Databases

The Androgen Receptor Gene Mutations Database (ARDB): 2004 Update

Bruce Gottlieb,1,4* Lenore K. Beitel,1,2 Jian Hui Wu,1 and Mark Trifiro1,2,3
1Lady Davis Institute for Medical Research, Sir Mortimer B. Davis-Jewish General Hospital, Montreal, Canada; 2Department of Medicine, McGill University, Montreal, Canada; 3Department of Human Genetics, McGill University, Montreal, Canada; 4Department of Biology, John Abbott College, Canada

Communicated by Alastair Brown

The current version of the androgen receptor (AR) gene mutations database is described. The total number of reported mutations has risen from 374 to 605, and the number of AR-interacting proteins described has increased from 23 to 70, both over the past 3 years. A 3D model of the AR ligand-binding domain (AR LBD) has been added to give a better understanding of gene structure–function relationships. In addition, silent mutations have now been reported in both androgen insensitivity syndrome (AIS) and prostate cancer (CaP) cases. The database also now incorporates information on the exon 1 CAG repeat expansion disease, spinobulbar muscular atrophy (SBMA), as well as CAG repeat length variations associated with risk for female breast, uterine endometrial, colorectal, and prostate cancer, as well as for male infertility. The possible implications of somatic mutations, as opposed to germline mutations, in the development of future locus-specific mutation databases (LSDBs) is discussed. The database is available on the Internet (www.mcgill.ca/androgendb/). Hum Mutat 23:527–533, 2004.

r 2004 Wiley-Liss, Inc.

KEY WORDS: databases; androgen receptor gene; AR; androgen insensitivity syndrome; AIS; prostate cancer; CaP; Kennedy disease; spinobulbar muscular atrophy; SBMA

INTRODUCTION

As we enter the 21st century, most genetic databases remain firmly entrenched in the 20th. Their function has been to act as passive repositories for listings of gene alterations that result in proteins whose putative altered structure presumably results in disease or nondisease phenotypes. While these databases have been relatively simple, they have been immensely useful in establishing many valuable links between specific mutations and disease. Meanwhile, the growth of the Internet has allowed this information to be readily accessible to almost all interested parties. In addition, it has allowed for online submission of data with the curator assuming responsibility for validating the data.

The effectiveness of these databases has been predicated on a number of assumptions. First, it is assumed that a particular mutation is indeed responsible for an altered phenotype. However, this clearly needs to be experimentally proven, by expressing the protein and then examining it. In most cases this is not done, based on the assumption that all mutations are likely to have a functional effect on the expressed protein. In addition, even when the gene is expressed, it is usually expressed, at least in the case of human genes, in a nonhuman cell line. Second, it is assumed that the mutation has been inherited, i.e., is germline, and will therefore be an excellent predictor of the inheritance pattern of a particular condition. However, it is being increasingly observed, particularly in diseases such as cancer, that the mutations are often somatic rather than germline in origin. The third assumption is that there is a clear distinction between mutations and polymorphisms (gene alterations that occur in >1% of a population). Finally, it is often assumed that once a locus-specific gene mutation database (LSDB) has been created, it is routinely updated and maintained.

Received 7 September 2003; accepted revised manuscript 23 January 2004.

Grant sponsor: Canadian Institutes of Health Research.
*Correspondence to: Bruce Gottlieb, Ph.D., Lady Davis Institute for Medical Research, Sir Mortimer B. Davis - Jewish General Hospital, 3755 Cote Ste. Catherine Road, Montreal, Quebec, Canada H3T 2E1.
E-mail: bruce.gottlieb@mcgill.ca
DOI 10.1002/humu.20044
Published online in Wiley InterScience (www.interscience.wiley.com).
However, in many cases this does not occur and the LSDBs gradually become obsolete [Stenson et al., 2003]. All of these assumptions have led to making most databases, in their present form, less than ideal for determining the significance of an alteration to a gene as the cause of a particular disease or condition. The androgen receptor gene mutation database (ARDB; www.mcgill.ca/androgendb) has confronted many of the same problems. In this article, we report on the developments that have occurred in the ARDB over the past few years, and the changes that we have introduced to make it both more relevant and more useful.

**DATABASE INFORMATION**

**Loss-of-Function Mutations**

**Androgen insensitivity syndrome.** The androgen receptor (AR; MIM# 313700) is a member of the superfamily of nuclear receptors that function as ligand-dependent transcription factors. Intracellular AR is essential for androgen action, whether of testosterone (T) or of its 5α-reduced derivative, 5α-dihydrotestosterone (DHT). Hence, the AR is essential for normal primary male sexual development before birth (mascu

The present version of the ARDB is based on GenBank reference sequence NM_000044.2, and contains over 500 entries of mutations causing androgen insensitivity syndrome (AIS), representing over 300 different AR gene (AR) mutations, from more than 600 patients with AIS. There has been a large increase in the number of reported AR mutations since the last published report on the database [Gottlieb et al., 1999]; the number of entries almost doubled, from 374 to 605. This increase has been partly attributed to the ease of sequencing the AR, but the increase might easily have been even larger; this is due to the fact many mutations are not reported in the literature unless they are unique. Further, even unique mutations may not be reported, as they may have been found as a result of routine blood tests that are administered solely to establish a diagnosis of AIS. Currently, it is still the policy of the curator not to accept submissions to the ARDB unless they have been accepted for publication, in order to ensure adequate quality control of the data.

There is still an unequal distribution of the mutations along the length of the AR, as shown in Figure 1. It has previously been suggested that these mutation-dense regions are hot spots that reflect the high density of mutable CpG sites in the region [Gottlieb et al., 1996]. It is also apparent that the types of mutations differ along the length of the AR. In particular, nearly all mutations in exon 1 (Fig. 1) cause complete AIS (CAIS), and nearly all are of the premature translation termination variety (Table 1), whether by direct mutation to a stop codon, or indirectly, by frameshifts after small deletions or insertions. To date, only 54 mutations have been reported in exon 1 of the AR in patients suffering from some form of AIS, despite the fact that it encodes more than half of the AR protein [Gottlieb et al., 1999], and even fewer have been reported in splicing and untranslated regions of the AR gene (Table 1). In the C-terminal ligand binding domain (LBD), there is a striking preponderance of missense mutations, with a significantly greater number of CAIS than partial AIS (PAIS) cases (Table 1).

The appearance of the database has been modified, and it is now presented in a portable document format (pdf; Adobe Systems, www.adobe.com) for ease of use. We have expanded the database by providing information on proven AR-interacting proteins; and a rough map of the region of the AR with which they interact (Fig. 2) is also now available on the database website, together with a table that lists details of the properties of each of these proteins. In the past few years, there has been a tremendous expansion in the number of these proteins, the number rising from 23 to 70 in just 3 years.

Of particular interest is the appearance of 12 silent mutations (nine in AIS, three in prostate cancer) in regions of exons not close to splice sites, which raised the possibility that mutations can cause problems at the mRNA level [Gottlieb et al., 1999]. Further, an increasing number of mutation entries (128 out of 605) now contain data proving the pathogenicity of the putative mutation by reconstituting the mutation and seeing its effect on a reporter gene, which greatly improves the quality of the data.

**Gain-of-Function Mutations**

**Prostate cancer.** To date, 85 AR mutations have been found in prostate cancer (CaP) tissue (MIM# 176807), almost all being single-base substitutions due to somatic mutations, rather than germ line mutations. These are now indicated in the database by being color-coded orange. As can be seen in Figure 1 and Table 1, the majority of somatic mutations occur in the LBD (≈45%), with a substantial number occurring in exon 1 (≈30%). Originally, it was thought that AR was not expressed in CaP tissues, but this does not appear to be the case [Edwards et al., 2003]. Considerable controversy has revolved around conflicting studies that only sometimes have found a significant number of AR mutations in CaP tissues [Culig et al., 2002]. It has been argued that AR mutations only appear during the latter stages of CaP and, in addition, some studies have indicated that antiandrogen treatments have resulted in AR mutations [Hyytinen et al., 2002]. This data is now incorporated into the database.

A considerable limiting factor in the value of this data is that, with a few exceptions, experiments to prove pathogenicity have not been done for CaP mutations, which considerably reduces the value of the data. In a recent functional analysis of disease-associated mutations of the AR, we have shown that most of
the cancer-associated mutations reported in the database have a significantly lower degree of base conservation when compared to the few cases where the pathogenicity of the mutation has been proven [Mooney et al., 2003].

**Kennedy disease.** Kennedy disease, or spinobulbar muscular atrophy (SBMA; MIM# 313200), a spinobulbar motor neuronopathy associated with mild AIS (MAIS), is one of the classic trinucleotide repeat expansion diseases that cause inherited neurogenerative disorders. It is caused by expansion of the glutamine-coding (CAG)₈₋₃₅ CAA tract in exon 1 of the AR to a total number of at least 38 trinucleotide repeats [Pinsky et al., 2001]. The MAIS component of SBMA may reflect a loss of AR transcriptional regulatory activity by virtue of a pathologically-expanded polyglutamine (polyGln) tract. It should be noted that in SBMA, the AI phenotype is quite variable. Since subjects with CAIS, including those with complete AR deletions, do not develop SBMA, this knowledge mandated the logic that the polyCAG-expanded AR or the polyGln-expanded AR protein is somehow motor neuronotoxic by a gain, not a loss, of function.

The biochemical, histopathologic, and neurophysiologic features of SBMA are, unremarkably, those secondary to motor denervation. A number of possible causes for this gain of function are now listed on the database website (Table 2).

**AR Protein Structure–Function Relationships**

In an effort to better understand the structure–function relationship of how specific mutations in the LBD cause AIS, the 3D structure of the AR LBD has been added to the database. We have produced this model with reference to X-ray crystallography data [Matias et al., 2000]. This is particularly important in the case of mutations that alter amino acids, but clearly less so for mutations that cause premature terminations. The model revealed a structure that consisted of 12 α-helices (Fig. 3). To understand the possible affect of each specific gene alteration, we have labeled our model to show the exact position of all 12 α-helices (Fig. 3). Thus any mutation can be physically placed in the actual 3D protein structure. It can be seen that the putative crystal
structure shows very few residues that are in close contact with the ligand-binding pocket when modeled with the synthetic ligand R1881 (Fig. 3). The availability of this model allows researchers to realize that the position of the mutant residue near the ligand-binding pocket is not necessary to produce severe ligand binding problems. Therefore, to try to elucidate how such mutations could potentially affect the structure of the ligand-binding pocket, and so explain the lack of ligand binding in these mutants, we have recently used molecular dynamic modeling techniques over extended periods of time (up to 4 nsec) to, in effect, create 4D structures of AR mutants [Wu et al., 2003]. In this study, a CAIS mutation some distance from the ligand-binding pocket produced a local structural distortion that also affected the ligand-binding pocket conformation. It is our intention to incorporate this type of information into the database in the near future. We believe that this will be particularly important where the pathogenicity of the mutation has not been proven.

Another aspect of the structure–function relationship is variable phenotypic expression, in which identical

### Table 1. Nature and Distribution of Unique AR Mutations That Cause Disease

<table>
<thead>
<tr>
<th>Loss of function disease</th>
<th>Type of mutation</th>
<th>N-terminal domain*</th>
<th>DNA-binding domain*</th>
<th>Hinge region*</th>
<th>Ligand-binding domain*</th>
<th>Splice site</th>
<th>Intron</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAIS</td>
<td>Single base substitution</td>
<td>13</td>
<td>17</td>
<td>1</td>
<td>100</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Complete gene deletion</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Partial gene deletion</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deletion (1-6 bases)</td>
<td>8</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Insertion</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Duplication</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAIS</td>
<td>Single base substitution</td>
<td>4</td>
<td>18</td>
<td>1</td>
<td>60</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multiple base substitution</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deletion (1-6 bases)</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAIS</td>
<td>Single base substitution</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Partial gene deletion</td>
<td></td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deletion (1-6 bases)</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*aa 1-534; *baa 559-624; *aa between DBD and LBD 625-663; *aa 664-919; *Somatic mutation.
mutations produce different phenotypes. The mutations showing some degree of variable expressivity now numbers 28; these are indicated in the database by being shown in green. Further, it is highly probable that many more cases of variable expressivity exist but, due to the limited phenotype descriptions available, they are not always apparent. A number of possible causes of variable phenotype expression including somatic mosaicism have been discussed previously [Gottlieb et al., 2001].

**AR CAG Tract Length Variation as a Risk Factor for Disease**

The database now lists the length of the CAG (polyglutamine) and GGC (polyglycine) tracts that are present in exon 1. At the present time, there are 58 entries in the database that show the lengths of these two tracts. The mean value of the CAG tract length is 22.34, which is significantly longer than in controls [Elhaji et al., 2001], though the number of database entries is too small to draw any specific conclusions. However, as more data becomes available, it will be interesting to see if increases in CAG repeat length, which reduce the efficacy of the AR [Mhatre et al., 1993], play any role in determining the AIS or CaP phenotype.

In the past few years, a number of studies have examined possible relationships between the length of the CAG repeat and the risk of getting certain diseases and conditions. These include: female breast cancer (MIM# 114480) [Elhaji et al., 2001]; male infertility [Casella et al., 2003]; prostate cancer, reviewed by Ferro et al. [2002]; uterine endometrial cancer (MIM# 608089) [Sasaki et al., 2000]; colorectal cancer [Ferro et al., 2002]; and esophageal cancer (MIM# 133239) [Dietzsch et al., 2003]. Table 2 has been added to the database that lists details of all of these diseases.

### DISCUSSION

**Germline Vs. Somatic Mutations**

A most significant issue facing LSBD curators and designers in the future is likely to be the calculating the relative genetic significance of somatic mutations, as opposed to germline mutations, particularly in so called “late-onset” diseases. At the present time, the only databases that have found it necessary to distinguish between somatic and germline mutations are those of certain cancer-associated genes, such as adenomatous polyposis coli (APC) [Laurent-Puig et al., 1998].

### TABLE 2. Diseases Associated With AR CAG Tract Length Variation

<table>
<thead>
<tr>
<th>Disease</th>
<th>CAG tract</th>
<th>Androgen sensitivity</th>
<th>Gain of function—possible causes</th>
<th>Symptoms</th>
<th>Reference</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Indirect association</th>
<th>Relative length of tract*</th>
<th>Associated risk factors</th>
<th>Comments</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate cancer</td>
<td>Shorter</td>
<td>Increased</td>
<td>1. Ethnicity</td>
<td>Inconclusive studies—possible somatic alterations</td>
</tr>
<tr>
<td>Male infertility</td>
<td>Longer</td>
<td>Reduced</td>
<td>2. Family history</td>
<td>Inconclusive studies—possible somatic alterations</td>
</tr>
<tr>
<td>Female breast cancer</td>
<td>Longer</td>
<td>Reduced</td>
<td>Ethnicity</td>
<td>BRCA1 mutation carriers</td>
</tr>
<tr>
<td>Endometrial cancer</td>
<td>Longer</td>
<td>Reduced</td>
<td>BRCA1 mutation carriers</td>
<td>Somatic alterations</td>
</tr>
<tr>
<td>Colon cancer</td>
<td>Shorter</td>
<td>Increased</td>
<td>Selective growth advantage—somatic alterations</td>
<td>Inconclusive studies—somatic alterations</td>
</tr>
<tr>
<td>Esophageal cancer</td>
<td>Shorter</td>
<td>Increased</td>
<td>Ethnicity</td>
<td>Inconclusive studies—except in African males</td>
</tr>
</tbody>
</table>

*Relative length of CAG tract compared with control populations.*

**FIGURE 3.** The AR ligand-binding domain as defined by the X-ray crystal structure (1e3g). Note that the α helices are numbered and the codon number at the start and finish of each helix is given. Spheres illustrate the bound ligand, R1881.
significance of these observations is primarily limited to whether or not a particular mutation is inherited.

In discussing the significance of somatic mutations in LSDBs, one possible consideration is that their effect on phenotype expression will be different from germline mutations. As reported in the ARDB, the somatic mutations found in CaP tissues seem to elicit a gain of function, as opposed to a loss of function. This results in individuals that have a normal male phenotype, except that they suffer from CaP. Thus, somatic mutations produce a degree of genetic heterogeneity within an individual organism. In fact, in the case of CaP tumors, which (like most tumors) consist of heterogeneous tissue, there is often a wide range of cell types within each tumor, ranging from normal to advanced cancerous, possibly indicating a concomitant genetic heterogeneity within the tumors. Thus mutation databases will have to reflect such possible genetic variation, by much more closely identifying the tissue phenotype associated with a specific gene alteration, whether mutation or polymorphism. Fortunately, the arrival of techniques such as laser capture microdissection (LCM) have allowed us to examine genes in very specific cell and tissue types. Recently we have used LCM in a study into variation in AR CAG repeat length in prostate cancer tissue [Gottlieb et al., 2002], and it is our intention to eventually expand the database to include specific tissue phenotypes when listing patients with somatic mutations.

In examining the relationship between somatic mutations and disease, it is not insignificant that CaP is considered a late onset disease. This might be expected, as studies have shown that the rate of somatic mutations may increase with age [Jackson and Loeb, 1998], possibly due to genomic instability. This suggests that by incorporating data into the ARDB about changes in AR CAG repeat length, which is known to be quite unstable over time [Zhang et al., 1994], it might be possible to provide important insights into how the disease progresses. Further, age considerations could also be significant in other late-onset diseases, including other cancers. Indeed, in a recent study, we have examined the possible association between CAG repeat length changes and the onset and progression of female breast cancer [Elhaji et al., 2001].

Recently, polymorphisms in DNA repair genes have been identified as a risk factor for a number of cancers [Goode et al., 2002]. What is particularly interesting for the ARDB, is that recent studies of DNA mismatch repair enzymes (MMR) in CaP tissues showed a decrease in the activity of these enzymes, with low expression of some of the MMR proteins [Yeh et al., 2001; Chen et al., 2001]. Incorporating information into the ARDB on MMR activity in CaP tissues that have AR gene alterations, could possibly lead to a further understanding into the genetic events that may lead to the initiation and progression of CaP. Clearly, the more information of this type that can be accumulated in LSDBs, the more likely it will be that clearer associations can be identified between the genetic events that lead to disease phenotypes such as cancers. This is particularly true in the age of genomics, in which mass screenings entail looking for many different genes associated with a particular condition. Perhaps it makes more sense to look for associations that might exist between genes that have already been identified with a particular condition, rather than looking at all possible random associations. Thus, in the case of genes having LSDBs, the genetic and phenotypic information contained within these databases would make such genes natural association partners for any putative genes that have been found by genomic screening to be associated with a similar disease phenotype.

**CONCLUSION**

At the present time, most LSDBs have been created with the intention of listing gene mutations and their phenotypic expression, in order to correlate specific phenotypes with specific genotypes. The emergence of somatic mutations as genetically significant events indicates that it may be necessary to adapt LSDBs to incorporate additional data on somatic mutations, genetic heterogeneity, and variable expressivity. It is becoming clear that the role of LSBD curators is becoming more, rather than less important, particularly as it relates to using LSDBs as tools to help in understanding specific structure–function relationships. In a previous discussion on the role of LSDB curators [Gottlieb et al., 1999], we stated that it was time for LSDB curators to seize the initiative and start to determine the nature of the data that needs to be part of their LSDBs. If curators set the criteria, it seems reasonable to believe that researchers will ultimately follow them. It has often been said that in any database, the quality of the data that you get out of it is only as good as the quality of the data that you put into it. Up until now, the quality control most curators have undertaken has been aimed at ensuring the accuracy of the data they have entered. However, perhaps it is now time for curators to take a more proactive role, particularly in light of the likelihood that, in the future, mutation databases are going to be playing a much more significant role in both biological and medical research.

**ACKNOWLEDGMENT**

This work is funded by grants to B.G. from the Canadian Institutes for Health Research.

**REFERENCES**


Goode ETAL.

ACKNOWLEDGMENT

This work is funded by grants to B.G. from the Canadian Institutes for Health Research.

REFERENCES


