BMC Genomics

BioMed Central

Software

Open Access

GeneLink: a database to facilitate genetic studies of complex traits

Elizabeth M Gillanders[1], Anthony Masiello[2], Derek Gildea[1], Lowell Umayam[2], Priya Duggal[3], Mary Pat Jones[2], Alison P Klein[3], Diana Freas-Lutz[1], Grace Ibay[3], Ken Trout[2], Tyra G Wolfsberg[2], Jeffrey M Trent[4], Joan E Bailey-Wilson[2] and Andreas D Baxevanis*[2]

Address: [1]Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892-8000, USA, [2]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892-8002, USA, [3]Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, MD 21224, USA and [4]Translational Genomics Research Institute, Phoenix, AZ 85004, USA

Email: Elizabeth M Gillanders - lgilland@mail.nih.gov; Anthony Masiello - ant@masiello.org; Derek Gildea - dgildea@mail.nih.gov; Lowell Umayam - umayamla@mail.nih.gov; Priya Duggal - pduggal@nhgri.nih.gov; Mary Pat Jones - mpjones@nhgri.nih.gov; Alison P Klein - aklein@nhgri.nih.gov; Diana Freas-Lutz - dlutz@nhgri.nih.gov; Grace Ibay - gibay@nhgri.nih.gov; Ken Trout - nezumi@fred.net; Tyra G Wolfsberg - tyra@nhgri.nih.gov; Jeffrey M Trent - jtrent@tgen.org; Joan E Bailey-Wilson - jebw@mail.nih.gov; Andreas D Baxevanis* - andy@nhgri.nih.gov

* Corresponding author

Published: 18 October 2004

BMC Genomics 2004, **5**:81 doi:10.1186/1471-2164-5-81

Received: 12 July 2004
Accepted: 18 October 2004

This article is available from: http://www.biomedcentral.com/1471-2164/5/81

Abstract

Background: In contrast to gene-mapping studies of simple Mendelian disorders, genetic analyses of complex traits are far more challenging, and high quality data management systems are often critical to the success of these projects. To minimize the difficulties inherent in complex trait studies, we have developed GeneLink, a Web-accessible, password-protected Sybase database.

Results: GeneLink is a powerful tool for complex trait mapping, enabling genotypic data to be easily merged with pedigree and extensive phenotypic data. Specifically designed to facilitate large-scale (multi-center) genetic linkage or association studies, GeneLink securely and efficiently handles large amounts of data and provides additional features to facilitate data analysis by existing software packages and quality control. These include the ability to download chromosome-specific data files containing marker data in map order in various formats appropriate for downstream analyses (e.g., GAS and LINKAGE). Furthermore, an unlimited number of phenotypes (either qualitative or quantitative) can be stored and analyzed. Finally, GeneLink generates several quality assurance reports, including genotyping success rates of specified DNA samples or success and heterozygosity rates for specified markers.

Conclusions: GeneLink has already proven an invaluable tool for complex trait mapping studies and is discussed primarily in the context of our large, multi-center study of hereditary prostate cancer (HPC). GeneLink is freely available at http://research.nhgri.nih.gov/genelink.

Background

In the past decade, hundreds of genes involved in the etiology of simple Mendelian disorders such as cystic fibrosis and Huntington's disease have been identified [1-3]. The genetic localization of these disorders, primarily through positional cloning approaches, has been highly successful because of the relatively simple model underlying disease pathogenesis. In the majority of these cases, a single mutated disease gene is both necessary and sufficient to cause the observed trait. In contrast, susceptibility to complex traits is heterogeneous, involving both multiple genetic and environmental risk factors, acting either independently or together.

Efforts to identify susceptibility genes involved in complex traits such as cancer, diabetes, hypertension, or Alzheimer's disease are complicated by genetic heterogeneity, incompplete penetrance, phenocopies, and the later age of onset of disease (thus unavailable DNA samples for parents of affected individuals). Each of these factors results in a significant reduction in power for any given study. Therefore, gene-mapping studies of complex traits require high-throughput genotyping performed on large collections of DNA samples using hundreds to thousands of polymorphic markers. The significant amounts of data generated during these genome surveys pose numerous data management challenges. In order to address these challenges, which are inherent in any large, collaborative genotyping study, we have developed a robust, easy-to-use database system named GeneLink.

GeneLink was initially developed to facilitate our studies of genetic susceptibility to prostate cancer, whose aims are to identify novel high- and moderate- penetrance genes involved in hereditary prostate cancer risk. These studies are multi-center collaborative efforts involving researchers from the United States, Finland, and Sweden [4-7]. The project included 496 families containing 5,247 individuals; DNA on 2,374 of these individuals was available for genotyping. We genotyped over 400 microsatellite markers for these individuals generating close to one million genotypes. This number is large but not atypical in gene mapping studies of complex traits. Given the considerable number of genotypes requiring analysis, it was obvious that we needed to develop a database management system that could handle such large quantities of data, as well as address data management issues unique to complex trait genetic analysis.

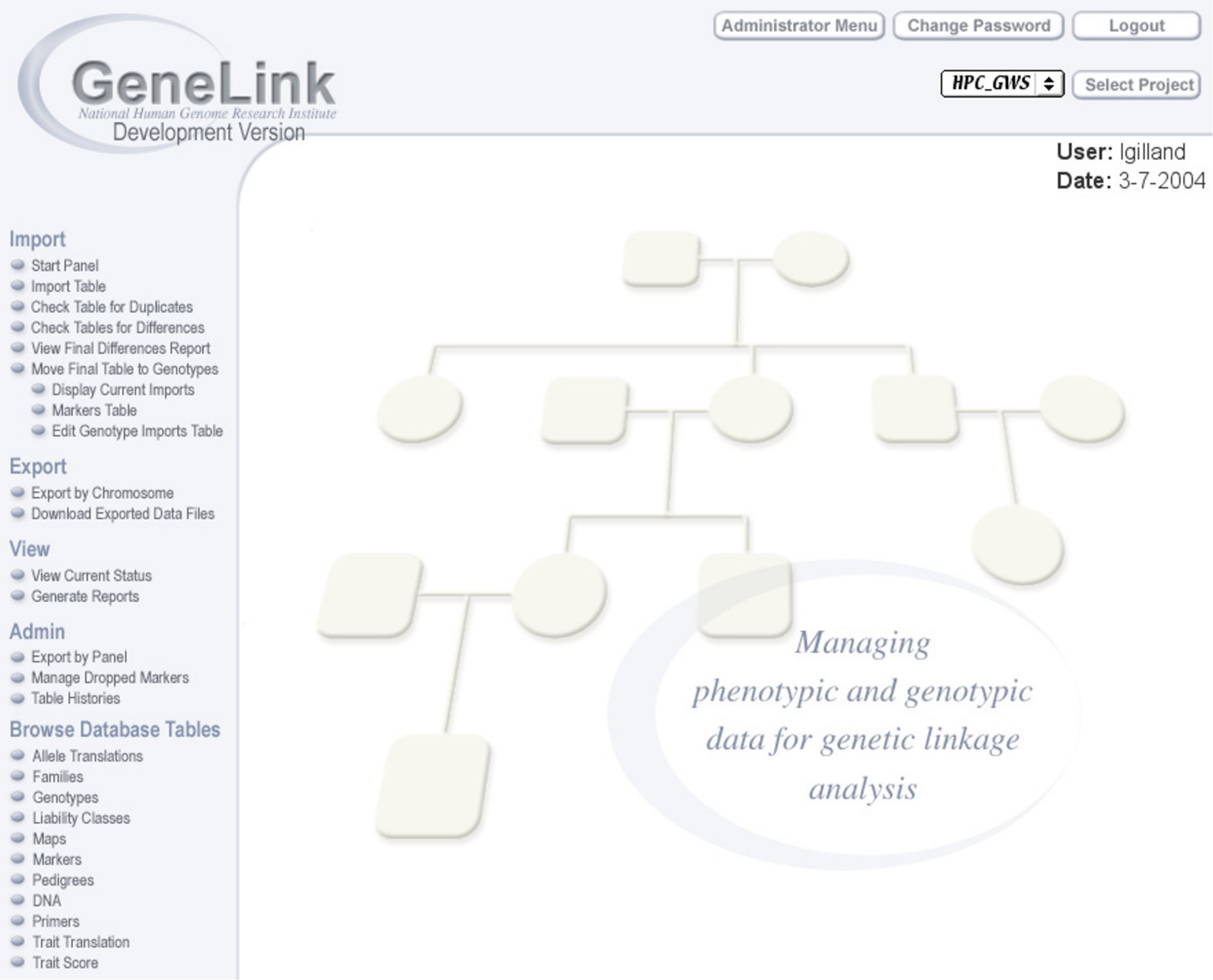
Implementation

GeneLink is a platform-independent, Web-accessible Sybase database that can manage complete genotypic, phenotypic and pedigree data for genetic linkage or association studies. Figure 1 shows the comprehensive GeneLink user's menu available following login. Access to the GeneLink database can be limited using two independent mechanisms. First, users can be granted one or more activity or privilege levels. The *admin* (administrator) privilege provides a user the ability to view data as well as manage access to the data. Users without *admin* privileges can be assigned the following privileges by the administrator: *export*, *import*, *modify* and *view*, with obvious permissions. Project- or user-specific reports summarizing assigned privileges can be generated. Second, as GeneLink provides the ability to associate a "group" (of users) with a collection site-specific set of families, collaborators can be provided the ability to view and/or manipulate all or only a subset of data.

Currently, GeneLink's database uses Sybase SQL server ASE 12.5.1, which runs on a Sun V880 computer running Unix. GeneLink's Web scripts to access the database require Perl version 5.6.1 or greater. The necessary CPAN Perl modules required by GeneLink are DBI, DBD::Sybase, CGI, and Carp. These modules are usually included in standard Perl 5 releases. A Web server such as Apache 1.3.29 is also required to run the GeneLink Web scripts. GeneLink can operate on a Sun Enterprise 6500 or similar machine configured to operate as a Web server.

Database structure

Figure 2 shows the relational design of the GeneLink database in detail. Data within GeneLink is stored in 11 primary tables: **Families, Pedigrees, Trait Score, Traits Translations, Markers, Primers, Maps, Genotypes, Allele Translations, Liability Classes** and **DNA.** Each table can be populated either by importing multiple records from a delimited text file or by inserting a single record at a time through a Web interface. The **Families** table stores pertinent clinical information regarding each family as a whole; each record reflects a single family included in the linkage or association study. For example, in our HPC study, we used the **Families** table to store information regarding evidence of male-to-male transmission of disease as well as the occurrence of other cancer types in the family. The **Pedigrees** table stores one record per individual within a family and contains the biological relationships (*FatherID*, *MotherID*) for each individual. Also included in the **Pedigrees** table is age, sex, whether the individual has been or will be genotyped (*Age*, *Sex*, *DNA*), two ways to store qualitative phenotype information (*StatusBroad*, *StatusNarrow*), and individual liability class information. In our HPC study, we defined prostate cancer affection status in two ways; specifically, we used the *StatusBroad* field to classify individuals as affected, unaffected or unknown, while the *StatusNarrow* definition was used for an affecteds-only analysis in which individuals were coded either as affected or unknown. More extensive trait or covariate information can also be stored in the **Trait Score** table. This table stores an unlimited number of qualitative or quantitative phenotypes as trait 1, trait 2, trait 3, and so forth. Definitions of the traits are stored in the **Traits Translation** table.



User: lgilland
Date: 3-7-2004

Managing phenotypic and genotypic data for genetic linkage analysis

Figure 1
Main menu GeneLink's main menu available following login. GeneLink is a platform-independent, Web-accessible database. Access to GeneLink is password-protected. Here the user (lgilland) has been given *import*, *export*, *view*, and *admin*, permissions.

The **Markers** table stores information regarding all markers typed in a given project, including the panel in which the marker was run. A panel is defined as a group of microsatellite or SNP markers which can be electrophoresed simultaneously by taking advantage of different fluorescent dye labels and varying amplicon sizes. Also stored in the **Markers** table is the allele size range (*ASR*) of the marker, the fluorescent dye used to label the forward primer, and the marker-specific genotype for a CEPH control individual (e.g., CEPH *1347-02* was used in our prostate cancer study). The **Primers** table provides additional information for each specific marker. This information includes UniSTS ID, GenBank accession number, forward primer sequence, reverse primer sequence, and primer purchasing and inventory information. The **Primers** table also contains comprehensive genetic map information (*deCODE*, *Généthon*, and *Marshfield positions*) as well as physical location (*Build*, *physical start position*, and *physical stop position*). The **Maps** table stores the genetic map location of a marker in the genome, as well as the relative order of markers along a chromosome and the distance between adjacent markers. The markers typed thus far in our HPC study are di-, tri-, or tetra- nucleotide repeats; however GeneLink is capable of handling any combination of microsatellite and single nucleotide polymorphisms (SNP) data.

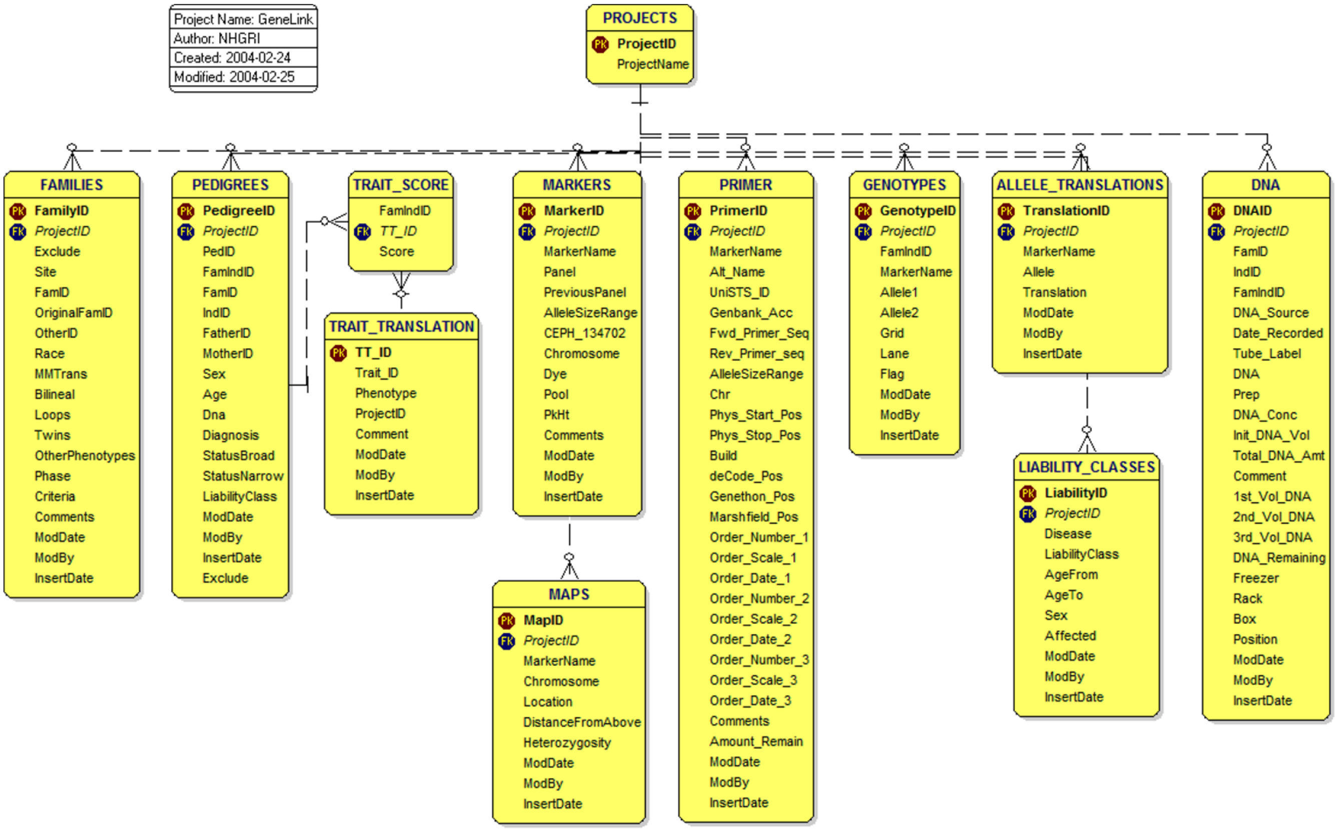
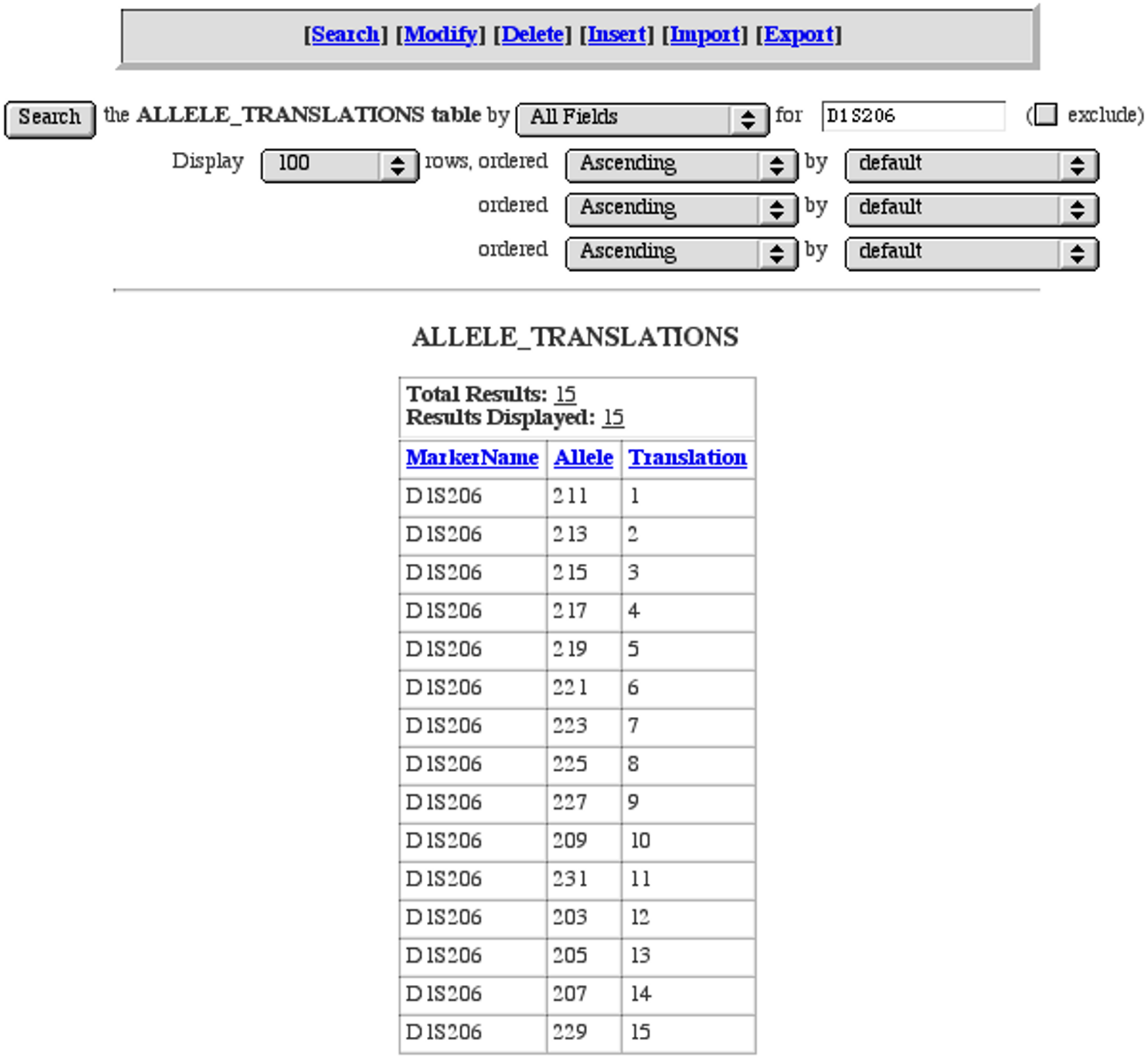


Figure 2

Relational design Relational design of GeneLink's 11 primary tables. Primary keys are indicated in red and foreign keys in blue. GeneLink enables pedigree information stored in the **Families** and **Pedigrees** tables to be easily merged with genotypic data stored in the **Genotypes** table. In the **Pedigrees, Liability Classes, Trait Score**, and **Trait Translation** tables GeneLink also manages extensive phenotypic data. The **Markers** and **Primer** tables store information regarding polymorphisms being genotyped and the **Maps** table stores genetic or physical map information, which determines the order in which data is exported. GeneLink's **Primer** and **DNA** tables provide labs with an easily implemented inventory system.

The **Genotypes** table is where final genotype data are stored. Each record represents an individual/marker combination and alleles are stored as size in base pairs. The **Genotypes** table also holds information regarding the laboratory run (*Grid*, *Lane*). When exporting data in LINKAGE or RelCheck format, GeneLink systematically recodes genotype labels (1, 2, 3,...) to provide properly-formatted data for analysis by these programs. The **Allele Translations** table provides a key or legend linking the newly "translated" genotype to the original size in base-pairs score. Future exports use identical "allele translation codes." New alleles identified after the first export will be added to the end of a marker's allele translation ensuring consistent recoding of alleles across exports (Figure 3). The **Liability Class** table stores information regarding specific liability classes, which can be incorporated in each LINKAGE export for analysis. Liability classes can currently be defined using any combination of age, sex, and affection status (based on *StatusBroad*, or *StatusNarrow* from the pedigrees table). For example, in our HPC study we used five different liability classes (Figure 4). Specifically, we were able to assign older, unaffected men (more likely not to be gene carriers) to a different liability class than younger, unaffected men who may be gene carriers but who have not yet presented with the disease. Finally, the **DNA** table stores information regarding all available DNA samples in the laboratory. This includes specifics regarding date received, concentration, quantity, and storage location.

Each of GeneLink's 11 tables has measures built-in for quality control purposes. First, all changes (import, modification or delete) to GeneLink records are stamped with the date, time, and USER ID of the individual doing the



Search the **ALLELE_TRANSLATIONS table** by All Fields for D1S206 (□ exclude)

Display 100 rows, ordered Ascending by default

ordered Ascending by default

ordered Ascending by default

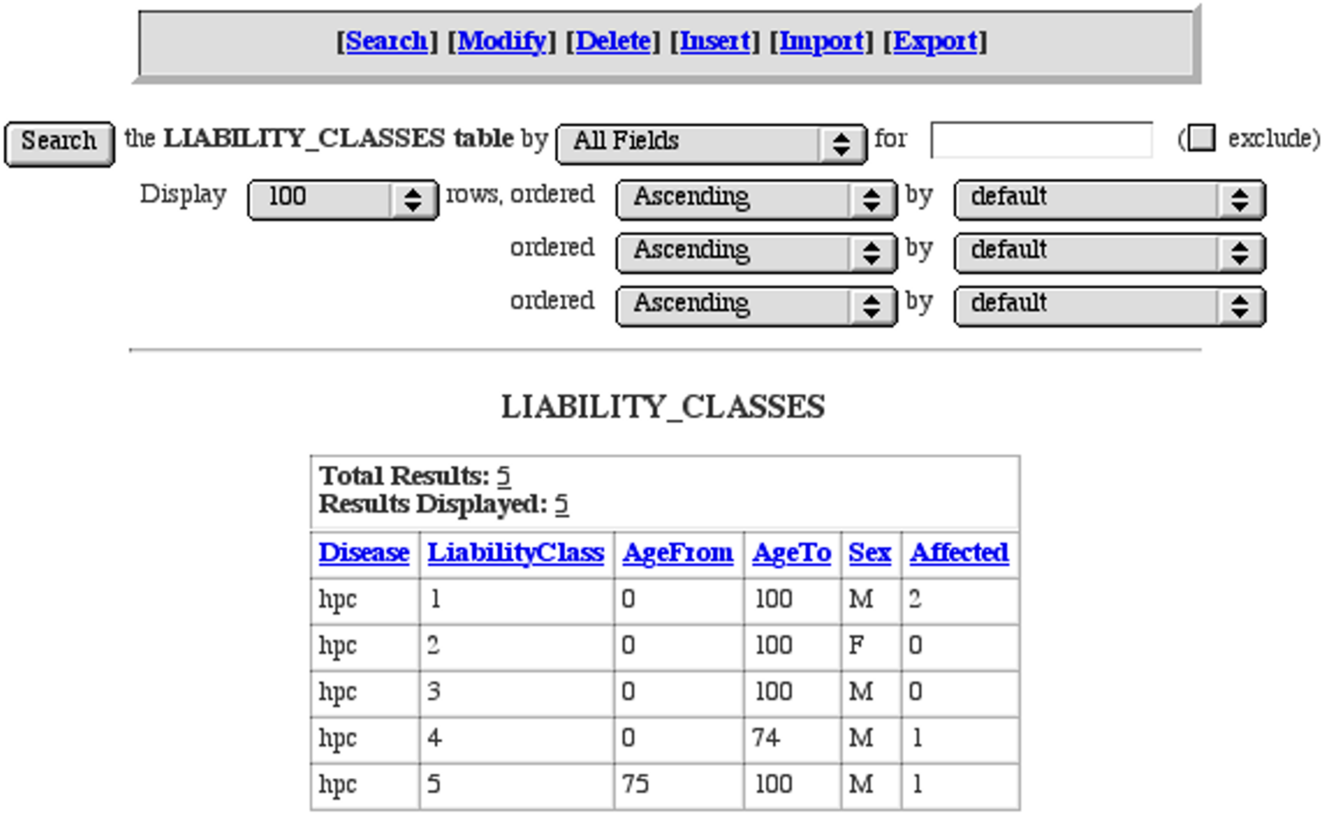
ALLELE_TRANSLATIONS

Total Results: 15
Results Displayed: 15

MarkerName	Allele	Translation
D1S206	211	1
D1S206	213	2
D1S206	215	3
D1S206	217	4
D1S206	219	5
D1S206	221	6
D1S206	223	7
D1S206	225	8
D1S206	227	9
D1S206	209	10
D1S206	231	11
D1S206	203	12
D1S206	205	13
D1S206	207	14
D1S206	229	15

Figure 3
Allele translation The Allele Translations table provides a key linking the "translated" allele (*Translation*) to the original size in base-pairs score (*Allele*). All LINKAGE or RelCheck exports will use identical "allele translation codes." In this example, 15 alleles have been identified for marker D1S206. New alleles identified after the first export (in this example alleles 10 to 15) will be added to the end of a marker's allele translation ensuring consistent recoding of alleles across families.

editing. Changes to the families, pedigrees, or genotypes tables can be easily reviewed in a **Histories** table. Second, contents of every field are verified on import, and users are warned of any failures, such as invalid format or duplicate records. Within the **Pedigrees** table specifically, checks were designed to confirm the presence of all individuals designated as parents within the families, as well as confirm that all fathers are male and all mothers are female. In addition, when genotypes are imported into the **Genotypes** table, GeneLink confirms that each individual included in the import is designated as having DNA in the **Pedigrees** table. Next, GeneLink checks that



the **LIABILITY_CLASSES table** by All Fields for (□ exclude)

Display 100 rows, ordered Ascending by default

ordered Ascending by default

ordered Ascending by default

LIABILITY_CLASSES

Total Results: 5
Results Displayed: 5

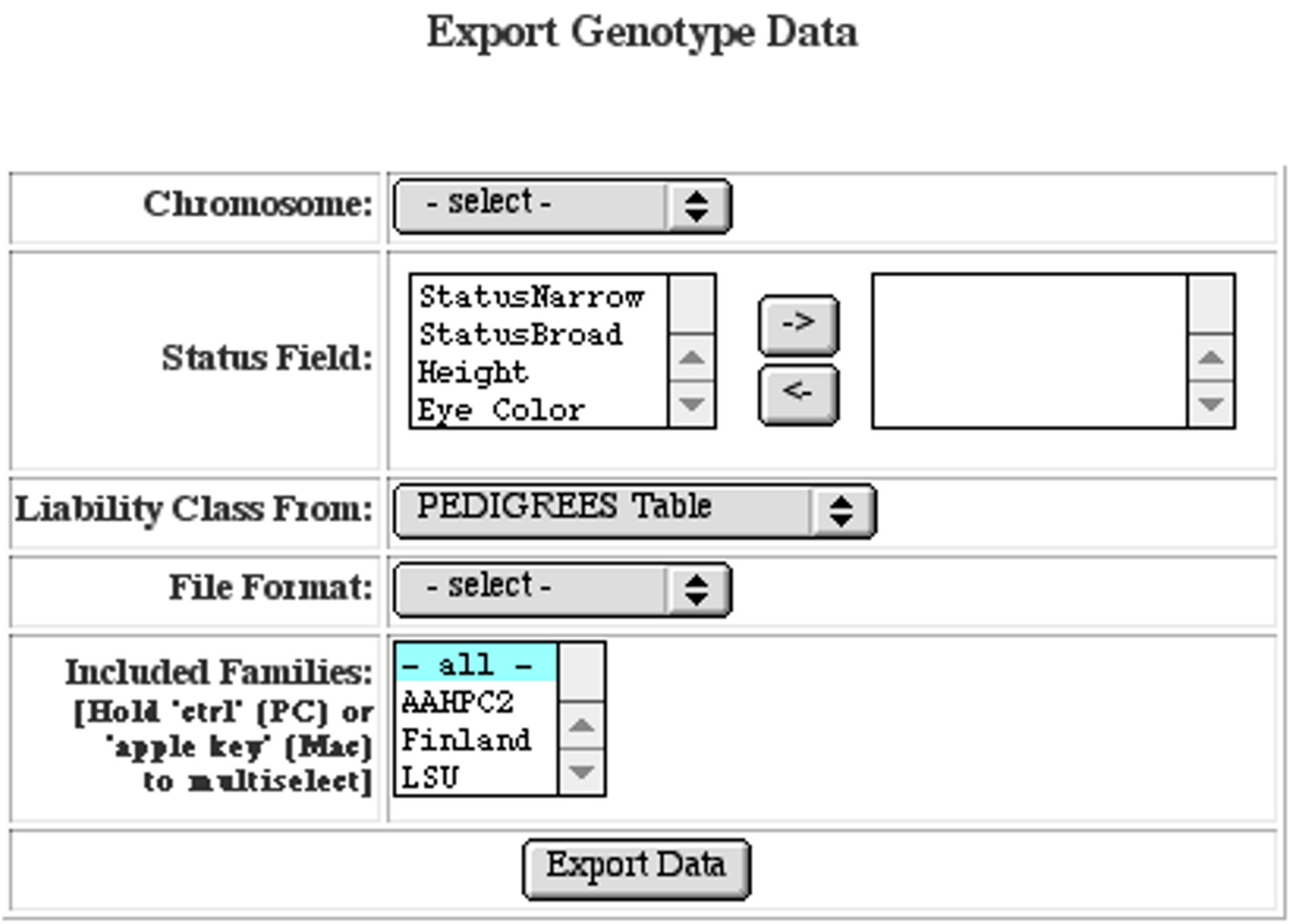
Disease	LiabilityClass	AgeFrom	AgeTo	Sex	Affected
hpc	1	0	100	M	2
hpc	2	0	100	F	0
hpc	3	0	100	M	0
hpc	4	0	74	M	1
hpc	5	75	100	M	1

Figure 4
Liability classes Within GeneLink liability classes can be defined using any combination of age, sex, and affection status.

each allele falls within the marker's designated allele size range (**Markers** table) and that the genotype for the control individual (e.g., 1347-02) matches what is expected (**Markers** table).

Exporting genotypes

We designed GeneLink to facilitate the merging of our genotype data with pedigree and phenotype information (from either the **Pedigrees** table or the **Trait Score** table). This facilitates exporting data in formats commonly used in downstream analyses (e.g., GAS or LINKAGE format). There are three ways data can be exported. First, any user with administrator (*admin*) privileges can export data by Panel in LINKAGE or GAS formats. We use this option when checking our data for Mendelian inconsistencies because most laboratory errors could be detected if multiple markers within a panel showed up as inconsistent. Next, other users with *export* privileges can export data by chromosome. All data are exported by chromosome in the map order specified in the maps table. Chromosomes are only available for exporting once data for all markers on that chromosome have been imported, though we did design GeneLink to accommodate the possibility that all markers on a chromosome may not have been typed for all sites. The export genotype data screen (Figure 5), from the ***Export by chromosome*** menu link, prompts users to specify which chromosome to export, which trait(s) to export, how to define liability classes if necessary, what file format is desired and which families to include in the export. Only the family collection sites to which a user's group has access will appear in this screen. The final way data can be exported by users with *export* privileges is directly from the **Families** table. Using the ***Export by chromosome*** option described above means that all families from the selected site(s) will be exported. Alternatively, exporting directly from the **Families** table makes it possible to export only a subset of families from a collection site or a subset of families from across sites. Regardless of the exporting mechanism employed, GeneLink generates an **Export Report**, which summarizes all the pertinent information regarding the export, including families, chromosome (exported markers and distance between



Export Genotype Data

Chromosome: - select -

Status Field: StatusNarrow, StatusBroad, Height, Eye Color

Liability Class From: PEDIGREES Table

File Format: - select -

Included Families: [Hold 'ctrl' (PC) or 'apple key' (Mac) to multiselect]
- all -
AAHPC2
Finland
LSU

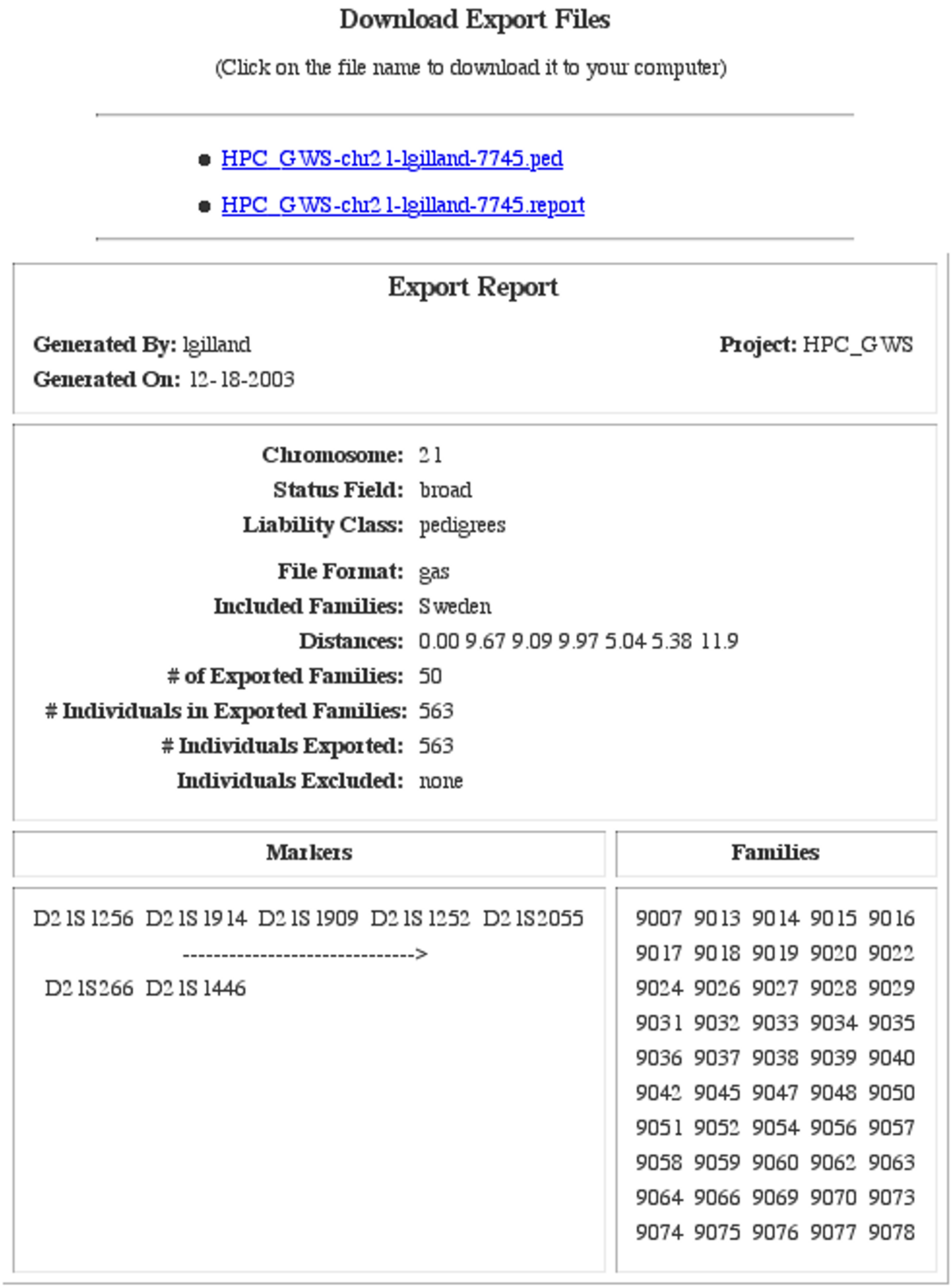
Figure 5
Export by chromosome The **Export Genotype Data** screen prompts users to specify 1) which chromosome to export, 2) which trait(s) to export (*Status Field*), 3) how to define liability classes, 4) what file format is desired and 5) which families to include in the export.

them), phenotypes included, liability class definition, and file format (Figure 6). Each export is given a name, which includes the project ID, chromosome exported, user ID, and a random number. This naming convention was designed to facilitate file management.

Results and discussion

When faced with the challenge of studying 496 hereditary prostate cancer families and a total of 5,247 individuals, we sought a publicly available database management system capable of handling the unique challenges that accompany a large-scale, multi-center genetic linkage study of a complex trait. Although data management systems have been developed [8-12], none could securely and efficiently handle a very large amount of data, as well as provide additional features to facilitate quality control and analysis of data generated. Therefore, we developed GeneLink, a database with unique features, to address these needs.

We designed GeneLink to use a Sybase database backend to take advantage of Sybase's ability to process large amounts of data. Currently, GeneLink is the only publicly available freeware database capable of efficiently storing millions of genotypes. The need for efficient data management will grow in importance as researchers explore



Download Export Files

(Click on the file name to download it to your computer)

- HPC_GWS-chr21-lgilland-7745.ped
- HPC_GWS-chr21-lgilland-7745.report

Export Report

Generated By: lgilland **Project:** HPC_GWS

Generated On: 12-18-2003

Chromosome:	21
Status Field:	broad
Liability Class:	pedigrees
File Format:	gas
Included Families:	Sweden
Distances:	0.00 9.67 9.09 9.97 5.04 5.38 11.9
# of Exported Families:	50
# Individuals in Exported Families:	563
# Individuals Exported:	563
Individuals Excluded:	none

Markers	**Families**

D21S1256 D21S1914 D21S1909 D21S1252 D21S2055
------------------------------>
D21S266 D21S1446
	9007 9013 9014 9015 9016
	9017 9018 9019 9020 9022
	9024 9026 9027 9028 9029
	9031 9032 9033 9034 9035
	9036 9037 9038 9039 9040
	9042 9045 9047 9048 9050
	9051 9052 9054 9056 9057
	9058 9059 9060 9062 9063
	9064 9066 9069 9070 9073
	9074 9075 9076 9077 9078

Figure 6

Export report GeneLink's **Export Report** provides a summary of all pertinent information relating to an export. This includes the file name (which incorporates the project ID, chromosome exported, user ID, and a random number), date the file was created and by whom. The **Export Report** also records which chromosome, phenotypes (*status field*) and families, were exported and in what file format. Finally, the **Export Report** also provides the distance between markers being exported.

genome-wide SNP association studies that may generate close to one billion genotypes (500 cases, 500 controls and 500,000 to 1,000,000 SNPs) [13]. We are currently updating GeneLink so it can run using either Sybase or Oracle. Furthermore, GeneLink was designed to avoid database-specific code and therefore should be portable to other open access DB engines, such as PostgreSQL, without too much difficulty.

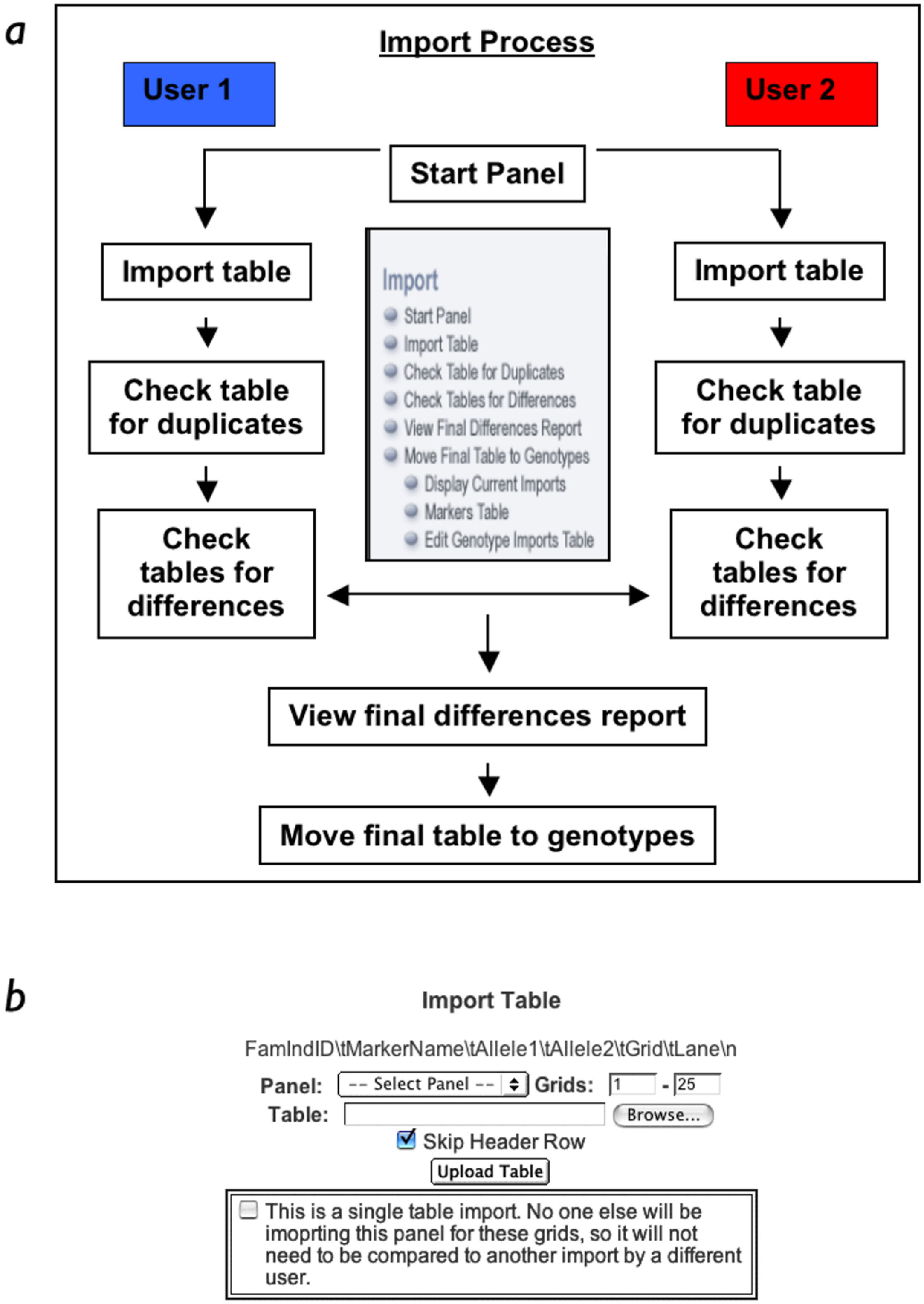
To collect the necessary number of DNA samples needed to provide sufficient power to detect linkage or association, collaborative efforts are almost always required. The Web-based interface of GeneLink facilitates multi-center collaborations, as data can easily be accessed via the Internet. GeneLink's Web-based interface also makes it platform-independent, a feature that was essential given the number of researchers who would be accessing it using various hardware-browser combinations. Other publicly-available databases described in the literature do not have this advantage. In this paper, we have presented GeneLink in the context of a collaborative effort in which multiple sites will need access to data generated in a single laboratory. However, GeneLink would also be valuable in the context of a meta-analysis of data generated in more than one laboratory. Making data access easier for our collaborators translated into the need for a sophisticated security system. Specifically, in our study of hereditary prostate cancer, researchers are permitted access to only their own set of data. This is important because, in some cases, a site's internal review board (IRB) protocol may not allow for raw data to be shared with other analysts.

GeneLink provides several other advantages for investigators performing linkage or linkage disequilibrium studies of complex traits. For example, the process by which genotypes can be imported into GeneLink was designed to be flexible enough to handle data from laboratories like our own which employ duplicated samples and double-scoring methods for quality control purposes. Using duplicated samples and double scoring aids us in keeping our genotyping error rate low (< 1%). In our HPC study, we included 92 duplicated samples (~ 4% of total samples) in order to evaluate our genotyping error rate. The entire import process is outlined in Figures 7A and 7B. After each of the initial steps (the import (step 1), duplicates check (step 2), and check for differences (step 3)), GeneLink produces a summary report (Figures 8A,8B, and 8C). The "Import Report" summarizes the details of the import, including the date, user ID, number of records imported, and file name of the uploaded flat file containing the genotype data (Figure 8A). Examples of the duplicates and differences reports are shown in Figures 8B and 8C.

Another challenge of complex trait linkage or association studies is formatting data appropriately for analysis by existing software packages. Chromosome-specific LINKAGE, GAS, and RelCheck format files can easily be exported by GeneLink. By design GeneLink's exporting capabilities also provide several additional advantages. First, GeneLink is capable of exporting multiple traits at the same time, thus facilitating analyses in which covariate information will be included. Second, by taking advantage of GeneLink's ability to generate liability classes defined by age, sex, and affection status, researchers can maximize power in the investigation of complex traits, which often exhibit reduced penetrance and phenocopies. Third, GeneLink's **Allele Translation** table allows comparison of alleles across families or across analyses, as each allele for each polymorphic marker will only be recoded once. This is particularly important as linkage disequilibrium or association studies become more common. Fourth, GeneLink's ability to export only a subset of families is critical, as genetic heterogeneity is a significant factor contributing to the difficulty of mapping genes involved in many complex traits. Multiple genes (RNASEL, ELAC2, and MSR1, among others; [14-16]) have been implicated in hereditary prostate cancer susceptibility, suggesting that genetic heterogeneity is likely to be a complicating factor in the gene mapping of HPC risk alleles regardless of the analysis method. Finally, GeneLink maintains a list of previously exported files, which eliminates redundant generation of data files by collaborators and functions as an archive of data files used for analyses.

Additional quality control measures were included in GeneLink's design. First, all changes to the database are recorded. As genetic studies of complex traits can be spread over many years, it is important to keep a detailed log of any changes made to the data. For example, an individual's affection status may change during the course of a study; therefore it is critical to track when this information was updated in the database. Second, in order to monitor data quality, GeneLink was also designed to perform several built-in checks, as described above.

Given that genetic studies of complex traits will generate millions (or even billions) of genotypes, it is essential to have appropriate mechanisms in place to ensure data integrity. In our study of hereditary prostate cancer, these checks immediately discovered a typographical error, which, if left undiscovered, would have resulted in data from an affected individual never being exported or analyzed. Finally, GeneLink generates detailed reports storing pertinent information regarding all imports and exports (Figures 6 and 8A), the status of projects (Figure 9), statistical information about markers (success rates and heterozygosity; Figure 10A), and DNA samples tested (Figure 10B). These reports are helpful in maintaining data quality. For example, in our HPC study with over 2,500 DNA


a
Import Process
User 1
User 2
Start Panel
Import table
Import table
Import
Start Panel
Import Table
Check Table for Duplicates
Check Tables for Differences
View Final Differences Report
Move Final Table to Genotypes
Display Current Imports
Markers Table
Edit Genotype Imports Table
Check table for duplicates
Check table for duplicates
Check tables for differences
Check tables for differences
View final differences report
Move final table to genotypes
b
Upload Table
Browse...

Import Table

FamIndID\tMarkerName\tAllele1\tAllele2\tGrid\tLane\n

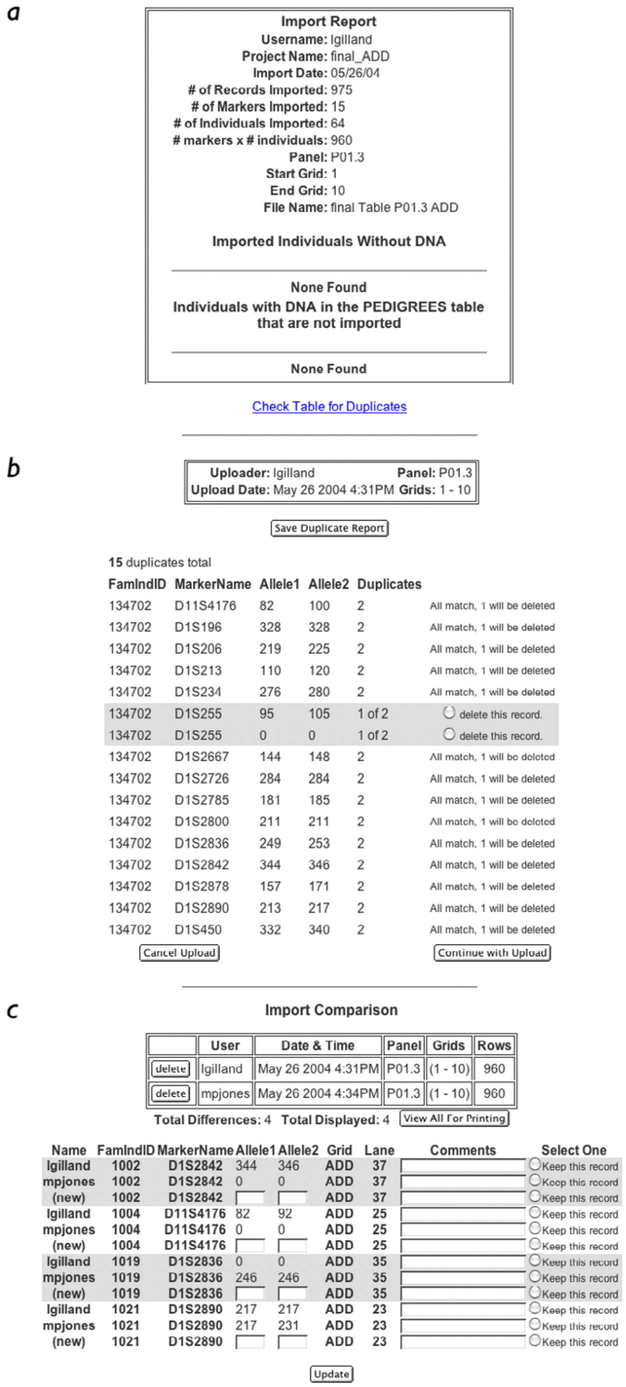
Panel: -- Select Panel -- Grids: 1 - 25

Table:

☑ Skip Header Row

□ This is a single table import. No one else will be imoprting this panel for these grids, so it will not need to be compared to another import by a different user.

Figure 7
A, and B. Import process Outline of Import process illustrates GeneLink's ability to be used in laboratories that include duplicated samples and double scoring for quality control purposes. Import process includes within table duplicate check and across table differences check. Using the Single table import allows the differences step to be skipped.



Import Report

Username: lgilland
Project Name: final_ADD
Import Date: 05/26/04
# of Records Imported: 975
# of Markers Imported: 15
# of Individuals Imported: 64
# markers x # individuals: 960
Panel: P01.3
Start Grid: 1
End Grid: 10
File Name: final Table P01.3 ADD

Imported Individuals Without DNA

None Found

Individuals with DNA in the PEDIGREES table that are not imported

None Found

Check Table for Duplicates

Uploader: lgilland **Panel:** P01.3
Upload Date: May 26 2004 4:31PM **Grids:** 1 - 10

Save Duplicate Report

15 duplicates total

FamIndID	MarkerName	Allele1	Allele2	Duplicates	
134702	D11S4176	82	100	2	All match, 1 will be deleted
134702	D1S196	328	328	2	All match, 1 will be deleted
134702	D1S206	219	225	2	All match, 1 will be deleted
134702	D1S213	110	120	2	All match, 1 will be deleted
134702	D1S234	276	280	2	All match, 1 will be deleted
134702	D1S255	95	105	1 of 2	○ delete this record.
134702	D1S255	0	0	1 of 2	○ delete this record.
134702	D1S2667	144	148	2	All match, 1 will be deleted
134702	D1S2726	284	284	2	All match, 1 will be deleted
134702	D1S2785	181	185	2	All match, 1 will be deleted
134702	D1S2800	211	211	2	All match, 1 will be deleted
134702	D1S2836	249	253	2	All match, 1 will be deleted
134702	D1S2842	344	346	2	All match, 1 will be deleted
134702	D1S2878	157	171	2	All match, 1 will be deleted
134702	D1S2890	213	217	2	All match, 1 will be deleted
134702	D1S450	332	340	2	All match, 1 will be deleted

Cancel Upload Continue with Upload

Import Comparison

	User	Date & Time	Panel	Grids	Rows
delete	lgilland	May 26 2004 4:31PM	P01.3	(1 - 10)	960
delete	mpjones	May 26 2004 4:34PM	P01.3	(1 - 10)	960

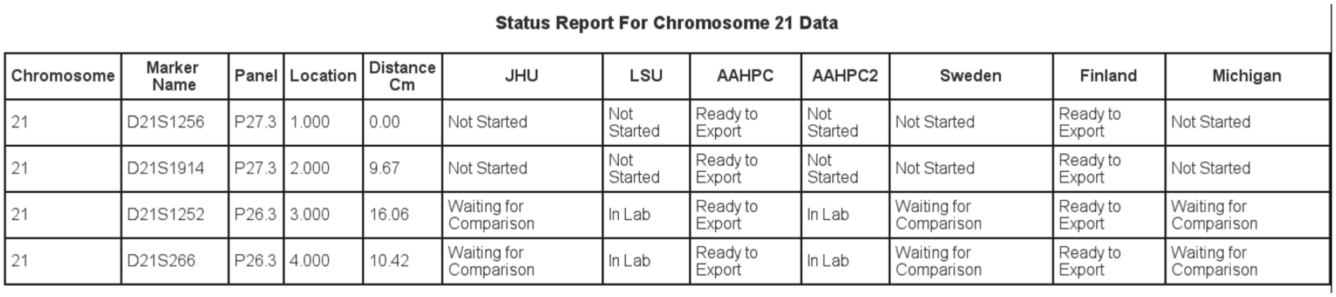
Total Differences: 4 **Total Displayed:** 4 View All For Printing

Name	FamIndID	MarkerName	Allele1	Allele2	Grid	Lane	Comments	Select One
lgilland	1002	D1S2842	344	346	ADD	37		○ Keep this record
mpjones	1002	D1S2842	0	0	ADD	37		○ Keep this record
(new)	1002	D1S2842			ADD	37		○ Keep this record
lgilland	1004	D11S4176	82	92	ADD	25		○ Keep this record
mpjones	1004	D11S4176	0	0	ADD	25		○ Keep this record
(new)	1004	D11S4176			ADD	25		○ Keep this record
lgilland	1019	D1S2836	0	0	ADD	35		○ Keep this record
mpjones	1019	D1S2836	246	246	ADD	35		○ Keep this record
(new)	1019	D1S2836			ADD	35		○ Keep this record
lgilland	1021	D1S2890	217	217	ADD	23		○ Keep this record
mpjones	1021	D1S2890	217	231	ADD	23		○ Keep this record
(new)	1021	D1S2890			ADD	23		○ Keep this record

Update

Figure 8

A, B, and C. Example of import (A), duplicates (B), and differences (C) reports Import report stores all information regarding genotypes imported into GeneLink. This information includes number of records imported (how many unique individuals, how many markers) as well as the name of the file in which records are stored. The duplicates process checks for duplicate genotypes within a table. A duplicate is defined as records with the same FamInd ID and marker. All duplicate records are reported. If duplicate records have the same two alleles then one record is deleted. If the two duplicate records do not have matching alleles than the user is prompted to select which record to delete. The differences process looks for differences in genotypes compared across tables (independently scored by two researchers). All differences are reported and the user is prompted to resolve each appropriately. The user is given the option to save either record or if either score isn't acceptable then new genotype can be indicated.

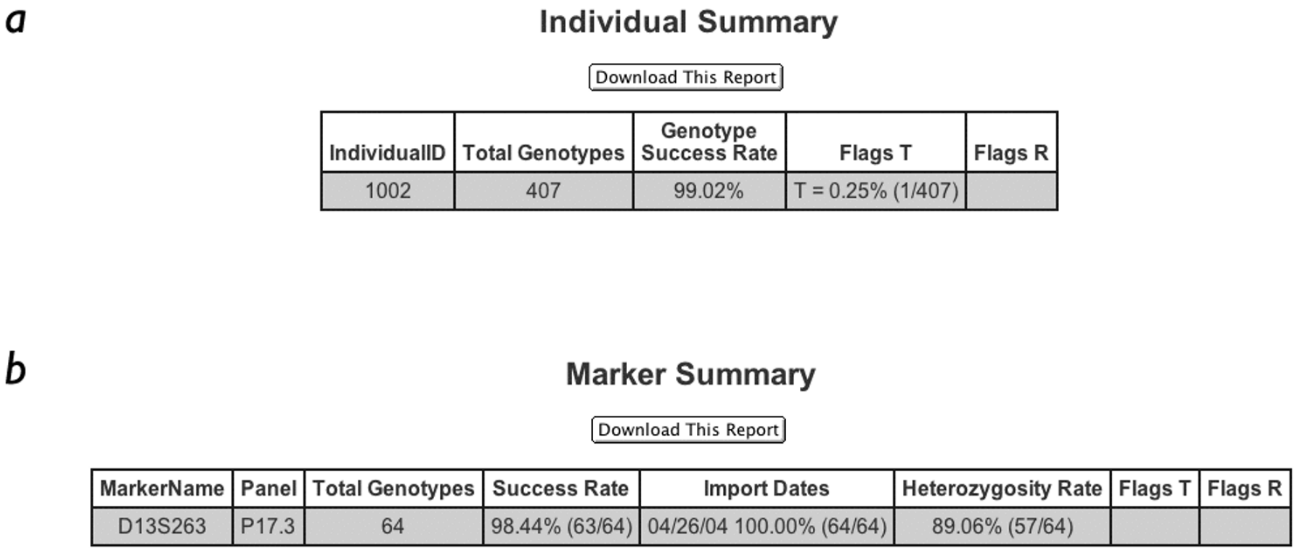


Status Report For Chromosome 21 Data

Chromosome	Marker Name	Panel	Location	Distance Cm	JHU	LSU	AAHPC	AAHPC2	Sweden	Finland	Michigan
21	D21S1256	P27.3	1.000	0.00	Not Started	Not Started	Ready to Export	Not Started	Not Started	Ready to Export	Not Started
21	D21S1914	P27.3	2.000	9.67	Not Started	Not Started	Ready to Export	Not Started	Not Started	Ready to Export	Not Started
21	D21S1252	P26.3	3.000	16.06	Waiting for Comparison	In Lab	Ready to Export	In Lab	Waiting for Comparison	Ready to Export	Waiting for Comparison
21	D21S266	P26.3	4.000	10.42	Waiting for Comparison	In Lab	Ready to Export	In Lab	Waiting for Comparison	Ready to Export	Waiting for Comparison

Figure 9

Status report GeneLink's status reports allow collaborators to easily tract the project's progress by site. Reports show markers by chromosome (in map order) and the status of each marker for each site. By site, markers can be *Not started*, *In Lab*, *Genotypes Imported*, *Single Table Imported*, *Waiting for Comparison*, *Compared and Ready*, *Ready to Finalize* and *Ready to Export*. Data can be exported only after all markers on a given chromosome for a given site are *Ready to Export*. In this example we are looking at markers D21S1256, D21S1914, D21S1909, D21S1252, D21S2055, D21S266, and D21S1446 on chromosome 21 for sites JHU, AAHPC, AAHPC2, Sweden, Finland and Michigan. Here data is ready to export for family sites AAHPC2 and Sweden.



Individual Summary

IndividualID	Total Genotypes	Genotype Success Rate	Flags T	Flags R
1002	407	99.02%	T = 0.25% (1/407)	

Marker Summary

MarkerName	Panel	Total Genotypes	Success Rate	Import Dates	Heterozygosity Rate	Flags T	Flags R
D13S263	P17.3	64	98.44% (63/64)	04/26/04 100.00% (64/64)	89.06% (57/64)		

Figure 10

A, and B. Marker (A) and individual (B) summaries GeneLink's **Marker Summary** provides success rates and heterozygosity for individual markers typed in the study. The **Marker Summary** also provides information regarding when the genotype records for this marker were imported (*Import Dates*). Marker quality can also be evaluated using the Flags column. Genotypes can be flagged with a T to temporarily blank scores from analyses. This is used for un-resolvable Mendelian inconsistencies. The R flag can be used for replaced DNA samples until the new DNA sample is evaluated. Neither T nor R flagged genotypes are exported. **Individual summaries** also provide global quality assurance information such as success or flag rates.

samples, it would have been very easy to miss that any single individual was failing for greater than 95% of markers if we were not using GeneLink. We were able to request new DNA samples for such individuals, as well as flag the stored data as potentially problematic.

GeneLink was designed primarily in the context of family-based studies of complex traits. It is capable of handling both linkage and association data, and can be used for both whole genome scans and/or candidate-gene studies. Further development of GeneLink will focus on extending its capabilities in regard to the case-based design. We recognize that both the family-based and case-based study designs have unique advantages, so we see it as critical to make GeneLink flexible enough to accommodate a case-based design. Currently, there is no limitation in storing case-based data however changes to GeneLink's exporting mechanisms should be made. Finally, in the same way that GeneLink is capable of storing "exported" data input files, future work will center on the storage of analysis results. Again, this would be helpful for multi-center collaborative studies, which will continue to be critical to successful efforts to identify genes important in complex trait etiology.

Conclusions

In summary, GeneLink was designed specifically to ease the data management burden of mapping complex traits. It provides many functions that make it a uniquely powerful tool for use in genetic linkage or association studies. GeneLink simplifies merging genotypic data with pedigree, phenotype, and genetic or physical map information. Specifically, GeneLink's design makes it ideal for large-scale, multi-center studies, which will become more and more common in efforts to dissect the genetic factors contributing to complex trait etiology.

Availability and requirements

Project name: GeneLink

Project home page: http://research.nhgri.nih.gov/genelink

Operating system(s): Platform independent

Programming language: Perl

Other requirements: Sybase SQL server ASE 12.5.1, Perl version 5.6.1 or greater, CPAN Perl modules DBI, DBD::Sybase, CGI, and Carp, Web server such as Apache 1.3.29

License: Sybase SQL server ASE 12.5.1

Any restrictions to use by non-academics: none

Author's contributions

EG, JT, JEBW and AB participated in database's design. AM, LU, KT and TW did all of the programming. EG, DG, PD, APK, MJ, DFL, and GI performed extensive testing of the database. EG, PD, TW, JEBW and AB drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The authors would like to thank Dr. Scott Diehl for early discussions of data management needs and database design.

References

1. Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Melmer G, Dean M, Rozmahel R, Cole JL, Kennedy D, Hidaka N, Zsiga M, Buchwald M, Riordan J, Tsui L, Collins FS: **Identification of the cystic fibrosis gene: chromosome walking and jumping.** *Science* 1989, **245(4922):**1059-1065.
2. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL, Drumm ML, Iannuzzi MC, Collins FS: **Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA.** *Science* 1989, **245(4922):**1066-1073.
3. The Huntington's Disease Collaborative Research Group: **A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes.** *Cell* 1993, **72(6):**971-983.
4. Lange EM, Gillanders EM, Davis CC, Brown WM, Campbell JK, Jones M, Gildea D, Riedesel E, Albertus J, Freas-Lutz D, Markey C, Giri V, Dimmer JB, Montie JE, Trent JM, Cooney KA: **Genome-wide scan for prostate cancer susceptibility genes using families from the University of Michigan prostate cancer genetics project finds evidence for linkage on chromosome 17 near BRCA1.** *Prostate* 2003, **57(4):**326-334.
5. Schleutker J, Baffoe-Bonnie AB, Gillanders E, Kainu T, Jones MP, Freas-Lutz D, Markey C, Gildea D, Riedesel E, Albertus J, Gibbs KD Jr, Matikainen M, Koivisto PA, Tammela T, Bailey-Wilson JE, Trent JM, Kallioniemi OP: **Genome-wide scan for linkage in finnish hereditary prostate cancer (HPC) families identifies novel susceptibility loci at 11ql4 and 3p25-26.** *Prostate* 2003, **57(4):**280-289.
6. Wiklund F, Gillanders EM, Albertus JA, Bergh A, Damber JE, Emanuelsson M, Freas-Lutz DL, Gildea DE, Goransson I, Jones MS, Jonsson BA, Lindmark F, Markey CJ, Riedesel EL, Stenman E, Trent JM, Gronberg H: **Genome-wide scan of Swedish families with hereditary prostate cancer: suggestive evidence of linkage at 5q11.2 and 19pl3.3.** *Prostate* 2003, **57(4):**290-297.
7. Xu J, Gillanders EM, Isaacs SD, Chang BL, Wiley KE, Zheng SL, Jones M, Gildea D, Riedesel E, Albertus J, Freas-Lutz D, Markey C, Meyers DA, Walsh PC, Trent JM, Isaacs WB: **Genome-wide scan for prostate cancer susceptibility genes in the Johns Hopkins hereditary prostate cancer families.** *Prostate* 2003, **57(4):**320-325.
8. Seuchter SA, Skolnick MH: **HGDBMS: a human genetics database management system.** *Comput Biomed Res* 1988, **21(5):**478-487.
9. Adams P: **LABMAN and LINKMAN: a data management system specifically designed for genome searches of complex diseases.** *Genet Epidemiol* 1994, **11(1):**87-98.
10. Cheung KH, Nadkarni P, Silverstein S, Kidd JR, Pakstis AJ, Miller P, Kidd KK: **PhenoDB: an integrated client/server database for linkage and population genetics.** *Comput Biomed Res* 1996, **29(4):**327-337.
11. McMahon FJ, Thomas CJ, Koskela RJ, Breschel TS, Hightower TC, Rohrer N, Savino C, McInnis MG, Simpson SG, DePaulo JR: **Integrating clinical and laboratory data in genetic studies of complex phenotypes: a network-based data management system.** *Am J Med Genet* 1998, **81(3):**248-256.
12. Li JL, Deng H, Lai DB, Xu F, Chen J, Gao G, Recker RR, Deng HW: **Toward high- throughput genotyping: dynamic and automatic software for manipulating large-scale genotype data using fluorescently labeled dinucleotide markers.** *Genome Res* 2001, **11(7):**1304-1314.
13. Botstein D, Risch N: **Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease.** *Nat Genet* 2003, **33(Suppl):**228-237.
14. Carpten J, Nupponen N, Isaacs S, Sood R, Robbins C, Xu J, Faruque M, Moses T, Ewing C, Gillanders E, Hu P, Bujnovszky P, Makalowska I, Baffoe-Bonnie A, Faith D, Smith J, Stephan D, Wiley K, Brownstein

M, Gildea D, Kelly B, Jenkins R, Hostetter G, Matikainen M, Schleutker J, Klinger K, Connors T, Xiang Y, Wang Z, De Marzo A, Papadopoulos N, Kallioniemi OP, Burk R, Meyers D, Gronberg H, Meltzer P, Silverman R, Bailey-Wilson J, Walsh P, Isaacs W, Trent J: **Germline mutations in the ribonuclease L gene in families showing linkage with HPC1.** *Nat Genet* 2002, **30(2):**181-184.

15. Tavtigian SV, Simard J, Teng DH, Abtin V, Baumgard M, Beck A, Camp NJ, Carillo AR, Chen Y, Dayananth P, Desrochers M, Dumont M, Farnham JM, Frank D, Frye C, Ghaffari S, Gupte JS, Hu R, Iliev D, Janecki T, Kort EN, Laity KE, Leavitt A, Leblanc G, McArthur-Morrison J, Pederson A, Penn B, Peterson KT, Reid JE, Richards S, Schroeder M, Smith R, Snyder SC, Swedlund B, Swensen J, Thomas A, Tranchant M, Woodland AM, Labrie F, Skolnick MH, Neuhausen S, Rommens J, Cannon-Albright LA: **A candidate prostate cancer susceptibility gene at chromosome 17p.** *Nat Genet* 2001, **27(2):**172-180.

16. Xu J, Zheng SL, Komiya A, Mychaleckyj JC, Isaacs SD, Hu JJ, Sterling D, Lange EM, Hawkins GA, Turner A, Ewing CM, Faith DA, Johnson JR, Suzuki H, Bujnovszky P, Wiley KE, DeMarzo AM, Bova GS, Chang B, Hall MC, McCullough DL, Partin AW, Kassabian VS, Carpten JD, Bailey-Wilson JE, Trent JM, Ohar J, Bleecker ER, Walsh PC, Isaacs WB, Meyers DA: **Germline mutations and sequence variants of the macrophage scavenger receptor 1 gene are associated with prostate cancer risk.** *Nat Genet* 2002, **32(2):**321-325.