Microsat_errcalc

User manual

Honka & Merikanto 2020

Microsat_errcalc, an error rate calculation program for microsatellite data. Designed by Johanna Honka. Code written by Kari Merikanto.

Microsat_errcalc source files are available from GitHub repository: https://github.com/karimerikanto/microsat_errcalc. An executable file is available from the authors; write an email request to johanna.honka@oulu.fi.

Contact information: Johanna Honka, PhD, University of Oulu
email: johanna.honka@oulu.fi

Cite as:

Honka, J. & Merikanto, K. 2020. Microsat_errcalc. GitHub repository. Available: https://github.com/karimerikanto/microsat_errcalc (*date*).

Introduction

Microsat_errcalc is a program that automates the calculation of genotyping error rate in microsatellite data. It is especially suitable for noninvasively collected samples, which may include several samples from the same individual and individuals sampled only once. Microsat_errcalc automatically omits individuals genotyped only once, only estimating error rates in individuals sampled twice or more, thus reducing data handling. Microsat_errcalc uses a simple excel-file as an input. In addition, the program can handle very large datasets (several thousands of samples). Microsat_errcalc can also estimate allele dropout rate and rate of false alleles or other errors.

This program performs direct-count error estimates, that is errors observed in the data. This method assumes that the correct genotype is present in the data, which might not always be true especially for solely non-invasively collected data. Microsat_errcalc does not perform iterative error rate estimation. For such purposes, we recommend program called Pedant (Johnson & Haydon 2007), which performs a maximum-likelihood-based method for error rate estimation from two replicates of single individual. Unlike other error rate calculation programs for microsatellite data, Microsat_errcalc can handle multiple replicates of a single individual and missing data. Other error rate calculation programs such as Pedant require a single replicate of an individual or are unable to handle large dataset such as Gimlet (Valière 2002).

Microsat_errcalc does not perform identification of null alleles (nonamplified alleles), short allele dominance (large-allele dropout) or identification of stutter peaks, which can also be classified as genotyping errors. For checking the data for such errors using Hardy-Weinberg equilibrium and for checking typographic errors (allele was typed in wrongly by the researcher), we recommend program Micro-Checker (Van Oosterhout et al. 2004) and for null allele estimation program FreeNA (Chapuis & Estoup 2007). Also, other programs are suitable for estimating genotyping errors using pedigree or parentage data such as program Cervus (Kalinowski et al. 2007) or program Colony (Jones & Wang 2010).

Microsatellite genotyping error rate

Microsatellite markers are inherently prone to genotyping errors, which occur when the real genotype does not match with the observed genotype (Bonin et al. 2004). Genotyping errors are problematic especially with non-invasively collected samples that contain low quantity of DNA and/or low quality DNA and genotyping errors can lead to false interpretations of the biolical data (Tabelet et al. 1996, Taberlet & Luikart 1999, Taberlet et al. 1999, Bonin et al. 2004, Hofmann & Amos 2005, Pompanon et al. 2005). As the genotype needs to be assessed with molecular analyses that are not error free, the real genotype cannot be directly assessed. Thus, genotyping errors are in practice defined as differences between two or more genotypes, genotyped independently from the same sample (Bonin et al. 2004). With non-invasively collected data, the same individual can be genotyped two or more times from independent samples (individual has shed multiple feathers, visited several hair snag sites or visited repeatedly the same snag site, left behind multiple faeces etc.). Such replicates pro-

duced by the non-invasive sampling process can also be used to asses genotyping errors. Such replicates present truly blind samples as the researcher does not initially know which samples belong to a same individual.

Microsatellite datasets nearly always contain genotyping errors and thus should be accounted for in order not to bias the final results (Bonin et al. 2004). One way of accounting for genotyping errors is to quantify the amount of genotyping errors (Bonin et al. 2004). This can also help to identify unreliable markers (unstable markers or markers which are difficult to score) or unreliable genotypes (potentially samples with too low-quality DNA for reliable analyses) which can be removed from the dataset. When the error rate is known, the researcher can make the decision if the data is trustworthy (Bonin et al. 2004).

Causes of genotyping errors can be assigned to four classes: variation in DNA sequence, low quantity or quality of DNA, biochemical artefacts or human errors (Pompanon et al. 2005). Error in variation in DNA sequence can be generated through a mutation in the flanking sequence of the microsatellite marker or within the primer site (for example deletion or insertion within the priming site). This manifests as an absence of PCR product and are called as null alleles (Callen et al. 1993). Null alleles lead to the absence of that allele and if the other allele amplifies, to false homozygosity. Low quantity and/or quality of DNA lead to allelic dropouts or false alleles (Taberlet et al. 1996). Allele dropout is a stochastic non-amplification of an allele i.e. only one allele amplifies in heterozygous locus (Taberlet et al. 1996). Allele dropout can occur in homozygous locus too, but the consequences of such allele dropout are indistinguishable from homozygote (Creel et al. 2003). False alleles on the other hand are PCR artefacts that resemble alleles i.e. heterozygous locus in a locus that is homozygous in reality. Of these types of errors caused by low quantity or quality DNA, allele dropouts are far more common than false alleles (Bonin et al. 2004). Biochemical artefacts are caused by the tendency of the *Taq* polymerase to add an adenine to the end of the 3' sequence, which is a rather common artefact (Pompanon et al. 2005). Often genotyping errors have a human cause due to scoring errors, data input mistakes, allelic dropouts, sample mix-up, pipetting error or contamination (Hofmann & Amos 2005). Of these errors, Microsat_errcalc can calculate the total error (caused by all factors), allele dropout rate (homozygote in otherwise heterozygous locus) and false allele/other error rate. Microsat_errcalc cannot distinguish errors caused by false alleles from human caused errors and these are thus not separated. Also, note when an individual is genotyped twice, Microsat_errcalc assumes that the heterozygous genotype is the correct genotype and calls this error as an allele dropout (allele dropouts are more common than false alleles). In order to verify the heterozygous nature of the locus, a third genotyping should be performed.

How the program works

Genotyping error rate is calculated based on replicated individuals as the number of errors divided by the number of alleles in which an error could have been detected (Creel et al. 2003) i.e. unsuccessful PCR-attempts were excluded. In addition, the error rate per locus is also calculated as this is the most commonly used estimate of error rate (Pompanon et al.

2005). The error rate per locus was calculated as the number of loci containing an error divided by the number of successfully amplified loci.

In order to estimate allele dropouts (ADO, amplification of only one allele in heterozygous loci) and false alleles (FA, PCR-generated) and other errors we used only samples that we genotyped in triplicates or more and thus having more reliable consensus genotype. An error in duplicate genotype could either be from an allele dropout or a false allele but with three or more genotypes we can differentiate between these categories with much higher confidence. The allele dropout rate (ADO, amplification of only one allele in heterozygous loci) is calculated as the number of amplifications containing a loss of one allele divided by the number of successfully amplified heterozygous loci as ADO can be detected only in heterozygous genotypes (Broquet & Petit 2004). ADO can occur also in homozygous genotypes but the result is indistinguishable from a true homozygote and thus undetected (Creel et al. 2003). FA and other allele rate is calculated as the number of amplifications containing a false allele divided by the number of successfully amplified loci as the false alleles can occur in both homo- and heterozygous genotypes (Broquet & Petit 2004). The ADO and FA/other error rates per alleles is calculated by dividing the number of ADO or FA/other error divided by the number of alleles in which an error could have been detected.

Input file

The input file is .csv file in which each individual is separated by a line brake (Fig. 1). Missing data is coded as "?" . NOTE, samples replicated only once do not need to be removed as the program calculates error rate only in samples replicated twice or more.

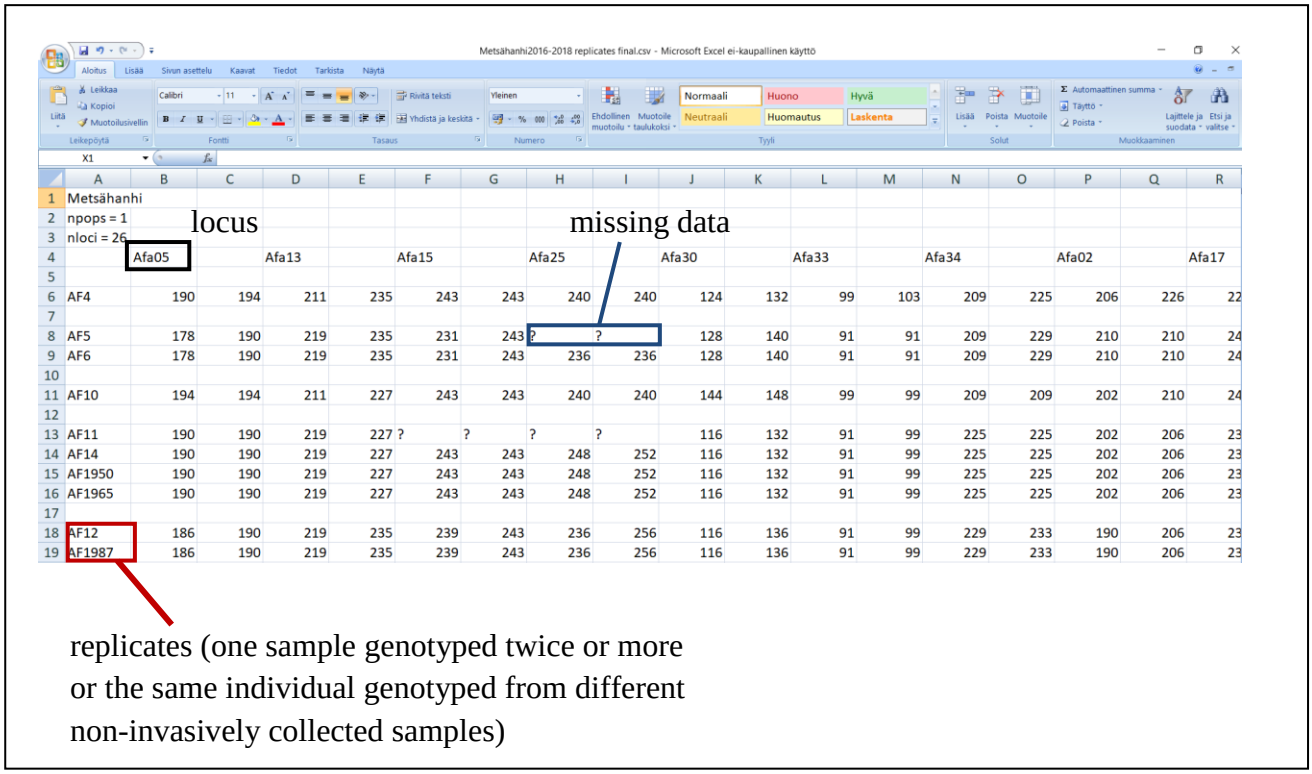
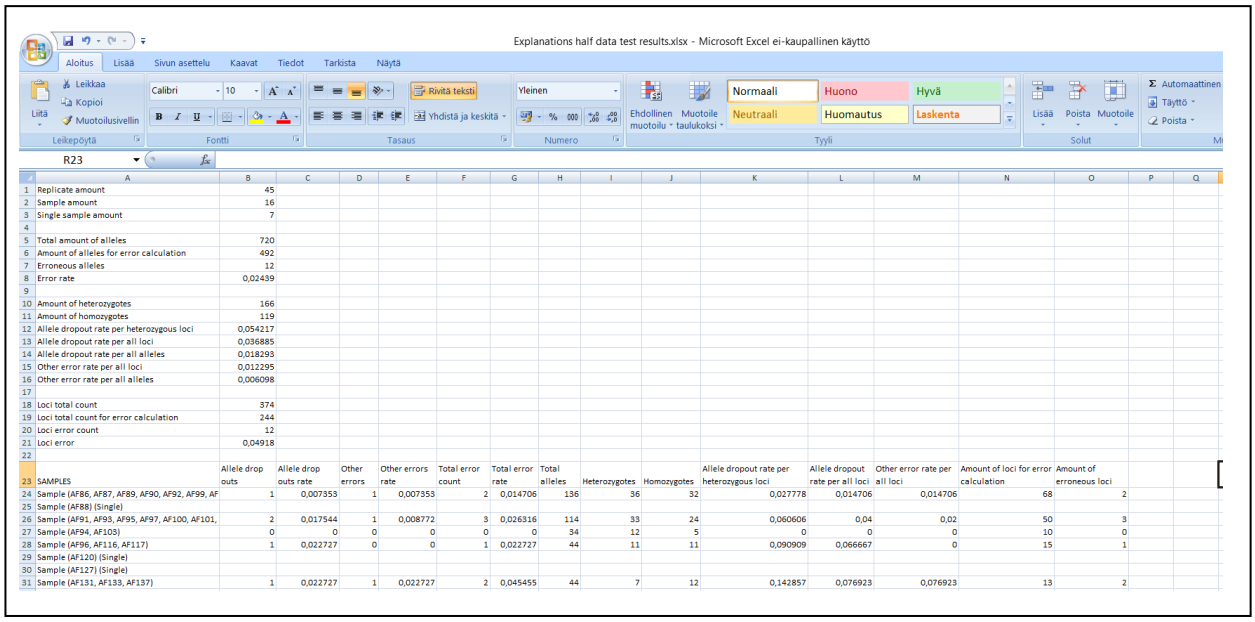


Fig. 1 Example of an input file

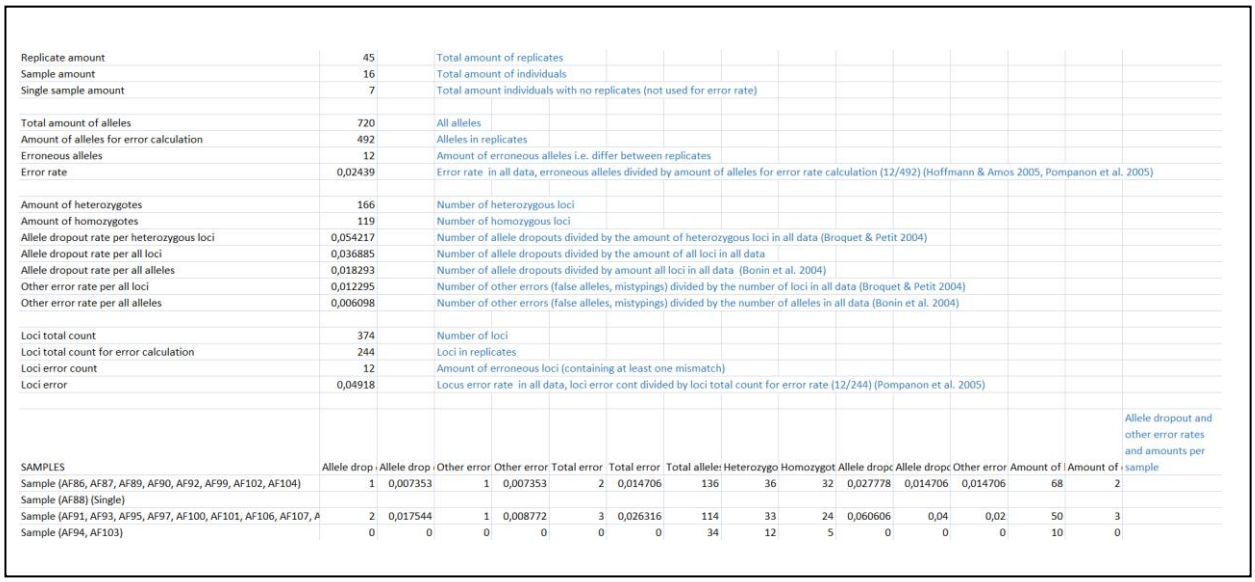
Output file

Example of an output file (Fig. 2). Explanation of the results in Figure 3.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Replicate amount	45															
2	Sample amount	16															
3	Single sample amount	7															
4																	
5	Total amount of alleles	720															
6	Amount of alleles for error calculation	492															
7	Erroneous alleles	12															
8	Error rate	0,02439															
9																	
10	Amount of heterozygotes	166															
11	Amount of homozygotes	119															
12	Allele dropout rate per heterozygous loci	0,054217															
13	Allele dropout rate per all loci	0,036885															
14	Allele dropout rate per all alleles	0,018293															
15	Other error rate per all loci	0,012295															
16	Other error rate per all alleles	0,006098															
17																	
18	Loci total count	374															
19	Loci total count for error calculation	244															
20	Loci error count	12															
21	Loci error	0,04918															
22																	
23	SAMPLES	Allele drop outs	Allele drop outs rate	Other errors	Other errors rate	Total error count	Total error rate	Total alleles	Heterozygotes	Homozygotes	Allele dropout rate per heterozygous loci	Allele dropout rate per all loci	Other error rate per all loci	Amount of loci for error calculation	Amount of erroneous loci		
24	Sample (AF86, AF87, AF89, AF90, AF92, AF99, AF	1	0,007353	1	0,007353	2	0,014706	136	36	32	0,027778	0,014706	0,014706	68	2		
25	Sample (AF88) (Single)																
26	Sample (AF91, AF93, AF95, AF97, AF100, AF101,	2	0,017544	1	0,008772	3	0,026316	114	33	24	0,060606	0,04	0,02	50	3		
27	Sample (AF94, AF103)	0	0	0	0	0	0	34	12	5	0	0	0	10	0		
28	Sample (AF96, AF116, AF117)	1	0,022727	0	0	1	0,022727	44	11	11	0,090909	0,066667	0	15	1		
29	Sample (AF120) (Single)																
30	Sample (AF127) (Single)																
31	Sample (AF131, AF133, AF137)	1	0,022727	1	0,022727	2	0,045455	44	7	12	0,142857	0,076923	0,076923	13	2		

Fig. 2 Example of an output file (.csv)



Replicate amount	45	Total amount of replicates
Sample amount	16	Total amount of individuals
Single sample amount	7	Total amount individuals with no replicates (not used for error rate)
Total amount of alleles	720	All alleles
Amount of alleles for error calculation	492	Alleles in replicates
Erroneous alleles	12	Amount of erroneous alleles i.e. differ between replicates
Error rate	0,02439	Error rate in all data, erroneous alleles divided by amount of alleles for error rate calculation (12/492) (Hoffmann & Amos 2005, Pompanon et al. 2005)
Amount of heterozygotes	166	Number of heterozygous loci
Amount of homozygotes	119	Number of homozygous loci
Allele dropout rate per heterozygous loci	0,054217	Number of allele dropouts divided by the amount of heterozygous loci in all data (Broquet & Petit 2004)
Allele dropout rate per all loci	0,036885	Number of allele dropouts divided by the amount of all loci in all data
Allele dropout rate per all alleles	0,018293	Number of allele dropouts divided by amount all loci in all data (Bonin et al. 2004)
Other error rate per all loci	0,012295	Number of other errors (false alleles, mistypings) divided by the number of loci in all data (Broquet & Petit 2004)
Other error rate per all alleles	0,006098	Number of other errors (false alleles, mistypings) divided by the number of alleles in all data (Bonin et al. 2004)
Loci total count	374	Number of loci
Loci total count for error calculation	244	Loci in replicates
Loci error count	12	Amount of erroneous loci (containing at least one mismatch)
Loci error	0,04918	Locus error rate in all data, loci error cont divided by loci total count for error rate (12/244) (Pompanon et al. 2005)

SAMPLES	Allele drop	Allele drop	Other error	Other error	Total error	Total error	Total allele	Heterozygo	Homozygot	Allele dropc	Allele dropc	Other error	Amount of	Amount of	Allele dropout and other error rates and amounts per sample
Sample (AF86, AF87, AF89, AF90, AF92, AF99, AF102, AF104)	1	0,007353	1	0,007353	2	0,014706	136	36	32	0,027778	0,014706	0,014706	68	2	
Sample (AF88) (Single)															
Sample (AF91, AF93, AF95, AF97, AF100, AF101, AF106, AF107, A	2	0,017544	1	0,008772	3	0,026316	114	33	24	0,060606	0,04	0,02	50	3	
Sample (AF94, AF103)	0	0	0	0	0	0	34	12	5	0	0	0	10	0	

Fig. 3 Explanation of the output file.

References:

Bonin, A., Bellemain, E., Bronken Eidesen, P., Pompanon, F., Brochmann, C., & Taberlet, P. (2004). How to track and assess genotyping errors in population genetics studies. *Molecular Ecology, 13*(11), 3261-3273.

Broquet, T., & Petit, E. (2004). Quantifying genotyping errors in noninvasive population genetics. *Molecular Ecology, 13*(11), 3601-3608.

Callen, D. F., Thompson, A. D., Shen, Y., Phillips, H. A., Richards, R. I., Mulley, J. C., & Sutherland, G. R. (1993). Incidence and origin of " null" alleles in the (AC) n microsatellite markers. *American Journal of Human Genetics, 52*(5), 922.

Chapuis, M. P., & Estoup, A. (2007). Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution, 24*(3), 621-631.

Creel, S., Spong, G., Sands, J. L., Rotella, J., Zeigle, J., Joe, L., ... & Smith, D. (2003). Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. *Molecular Ecology, 12*(7), 2003-2009.

Hoffman, J. I., & Amos, W. (2005). Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology, 14*(2), 599-612.

Johnson, P. C., & Haydon, D. T. (2007). Maximum-likelihood estimation of allelic dropout and false allele error rates from microsatellite genotypes in the absence of reference data. *Genetics, 175*(2), 827-842.

Jones, O. R., & Wang, J. (2010). COLONY: a program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources, 10*(3), 551-555.

Kalinowski, S. T., Taper, M. L., & Marshall, T. C. (2007). Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology, 16*(5), 1099-1106.

Pompanon, F., Bonin, A., Bellemain, E., & Taberlet, P. (2005). Genotyping errors: causes, consequences and solutions. Nature Reviews Genetics, 6(11), 847-859.

Taberlet, P., Griffin, S., Goossens, B., Questiau, S., Manceau, V., Escaravage, N., ... & Bouvet, J. (1996). Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research, 24*(16), 3189-3194.

Taberlet, P., & Luikart, G. (1999). Non-invasive genetic sampling and individual identification. *Biological Journal of the Linnean Society, 68*(1-2), 41-55.

Taberlet, P., Waits, L. P., & Luikart, G. (1999). Noninvasive genetic sampling: look before you leap. *Trends in Ecology & Evolution, 14*(8), 323-327.

Valière, N. (2002). GIMLET: a computer program for analysing genetic individual identification data. *Molecular Ecology Notes, 2*(3), 377-379.

Van Oosterhout, C., Hutchinson, W. F., Wills, D. P., & Shipley, P. (2004). MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes, 4*(3), 535-538.