 \quad (56)$$

where

p_i is the frequency of the i^{th} allele in the given population,

n is the number of persons examined for this marker and

i attains values from 1 to n .

For markers that have only one copy per cell (Y-chromosomal markers and X-chromosomal markers in men), one uses another formula:

$$PD = 1 - \sum_{i=1}^n p_i^2 \quad (57)$$

where

p_i is the frequency of the i^{th} allele in the given population,

n is the number of persons examined for this marker and

i attains values from 1 to n .

In addition to the power of discrimination there is **combined power of discrimination** CPD. The CPD requires that all markers of the study are inherited independently.

$$CPD = PD_1 \cdot PD_2 \cdots PD_n \quad (58)$$

where

n is the number of examined markers and

$PD_{1...n}$ is the power of discrimination PD (formula (57)) of the markers 1 to n .

4.8 Mean Exclusion Chance (MEC)

The **mean exclusion chance MEC** indicates the probability with which the population data excludes non-fathers and non-mothers from paternity and maternity. There are different formulas to calculate the MEC depending on the marker type.

The MEC for autosomal markers is calculated as follows (after KRÜGER, 1968):

$$MEC_{Krüger} = \sum_{i=1}^n p_i^3 \cdot (1-p_i)^2 + \sum_{i=1}^n p_i \cdot (1-p_i)^3 + \sum_{i < j}^n p_i p_j \cdot (p_i - p_j) \cdot (1-p_i - p_j) \quad (59)$$

where

p_i is the frequency of the i^{th} allele in the given population,

p_j is the frequency of the j^{th} allele in the given population,

n is the number of persons examined for this marker,

i attains values from 1 to n and

j attains values from 1 to n .

An MEC for Y-chromosomal markers and male children has been developed by CHAKRABORTY (1985). The value is calculated with the same formula like the PD for a marker with one copy per cell (formula (57)).

The MEC for X-chromosomal markers and female children (after KISHIDA, 1997) is calculated as follows:

$$MEC_{Kishida} = \sum_{i=1}^n p_i^3 \cdot (1-p_i)^2 + \sum_{i=1}^n p_i \cdot (1-p_i)^2 + \sum_{i < j}^n p_i p_j \cdot (p_i - p_j) \cdot (1-p_i-p_j) \quad (60)$$

where

p_i is the frequency of the i^{th} allele in the given population,

p_j is the frequency of the j^{th} allele in the given population,

n is the number of persons examined for this marker,

i attains values from 1 to n and

j attains values from 1 to n .

The MEC according to DESMARAIS (1998) is calculated as follows:

$$MEC_{Desmarais-Trio} = 1 - \sum_{i=1}^n p_i^2 + \sum_{i=1}^n p_i^4 - \left(\sum_{i < j}^n p_i^2 \right)^2 \quad (61)$$

$$MEC_{Desmarais-Duo} = 1 - 2 \cdot \sum_{i=1}^n p_i^2 + \sum_{i=1}^n p_i^3 \quad (62)$$

where

p_i is the frequency of the i^{th} allele in the given population,

n is the number of persons examined for this marker and

i attains values from 1 to n .

4.9 Gene Diversity (GD)

The **gene diversity GD** is only determined for Y-chromosomal markers. The gene diversity indicates the expected heterozygosity meaning the probability of two randomly selected alleles being different:

$$GD = \frac{n}{n-1} \cdot \left(1 - \sum_{i=1}^n p_i^2 \right) \quad (63)$$

where

p_i is the frequency of the i^{th} allele in the given population,

n is the number of persons examined for this marker and

i attains values from 1 to n .

4.10 Hardy-Weinberg Equilibrium (HWG)

A Hardy-Weinberg test investigates whether a population meets the conditions of a Hardy-Weinberg equilibrium. In a population in Hardy-Weinberg equilibrium one finds a certain ratio between homozygous and heterozygous genotypes (depending on the allele frequencies) which remains constant from generation to generation.

The mathematical model assumes an ideal population. That means a population without mutation, selection and migration in which all parent combination are the same likely and which is so large that the frequency distributions of genotypes and alleles are basically unaffected by random losses of an individuals are gene drift.

Heavy deviations from the Hardy-Weinberg equilibrium can be indicators of inbreeding of silent alleles (chapter 3.7.4) if there are more homozygous genotypes than expected.

The Standard Hardy-Weinberg test is a specialized chi-square test investigating whether a population is in Hardy-Weinberg equilibrium. The test is performed for single markers. In doing so, one first determines the chi-square value using the formula:

$$\chi^2 = \sum_{i=1}^k \frac{(g_o - g_e)^2}{g_e} \quad (64)$$

where

g_o is the observed frequency of genotype i in the population,

g_e is the expected frequency of genotype i in the population (formula (7)),

k is the number of different genotypes for this marker in the population and

i attains values from 1 to k .

The chi-square value is then compared to a contingency table. The null hypothesis ('The population meets the conditions of the Hardy-Weinberg equilibrium') is accepted if the chi-square is lower or equal to the value indicated by the contingency for the given number of degrees of freedom and the desired significance level.

There are no fixed significance levels in GenoProof®. Instead, the program calculates the highest possible significance level that allows accepting the null hypothesis. The significance level (which is the probability of rejecting a correct null hypothesis) is then subtracted from 100% and presented as probability. In other words: The greater the probability the lesser the chance of error.

Always check the probability when performing a Hardy-Weinberg test!

The Hardy-Weinberg test described above is a so-called asymptotic test. Asymptotic tests produce unreliable results if the sampling number is too low. Hence, they can only be applied if you have a certain minimum number of samples.

The results of exact tests in contrast are also reliable if you have only a small number of results. Exact tests required many calculations and their conduction might take several minutes. Exact tests are not performed automatically for this reason. Instead you have to start them manually in on the **Statistical Parameters tab** of the **editor window** for population study marker results.

The Hardy-Weinberg exact test implemented in GenoProof® is based on a Monte-Carlo approach published by Guo in 1992. This method compares the observed distribution of genotypes with the distribution of simulated populations that are similar to the observed one. 'A similar population' is here defined as a population based on the same number of samples and the same allele counts like the observed population but with genotypes that are assembled randomly.

Now one determines for each simulated population how likely its genotype distribution would be if you assume the Hardy-Weinberg equilibrium to be met. The probability of the observed population R is compared to the probability of each simulated population. One then determines a value P according to the formula:

$$P = \frac{K}{n} \quad (65)$$

where

K is the count of simulated populations having a probability that is lower or equal to the probability of the population R (the observed population) and

n is the total count of simulated populations.

GenoProof® simulates 17,000 populations for the comparison as suggested by Guo. The P value one achieves using this number of simulations has a 99%-chance of being situated within an interval of ± 0.01 units around the real P value.

5 Chimerism Analyses

Chimerism analyses investigate the success of bone marrow transplantations. Shortly after transplantation, the patient will have both cells with the original genotype P of the patient and cells with the genotype D of the donor. Thus, the patient shows the characteristics of a chimera. In an ideal case, the fraction of cells with the original patient's genotype P will reduce to zero within time, so that finally one finds only cells with the donor's genotype D.

A chimerism analysis determines the fraction of donor genotype F(D) within a sample. In order to conduct a chimerism analysis, one needs to know the genotypes D and P (the genotype of the donor and the genotype of the patient before transplantation). Only alleles with an unambiguous assignment to either D or P are considered in the calculation. Every marker, differing in at least one allele, can be included into the analysis.

The analysis can be performed on basis of either peak areas (recommended) or peak heights.

Following formula calculates the fraction of donor genotype in %:

$$F(D) = 100 \cdot \frac{A(D)}{A(D) + A(P)} \quad (66)$$

$$F(D) = 100 \cdot \frac{h(D)}{h(D) + h(P)}$$

where

F(D) is the fraction of cells with the genotype D of the donor,

A(D) is the total area of the peaks of the donor alleles,

A(P) is the total area of the peaks of the original alleles of the patient,

h(D) is the height of the peaks of the donor alleles,

h(P) is the height of the peaks of the original alleles of the patient.

An overall evaluation of the sample is achieved by calculation of the arithmetic average of all marker percentages that could be included into the analysis.

The program does not only calculate the fraction of cells with the genotype of the donor but also a limit of detection corresponds with a noise-to-signal-ratio. With that, the ratio of noise signals to allele signals is suggested. It is calculated by ratio of the biggest peak without allele assignment of the panel to the weakest allele peak of the marker.

Note: Sometimes you will not obtain a detection limit. Possible reasons could be that markers could not be included into the analysis or that noise levels were so low, that no peaks have been detected in the baseline at all.

Following formula is used for calculation of the detection limit:

$$DL = \frac{A_{\max}(\text{Panel})}{A_{\min}(\text{Marker})} \quad (67)$$

$$DL = \frac{h_{\max}(\text{Panel})}{h_{\min}(\text{Marker})}$$

where

DL is the limit of detection, ratio of noise signals to allele signals,

A_{max}(Panel) is the peak area of the biggest peak without allele assignment in the panel area,

A_{min}(Marker) is the peak area of the allele peak with the smallest area of the respective marker,

h_{max}(Panel) is the highest peak without allele assignment in the panel area and

h_{min}(Marker) is the lowest allele peak of the respective marker.

Following, the detection limit values suggest following conclusions:

detection limit > 1 The highest peak without allele assignment is bigger than the weakest allele peak.

detection limit = 1 The highest peak without allele assignment is as big as the weakest allele peak.

detection limit < 1 The highest peak without allele assignment is smaller than the weakest allele peak.

Note: In some circumstances it is possible that no detection limit will be displayed. That can be the case whenever a marker could not be considered or whenever the background noise is too low so that no peak could be detected for the base line.

6 Forensic Examinations

In the section Forensics you can create forensic examinations to calculate the probability of identity.

The **probability of identity** (match probability) is the likelihood, by which one can exclude the inadvertent mistaking of two persons having the same combination of attributes. This probability is calculated for sample genotypes according to Hummel (1997). Thereby an integrated population database containing the allele frequencies of all commercial markers for many ethnic groups is used.

The following examples demonstrate how probabilities of identity are calculated in GenoProof®:

- Let us assume a person with following genotype and frequencies:
 - Marker 1:** Alleles aa with $p(a)=0.6$; according to the Hardy-Weinberg equilibrium (formula (7)) it is $p(a-a)=a^2=0.36$
 - Marker 2:** Alleles bc with $p(b)=0.5$ and $p(c)=0.4$; according to the Hardy-Weinberg equilibrium (formula (7)) it is $p(b-c)=2 \cdot p(b) \cdot p(c)=0.4$
 - Marker 3:** Alleles dd with $p(d)=0.02$; according to the Hardy-Weinberg equilibrium (formula (7)) it is $p(d-d)=d^2=0.004$
- The **probability of exclusion in case of identity** is calculated by the following equation:

$$A=1-(p(a-a) \cdot p(b-c) \cdot p(d-d)) \quad (68)$$

Hence, it is $1-(0.36 \cdot 0.4 \cdot 0.004)=0.9994$;

an accidental mistaking is recognized with a chance of 99.94%.

- The **probability in case of non-identity** A^- is calculated in the following way:

$$A^-=1-A \quad (69)$$

This is $1-0.9994=0.0006$ that is 0.06%.

- The **probability of identity** is calculated as following:

$$W=\frac{1}{1+A^-/A} \quad (70)$$

With it, it is $W=\frac{1}{1+0.0006/0.999}=0.9994$ that is 99.94%.

- The **probability of non-identity** is:

$$W^-=1-W \quad (71)$$

that is 0.06% in our example.

Note: Whenever a homozygote is observed in a DNA profile, **allele dropout** is an important consideration. Allele dropout can take place due to silent alleles or low DNA concentration.

Therefore according to NRC II recommendations $2p$ instead of p^2 should be used for calculation of genotype frequencies.

In GenoProof® the **2p rule** is used if an only allele in a marker is not marked as homozygote and therefore is considered to be an allele dropout. It is not used for X- and Y-chromosomal marker in samples of male persons.

7 Index of Tables and Figures

Table 1: Predicates of paternity testing according to Hummel (1997).....	13
Table 2: Formula for trio cases	20
Table 3: Formula for duo cases	21
Table 4: Formula for calculation of the paternity index for paternity test in case of incest	23
Table 5: IBD coefficients for autosomal marker	24
Figure 1: Null hypothesis and alternative hypothesis of a paternity test	9
Figure 2: Trio cases for paternity tests (on the left) and maternity tests (on the right).....	16
Figure 3: Duo cases for paternity tests (on the left) and maternity tests (on the right)	17
Figure 4: Null hypothesis $H(X)$ in a paternity test	18
Figure 5: Alternative hypothesis $H(Y)$ in a paternitiy test.....	19
Figure 6: Complex parentage case.....	26

8 References

- American Association Blood Banks AABB (2004): „Annual Report Summary for Testing In 2003“, prepared by the Parentage Testing Standards Program Unit.
- Baur MP (1993): “Biostatistical evaluation of DNA polymorphisms and PCR results in paternity cases”, ESTM course, 1993, Venice.
- Butler JM (2005): “Forensic DNA Typing - Biology, Technology and Genetics of STR Markers”, 2. Ed, Elsevier Academic Press, London.
- Buckleton J, Triggs C & Walsh S (2005): “Forensic DNA Evidence Interpretation”, CRC Press, Boca Raton, Florida.
- Chakraborty R (1985): “Paternity testing with genetic markers: are Y-linked genes more effective as autosomal ones?”, *Am J Med Genet.* 1985 Jun;21(2):297-305.
- Chakraborty R (1992): “Sample size requirements for addressing the population genetic issues of forensic use of DNA typing”, *Hum Biol.* 1992 Apr; 64(2):141-59.
- Conradt J (1983): “Serostatistische Absammungsbegutachtung: Ein Algorithmus für Verwandtschaftsfälle und das Daten- und Programmsystem PAPS”, PhD Thesis, Institut für medizinisch-biologische Statistik und Dokumentation der Philipps-Universität Marburg/Lahn. 1983.
- Desmarais D, Zhong Y, Chakraborty R, Perreault C, Busque L (1998): „Development of a highly polymorphic STR marker for identity testing purposes at the human androgen receptor gene (HUMARA)“. *J Forensic Sci.* 1998 Sep;43(5):1046-9.
- Essen-Möller E (1938): „Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis – Theoretische Grundlagen“, *Mitt. Anthropol. Ges. Wien*, 68, 9-53.
- Essen-Möller E, Quensel CE (1939): „Zur Theorie des Vaterschaftsnachweises aufgrund von Ähnlichkeitsbefunden“, *Dt. Z. ges. gerichtl. Med.* 31, 70-96.
- Evetts IW & Gill P (1991): “A discussion of the robustness of methods for assessing the evidential value of DNA single locus profiles in crime investigations”, *Electrophoresis.* 1991 Feb-Mar;12(2-3):226-30.
- Fisher RA (1951): „Standard calculations for evaluating a blood-group system“, *Heredity.* 1951 Apr;5(1):95-102.
- Foreman LA & Evetts IW (2001): „Statistical analyses to support forensic interpretation for a new ten-locus STR profiling system“, *Int J Legal Med.* 2001;114(3):147-55.

- Fung WK (2003): „User-friendly programs for easy calculations in paternity testing and kinship determinations“, *Forensic Science International* 136 (2003) 22–34.
- Gill P, Brenner CH, Buckleton JS, Carracedo A, Krawczak M, Mayr WR, Morling N, Prinz M, Schneider PM, Weir BS (2006): „DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures“, *Forensic Science International* 160 (2006) 90–101.
- Gjertson DW, Brenner CH, Baur MP, Carracedo A, Guidet F, Luque JA, Lessig R, Mayr WR, Pascali VL, Prinz M, Schneider PM, Morling N (2007): „ISFG: Recommendations on biostatistics in paternity testing“, *Forensic Sci Int Genet.* 2007 Dec;1(3-4):223-31. Epub 2007 Aug 6.
- Guo SW, Thompson EA (1992): „Performing the exact test of Hardy-Weinberg proportion for multiple alleles“, *Biometrics.* 1992 Jun;48(2):361-72.
- Henke J, Fimmers R, Baur MP, Henke L (1993): „DNA-minisatellite mutations: recent investigations concerning distribution and impact on parentage testing“, *Int J Legal Med.* 1993;105(4):217-22.
- Hoppe JD, Kurth R & Sewing KF (2002): „Richtlinien für die Erstattung von Abstammungsgutachten“, *Bekanntmachung der Bundesärztekammer in Deutsches Ärzteblatt*, 8th March 2002
- Hummel K (1997): „Erblich-polymorphe Eigenschaften des Blutes zur Klärung strittiger Blutverwandtschaften und fraglicher Identitäten“, Verlag Dr. Kovac.
- Kayser M, Sajantila A (2001): „Mutations at Y-STR loci: implications for paternity testing and forensic analysis“, *Forensic Sci Int.* 2001 May 15;118(2-3):116-21.
- Kishida T, Tamaki Y (1997): „Japanese population data on X-chromosomal STR locus AR“. *Nihon Hoigaku Zasshi (Japanese Journal of Legal Medicine).* 1997 Oct;51(5):376-9.
- Krüger J, Fuhrmann W, Lichte KH, Steffens C (1968): „Zur Verwendung von sauren Erythrocyten-Phosphatasen bei der Vaterschaftsbegutachtung“, *Dtsch Z Gesamte Gerichtl Med.* 1968;64(2):127-46.
- Nyholt DR (2005): „On the Probability of Dizygotic Twins Being Concordant for Two Alleles at Multiple Polymorphic Loci“, *Twin Research and Human Genetics*, Volume 9 Number 2 pp. 194–197
- NRCII (National Research Council Committee on DNA Forensic Science) (1996): „The Evaluation of Forensic DNA Evidence“, National Academy Press, Washington DC.
- Minataka K, Ishitani A, Ito N, Nagaike C, Morimura Y, Hatake K, Suzuki O (1996): „[Paternity probability in the cases of incest]“, *Nihon Hoigaku Zasshi.* 1996 Jun;50(3):149-55. Japanese.
- NIST (2005): <http://www.cstl.nist.gov/biotech/strbase/mutation.htm>.
- Weir BS (1996): „Genetic Data Analysis II: Methods for Discrete Population Genetic Data“, Sunderland, Massachusetts, Sinauer Associates.

