

SNP discovery with the *Finch-Server*

Todd M. Smith and Sandra Porter Geospiza, Inc. 3939 Leary Way, NW, Seattle, WA, 98107

Visit us at booth 409
www.geospiza.com

Introduction

Single-nucleotide polymorphisms (SNPs) hold promise in elucidating subtle relationships between genetic variation and human disease. True SNPs are defined as occurring at a frequency of at least 1% in a given population. However, the scale of the human genome, and the cost to validate potential SNPs by genotyping sufficient numbers of individuals, make it more cost effective to scan for SNPs by computational methods which can be later validated by genotyping appropriate populations.

In this project, we planned to test a variation on a commonly used pipeline for SNP discovery. Phred, Phrap and Consed have often been cited by other researchers as tools in a SNP discovery pipeline. Phred serves to differentiate between potential polymorphisms and sequencing errors by assigning a probability to the likelihood of error. Phrap has been used as a tool to assemble sequences and generate alignments that can be viewed in Consed. The ability to view alignments by eye, together with information about sequencing quality, aids in identifying potential SNPs.

We have noted, however, in scanning the literature that one of Phrap's features in SNP discovery has been overlooked. When Phrap assembles reads into a contig, a sequence is produced that's a mosaic of all bases with the highest quality score. Phrap quality scores are calculated from Phred scores and additional information such as read orientation and sequencing chemistry. Phrap also calculates error probabilities for discrepancies between reads that are based on the product of the error probabilities for the two reads. These discrepancy values are not available from Consed and can only be obtained from Phrap output or from the *Finch-Assembly Manager*.

Geospiza's *Finch-Server* provides graphical reports that plot the Phrap discrepancy values vs. the location of the discrepancy within the contig. Also provided for each contig, are the Phrap qualities and the depth of coverage. This information is helpful in visually scanning the results of an assembly and identifying potential SNPs.

Methods

Trace files for CGAP sequences (table below) were obtained from the Washington University Genome Center's web site. Other CGAP trace files and SNP consortium files were unavailable. The trace files were imported into the Finch Chromatogram Manager and processed with Phred. The Finch Assembly Manager was used to assemble the sequences with Phrap into contigs and locate positions with high quality discrepancies. Contig positions with different discrepancy values were compared with SNPs listed in CGAP (<http://cgap.nci.nih.gov>) and dbSNP (<http://www.ncbi.nlm.nih.gov>). Fasta formatted contig sequences were exported from the Finch Server and used to query BLAST dbSNP and generate multiple alignments. These sequences were compared to CGAP sequences in the CGAP SNP viewer. The position of the discrepancy was confirmed by locating the discrepant bases in consed. In some instances differences in alignment spacing between the CGAP viewer and consed made it difficult to compare the location of the SNP with the numerical position in the contig. Although we used many of the same trace files, the parameters for the CGAP assemblies were unavailable and our assemblies produced contigs that differed from CGAP. Contig positions with high quality discrepancies were considered confirmed if the discrepancies could be located with confidence in CGAP or dbSNP.

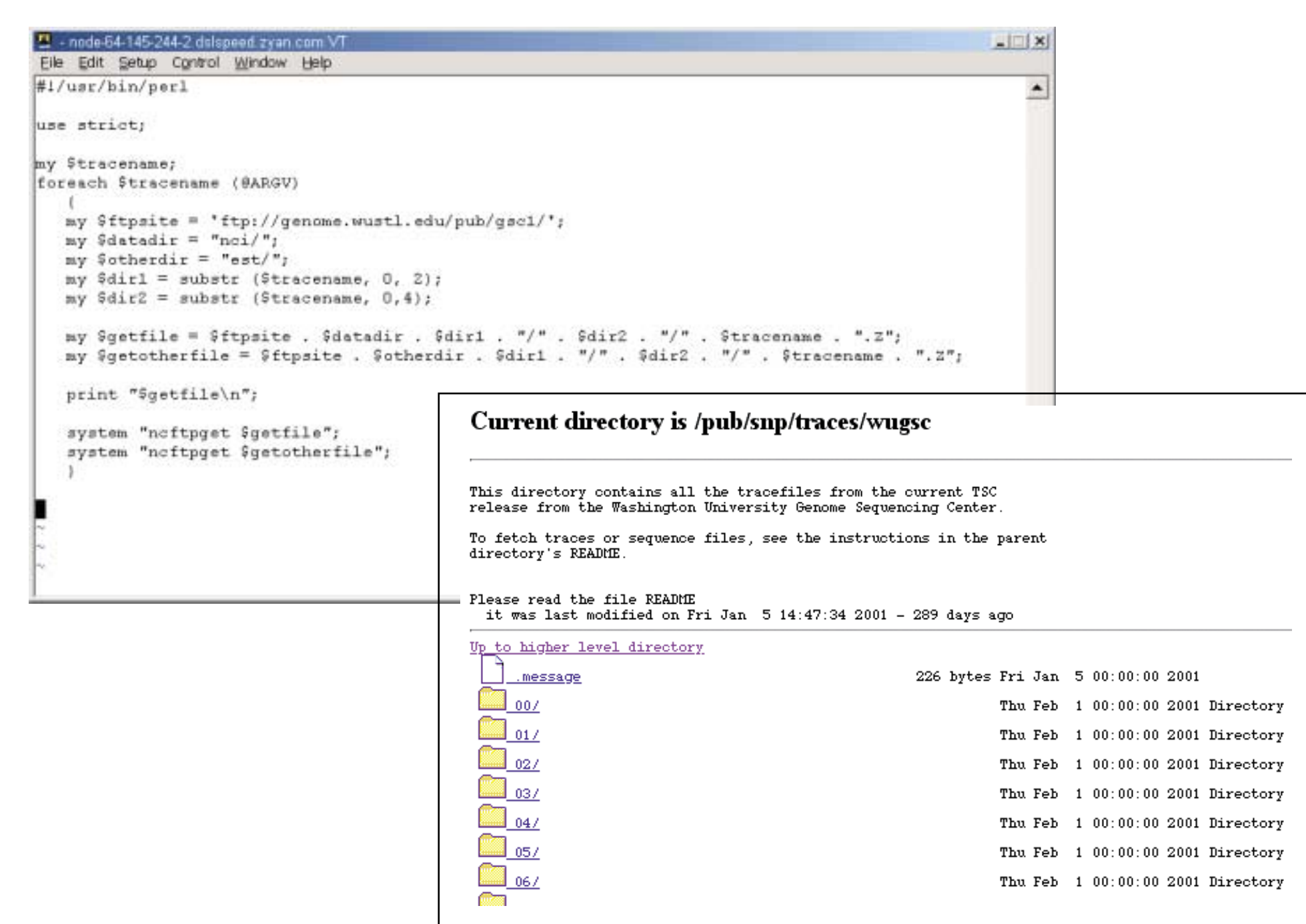
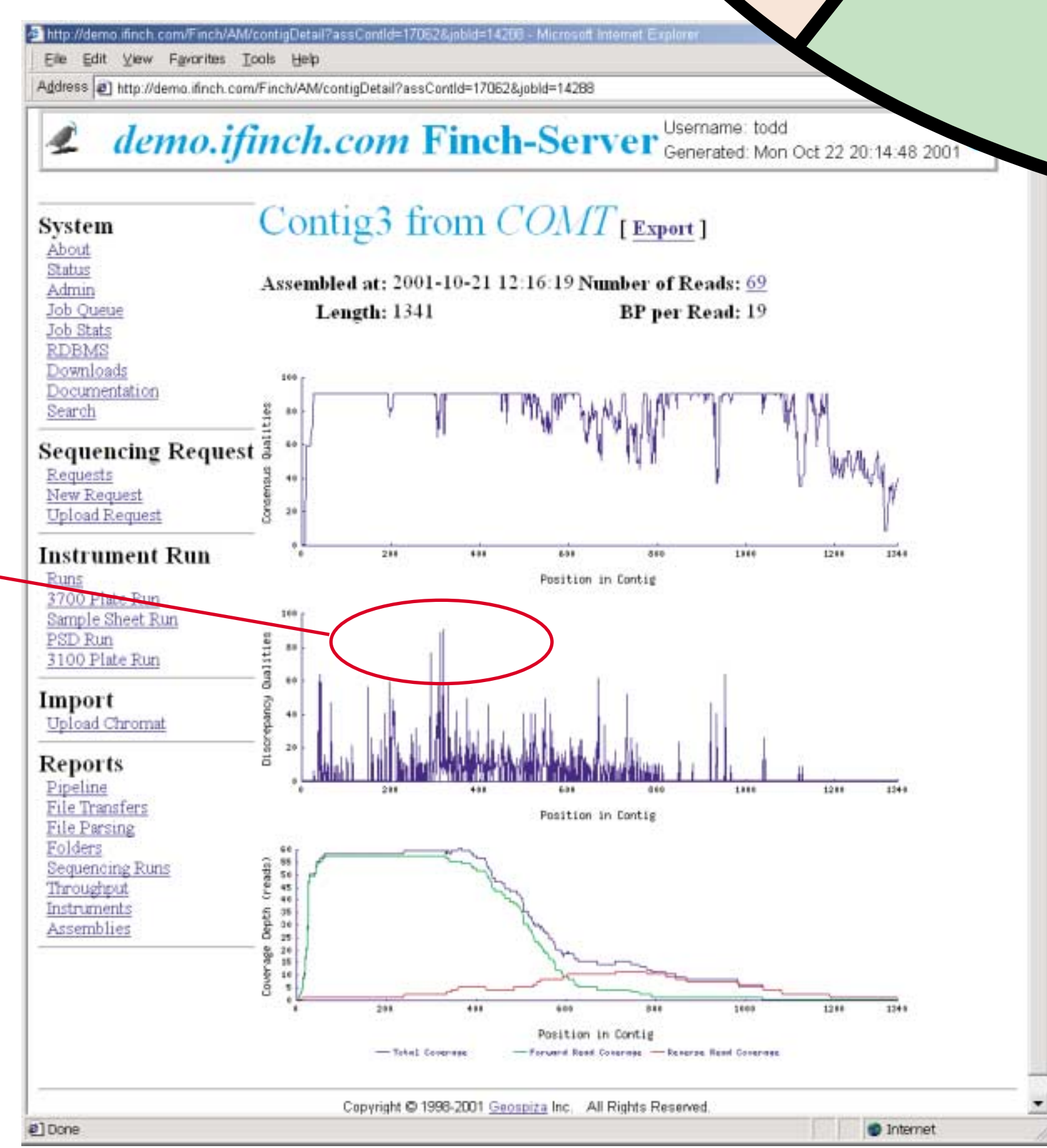
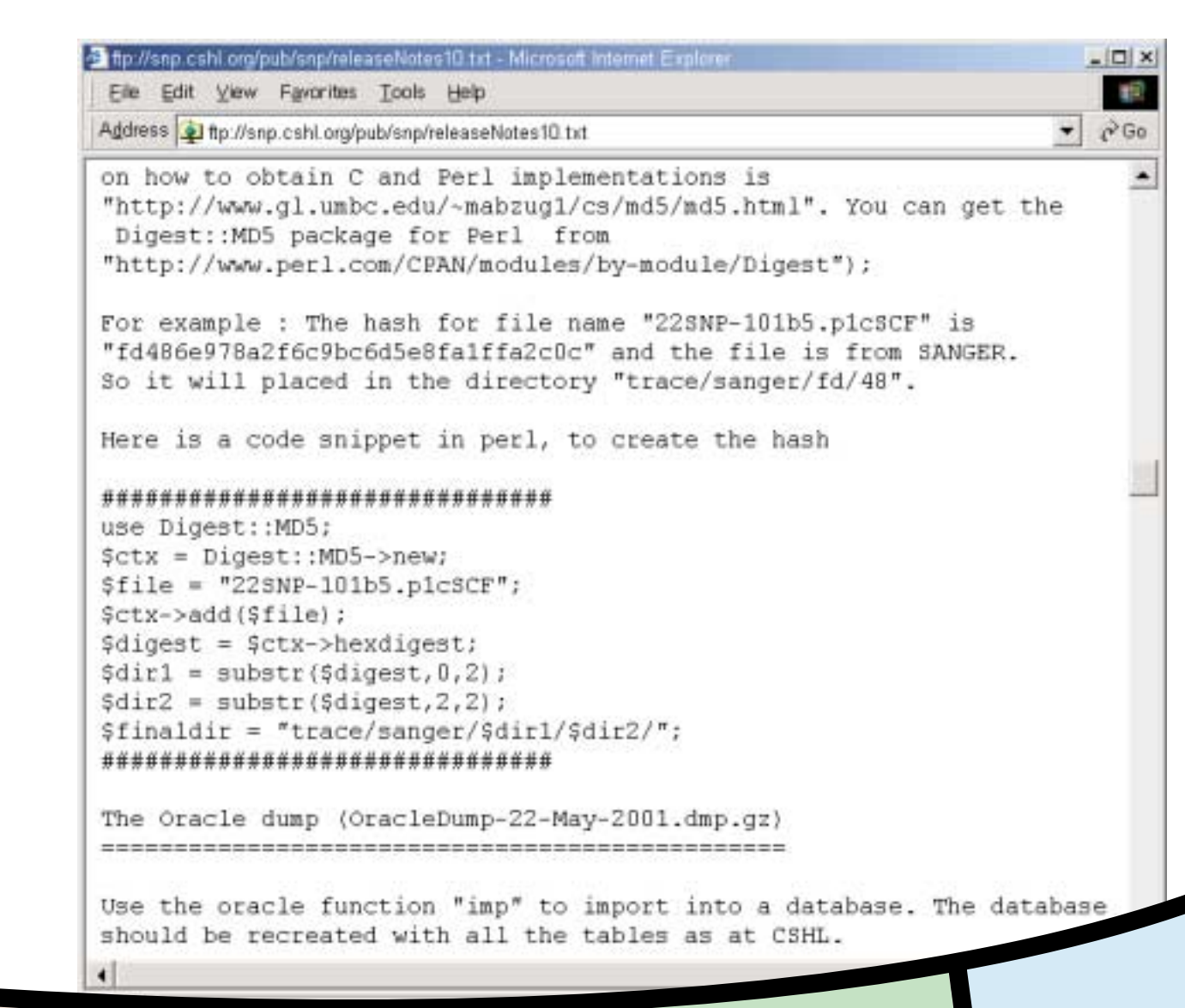
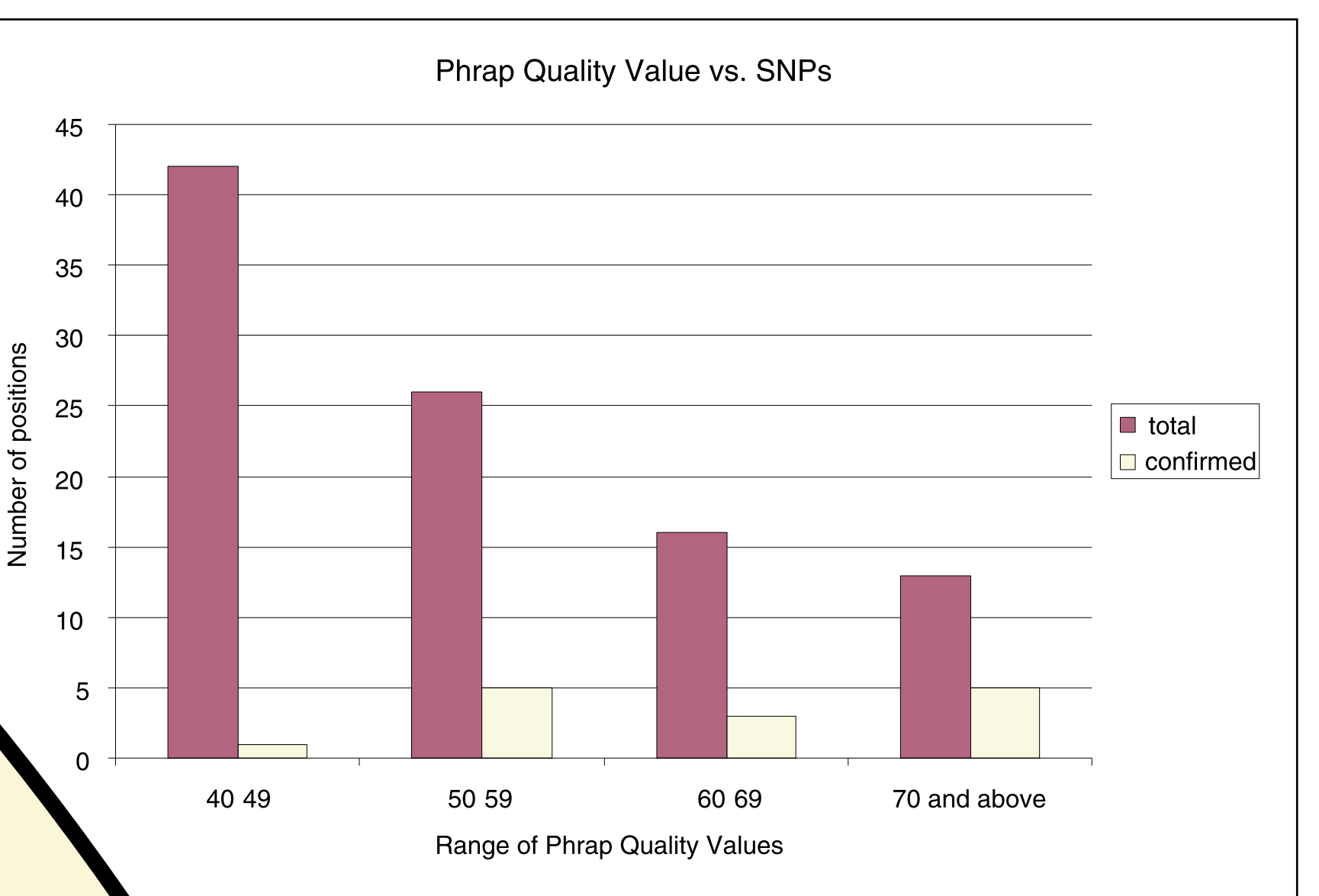
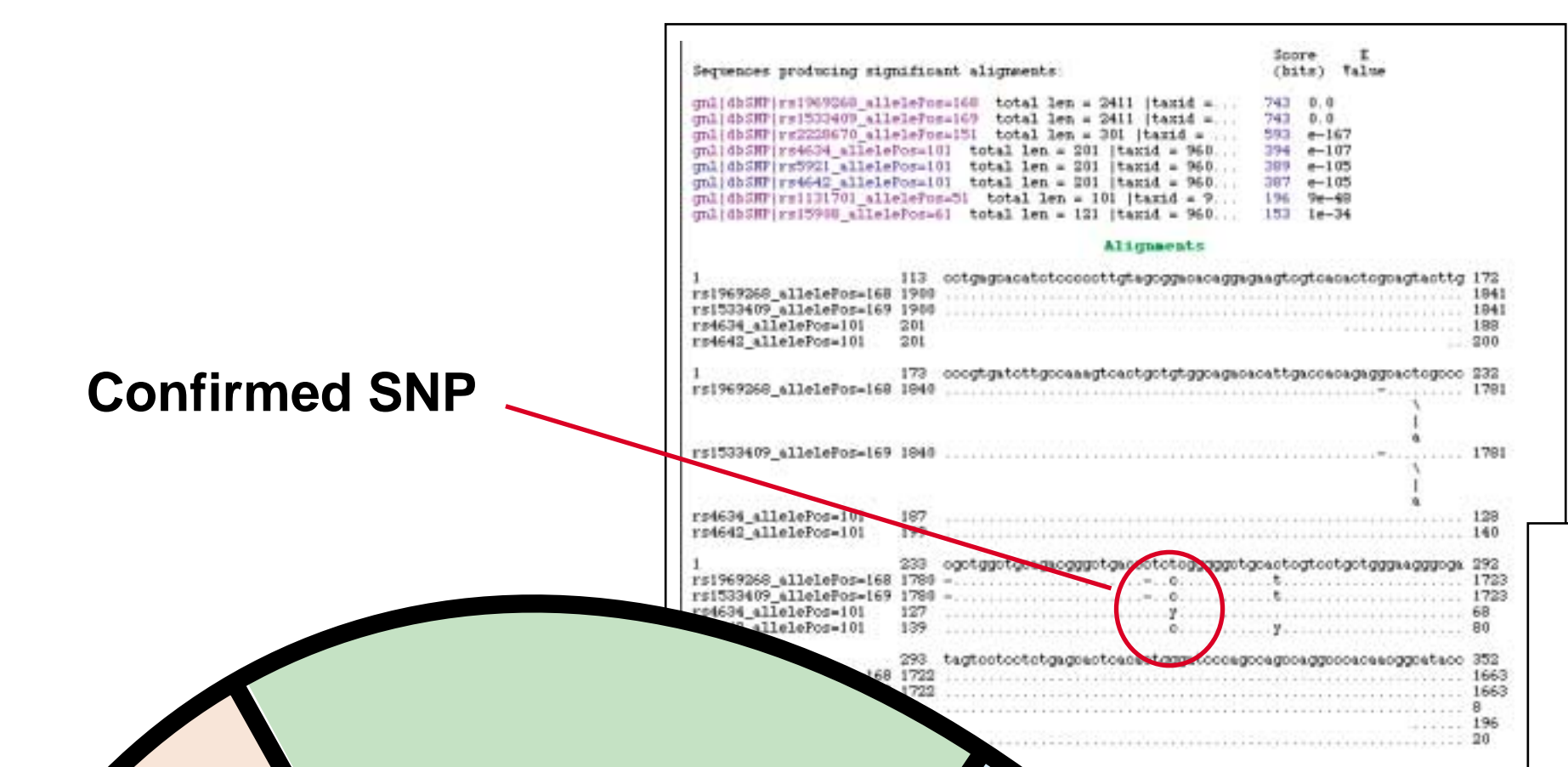
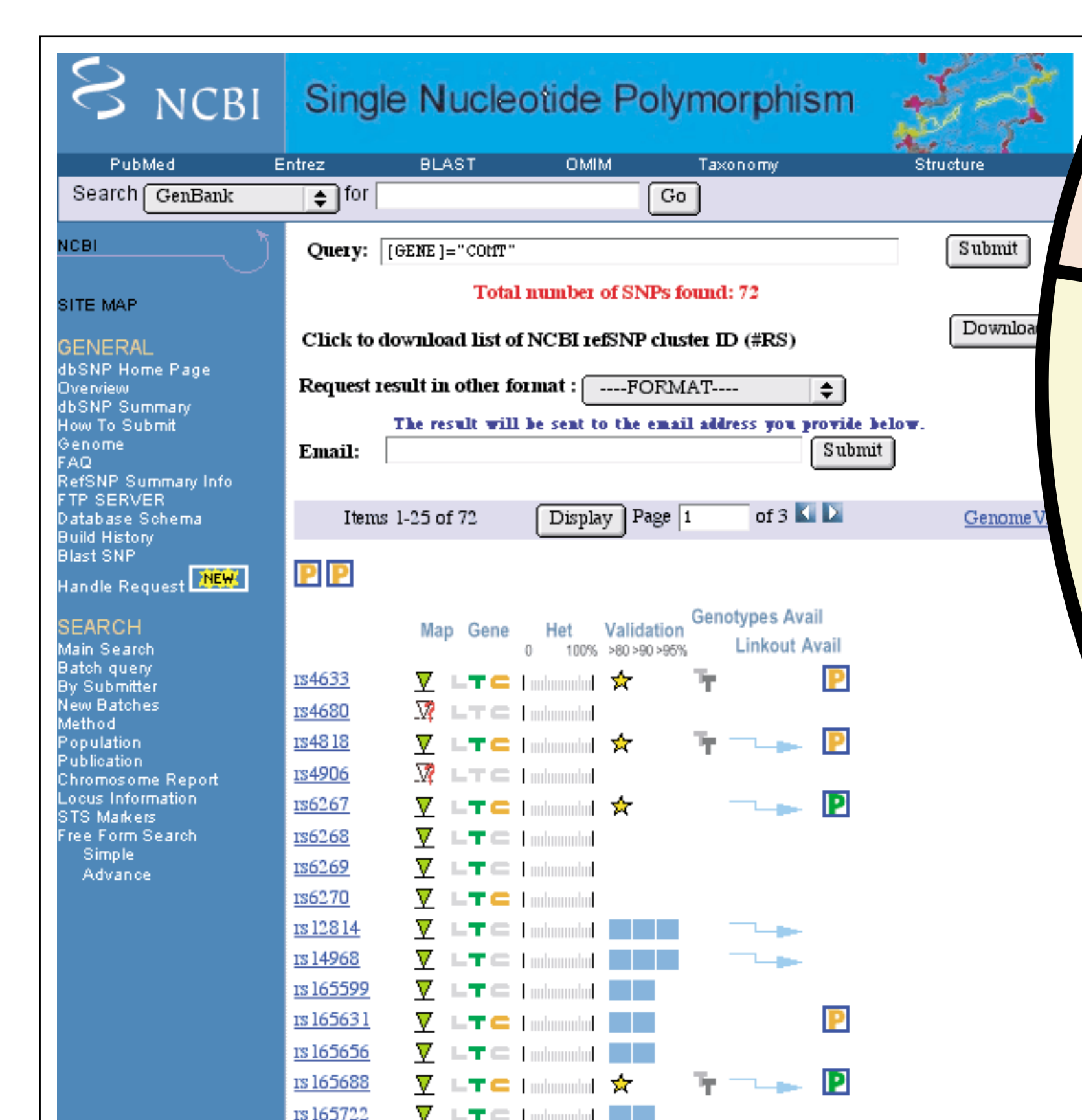
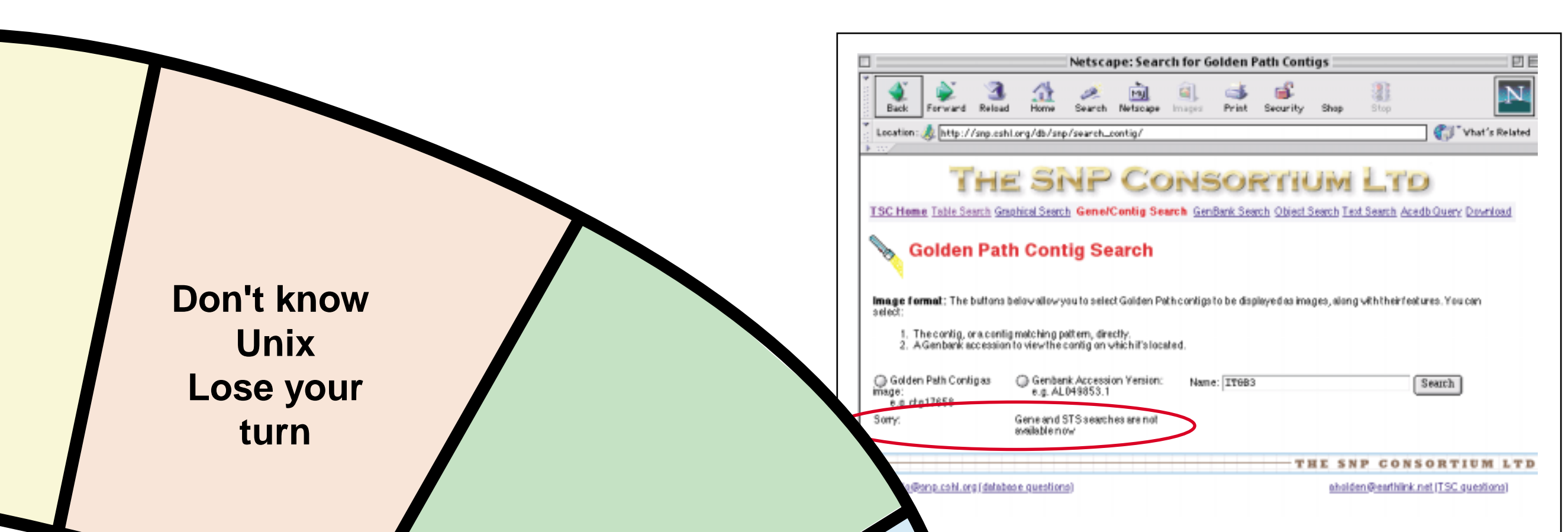
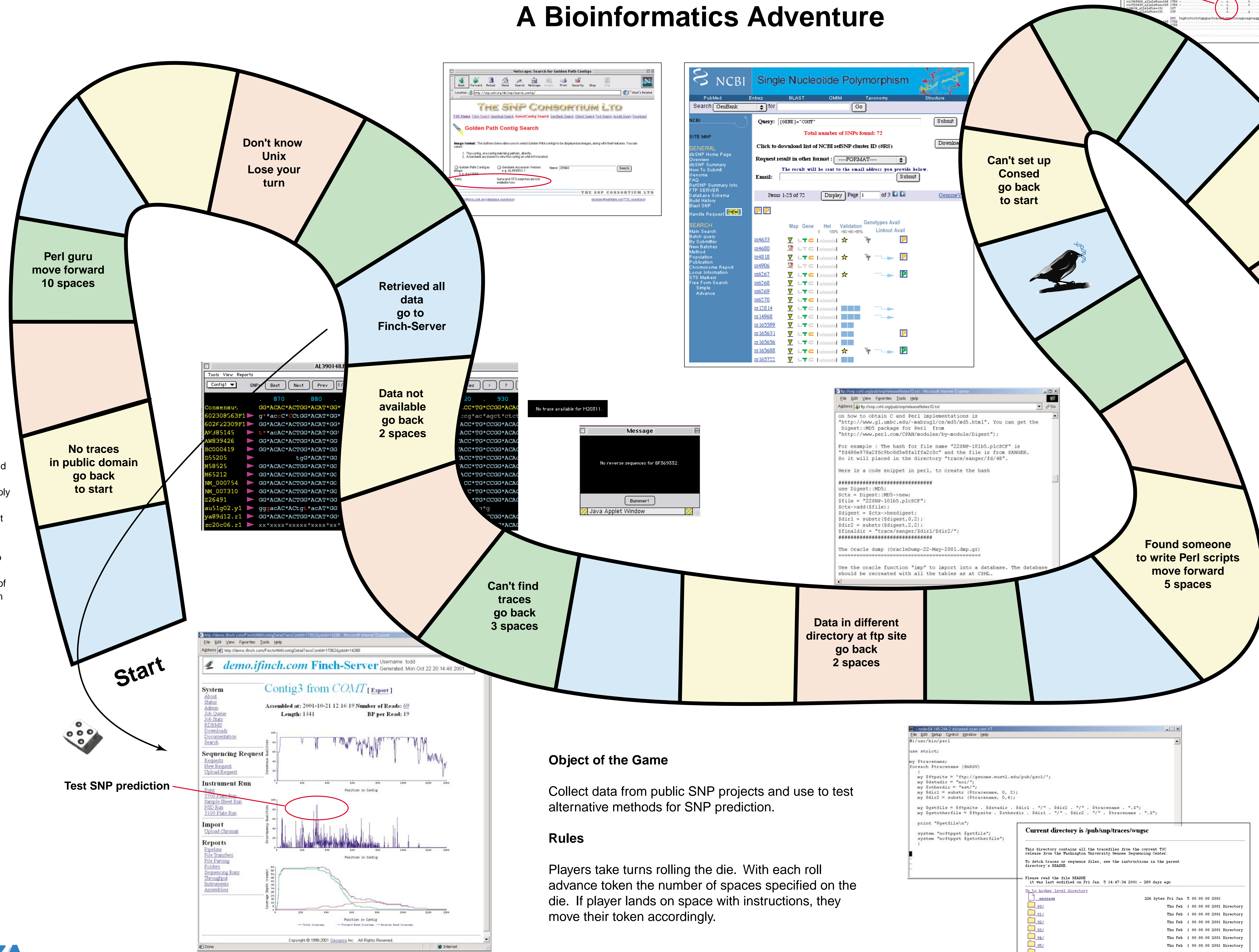
Genes Tested

Gene description	Function	UCLA ^a	CGAP SNPs ^b
COMT	Catechol-o-methyltransferase	3	6
F13A1	Coagulation Factor VIII, A1 peptide	2	2
FGB	Fibrinogen, B beta peptide	5	4
HBB	Hemoglobin beta	6	4
ITGB3	Integrin beta 3	3	3
LIPC	Lipase, hepatic	3	1
NFATC4	nuclear factor of activated T-cells	2	1

a. Izarry, et al. 2000 Nature Genetics 26, pp. 233
b. Buétow, Edmonson, and Cassidy, 1999 Nature Genetics Vol. 21, pp. 323
<http://cgap.nci.nih.gov>



Genome Land A Bioinformatics Adventure



Conclusions

Phrap's discrepancy values provide a simple and rapid method to identify potential SNPs. Five of 12 SNPs predicted for discrepancy values above 70 were confirmed in public SNP resources. Some of the unconfirmed SNPs are potentially real because they were found in both forward and reverse directions or were observed in multiple reads of the same orientation.

Tremendous investments are being made by government and industry to develop SNP resources for the scientific community. Most however, are too user unfriendly to be used as a lasting resource. Problems range from data not being in expected directories at FTP sites, to data missing at sites, to sites not making their data available.

The consequence is that it is difficult to independently validate alternative computational methods or for one group to test another group's observation. Further, undocumented methods and algorithms make it challenging to reproduce results presented at various web-sites and in the scientific literature. Thus, researchers expecting to validate SNP predictions will have to re-collect large amounts of data and will need robust commercially supported software to manage and analyze their data.

References

Altschuler, D., Pollara, V., Cowles, C., Van Etten, W., Baldwin, J., Linton, L., and E. Lander. A SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407: 513-516 (2000).

Buétow, K., Edmonson, M. and A. Cassidy. Reliable identification of large numbers of candidate SNPs from public EST data. *Nature Genetics* 21: 323-325 (1999).

Izarray, K., Kustanovich, V., Li, C., Brown, N., Nelson, S., Wong, W. and C. Lee. Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nature Genetics* 26: 233-236 (2000).

Marth, G., Yeh, R., Minton, M., Donaldson, R., Li, O., Duan, S., Davenport, R., Miller, R., Pu., et al. Single-nucleotide polymorphisms in the public domain: how useful are they? *Nature Genetics* 27: 371 - 372 (2001).

Picoult-Newberg, L., Ideker, T., Polh, M., Taylor, S., Donaldson, M., Nickerson, D., and M. Boyce-Jacino. Mining SNPs from EST databases. *Genome Research* 9: 167-174 (1999).

Taillon-Miller, P., Gu, Z., Li, Q., Hillier, L., and P. Kwok. Overlapping genomic sequences: A treasure trove of single-nucleotide polymorphisms. *Genome Research* 8: 748-754 (1998).

The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409: 298-933 (2001).

Geospiza, Inc., the Finch logo, and Finch, Finch-Server, Finch-Server (patent pending), are registered trademarks of Geospiza Incorporated. All other trademarks, service marks and registered trademarks appearing herein are the properties of their respective owners and are hereby acknowledged. Printed at GSAC08/2001, San Diego CA.