Using Raw Sequence Data to Analyze Tumor Suppressor Mutations

James D. Fackenthal

1Department of Biological Sciences, Benedictine University

Abstract

Learning to identify DNA sequence variants from raw sequencing data (electropherograms) is a basic molecular biology skill with applications to basic research and clinical practice. Yet learning to map DNA sequence variants to functional regions of genes requires knowledge of gene structure and function, bioinformatics, and sequence analysis software. This exercise uses real-world examples of DNA sequence variants seen in tumor suppressor genes from families with various inherited cancer syndromes.

After seeing a practice example of mutation characterization, students are given information about a hypothetical cancer patient including a description of the disease, family history, and a short segment of raw DNA sequence from a relevant tumor suppressor gene showing a heterozygous mutation. The student's responsibility is first to determine the structure of the gene using bioinformatics resources including GeneCards, Online Mendelian Inheritance in Man (OMIM), the Unified Protein Knowledge Base (UniProtKBase), Ensembl, GenBank, and RefSeq. Second, the student will use the UniPro UGENE genomic analysis software to determine the location of the mutation within the gene, the sequence of the mutant allele, and predict the functional impact of the mutation. This exercise allows student to practice working with genomic databases and genomic sequence analysis software, introduces important topics in clinical genetics diagnostics, and reinforces classroom concepts relating to the functional roles of parts of the gene and pre-mRNA.

Learning Goals

Students will:

◊ learn how DNA sequence information is used to identify and characterize mutations in cancer syndrome families.
◊ learn how to navigate databases to understand annotated gene structures, functions, and universal identifier codes.
◊ learn how to use genomic visualization software to visualize and manipulate DNA sequences, understand the functional consequences of mutations with respect to gene structure, and report mutations in a standard format to clinicians and researchers.
◊ From the Bioinformatics Learning Framework:
  » What is the role of computation in hypothesis-driven discovery processes within the life sciences?
◊ From the Genetics Learning Framework:
  » How do different types of mutations affect genes and the corresponding mRNAs and proteins?
  » What experimental methods are commonly used to analyze gene structure and gene expression?

Learning Objectives

Students will be able to:

◊ explain why genetic counselors recommend DNA sequencing of certain tumor suppressor genes for patients from families with familial cancer syndromes.
◊ describe how dideoxy sequencing is performed and how electropherograms are interpreted.
◊ characterize their own "mystery" mutation with respect to their gene's chromosomal location, structure, functions, and reference numbers for other databases, while navigating GeneCards and ReSeq sequences in the NCBI GenBank.
◊ use Unipro's UGene genome visualization software to visualize their gene's sequence in the context of gene structure.
◊ localize, interpret, and name their aberrant DNA sequences according to the HGNC reporting conventions.
INTRODUCTION

There are many inherited disorders that are routinely screened upon recommendation by genetic counselors. Many of these are caused by multiple chromosomes, missing chromosomes, chromosomal duplications, deletions, translocations, and inversions, all of which may be identified by cytogenetic analysis. Others, however, result from point mutations that can only be characterized by DNA sequence analysis (1, 2). Such disorders include cystic fibrosis, Tay-Sachs disease, and hemophilia (3–5). In this exercise we focus on mutations in known tumor suppressor genes that result in increased cancer risk. Patients with a strong family history of early-onset cancer or multiple occurrences among first-degree relatives are often referred for gene sequencing. Understanding whether a cancer case is associated with an inherited mutation may inform affected individuals about the likelihood of developing a second primary cancer and help family members understand their own risks and the appropriate screening regimens. Several large companies provide sequencing services that not only provide clinical information to patients but also participate in research that contributes to our understanding of cancer risk and progression.

Learning how to interpret the functional consequences of germline DNA sequence variation can be challenging in both educational and professional settings. Students are taught the regulatory and coding nucleic acid sequences required for transcription, mRNA processing, and protein translation, but are rarely given the opportunity to test their understanding of these concepts with real-world clinical applications. Likewise, researchers training to perform whole-genome sequencing to identify mutations in disease-associated genes in clinical samples require hands-on experience to predict the effects of the DNA sequence variations they identify. In the exercise described here, students will learn to understand dideoxy DNA sequencing analysis, practice navigating genomic data sources, learn the nomenclature conventions for human genes and messenger RNAs, use sequence analysis software to distinguish wild type from variant DNA sequences, and determine whether the sequence variant affects the transcription, splicing, or translation. This, in turn, allows students to predict whether the sequence variants seen in a clinical sample may indicate an increased risk for developing a disease.

Dideoxy sequencing analysis: While advanced high-throughput technologies are the methods of choice for whole-genome sequencing and simultaneous analysis of large sample numbers, traditional dideoxy sequencing remains the most accessible method for identifying mutations in single genes from individual subjects. This method, developed by Sanger in 1977 (6) and modified to make use of fluorescently labeled molecules (7), uses fragments of DNA amplified by the polymerase chain reaction (PCR) and mixtures of fluorescent dideoxy nucleotides in in vitro DNA synthesis reactions to produce a population of single stranded DNA molecules, each terminating with a fluorescent base, which are size-separated and detected using capillary electrophoresis, generating a graphical image of colored peaks (an electropherogram) representing the DNA sequence of the original PCR fragment. For typical clinical analysis of an individual gene, sequence analysis is performed for each exon, some intronic regions near the intron-exon boundaries and, in many cases, the gene promoters. In regions where the subject is heterozygous for a DNA sequence variant, there will be two different colored peaks at the same position representing the different alleles of the two gene homologs. A single position will have two peaks when the alleles differ by a single base change, and a series of positions will have two peaks when the variant allele is a small insertion or deletion.

Searching bioinformatic databases: To distinguish the wild type from variant alleles and to locate the position of the potential mutation within the structure of the gene, it is necessary to learn about the gene structure and the reference nomenclatures used for the gene, its most common transcripts, and its most common protein isoforms. There are numerous bioinformatics resources containing this information, each with its own advantages and disadvantages. Several are devoted to human genes with known disease associations. The GeneCards site is a useful starting point for learning basic information about a clinically important gene and finding the various reference numbers for that gene used in other databases, including The Human Genome Organization (HUGO) Gene Nomenclature Committee (HGNC), Online Mendelian Inheritance in Man (OMIM), the Unified Protein Knowledge Base (UniProtKB), and the European Bioinformatic Institute’s Ensembl genome browser. The Entrez database and retrieval system is useful for retrieving information from large databases like GenBank and RefSeq, but results from searching these sources often yield too much information that can be difficult to sort through. One way to narrow down the identifiers for individual genes and major transcript variants is to use the Locus Reference Genomic (LRG) website. This site allows users to find NCBI GenBank reference numbers for genomic DNA, major mRNA isoform, and major protein isoforms for single human genes of interest without the clutter of irrelevant search results.

Bioinformatics software: Once unique identifiers for sequences of interest have been determined, they can be used to search and download sequences into analysis software. There are numerous software packages available for visualizing the structural features of genes aligned with DNA, RNA, and protein sequences. Some exist online while others can be downloaded to be used as desktop applications. The UGENE software (Unipro) is a powerful package of programs that allows sequence visualization and manipulation, as well as high-throughput/next-gen data analysis and various assembly/construction projects. UGENE can be downloaded and run on any platform free of charge. Using unique gene identifiers as search terms, students can download gene sequences, locate the wild type sequence in the region of their variant sequence, and determine the molecular nature of the sequence variant. As the wild type sequence is displayed in the context of the exon structure and translational reading frame, students can determine whether their mutant sequence affects correct splicing of the pre-mRNA or the coding region of the mature mRNA. These are essential skills for understanding modern basic biological and clinical research.

Intended Audience

This tutorial and worksheet were devised in the Fall of 2020 at Benedictine University, a university aimed at teaching with limited research emphasis, for use in a Genetics Laboratory
course. Since then it has been used in six sections over the course of four semesters with minor modifications, with classes typically consisting of Biology and Health Science majors from sophomore to senior levels. It is expected that students will have had freshman-level introductory biology courses that cover the basics of molecular genetics. The aim of the lesson is to complement a sophomore-level introductory course in genetics by providing an exercise in applied understanding of the structure of genes, the nature of mutations, and a clinical application of mutation detection, characterization, and reporting. It could also be used in courses designed for career advancement of genetic counselors or technicians working in clinical laboratories.

**Required Learning Time**

Before class, students will be required to download Unipro’s UGENE onto their laptops (8). This can take anything from minutes to an hour. The instructor will spend some time before class sending each student their own case to solve (Supporting Files S1, S2). The lesson was designed to be completed in a single 170-minute lab session, and the first half hour may be used for an introductory lecture.

**Prerequisite Student Knowledge**

Students are expected to understand the functional roles of each part of a eukaryotic gene, including (but not limited to) the promoter, the open reading frame, and the intronic sequences required for correct splicing. Students are also expected to understand how germline nuclear mutations are passed on to half the descendants of each carrier.

**Prerequisite Teacher Knowledge**

Instructors should be familiar with the biological concepts described above for students. Additionally, instructors should be familiar with: (i) the chemistry of fluorescence dideoxy DNA sequencing, (ii) the information available on GeneCards, (iii) the Human Genome Organization Gene Nomenclature Committee system for naming genes and their sequence variants, (iv) how to search for information about individual genes using various tools available from the National Library of Medicine, especially GenBank, (v) how to create, edit, and visualize gene and cDNA sequences using the UGENE genome visualization software.

**SCIENTIFIC TEACHING THEMES**

**Active Learning**

After the introductory lecture (Supporting Files S3, S4), the students work through the exercise on their own (Supporting Files S1, S2, S5, S6). The instructor circulates to check progress and answer questions. As the assignment may be completed any time prior to the subsequent meeting, students typically meet outside of class to answer each other’s questions.

**Assessment**

The worksheet is typically turned in at the beginning of class the following week. A sample grading scheme is provided (Supporting File S6), but may be adjusted. After the worksheet is turned in, a quiz is given testing the students’ knowledge of tumor suppressor biology, pedigree analysis, dideoxy sequencing, bioinformatics, and electropherogram interpretation. The course for which this exercise was developed includes a comprehensive midterm and final exam, which will also include material from this exercise.

**Inclusive Teaching**

Access to health care services that provide cancer screening, diagnostics, and advanced therapies is not discussed in the class. However, students will understand that the usefulness of a pedigree depends on relatives’ willingness to share their medical histories. There are widely varying cultural stigmas and taboos associated with cancer that may limit access to the information necessary for a recommendation of genetic testing, and some students may have troubling questions about their own risks (9). Additionally, there are varying cultural differences facing the challenges of informing patients’ first-degree relatives of the likelihood of carrying a high-risk mutation and the advisability of seeking genetic testing (10). Finally, different students may have different levels of personal experience with familial cancers, and there may be different levels of discomfort associated with clinical discussions of cancer risk. Instructors are advised to be aware of these different responses to personal and family challenges. It has been suggested that media coverage of celebrities who are open about their own cancer histories helps individuals deal with anxieties about their own situations (11–13).

**LESSON PLAN**

**Overview (Table 1)**

Prior to meeting in class, the instructor will distribute cases consisting of case histories, pedigrees, and electropherograms by some electronic means (Supporting File S1). Students will load Unipro UGENE software onto their laptops prior to class. During class, the lesson consists of a brief (~30 minute) presentation on how dideoxy DNA sequencing works and how electropherograms are interpreted (Supporting Files S3, S4). The rest of the lesson consists of students working the worksheet on their own while the instructor circulates in the classroom (Supporting File S5). The student will work through the following exercises first with a practice gene (NFI), then repeat the exercise with the gene they were given. Students will (i) acquire the GenBank reference numbers for both the genomic and cDNA files for their genes into UGENE, (ii) using the electropherogram they were given, search for the location of their mutations within the sequence of both the genomic DNA and the cDNA, (iii) determine the molecular defect associated with the mutant allele of their gene, (iv) give the mutation the correct name according to its location in the Human Genome Organization (HUGO) Gene Nomenclature Committee (HGNC), and (v) answer some brief questions relating to general knowledge of the gene function. The lesson is largely self-explanatory from the introductory slides and the worksheet. However, an instructor may wish to explore some material beyond what is required of students to do the exercise, especially the uses of the bioinformatics websites and desktop genomics software.
Using Raw Sequence Data to Analyze Tumor Suppressor Mutations

Introductory Slides (15–30 min)

Dideoxy Sequencing (Supporting Files S1, S2)

The slides and explanations in Supporting Files S3 and S4 may be shown in class before the students begin working on the worksheet. This part of the lesson typically requires multiple explanations, which is why it is reviewed in a slide presentation before students read the same material in the worksheet. In cases where instructors are not familiar with Sanger dideoxy chain termination nucleotide sequencing, it may be advisable to review the method in order to better explain the introductory figure in the introductory slides (Supporting Files S3, S4). Historical and review information may be found in Sanger et al. (6) and Shendure et al. (14).

Worksheet (~2 Hours; Supporting File S5)

GeneCards

This section is part of the worksheet that students complete in class or on their own. GeneCards is described in the worksheet as a “one-stop shopping” place for genetic information. It is included in this worksheet to introduce students to the wealth of information available about any gene, mRNA, or protein of interest. Examining the information available for any gene allows the student to become familiar with useful tools for future projects. It also introduces students to the challenges of organizing information different ways for different purposes. For example, geneticists interested in the location of a gene may have different criteria for classifying and organizing information from those interested in gene structure or gene function. Thus, the amount of information provided by GeneCards for any student may seem overwhelming. The worksheet encourages students to explore the resources found on GeneCards using the NF1 gene an example. These resources include the Online Mendelian Inheritance in Man (OMIM), the Unified Protein Knowledge Base (UniProtKB), Ensembl, Entrez, and RefSeq. However, the specific piece of information needed for the next step in the worksheet is the GenBank reference number for the NF1 example and the students’ assigned gene, and the worksheet directs students to the National Center for Biotechnology Information to acquire it. After becoming familiar with the information in GeneCards, students will return to the information therein at the end of the exercise to learn more about their gene.

National Library of Medicine, National Center for Biotechnology Information

As with GeneCards, the amount of information available from the National Institute of Health (NIH) National Library of Medicine (NLM) National Center for Biotechnology Information (NCBI) website is extensive and potentially overwhelming. It is advisable for the instructor to become familiar with many of the resources available on this site in order to provide guidance to students who wish to apply their skills to other projects. However, the only information students need from this site to complete the exercise is the GenBank reference numbers for NF1 (the practice gene) and the assigned genes. This number is used to load the relevant genetic information into the genomic visualization software (UGENE). The instructor may wish to practice locating this number as instructed in the worksheet.

Unipro UGENE

The UGENE bioinformatics tool kit is open-source freeware produced by Unipro (8). It is a very powerful suite of tools that allow students to visualize, manipulate, and interpret genomic, mRNA, and protein data for any given gene. Students will load the GenBank reference numbers for both genomic and cDNA sequences for NF1 (the practice gene) and their own gene to create new sequence files displaying the sequences of interest. Students will then use the sequence searching routine to find the location of their mutation and determine the exact molecular nature of the base change. With this information, students can give their mutation a HUGO HGNC name that will be universally understood by human geneticists.

Final Questions

At the end of the worksheet, students are asked to answer questions about the nature of the mutation they found, including the molecular changes, location within the gene, the likely functional impact. They are also asked to provide a brief description of the function of their gene, using GeneCards or any other bioinformatics source they may find on the NCBI website.

TEACHING DISCUSSION

According to the American Association for the Advancement of Science’s Vision and Change in Undergraduate Biology Education: A Call to Action (15), information flow, exchange, and storage is one of five core concepts for biological literacy. In this exercise, the textbook instructions on how the structure of the gene relates to the structure of a protein is embodied by the real-world clinical practice of identifying and characterizing mutations in tumor suppressor genes at the molecular level. Historically, students find the concept of a mutation challenging, both as process and an entity (16). Since the fall of 2020, 70 students at Benedictine University have performed evolving versions of this exercise, receiving a weighted average score of 81% based on responses to the questions at the end of the worksheet. While other published lessons focus on family histories and mutation effects at the cellular level, it is difficult to find lessons that give practical instruction on interpreting raw tumor suppressor mutation data. Still, this success rate compares favorably with the posttest results from another published lesson in which students were asked four questions about inheritance patterns and probabilities (41%–81% answered correctly) (17). As we currently use the exercise, the students’ mastery of this material is primarily evaluated by worksheet performance. However, we do ask questions about this material on midterm and final exams. Sample questions are provided in Supporting File S7.

When students ask questions about the assignment, the confusion typically stems from trying to navigate the genomic visualization software, which does take practice. In general, students successfully develop a practical understanding of the different parts of a gene, including splice sites and coding regions, and how different kinds of mutations disrupt gene functions. For example, students will determine whether a mutation results from a single base change, or an insertion/deletion event. They will be able to determine whether the base change results in an amino acid substitution, a disrupted splicing regulator, or a frameshifting mutation. Additionally, the students develop real-world skills in interpreting the kinds of data clinical geneticist routinely perform for patients seeking genetic counseling. One of the challenges that often requires extra thought on the part of the students is thinking critically about the literal interpretation of the electropherograms with...
double peaks representing heterozygous mutations. Specific challenges are (i) understanding that the reported sequence may be either the sense or the antisense DNA strand, requiring extra thought in interpreting the clinical impact of the mutation, (ii) understanding that heterozygous mutations may appear as individual double peaks in cases of single base substitutions and multiple double peaks in cases of insertions or deletions, and (iii) understand that sequences near mutant sequences used to identify the position of the mutation within a gene may lie within introns and will not be found searching cDNA sequences. Each sequence given to a student may present a different challenge, and it may be useful to give each student more than one case. Students and instructors interested in learning more about tumor suppressors, mutations, and inherited cancer risk may consult any of a number of general molecular biology text books, including *The Cell: A Molecular Approach* by Geoffrey Cooper et al. (18) and *Molecular Biology of the Cell* by Bruce Alberts et al. (19). Overviews of dideoxy sequencing can also be found in general molecular biology textbooks, including Alberts *et al.* (19).

**SUPPORTING MATERIALS**

- S1. Tumor suppressors – Student cases
- S2. Tumor suppressors – Case history solutions
- S3. Tumor suppressors – Introductory slides
- S4. Tumor suppressors – Slide explanations
- S5. Tumor suppressors – Worksheet
- S6. Tumor suppressors – Grading key
- S7. Tumor Suppressors – Sample quiz and exam questions

**Notes on the Supporting Materials:**

- Supporting File S1 contains case histories and drawings of pedigrees that were invented by the author. The “electropherograms” are drawings created by the author based loosely on images cited in Supporting File S2.
- All drawings in Supporting File S3 were created by the author. The “electropherogram” in the last slide is a drawing based loosely on an image appearing in Cai *et al.* (20).
- Supporting File S5 is a worksheet with numerous images. Figure 1 was drawn by the author. Figure 2 is a drawing based loosely on an image appearing in Cai *et al.* (20). Figures 3–8 are screenshots of publicly available websites as described in the text. Figures 9–11 are screenshots of images of UGENE open-source software windows as described in the text. Figures 12–14 are drawings by the author.
- The Primary Image is a drawing created by the author.

**ACKNOWLEDGMENTS**

I thank Benedictine University for allowing me to teach the Genetics Laboratory course. This work was otherwise unfunded.
Table 1. Teaching timeline table.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Description</th>
<th>Estimated Time</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Preparation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distribute cases</td>
<td>The instructor distributes cases to each student by email or other electronic means before class.</td>
<td>30–90 minutes</td>
<td>Sample cases are included with Supporting File S1</td>
</tr>
<tr>
<td>Load Unipro UGENE</td>
<td>Each student loads Unipro UGENE onto their laptop.</td>
<td>10–60 minutes</td>
<td></td>
</tr>
<tr>
<td><strong>In Class</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lecture</td>
<td>Provide a brief overview of how dideoxy sequencing works and how to interpret electropherograms from patients that are heterozygous for tumor suppressor mutations.</td>
<td>30 minutes</td>
<td>Sample slides are included with Supporting Files S3 and S4</td>
</tr>
<tr>
<td>Worksheet</td>
<td>Students work through the worksheet during class with assistance from the instructor or in groups after class.</td>
<td>1–3 hours</td>
<td>The worksheet is included with Supporting File S5</td>
</tr>
</tbody>
</table>
REFERENCES


