Standard Nomenclature in Bacterial Genetics

The genes of *E. coli* are given names following a standard system of nomenclature first published as a set of semi-formal guidelines by Demerec *et al.* (1966). Familiarity with the system is essential for reading the literature of bacterial genetics, and you are expected to follow the basic conventions in your lab reports. The standardized gene names are powerful tools for searching the literature and genomic databases.

The purpose of this exercise is to become conversant with the basic genetic nomenclature for *E. coli*, and to use it in exploring several features of *E. coli* genetics in a genome browser (EcoCyc).

**DESCRIBING GENES**

Each gene of the *E. coli* genome is designated by a 3-letter, lower case, italicized symbol. Ideally, the gene symbol serves as a mnemonic, or acronym, for the function of the gene or gene product. When there are several genes whose products have similar functions, the genes are distinguished from each other by an italicized capital letter immediately following the 3-letter symbol. Often these letters are assigned to the genes alphabetically in the sequence in which the genes were discovered. Occasionally, letters are used to specify a particular function. For example, the superscript “R” can be used to denote a gene for a regulatory protein and “O” is often used to designate operators.

**EXAMPLE:**  
*thrC*  
A gene whose product is one of several proteins involved with the threonine biosynthetic pathway.

Historically, many *E. coli* genes were discovered and named long before the function of the gene product was understood. So, in the early days of bacterial genetics the gene symbol was usually based on the *phenotype* associated with the expression of known mutant alleles of the gene. Many of these original gene names have been subsequently changed as the functions the gene products have been elucidated. This creates the unfortunate situation in which the same gene may be referred to by different names (synonyms) in the older and newer literature. (See the discussion of *rpsL* below.)

The completed *E. coli* genome sequence revealed hundreds of previously undiscovered genes with no known function. As an expedient, these genes were provisionally assigned 3-letter symbols beginning with "y". These provisional symbols have the form y_ _ _, in which the second and third letters are based on the location of the gene in the genetic map. Currently (2005) there are 1,063 genes of this form (34% of all *E. coli* genes) ranging from *yaaA* at 0.1 min. to *yjjY* at 100’ of the genetic map. These symbols will be changed when the function of the gene product is understood.

Specific mutant alleles of a given gene are designated by a numeral immediately after the symbol. Assignment of allele numbers is difficult to organize when several different labs are independently obtaining mutant alleles of the same gene. In this introductory class we largely dispense with allele number designations.

**EXAMPLE:**  
*xyzB1* a specific mutant allele of the wild-type gene *xyzB*

A few other terms that may crop up:

"Locus" refers to a site within a genome (often a gene) that has been identified and mapped by genetic analysis. Frequently, the term locus can be used interchangeably with the term "gene" without distorting the meaning of a statement. The term locus is falling into disuse among the current generation of microbial geneticists, but you will encounter it in the older literature.

"Open Reading Frame (ORF)" refers to a genetic sequence that is translated into protein. Again, the term ORF can often be replaced by the term gene without altering the essential meaning of a statement. ORF is also a quasi-synonym for "cistron", another formerly fashionable term that has become increasingly rare in the current literature.
DESIGNATING PHENOTYPES

Phenotype designations are given as a three-letter, non-italicized symbol with the first letter capitalized. Usually, the 3 letters are the same as those used to designate the corresponding genes, so it is easy to confuse gene designations with phenotype designations.

+/- superscripts are used to distinguish the wild-type phenotype (+) from a mutant, often a null phenotype (-).

EXAMPLE: \textit{Thr}^{+} The ability to synthesize the amino acid threonine (a phenotypic trait).

Wild-type \textit{E. coli} have a \textit{Thr}^{+} phenotype. (Sometimes we say that \textit{E. coli} is a threonine prototroph.) Null mutations in the gene \textit{thrC}, as well as several other \textit{thr} biosynthetic genes, produce a \textit{Thr}^{-} phenotype (i.e. Threonine must be provided in the growth medium of \textit{Thr}^{-} strains)

The superscripts R and S may be used to denote resistance or sensitivity to an antibiotic or other antimicrobial agent.

EXAMPLE: \textit{Rif}^{S} Sensitivity to the antibiotic Rifampicin.

\textit{Tsx}^{R} Resistant to infection by the bacterial virus T6.

DESIGNATING GENE PRODUCTS

Gene products (proteins) are designated by the same symbol as the gene, non-italicized, and with the first letter in upper case. These do not have superscripts.

EXAMPLE: \textit{ThrC} The enzyme O-phospho-L-homoserine phospho-lyase, product of the \textit{thrC} gene, and essential for the \textit{Thr}^{+} phenotype.

DESIGNATING STRAINS

Nearly all the \textit{E. coli} strains used laboratory research, and in this course, are derived from a strain designated \textit{E. coli} K-12 that was isolated in 1922 from a patient at Stanford University hospital and has been maintained in laboratory culture ever since. (BTW: The patient's medical condition had nothing to do with \textit{E. coli}.)

Strain designations also frequently cite the presence of accessory genetic elements such as plasmids or bacteriophage genomes. The original K-12 strain contained both a large plasmid called “F” and the bacteriophage \textit{\lambda} (Lambda) genome. Therefore it is properly written as \textit{E. coli K12 F}^{+} (\textit{\lambda}^{+}).

Each lab strain of \textit{E. coli} has a specific strain designation given to it by the laboratory that derived the strain. The derivation of the strain and its genotype are published the first time the strain appears in the literature. In all subsequent publications involving this strain, the publication describing its derivation is cited among the references. In principle, this system allows us to follow the genealogy (i.e. evolution) of any laboratory \textit{E. coli} strain back through the literature to the original strain isolated in 1922.
MORE EXAMPLES:

**lacZ**  Probably the most famous of all *E. coli* genes, **lacZ** is part of the "lac operon" and codes for the gene product **LacZ** (the enzyme β-galactosidase) which is required for growth on the sugar lactose (a **Lac⁺** phenotype). Null mutations in **lacZ** produce a **Lac⁻** phenotype, meaning the strain does not metabolize lactose (because there is no functional **LacZ** gene product.

**metB1**  **metB** is one of several genes whose products are enzymes that participate in the biosynthesis of the amino acid methionine. **metB1** is a specific mutant allele of **metB** that eliminates the function of the gene product (i.e. it is a null allele). Wild-type *E. coli* grows on media without methionine; i.e. it has a **Met⁺** phenotype. *E. coli* strains with the **metB1** allele have a **Met⁻** phenotype.

**rpsL**  The gene **rpsL** codes for one of the proteins in the small subunit of the *E. coli* ribosome. "**rps**" is a mnemonic for "ribosomal protein small subunit". Genes for other proteins in the small subunit are designated **rpsA**, **rpsB**, etc.

The first known mutant allele of **rpsL** led to *E. coli* that were resistant to the antibiotic streptomycin (phenotype=**StrR**). Streptomycin kills the streptomycin-sensitive (**StrS**) wild-type *E. coli*. Because this phenotypic property was the only thing known about the gene when it was originally discovered and mapped, **rpsL** was originally named **strA** and the original allele for streptomycin resistance was designated **strA1**. Thus, **strA** and **rpsL** are synonyms. Some authors, in a valiant attempt to overcome this confusion, will refer to the gene as **rpsL (=strA)**.

**ybcF**  This gene symbol refers to a gene with unknown function. 'b' (the second letter of the alphabet) indicates a genetic map position between 10 and 20 minutes. 'c' (the third letter of the alphabet) indicates a genetic map position between 12 and 13 minutes. "F" (the sixth letter of the alphabet) indicates this is the 6th gene of unknown function located in this region.

**MG1655**  Designation of the strain used in the *E. coli* genome sequencing project. This strain is very closely related to the original K1-2 strain isolated from nature, but without the F plasmid and Lambda bacteriophage. Therefore it would be designated *E. coli* **MG 1655 F⁻ (λ⁻)**.

**Summarizing:**

**pro**  symbol for a gene whose function has something to do with the cellular process of proline biosynthesis; implies that there is only a single gene of this class

**proB**  symbol for a gene whose function has something to do with the cellular process of proline biosynthesis; implies that there are other genes of this class

**proB3**  symbol for a specific mutant allele of the gene **proB**

**Pro⁺**  designation for the wild-type phenotype associated with function of the gene, the ability to synthesize proline

**Pro⁻**  designation of the phenotype associated with loss of function of a gene in the **pro** class (the **null** phenotype), inability to synthesize proline (i.e. If you want to grow a **Pro⁻** strain, you must provide proline in the culture medium.)

**ProB**  The protein product of the gene **proB**; the enzyme γ-glutamyl kinase [EC 7.7.2.11], responsible for the first committed step in the proline biosynthetic pathway.
REFERENCES (available on the course web site)

Demerec et al. (1966) Genetics, 54, 61.
A Proposal for a Uniform Nomenclature in Bacterial Genetics

- This is the original source of the system used for *E. coli* genetic nomenclature.
ASSIGNMENT (100 Points)

Questions #1-5 and #8 are 10 pts. each.
Questions #6 and #7 are 20 pts. each.

If you do the optional question (#9) you can get 10 points to make up for any you have missed on #s 1-7; but you can’t go above 90 pts. max.

1. What would each of the following suggest to you?

ara
araC
araC4
Ara+
Ara-
araC

Hint: Wild-type E. coli can metabolize the carbohydrate arabinose.

2. When I was a new graduate student, I mapped the location of the E. coli gene dnaC.

Go to the Resources page of the course web site and do an EcoCyc search on "dnaC" to get the following information:

The approximate position of this gene in the E. coli genetic map in minutes (= centisomes).
The identities of the genes on either side of dnaC.
The general function of the DnaC gene product.
The size (# amino acids) of DnaC protein.

3. How many E. coli genes have the 3-letter designation "rps"?

Try a search in the EcoCyc.

What do all of the RPS_ gene products of these genes have in common?

4. What about genes for proteins in the large (50S) ribosomal subunit? How many are there, and what are their 4-letter symbols?

(Hint: You might want to check that your answer for the number of genes is consistent with the known number of proteins in bacterial ribosomes. This information should be readily available in any standard text in genetics, molecular biology, or cell biology; not to mention the internet.)
5. What would be the genetic map location of the gene *yeaH*?

According to EcoCyc, is there really a gene with this designation in the *E. coli* genome?

6. This is a challenging one, but it will help you become more familiar with the EcoCyc database and perhaps lead you to a startling discovery!

In the EcoCyc database, search on the gene *yehU* in the genome of *E. coli* K-12 strain MG1655.

(Strain MG1655 is the standard "wild-type" laboratory *E. coli* strain and should be the default organism selected when the database opens.)

Record the position of the 5’ nucleotide position for the gene in the complete MG1655 sequence. (The direction of transcription is shown in the diagram further down the page, or in the genome browser.)

Scroll down the page and click the "Select Allowed Organism" button.

This brings up an extensive list of bacterial species for which complete genome sequences are available. Note that *E. coli* strain K-12 MG1655 is already shown as the default selected genome in the right-hand panel.

Among the available genomes find *E. coli* strain O157:H7 EDL933. EDL933 is a notorious human pathogenic *E. coli* strain responsible for deadly outbreaks of hemorrhagic colitis (bloody diarrhea). Add EDL933 to the list of selected genomes and click OK. This returns you to the *yehU* gene page.

Scroll down again and click the "Align in Multi-Genome Browser" button.

You are now getting a comparison of the two genomes centered on the region containing *yehU* in strain MG1655.

Does the EDL933 genome have an equivalent gene to *yehU* in MG1655?

If so, record the position of the 5’ nucleotide position for the gene in the complete EDL933 sequence. (Try zooming in far enough to read the positions directly from the scale in the browser.)

Now identify the gene that is immediately to the right of *yehU* in the MG1655 sequence.

Now use the multi-genome browser function again to compare the genomes in the region centered on this gene in MG1655.

You will find that ED933 has a gene equivalent to (orthologous to) the gene in MG1655, but it is annotated with a different name. (Hint: If you do a mouse rollover of a gene, a popup window provides information about a gene's functional annotation. This should help you conclude that the genes in the 2 strains are orthologous even though they have been given different names.)

What is the name of the EDL933 gene orthologous to *mlrA*?

Record the position of the 5’ nucleotide of this gene in the 2 strains.

What is the distance (in bp) between the 5’ ends of these two genes in their respective genomes?
Do an EcoCyc search for the gene *stx1* in the genomes of *E. coli* strains MG1655 and EDL933 and record the position of the gene.

Do a mouse rollover to look at the function of *stx1* and of the genes that flank it.

Does it seem to you there is something strange going on here?

Many strains that we work with in this course have a deletion of the wild type chromosome that is formally symbolized Δ(*gpt lac*). The Greek letter delta designates a deletion. The two genes in parentheses indicate the first and last genes known to lie within the deletion. What % of the *E. coli* genome (approximation) is deleted in Δ(*gpt lac*)?

Approximately how many genes lie within this deletion.

OPTIONAL for the cloning obsessed! (10 additional points available to make up for any that you missed above.)

*E. coli* strain DH5 is among the most widely used descendants of *E. coli* K-12 in modern molecular biology. It was developed as an optimal host for transformation by recombinant plasmids (i.e. cloning vector plasmids such as pUC19 carrying “foreign” DNA inserts). Three specific mutations have been introduced into the genotype of strain DH5 which make it exceptionally efficient in the job of acquiring and maintaining these recombinant plasmid vectors.

*hsdR17*

*recA1*

*endA1*

For any ONE of these 3 mutations, determine the function of the wild-type gene product and suggest briefly why (or how) a null mutation in that gene could affect transformation efficiency.

It is fairly straightforward to look up these genes and find out what the gene products do. It is more challenging to understand why mutations in these genes make strain DH5 a better host for recombinant plasmids. You may need to review the basic scenario for molecular cloning in *E. coli* plasmid vectors.