

# Tools for Analysis of Population Statistics

By Allan Tereba  
Promega Corporation  
e-mail: [genetic@promega.com](mailto:genetic@promega.com)

*Using this Microsoft® Excel workbook template, genotype data from the landscape view in Hitachi's STaRCall™ software and from the allele table view in Applied Biosystem's Genotyper® software can be pasted directly into the "PowerStats" workbook template to obtain statistics on the distribution of alleles within particular population subsets.*

## INTRODUCTION

Population statistics is an essential element for VNTR- and STR-based forensic and paternity testing. This information provides the basis for determining the probability of paternity or involvement of a suspect in a crime. While race-based population statistics have been calculated for commonly used VNTR and STR loci, regionally-based population statistics are sometimes desired. In addition, complete population statistics are needed for potentially useful new loci.

During the development of the tetranucleotide (1) and pentanucleotide STR systems (2), we needed a simple, quick and reliable method to analyze the race-based population statistics of potentially useful loci. We wanted a process that required minimal data entry and could be performed using either a Macintosh® or PC-compatible desktop computer and standard software.

## OVERVIEW

To this end, we developed a Microsoft® Excel workbook template, which we call "PowerStats". Genotype data from the landscape view in Hitachi's STaRCall™ software and from the allele table view in Applied Biosystem's Genotyper® software can be pasted directly into the "PowerStats" workbook template to obtain population statistics for each population subset. The ability to directly paste the results into this template saves time and eliminates transcription errors. When data is entered the template automatically determines the unique alleles observed, calculates population statistics for each race or subgroup and charts the results in a way that allows for easy identification of heterozygosity and racial comparisons. A summary worksheet displays the relevant information for all race or population sub-

groups. The template is compatible with Office 98 for the Macintosh®, Office 97 for PC-compatible computers and, with some charting restrictions, Excel 5/Office 95. Because there are a large number of calculations, computers with slow processors may take some time to display the results. Allocation of more memory for Excel may also be necessary. We provide free access to this template via our web site at: [www.promega.com/geneticidtools/](http://www.promega.com/geneticidtools/).

## TEMPLATE DETAILS

The "PowerStats" workbook template contains several worksheets. The "Instruction" worksheet (1) provides brief instructions for using the template, 2) summarizes the steps that are performed automatically, 3) lists some limitations of the template, and 4) gives the formulas and references used to derive the population statistics.

The genotype data for a locus is entered into the "Genotype" worksheet (Figure 1). This worksheet is divided into five race or population-specific sets, each with four columns. The set names can be changed if desired and will automatically change in the other worksheets. However, the worksheet tabs must be changed manually to correspond to the new names. The first column is for sample name and is optional. The second and third columns are for the two alleles found in each sample (either allele may be entered first) and are where data is pasted from STaRCall™ or Genotyper® spreadsheets, another spreadsheet or entered manually. All samples must have an allele number in both columns. Positive numbers up to 99 and microvariants with one decimal place (e.g., TH01 allele 9.3) are acceptable. Samples containing one allele or an allele with text (e.g., nc or 15.x) will be ignored in the statistical

analyses. Three-allele patterns are not supported. The fourth column automatically arranges the alleles with the largest allele first and separates the numbers with a comma. Data can be updated at any time to allow for correction of miscalled samples or the addition of more samples. Blank rows are allowed. The present template allows a maximum of 600 samples (rows) per population-specific set and a total of 50 different alleles or bins. Finally, the locus name can be entered in cell B1 if desired and will automatically appear on the other worksheets.

The genotype data is linked to six worksheets. The “Calc” worksheet scans and automatically determines the unique alleles found in the data for all race-specific sets. This feature eliminates the potential error of leaving out a rare allele and allows the race-specific allele frequencies to be compared in a composite graph. This worksheet is for calculation purposes only and does not contain any end results or outcome analysis.

The five race-specific worksheets, Caucasian, Black, Asian, Hispanic and Native (Figure 2), use the “Calc” and “Genotype” worksheet values to calculate the population statistics for each race or population subset. This includes the frequency of each allele, percent heterozygosity and homozygosity, polymorphism information content (PIC) (3), probability of match (4,5), power of discrimination (4,5), power of exclusion (5), and typical paternity index (5). The race-specific worksheets give a visual representation of the data and provide the statistics for individual races but are less suited for printing than the “Graph” and “Summary” worksheets.

The “Graph” worksheet summarizes the allele frequency for each of the 5 race- or population-specific groups and automatically plots allele versus frequency as a bar graph (Figure 3). This chart allows easy visualization of allele and race-specific distributions. In Office 98 for the Macintosh® and Office 97 for IBM®-compatible computers, the number of alleles graphed can be adjusted by clicking once on the chart. This places a colored box around the graphed data that is located to the right of the chart. The box can be adjusted by dragging the lower right corner to the appropriate row so that only observed alleles are charted.

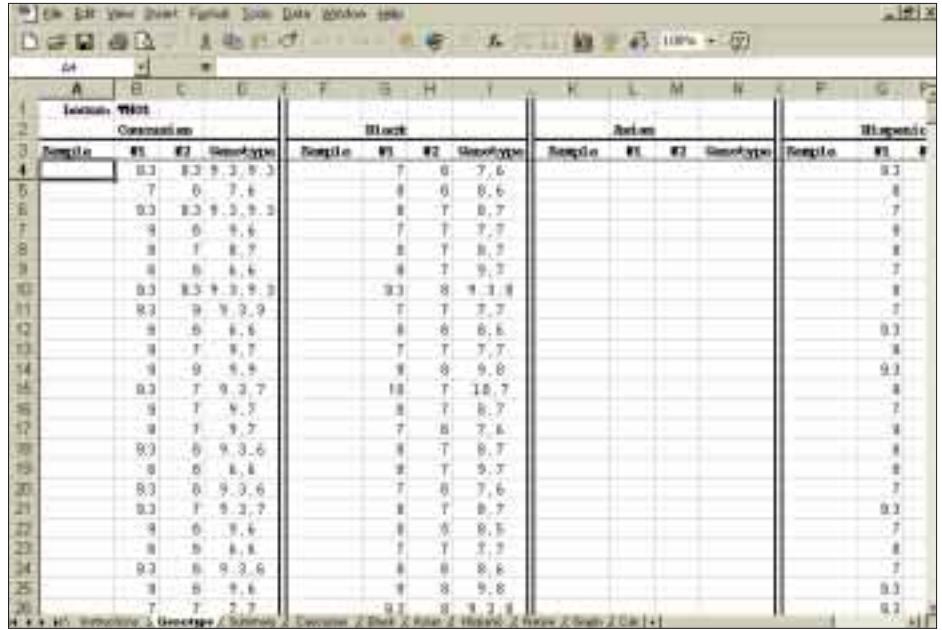


Figure 1. Screen shot of the “Genotype” worksheet with sample data. This worksheet is where all data is entered. Genotype data is entered starting with row 4 under Sample, #1 and #2. Sample name is optional (not included here) and blank rows are allowed. Locus name can be entered in cell B1 and race subsets (row 2) can be changed if desired.

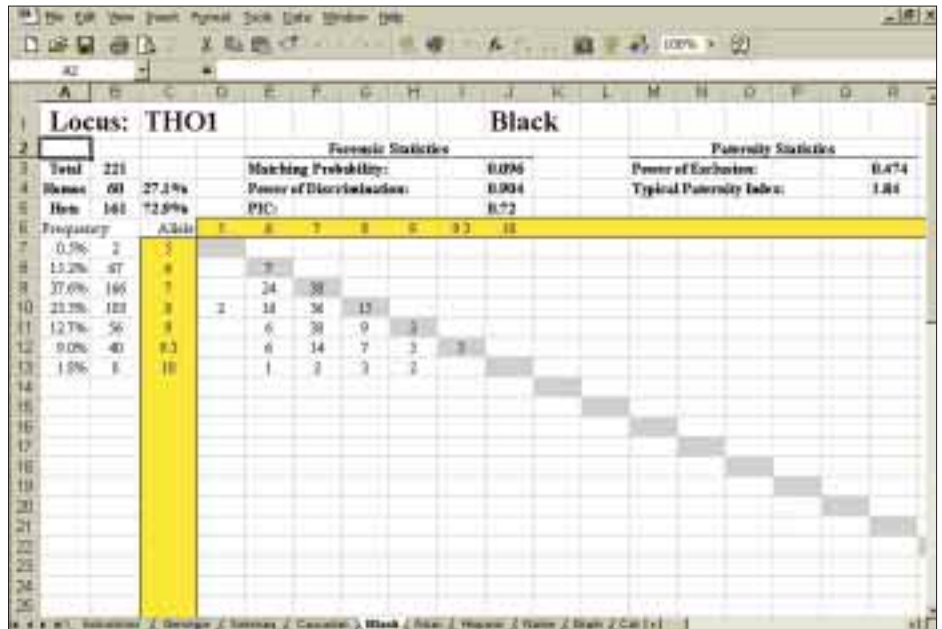


Figure 2. Screen shot of one race-specific worksheet with sample data. This information is also displayed in the “Summary” worksheet.

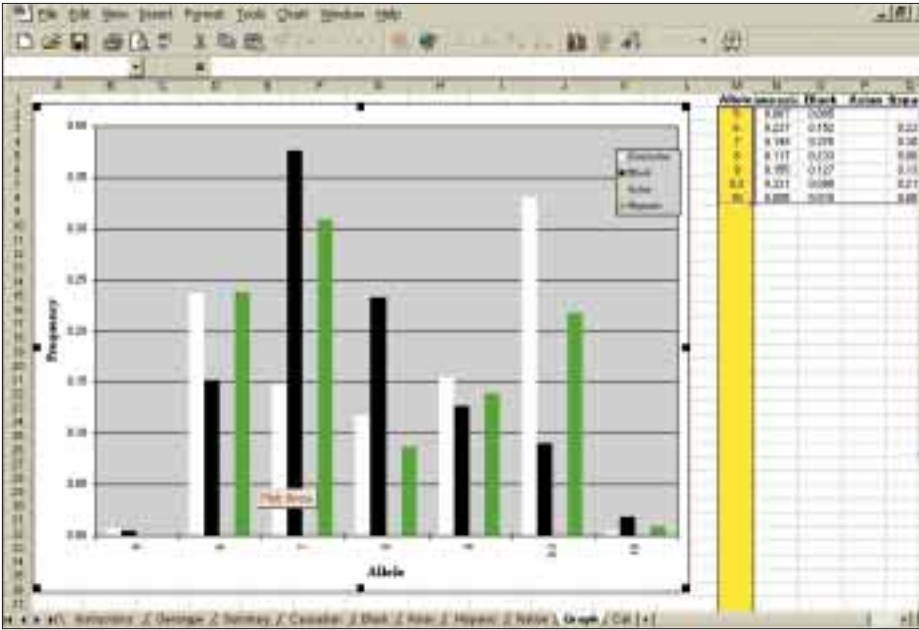


Figure 3. Screen shot of the “Graph” worksheet with sample data. The worksheet displays the race-based allele frequencies as a bar graph. A table of the data, to the right of the graph, can be used to select the data that is graphed in Office 97/98.

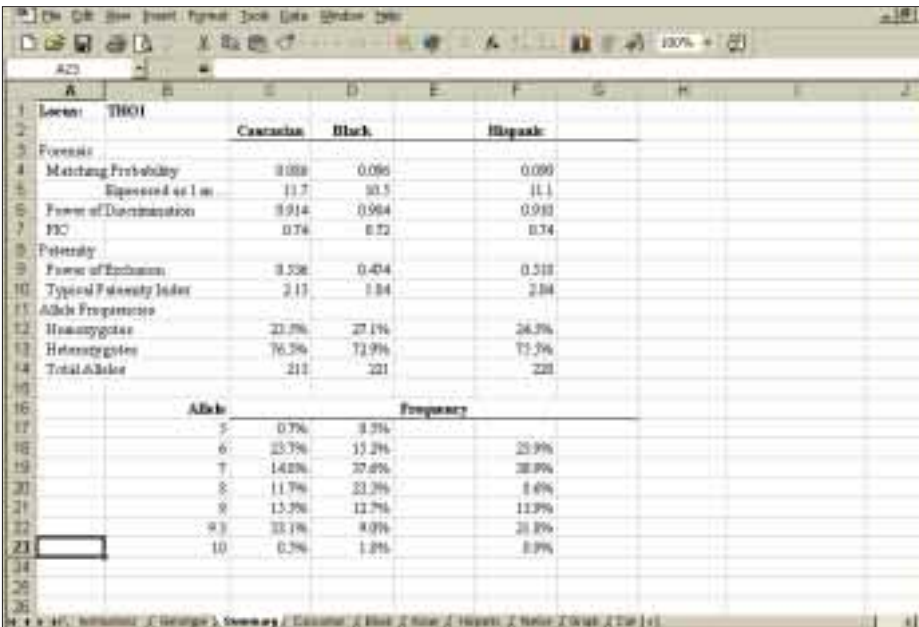


Figure 4. Screen shot of the “Summary” worksheet with sample data. This worksheet provides a summary of all the results for the different race-based subsets.

The “Summary” worksheet compiles all of the statistical data for the 5 different races including frequency values for each allele (Figure 4). For those individuals interested only in the final results, this worksheet provides an easy-to-read summary of the data. As with the other worksheets, this worksheet is write-protected to prevent accidental typing in one of the cells containing a formula. Since there are many links within and between spreadsheets, removing this protection should be done with caution.

This template has been used successfully by our research group and has given identical statistical results when compared with information generated by other means. However, because of the complexity of the template, the variability of data and the variety of computer systems and application versions in use, we are unable to test all configurations. If a problem develops with this template, please contact us at [genetic@promega.com](mailto:genetic@promega.com).

REFERENCES

1. Lins, A. *et al.* (1998) *J. Forensic Sci.* **43**, 1.
2. Bacher, J. and Schumm, J.W. (1998) *Profiles in DNA* **2**(2), 3.
3. Botstein, D., *et al.* (1980) *Am. J. Hum. Genet.* **32**, 314.
4. Jones, D.A. (1972) *J. Forensic Sci. Soc.* **12**, 355.
5. Brenner, C. and Morris, J. (1990) p.21-53 In: *Proceedings for the International Symposium on Human Identification 1989*. Promega Corporation, Madison, WI.

GenePrint and PowerPlex are trademarks of Promega Corporation.

FMBIO is a registered trademark of, and STaRCALL is a trademark of Hitachi Software engineering Company, Ltd. Genotyper is a registered trademark of The Perkin-Elmer Corporation. IBM is a registered trademark of International Business Machines Corporation. Macintosh is a registered trademark of Apple Computer, Inc. Microsoft is a registered trademark of Microsoft Corporation.