Question 4

A user wishes to find all the single nucleotide polymorphisms that lie between two sequence-tagged sites. Do any of these single nucleotide polymorphisms fall within the coding region of a gene? Where can any additional information about the function of these genes be found?

doi:10.1038/ng969

The starting point for this search would be the web site for the Database of Single Nucleotide Polymorphisms (dbSNP) at the NCBI[13], which is located at http://www.ncbi.nlm.nih.gov/SNP. There is a series of links on the page that allow the user to search using either information about the database submission itself or information regarding genes and gene loci.

For this particular search, assume that the region of interest is known and defined by two STS markers, RH70674 and G32133. Begin by scrolling to the section labeled *Between Markers* at the bottom of the page. Enter the STS marker names 'RH70674' and 'G32133' into the two text boxes, and click on *Submit STS Markers*. This will produce a display showing SNPs 1–25 out of the total of 81 within the region of interest. Go to page 3 of the display by entering '3' in the Page box and clicking *Display*.

The resulting page (Fig. 4.1) illustrates most of the possible types of result one would find on a typical dbSNP results page. In the table, starting from the left, the first column gives the individual dbSNP cluster IDs (all starting with '*rs*'). The second column, labeled *Map*, shows whether a particular SNP has been mapped to a unique position in the genome (illustrated by a single green arrow, as in the first row of the example) or to multiple positions (not shown here).

The next set of columns, labeled *Gene*, indicates whether these SNPs are associated with particular features, such as genes, mRNAs or coding regions. The three columns (*L*, *T* and *C*) are either lit up or appear gray in every row. Taking each in order:

If the *L* (for locus) appears in blue, part or all of the marker position lies either within 2 kilobases (kb) of the 5′ end of a gene feature or within 500 bases of the 3′ end of a gene feature.

If the *T* (for transcript) appears in green, part or all of the marker position overlaps with a known mRNA. This does not mean, however, that the SNP marker necessarily falls within a coding region.

If the *C* (for coding) appears in orange, part or all of the marker position overlaps with a coding region.

The next column, labeled *Het*, indicates the average heterozygosity observed for this marker, on a scale of 0–100%. A reading of zero means that no information is available for that particular marker, whereas the pink bars show a 95% confidence interval for the marker. The *Validation* column indicates whether the marker has been validated (shown by a star) or is unvalidated (shown by light blue boxes). Validated markers have been verified by independent re-analysis of the sequence. All of the unvalidated markers shown in Fig. 4.1 are denoted by three blue boxes, which, according to the scale at the top of the column, means that there is a >95% success rate in validation. This figure indicates the probability that this marker is real. (The success rate is defined as 1 – false-positive rate.)

In the penultimate column, the symbol *TT* (not shown here) indicates that individual genotypes are available for this marker. Finally, the *Linkout Avail* column indicates which markers are linked to other databases; a *P* in this column indicates that the variation has been mapped to a known protein structure. For a complete description of all the features within this display, click on any part of the header above the columns.

Returning to the original question, one of the SNPs displayed on this page does indeed fall within a coding region, as indicated by an orange *C*. To obtain more information on any particular SNP, simply click on the hyperlinked SNP Cluster ID. Clicking on *rs1059133*, for example, produces a new page, with all available information on that SNP (Fig. 4.2). Under the header marked *Submitter records for this RefSNP Cluster* is a list of the individual SNPs (in this case, only one SNP) that have been clustered together to form this single reference SNP. The sequence of the SNP is shown in the next header. Under the header marked *NCBI Resource Links* are GenBank and NCBI RefSeq entries that are associated with this SNP. Scrolling further down on the SNP page (Fig. 4.3), the gene whose coding region this SNP falls within is indicated on the *LocusLink Analysis* section (*ADAM2*, a disintegrin and metalloproteinase domain 2). The SNP allele is G/C, a non-synonymous change leading to replacement of the Asp residue in the reference sequence by a His residue. Links are also provided to the NCBI Map Viewer, Ensembl map and UCSC genome assembly in the section labeled *Integrated Maps*. The sections labeled *Variation Summary* and *Validation Summary* (not shown) give the raw data on this particular SNP.

To answer the final part of this question requires jumping from dbSNP to LocusLink[10]. To do so, click on the *ADAM2* link in the line marked *LocusLink* at the top of the page (Fig. 4.3). This brings the user to the LocusLink page for *ADAM2* and provides numerous jumping-off points to the NCBI and affiliated resources through the boxed links at the top of the page. More information on these resources can be found by following the LocusLink FAQ link in the left-hand column of the page. By simply examining the LocusLink page itself, one sees that the ADAM2 protein belongs to a family of membrane-anchored proteins that have been implicated in processes as diverse as fertilization, muscle development and neurogenesis.

One often-overlooked source of information on genes and gene products is OMIM[14]. This is an electronic version of the

Using the UCSC browser, users can retrieve the positions of genome annotations such as SNPs as a text file suitable for loading into a spreadsheet program. While looking at the browser for a defined chromosomal region, click on the *Tables* link (Fig. 1.6, upper blue bar). Similarly, to export a list of genome annotations in a defined chromosomal region at Ensembl, click on *Export* from any ContigView window (Fig. 1.14, center yellow bar).

catalog of human genes and genetic disorders developed by Victor McKusick at The Johns Hopkins University. OMIM provides the user with concise textual information from the published literature on most human disorders with a genetic basis, and links back to the primary literature as appropriate. Information comprising an OMIM entry includes the gene symbol, alternate names for the disease, a description of the disease (including clinical, biochemical and cytogenetic features), details of the mode of inheritance (including mapping information) and a clinical synopsis. These entries are manually curated, ensuring that the 'executive summary' is up to date and accurate. Although OMIM can be searched directly, many LocusLink entries also link to the OMIM record for the gene. The OMIM entry page for the ADAM2 protein is shown in Fig. 4.4. The page is fully hyperlinked to PubMed, GenBank and other related databases.

Figure 4.1

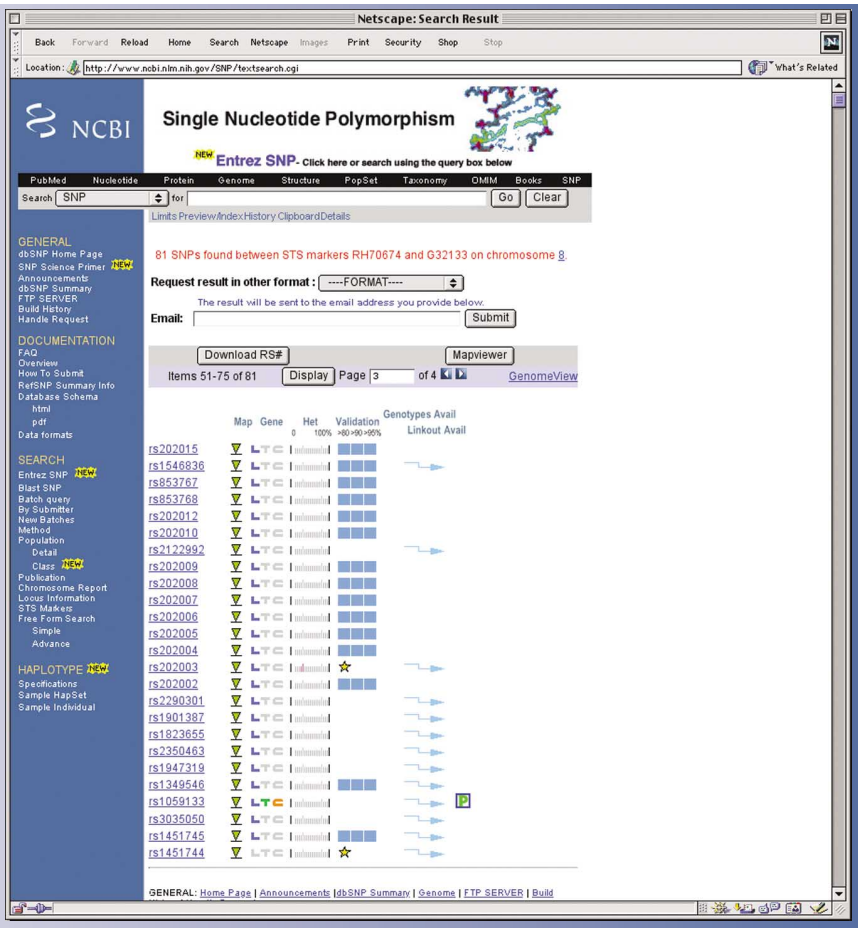
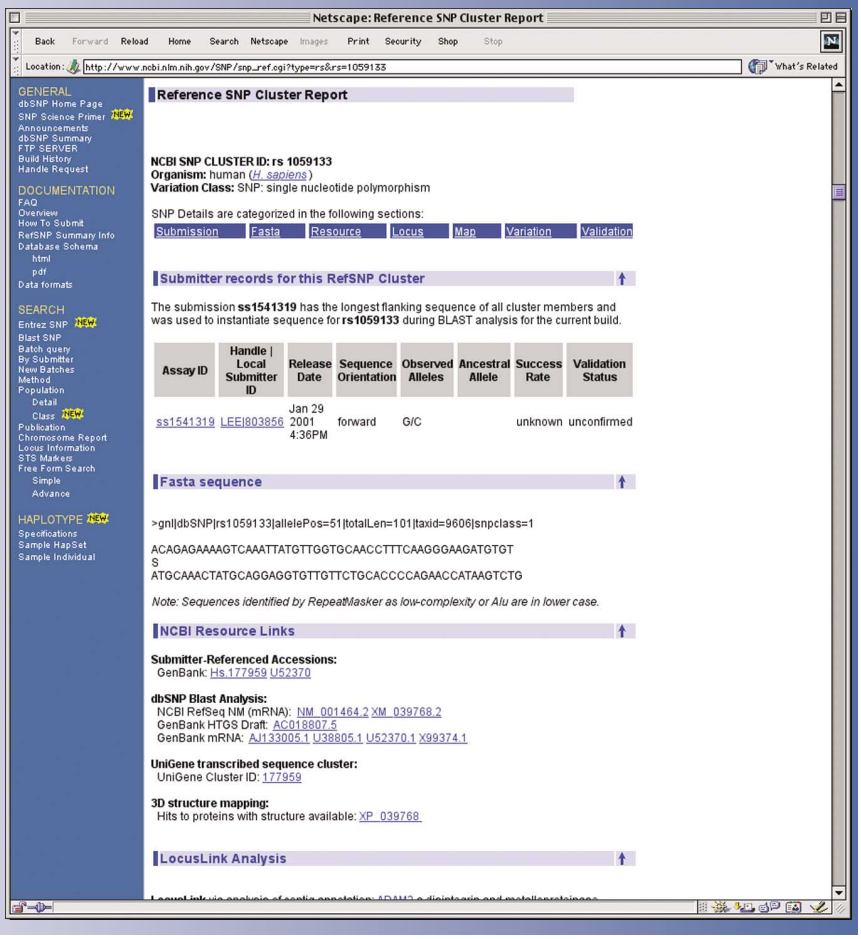


Figure 4.2



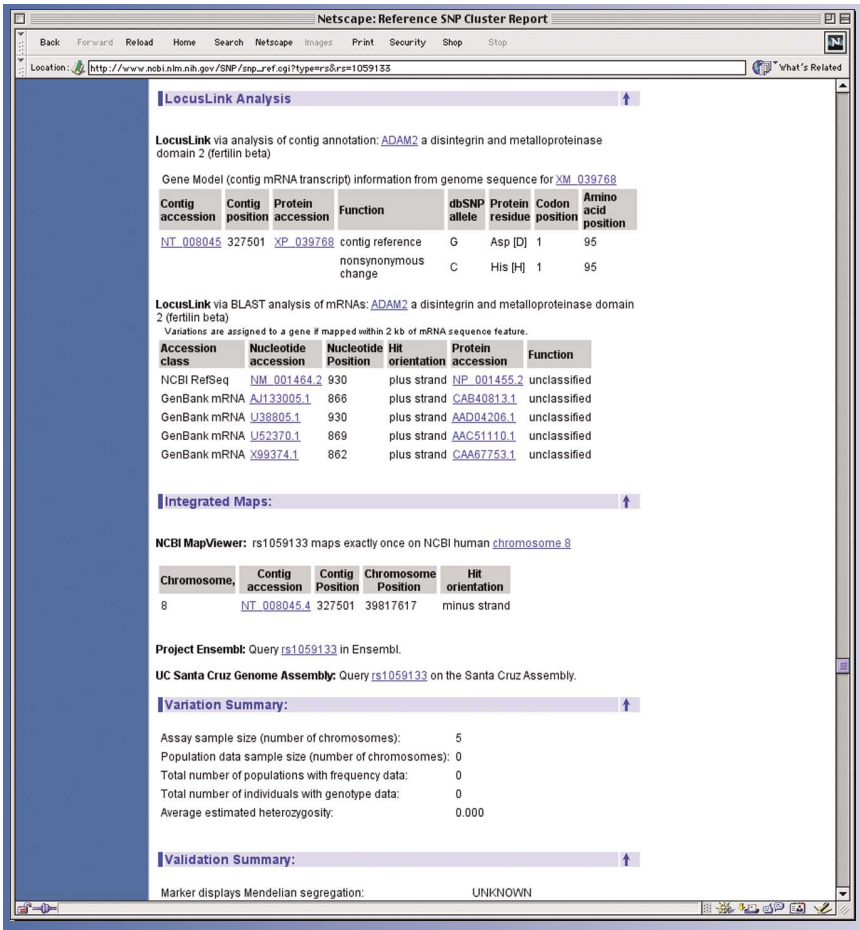


Figure 4.3

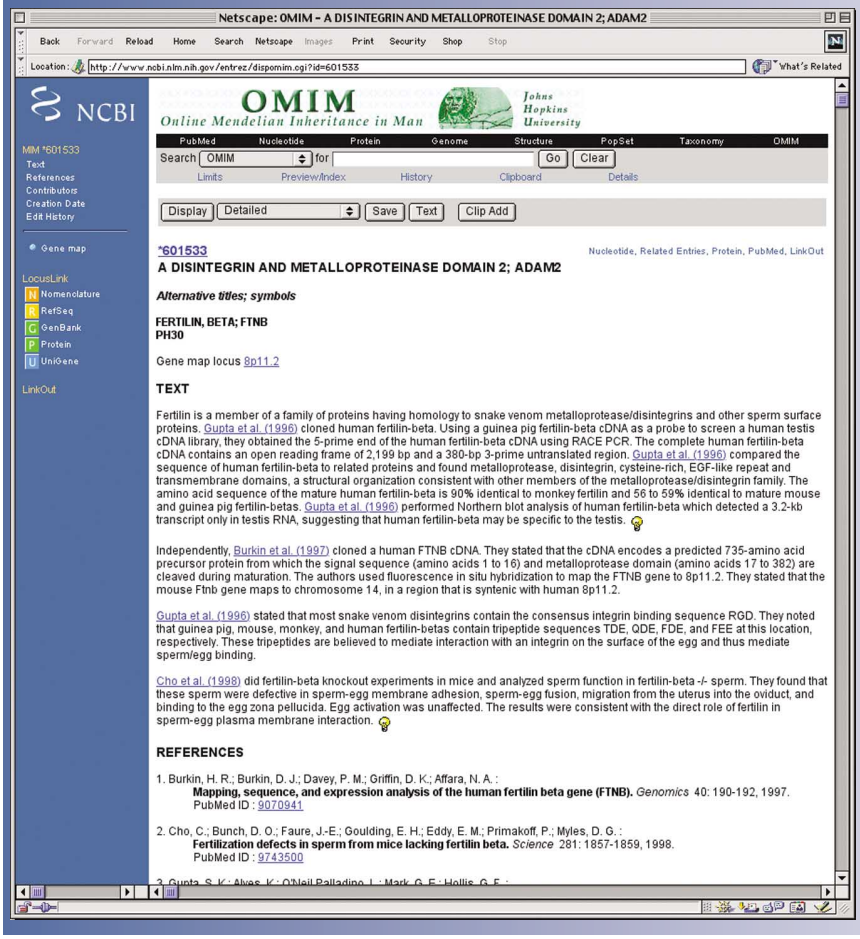


Figure 4.4