

Paulo A. Otto
Departamento de Genética e Biologia Evolutiva
Instituto de Biociências IB
Universidade de São Paulo USP
Rua do Matão 277
05508-090 São Paulo SP Brazil
otto@usp.br , otto@ib.usp.br

PROGRAM_PACK (POPULATION GENETICS, BIostatISTICS AND BASIC APPLIED MATHEMATICS WITH THE COMPUTER)

**A COLLECTION OF EXECUTABLE PROGRAMS /
SCRIPTS WRITTEN AND COMPILED IN LIBERTY
BASIC (© ShopTalk Systems), MATHEMATICA (©
Wolfram Res.), QB/QB45 (© Microsoft
Inc./QB64Team), R (© R Foundation for Statistical
Computing), AND VISUAL BASIC (© Microsoft Inc.)**

Artigo com link de acesso / Article with access link

<https://genoma.ib.usp.br/files/upload/261/programpack.pdf>

na seção / at the section Materiais Didáticos / Educação e
Difusão , Genoma IBUSB (<https://genoma.ib.usp.br>)

POPULATION GENETICS, BIOSTATISTICS AND BASIC APPLIED
MATHEMATICS WITH THE COMPUTER

Paulo A. Otto
Department of Genetics and Evolutionary Biology
Instituto de Biociências IB
Universidade de São Paulo USP
Rua do Matão 277 Room 344
05508-090 São Paulo SP Brazil
e-mail otto@usp.br , otto@ib.usp.br

This article describes a large set of standalone executable programs on population genetics, statistics and mathematics. The software is available free of charge at the site <https://sites.google.com/view/otto-usp/program-pack> .

Activities involved in the development of genetics projects with students and in collaborative studies with colleagues at the University of São Paulo over the years resulted in the production of a large number of programs written in various languages. The listings of many QBasic (Microsoft Inc.) and Mathematica (Wolfram Res.) programs were distributed to students attending graduate and post-graduate courses delivered at USP or at the Genetics Department of the National University of Misiones UNAM, Posadas, Misiones, Argentina). The material used at UNAM was rewritten/adapted as two e-books edited by Editorial Universitaria, UNAM (Otto, 2008a, 2008b).

Recently the author decided to make the programs available to a wider audience of users and during the last three years he worked in transforming some of the original QBasic programs into executable software by using the QB64 compiler (from Rob Galleon, QB64Team) downloaded from the internet; he also adapted and standardized practically all his Liberty Basic (ShopTalk Systems) programs (most of them rewritten from QBasic font listings) and compiled them to create their corresponding executable forms. Some other programs were rewritten (from original Mathematica and QBasic font listings) using the R language. The result of this project was the organization/production of a large program library together with PDF documents describing them, listing the source (non-compiled) programs and illustrating their use with suggested pre-recorded data or sets of sample data retrieved from data files, all with the corresponding printed results. Despite the fact that some of the programs included

here were rewritten/adapted from relatively old material, all software in this pack deals with standard procedures still in current use. The pack's contents should interest primarily students and professionals specialized in the teaching and research of population genetics. Some programs, however, are not capable of dealing with complex situations such as the analysis of very large data bank sets (a common situation in modern molecular genetics), which require the use of professionally designed, specialized software, obtained/loaded from the Internet or acquired from commercial sources.

The result of this project was the organization and production of a program library consisting of over 600 executable programs contained in five different folders together with PDF documents describing them and a variety of archives, listing the source (non-compiled) programs and illustrating their use with suggested pre-recorded data or sets of sample data retrieved from data files, all with the corresponding printed results; together, more than 5000 archives are housed in these folders.

The five folders just mentioned are identified by the names LIBERTY_BASIC, QBASIC_QB64, R, MATHEMATICA, and VISUAL_BASIC, which are the proprietary names of the companies ShopTalk Systems, Microsoft Inc./QB64Team, R Foundation for Statistical Computing, Wolfram Res. and Microsoft Inc. respectively.

Most programs contained in the folders MATHEMATICA and VISUAL_BASIC run with restrictions, working adequately only on (very) old Windows versions (3.1 and 95) and just a few on Windows XP. One solution to get them running into newer Windows versions is to obtain a program (such as the excellent software Virtual Box from the www) that creates a virtual partition into the computer, where older Windows versions can be installed in order to run the above-mentioned programs. Details on the contents of the folders MATHEMATICA and VISUAL_BASIC are described in the Program Pack page of the author's site and in the pdf files MATHEMATICA.PDF and VISUALBASIC.PDF. Programs developed with the Mathematica package (version 2.1 was used throughout) run perfectly only on older versions of Windows (3.1 and 95), using the MathReader software (Wolfram Inc.), which allows the user to run them without modifying or editing them. Due to the restrictions above, the listings corresponding to the programs in MA and MB files of the Mathematica v.2.1 package

are also presented in DOC format. The Visual Basic programs work well up to the XP version of Windows, with the exception of the programs StatDist and PenCalc included, which run perfectly on newer versions of Windows (both were tested on the Windows 7 version). In the lines that follow only the material contained in the three main folders (LIBERTY_BASIC, QBASIC_QB64, and R) is described in detail.

(1) LIBERTY_BASIC FOLDER

All programs contained in the subfolder POPULATION_GENETICS are applicable in the areas of theoretical population genetics or in the analysis of data collected from diploid species of organisms, with some emphasis on human and medical genetics. All STATISTICS programs are from the area of biostatistics, most of them applied to solving problems in the areas of formal, quantitative and population genetics, with an emphasis on the most common tests of parametric and non-parametric statistics, such as correlation/regression analysis, variance analysis and contingency table analysis. Most MATHEMATICS programs deal with basic methods commonly used in statistics and population genetics (with an emphasis on numerical analysis) or with standard miscellaneous procedures (calendar, string handling etc.). All programs can be executed directly from their folder location chosen by the user on his/her computer. DAT data files from Liberty Basic and Qbasic QB64, as well as csv data files from R programs must be transferred to the TEMP folder in the C (root) directory of the user's computer, location in which programs requiring recorded sample data will look for the files. The LIBERTY_BASIC folder contains five subfolders: (1) DATA_FILES (57 sample data files to be used as examples by some mathematics, statistics or population genetics programs). (2) LIBRARY_AUXILIARY_FILES (eight auxiliary and library files used in conjunction with executable programs). (3) MATHEMATICS (41 folders with a total of 581 bas, doc, txt, pdf, tkn and exe files, in addition to the set of eight auxiliary and library programs). (4) POPULATION_GENETICS (150 folders with a total of 2215 files, in addition to the set of eight auxiliary and library programs). (5) STATISTICS (96 folders with a total of 1262 files, in addition to the set of auxiliary and library programs). All executable programs compiled from Liberty Basic will work only in the environment where the corresponding TKN files (with the same name as the EXE file) and all eight auxiliary/library files are located; that was precisely the reason for duplicating these files in all folders with EXE programs from the Liberty Basic

language. The folder contains also the pdf archive LIBERTYBASIC.PDF detailing all programs and their font listings and use, with results, printouts and graphs of sample programs. The lines below show the printout of program HW_TESTING_01.EXE, with $D = N(AA)$, $H = N(Aa)$ and $R = N(aa)$ user sample inputs of 16, 45 and 39 respectively; in this list, $d = P(D) = N(AA)/N = 0.16$.

```
PROGRAM FILENAME HW_TESTING_01.EXE
COMPILED FROM    HW_TESTING_01.BAS
HARDY-WEINBERG TESTING [TWO AUTOSOMAL CODOMINANT ALLELES (A,a)]
                    BY MEANS OF LOG-LIKELIHOOD TEST
```

```
D = N(AA) = 16
H = N(Aa) = 45
R = N(aa) = 39
N          = 100
```

```
p = P(A) = (2D+H)/2N      = 0.3850
var1(p)  = pq/2N          = 0.001184
se1(p)   =                = 0.0344
var2(p)  = (p+d-2p^2)/2N = 0.001243
se2(p)   =                = 0.0353
```

MODEL	PARAMETER VALUES			2LOG(Pi)
(1) General	0.1600	0.4500	0.3900	-203.954
(2) Hardy-Weinberg	0.1482	0.4736	0.3782	-204.200

```
c.sq. (1) = 2log(P1/P2) = 2[log(P1)-log(P2)] = 0.247; P = 0.6195
c.sq. (2) = N{(H^2-4DR)/[(2D+H)(H+2R)]}^2 = 0.247; P = 0.6190
```

(2) QBASIC_QB64 FOLDER

The QBASIC_QB64 folder contains six subfolders: (1) DATA_FILES (36 sample DAT files). (2) GRAPHICS (24 folders with a total of 140 files, some with the corresponding images of the DOS-level precision graphics generated by EXE programs). (3) MATHEMATICS (16 folders with a total of 64 files). (4) MISCELLANEOUS (seven folders with a total of 29 files, consisting of a small collection of programs with general applications such as calendars, text manipulation, etc). (5) POPULATION_GENETICS (39 folders with a total of 158 files). (6) STATISTICS (50 folders with a total of 228 files). The QBASIC_QB64 folder contains also a pdf archive (QBASICQB64.PDF) detailing all programs and their listings and use. The lines below show the printout of program HLAHAPL1.EXE, which estimates the HLA system haplotype frequencies $P(A1/B8)$, $P(A1/-)$, $P(-/B8)$ and $P(-/-)$ in a random

population sample of individuals tested with two sera (anti-A1 and anti-B8) with user sample inputs of 376, 235, 91, and 1265:

NO. OF INDIVIDUALS A1(+)/B8(+)	=	376
NO. OF INDIVIDUALS A1(+)/B8(-)	=	235
NO. OF INDIVIDUALS A1(-)/B8(+)	=	91
NO. OF INDIVIDUALS A1(-)/B8(-)	=	1265
NO. OF INDIVIDUALS TESTED	=	1967

ALLELE FREQUENCIES

P(A1) = 0.1697
P(B8) = 0.1267

INFERRED HAPLOTYPE FREQUENCIES

P(A1/B8) = 0.0984
P(A1/ -) = 0.0713
P(- /B8) = 0.0283
P(- / -) = 0.8019

EXPECTED HAPLOTYPE FREQUENCIES

P(A1/B8) = 0.0215
P(A1/ -) = 0.1482
P(- /B8) = 0.1052
P(- / -) = 0.7251

LINKAGE DISEQUILIBRIUM VALUES

D(A1/B8) = 0.0769
D(A1/ -) = -.0769
D(- /B8) = -.0769
D(- / -) = 0.0769

CHISQ1 = $N(N1N4 - N2N3)^2 / [(N1+N2)(N3+N4)(N1+N3)(N2+N4)]$
CHISQ2 = $N(PA1B8 - P1P2)^2 / (P1Q1P2Q2)$

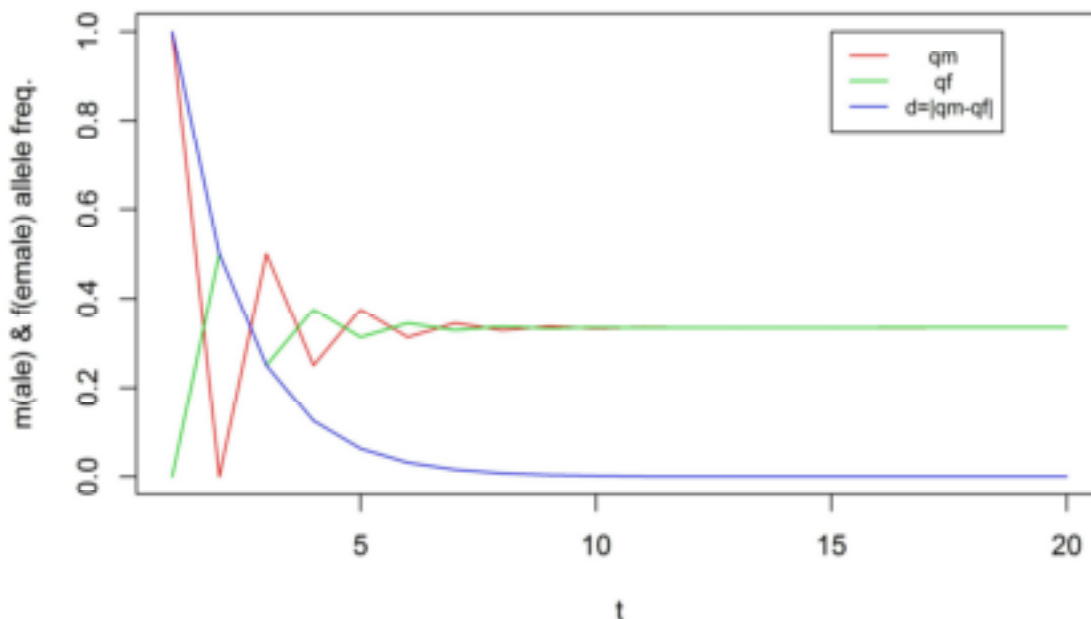
CHISQ1 = 699.350280
CHISQ2 = 745.600440

(3) R FOLDER

The R folder contains four subfolders: (1) DATA_FILES (26 csv files with sample data). (2) MATHEMATICS (18 folders with a total of 72 r, doc, pdf and txt files). (3) POPULATION_GENETICS (49 folders with a total of 214 files); (4) STATISTICS (64 folders with a total of 256 files). As in previous cases, the main folder contains also a pdf archive (R.PDF) detailing all programs, their listings and use. To run programs in R, the user must install the software and its specific routine packages on their computer, as well as the RStudio editor console, which can be obtained free of charge from the www. On the network itself, the user will find practical and simplified instructions for installing and using the Rstudio console (which should be used instead of

the company's R_3.5.1 or similar application). The lines below show, as an example, the font listing (script) and the printout of program sex_linked_1.r, which shows the evolution of x-linked bi-allelic frequencies among males (m) and females (f), over 20 discrete generations of random mating:

```
> # sex_linked_1.r
> # evolution of x-linked gene frequencies among
> #       males (m) and females (f)
> m <- c(1:20)
> f <- c(1:20)
> d <- c(1:20)
> # input initial values m0, f0 and d0 = |m0-f0| on line 8 below
> m[1] <- 1; f[1] <- 0; d[1] <- 1
> for (i in 2:20) {
+   m[i] <- f[i-1]
+   f[i] <- (f[i-1]+m[i-1])/2
+   d[i] <- abs(m[i]-f[i])
+ }
> plot (m, type = "l", col = 2, ylim=c(0,1), xlab = "t",
+       ylab = "m(ale) & f(emale) allele freq."
+       )
> lines (f, type = "l", col = 3)
> lines (d, type = "l", col = 4)
>
> legend(15, 1, cex = .75,
+       c("   qm", "   qf", "d=|qm-qf|"),
+       col = c(2,3,4), lty = 1)
```



INSTRUCTIONS FOR USING THE PROGRAMS

To get the programs, the user should access the page https://sites.google.com/view/otto-usp/PROGRAM_PACK from the site <https://sites.google.com/view/otto-usp/home>).

The PROGRAM_PACK directory is organized as follows:

PROGRAM_PACK

LIBERTY_BASIC

DATA_FILES

LIBRARY_AUXILIARY_FILES

MATHEMATICS

POPULATION_GENETICS

STATISTICS

LIBERTYBASIC.PDF

QBASIC_QB64

DATA_FILES

GRAPHICS

MATHEMATICS

MISCELLANEOUS

POPULATION_GENETICS

STATISTICS

QBASICQB64.PDF

R

DATA_FILES

MATHEMATICS

POPULATION_GENETICS

STATISTICS

R.PDF

MATHEMATICA

GRAPHICS

MATHEMATICS

MATHREAD

POPULATION_GENETICS

STATISTICS

MATHEMATICA.PDF

VISUAL_BASIC

GENETICS

MATHEMATICS

STATISTICS

VISUALBASIC.PDF

The user can get all the programs contained in any of the different sets (LIBERTY_BASIC, QBASIC_QB64, etc) or all the programs contained in the subfolders DATA_FILES, MATHEMATICS, POPULATION_GENETICS etc. To run the chosen programs, double-click the archives with the EXE extension

(in case of LIBERTY_BASIC and QBASIC_QB64 files) or r (in case of R files, after uploading the necessary software and the suggested console from the www). All other files contained in the folders are font listings or auxiliary/library/documentary archives that work only with the executable programs or the corresponding commercial software (QBASIC, Liberty Basic, QB64) properly installed in the user's computer.

All programs have been tested and work perfectly, respecting the restrictions detailed in the user instructions in the PDF documents that accompany them, which must be read carefully. As this is a material with unrestricted free access, the author cannot be held responsible for any problems arising from the use/misuse of the material. When running programs, the user must enter the data requested by the computer carefully and accurately, since most programs do not contain protections against common typing errors, included in commercial programs made by firms or professional programmers (which is not the case of the author, despite his life-long romance with computers). When an error occurs, the program (generally/hopefully) will simply stop running. Users are asked to report any difficulties they encounter or errors that occur while installing or running the programs. The use of the material in the folder PROGRAM_PACK should contain, when applicable, a reference to its author: P. A. Otto, Department of Genetics and Evolutionary Biology (Biosciences Institute, University of São Paulo, Brazil).

ACKNOWLEDGMENTS

Some mathematical programs (especially those dealing with numerical methods and miscellaneous applications) were partly adapted from/inspired by the excellent BASIC software font listings in the books by Meck (1984), Poole et al. (1982), and Schmidt and Stickler (1989). The many stipends and scholarships received from FAPESP, CNPq and CAPES (all from Brazil) granted (to the author and his undergraduate and graduate students) in many occasions during his career at the University of São Paulo are heartily acknowledged. I want to thank Dr. Pérsio Marcondes do Amaral for his help in organizing the pdf archives; Mr. Milton Brasileiro (from the Informatics Service, IB-USP) for his professional expertise and help in organizing all the material into a www site; Mrs. Lilian Dluhosch for her many comments and corrections on English language usage.

REFERENCES

Meck HR. Numerical Analysis. Prentice Hall Inc., Englewood Cliffs, 1984, 205 pp.

Otto PA. Métodos Clásicos y Modernos para el Análisis de Datos en Genética Humana (Classical and Modern Methods in Human Genetics Data Analysis). EdUNaM (Editorial Universitaria de la Universidad Nacional de Misiones, Argentina), Posadas, 2008, 266 pp. e-book (pdf format) available from <https://sites.google/view/otto-usp/population-genetics> or from <http://www.lacygh.com.ar/abajo.htm>

Otto PA. Genética de Poblaciones Humanas (Human Population Genetics). EdUNaM (Editorial Universitaria de la Universidad Nacional de Misiones, Argentina), Posadas, 2008, 214 pp. e-book (pdf format) available from <https://sites.google/view/otto-usp/population-genetics> or from <http://www.lacygh.com.ar/abajo.htm>

Poole L, Borchers M, Burke PM. Some common BASIC programs. McGraw-Hill Inc., New York, 1982, 225 pp.

Schmidt K, Stickler W. Problem, Algorithmus, Program. Ein BASIC-Lehrbuch. Verlag Harri Deutsch, Frankfurt am Main, 1989