**“Standards and guidelines for the interpretation of sequence variants”:**

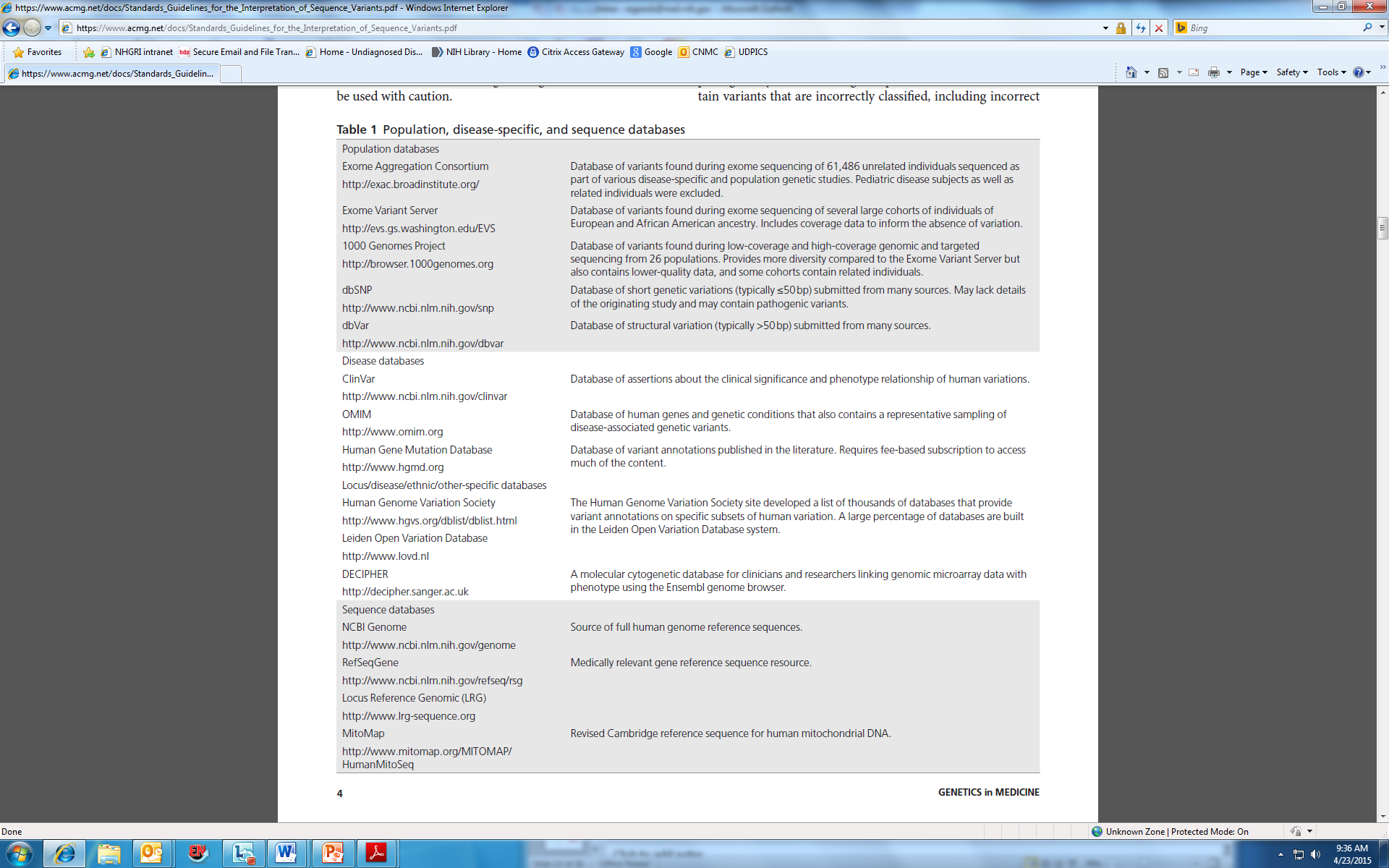
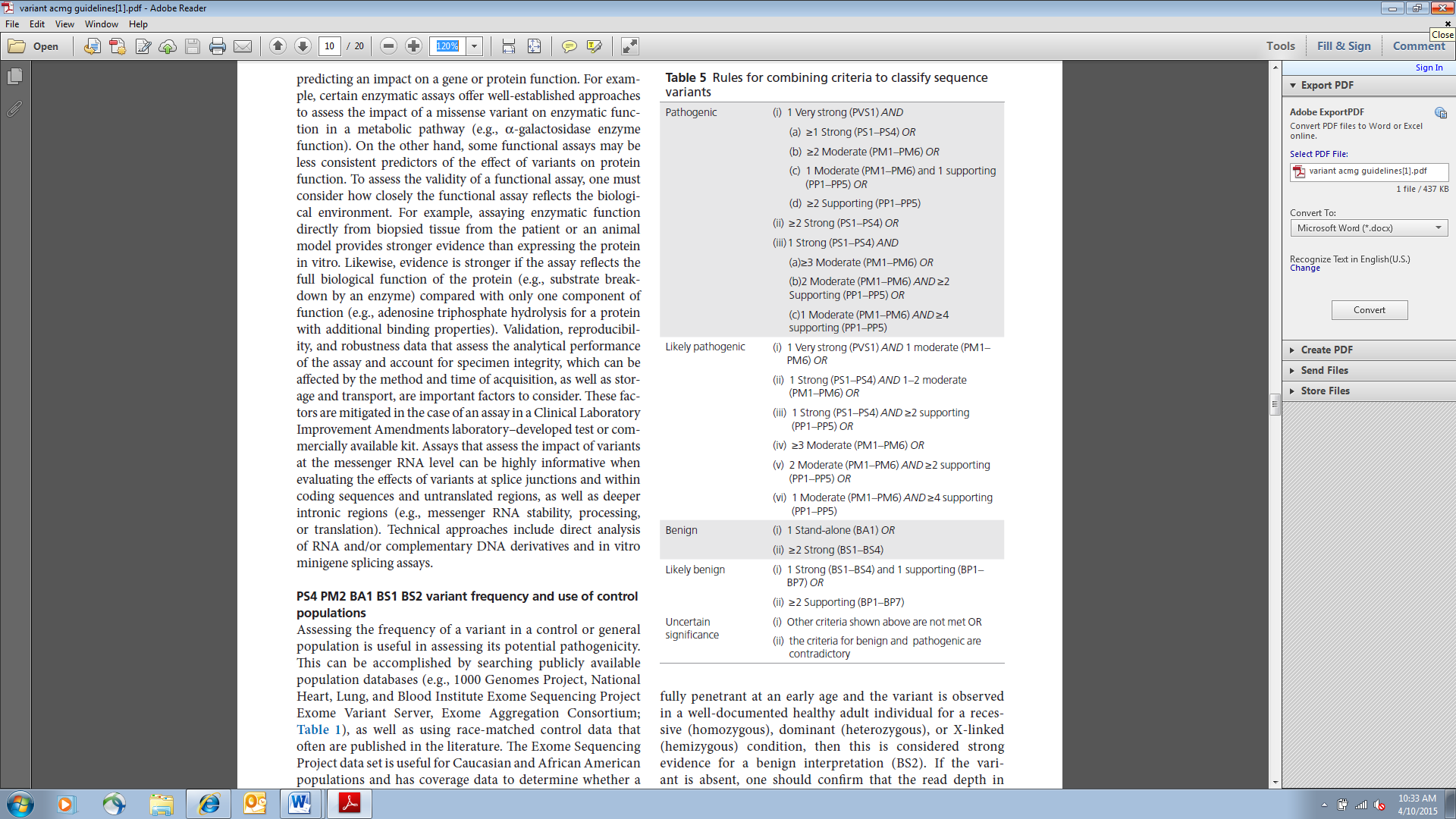
**A Starting “How To” Guide for Learners at All Levels.**

With the publication of the “Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology”, a need emerged among learners from undergraduates to tenured professors for a user-friendly guide to allow learners to access the information and databases efficiently needed to both interpret and re-interpret variants. While the initial population most concerned with this process are variant callers, with the increased need to re-analyze variants called in the past, the cohort of people with this skill set needs to expand to all geneticists to allow more reanalysis of patient variant calls. This document does not replace a personal mentor as you begin the journey into genomic medicine, but instead to help you feel more confident and to know the vast number of resources you have available to you. But if you need to know how to find a [reference sequence](#reference), find a [modeling program](#prediction), or just have an idea of how to find [known variants](#HGMD), this is for you!

Within the published guidelines is a list of online tools that can be used to determine classification (Table 1). This document is designed to further delineate how to access these tools. For each resource described, there is a hyperlink to a page below that will describe how to access and use that resource. Freely accessible resources are emphasized.

Table 5 shows the rules for combining the classification criteria. At the end of this experience, the hope is that learners will feel confident to propose a pathogenicity score for their variant.

And if it is a mitochondrial change, use [mitochondrial-specific tools](#mito) since they are unique for the mitochondrial genome.



**Pathogenic: Very Strong (PVS1)**

PVS1 null variant (nonsense, frameshift, canonical ±1 or 2 splice sites, initiation codon, single or multiexon deletion) in a gene where LOF is a known mechanism of disease

Caveats:

• Beware of genes where LOF is not a known disease mechanism (e.g., *GFAP*, *MYH7*)

• Use caution interpreting LOF variants at the extreme 3′ end of a gene

• Use caution with splice variants that are predicted to lead to exon skipping but leave the remainder of the

protein intact

• Use caution in the presence of multiple transcripts

*This guideline will require that you know the gene well, but is not that technically hard to find. For example, check* [*Pubmed*](#pubmed)*,* [*GeneReview*](#GeneReview)*, or* [*HGMD*](#HGMD) *for information to confirm if the pathogenicity is loss of function. If it is a gain of function mutation, then these changes would not follow the pathogenicity. The most difficulty of the “cautions” is determining if a loss of an exon leaves you out of frame for the remainder of the protein. For example, in Beckers MD, an in-frame deletion is not as deletirious as an out of frame seen in Duchenne MD.*

*For splice site changes it gets a bit harder. If it is a +1 or 2 site, check the* [*consensus sequence map*](#splice)*. For other splice site changes, you will need to access a* [*splice site predictor*](#splice)*. Ask your mentor which one they like or you might need to do some trial and error. A few of the splice predictors are listed in the article. In general, they are not that consistent and can be a bit tricky to use. For secondary calls (calls that are CERTAIN), consider only checking +/- 1 or 2 sites. Otherwise, the reliability drops pretty dramatically.*

**Pathogenic: Strong (PS)**

PS1 Same amino acid change as a previously established pathogenic variant regardless of nucleotide change

Example: Val→Leu caused by either G>C or G>T in the same codon

Caveat: Beware of changes that impact splicing rather than at the amino acid/protein level

*Using a combination of* [*HGMD*](#HGMD) *and* [*ClinVar*](#clinvar) *will likely be the fastest way to identify this answer.*

PS2 De novo (both maternity and paternity confirmed) in a patient with the disease and no family history

Note: Confirmation of paternity only is insufficient. Egg donation, surrogate motherhood, errors in embryo

transfer, and so on, can contribute to nonmaternity.

*This would be given to you on a clinical report or from your pipeline information. You should consider searching for the same de novo mutation in others and ensure that it is not a common change using frequency database evaluation.*

PS3 Well-established in vitro or in vivo functional studies supportive of a damaging effect on the gene or gene

product

Note: Functional studies that have been validated and shown to be reproducible and robust in a clinical

diagnostic laboratory setting are considered the most well established.

*Using* [*Pubmed*](#pubmed)*,* [*HGMD*](#HGMD)*, and/or* [*ClinVar*](#clinvar) *would likely be the fastest way to do this evaluation. If there is a* [*GeneReview*](#GeneReview)*, they would likely mention the accepted clinical diagnostic testing developed for the disorder. Especially for metabolic conditions, clinical testing may be easily available for your case.*

*While a crystal image of the Human Protein is the best resource, there are in silico functional* [*domain mappers*](#functional_domains) *that can be very helpful.*

PS4 The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls. Note 1: Relative risk or OR, as obtained from case–control studies, is >5.0, and the confidence interval around the estimate of relative risk or OR does not include 1.0. See the article for detailed guidance. Note 2: In instances of very rare variants where case–control studies may not reach statistical significance, the prior observation of the variant in multiple unrelated patients with the same phenotype, and its absence in controls, may be used as moderate level of evidence.

*This is very hard data to find. One of the ways to try is to see if the GeneReview has references that might be helpful since this type of data would be unlikely in the GeneReview. Family reports would also be a good place to start with this. Those might be best found through ClinVar or HGMD. The article has a few examples if you have not thought about the OR before (page 11)*

**Pathogenic: Moderate**

PM1 Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of

an enzyme) without benign variation

*Using* [*HGMD*](#HGMD) *or* [*GeneReview*](#GeneReview)*, check this location or functional domain region. Also, consider checking for a HUMAN crystal structure report to map the gene change to a domain. While a crystal image of the Human Protein is the best resource, there are in silico functional* [*domain mappers*](#functional_domains) *that can be very helpful.*

PM2 Absent from controls (or at extremely low frequency if recessive) (**Table 6**) in Exome Sequencing Project,

1000 Genomes Project, or Exome Aggregation Consortium. Caveat: Population data for insertions/deletions may be poorly called by next-generation sequencing.

[*ExAC*](#exac) *is the combined information from other projects and can be the easiest place to start*.

PM3 For recessive disorders, detected in *trans* with a pathogenic variant

Note: This requires testing of parents (or offspring) to determine phase.

*It is COMMON for this testing to NOT be performed in patients. You may need to contact a home provider/clinician to determine if parental testing was performed to determine the phase.*

PM4 Protein length changes as a result of in-frame deletions/insertions in a nonrepeat region or stop-loss variants

*This information is usually given in the report, if you are confirming a change from a report. If this information is not available, then, use one of the mutation analysis programs to help you in this decision.*

PM5 Novel missense change at an amino acid residue where a different missense change determined to be

pathogenic has been seen before

Example: Arg156His is pathogenic; now you observe Arg156Cys

Caveat: Beware of changes that impact splicing rather than at the amino acid/protein level.

*Either HGMD or ClinVar will be helpful here. Just double check any HGMD published reports or ClinVar pathogenicity to be sure that it is truly pathogenic. This is a bit of a “cyclic” finding, so don’t give your variant a point for pathogenicity unless you believe/trust the report that stated it was pathogenic!*

PM6 Assumed de novo, but without confirmation of paternity and maternity

*No parental testing=no proof of this finding!*

**Pathogenic: Supporting**

PP1 Cosegregation with disease in multiple affected family members in a gene definitively known to cause the

disease

Note: May be used as stronger evidence with increasing segregation data

*This finding will require multiple family members and testing to be supportive. Be leary of common variants that may be co-segregating WITH your gene of interest or in close proximity to your gene of evidence.*

*Also, if you have a gene co-segregating, consider checking a database for other genes in close proximity to see if another close gene may be your true culprit. The easiest way to do such a search would be to use the* [*UCSC Gene Browser*](#ucsc)*.*

PP2 Missense variant in a gene that has a low rate of benign missense variation and in which missense variants

are a common mechanism of disease

*Using* [*HGMD*](#HGMD) *or* [*ClinVar*](#clinvar) *will give you an estimate of the number of missencene variations described as pathogenic in the past. As always, blind belief in the HGMD/published data can lead to over calling of variants.*

PP3 Multiple lines of computational evidence support a deleterious effect on the gene or gene product

(conservation, evolutionary, splicing impact, etc.)

Caveat: Because many in silico algorithms use the same or very similar input for their predictions, each

algorithm should not be counted as an independent criterion. PP3 can be used only once in any evaluation of

a variant.

[*Computational programs*](#prediction) *and their mechanisms for making determination are listed in the article and a brief description of two commonly used ones are shown.*

PP4 Patient’s phenotype or family history is highly specific for a disease with a single genetic etiology

*This will require access to private patient records or a good clinical description on the documentation if you are evaluating clinical laboratory sample. Be very cautious that the phenotype being used is specific and consider the sensitivity. For example, hypotonia is very sensitive for Down syndrome, but not at all specific. So, phenotypes like “dysmorphic, delay, hypotonia, dysmorphic” may not be helpful and may give you a false sense of support for this category.*

PP5 Reputable source recently reports variant as pathogenic, but the evidence is not available to the laboratory

to perform an independent evaluation

[*GeneReview*](#GeneReview) *is notorious for stating a mutation is pathogenic but not why. However, this PP5 gives you the right to trust the experts, at least partially!*

**Benign: Stand-alone**

BA1 Allele frequency is >5% in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium

[*ExAC*](#exac) *is a database that will allow you to find the allele frequency. The database shows ethnicity profiles, which can be important in determining if your change of interest is clinically important.*

**Benign: Strong**

BS1 Allele frequency is greater than expected for disorder

*Based on the findings in BA1, allele frequency in databases, you will then need to determine the expected allele frequency. For an autosomal dominant disorder, the allele and disease frequency are the same. For an X-linked recessive disorder, the expectation would be that the rate of disease in men would be equal to the allelic frequency. However, in woman it would be equal to the autosomal recessive allele frequency. For autosomal recessive disorders the allele frequency equals 2pq, where p is the frequency of the mutant allele, q is the frequency of the normal allele (in this case p=1-q, and if q is very small you can approximate p as 1). The disease frequency is q2. For example, if the disease you are studying is 1/10,000. Then q=1/100 and the carrier frequency (allele frequency) is 2pq=2x(1-1/00=1)(1/100)=1/50.*

BS2 Observed in a healthy adult individual for a recessive (homozygous), dominant (heterozygous), or X-linked

(hemizygous) disorder, with full penetrance expected at an early age

BS3 Well-established in vitro or in vivo functional studies show no damaging effect on protein function or splicing

*This will require both evaluation by checking* [*HGMD*](#HGMD) *and* [*ClinVar*](#clinvar) *but might also require searching for the change on* [*PubMed*](#pubmed)*.*

BS4 Lack of segregation in affected members of a family

Caveat: The presence of phenocopies for common phenotypes (i.e., cancer, epilepsy) can mimic lack of segregation among affected individuals. Also, families may have more than one pathogenic variant contributing to an autosomal dominant disorder, further confounding an apparent lack of segregation.

*This would be a possible case where you have results from other family members showing the same change in a phenotypically normal family member. Remember, this can get complicated if phenotyping of the unaffected individual is in question or not complete. Also, for non-fully penentrant changes this can be a very complicated issue to consider.*

**Benign: Supporting**

BP1 Missense variant in a gene for which primarily truncating variants are known to cause disease

*By checking* [*HGMD*](#HGMD) *and* [*ClinVar*](#clinvar)*, you can determine if there are different types of mutations. Also, an easy place to start would be a* [*GeneReview*](#GeneReview) *since they often have a chart of the types of genetic changes that are disease causing.*

BP2 Observed in *trans* with a pathogenic variant for a fully penetrant dominant gene/disorder or observed in *cis* with a pathogenic variant in any inheritance pattern

*This would often require either next generation sequencing to allow for allelic segregation or family studies to identify if changes are in cis or trans. If a “more” pathogenic change is cis with your suspected change, your change probably isn’t pathogenic (or not affecting pathogenicity here). One case where this would not be true is if you are evaluating a frameshift mutation leading to a premature stop codon that would prevent a known missense/stop codon from being read in the gene you are evaluating.*

BP3 In-frame deletions/insertions in a repetitive region without a known function

*If the gene you are evaluating has a HUMAN crystal structure, look at the paper and see if your region of change is in a known area of structure. For example, adding an extra or removing an amino acid passing through a membrane can be very damaging; however, in a loop region it may not have any affect at all. By looking at the crystal structure this might give you more input.*

*To find a crystal structure, I’d recommend using pubmed. Most crystal papers have a list of mutations and which domains they fall into. They also have a link to the structure for 3D reconstruction (recent papers). BEWARE: don’t believe non-human crystal structures or if the structure is not with co-factors or associated proteins, since both can have structures that are not* in vivo *representations.*

BP4 Multiple lines of computational evidence suggest no impact on gene or gene product (conservation, evolutionary, splicing impact, etc.)

Caveat: Because many in silico algorithms use the same or very similar input for their predictions, each algorithm cannot be counted as an independent criterion. BP4 can be used only once in any evaluation of a variant.

The computational evidence should be used in following with Table 2 from the paper to ensure that the same criteria are not used multiple times.

[*Computational programs*](#prediction) *and their mechanisms for making determination are listed in the article and a brief description of two commonly used ones are shown.*

BP5 Variant found in a case with an alternate molecular basis for disease

*In this example, GeneReviews would be helpful since they list known forms of inheritance and the gene changes associated. For example, if a gene can have autosomal dominant or autosomal recessive forms of inheritance, a change previously associated with autosomal recessive disease would not be an autosomal dominant-causing change. Usually, dominant disorders are gain of function or form a complex, have abnormal regulation, are happlo-insufficient, or a dominant negative mechanism. Autosomal recessive is loss of function mutations.*

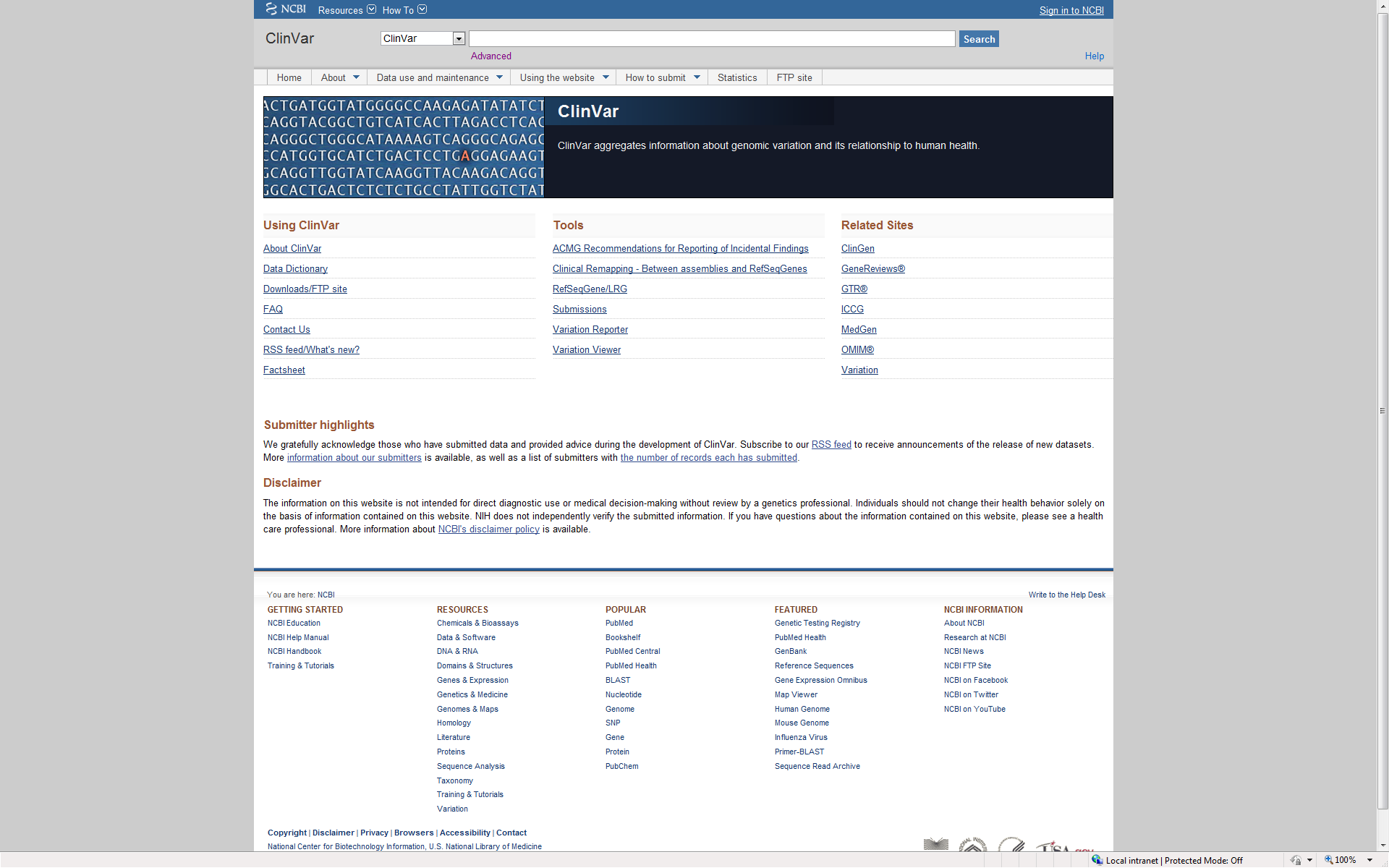
BP6 Reputable source recently reports variant as benign, but the evidence is not available to the laboratory to perform an independent evaluation

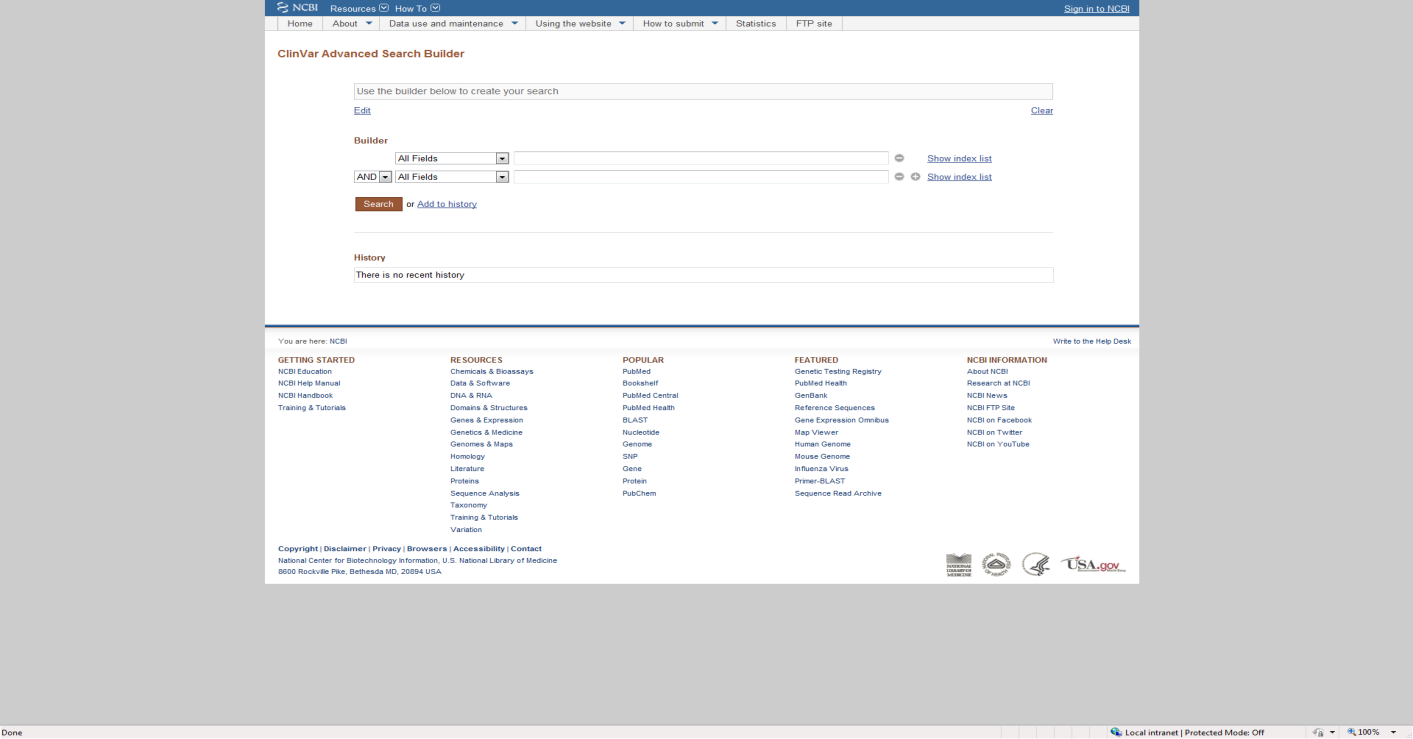
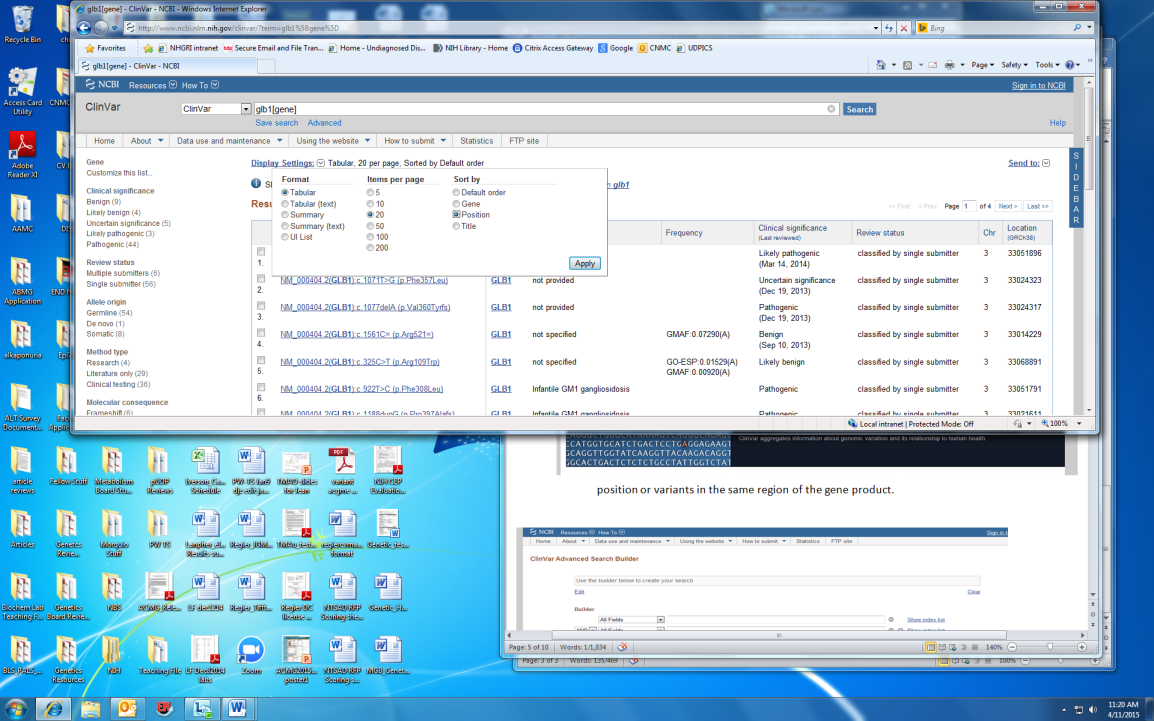
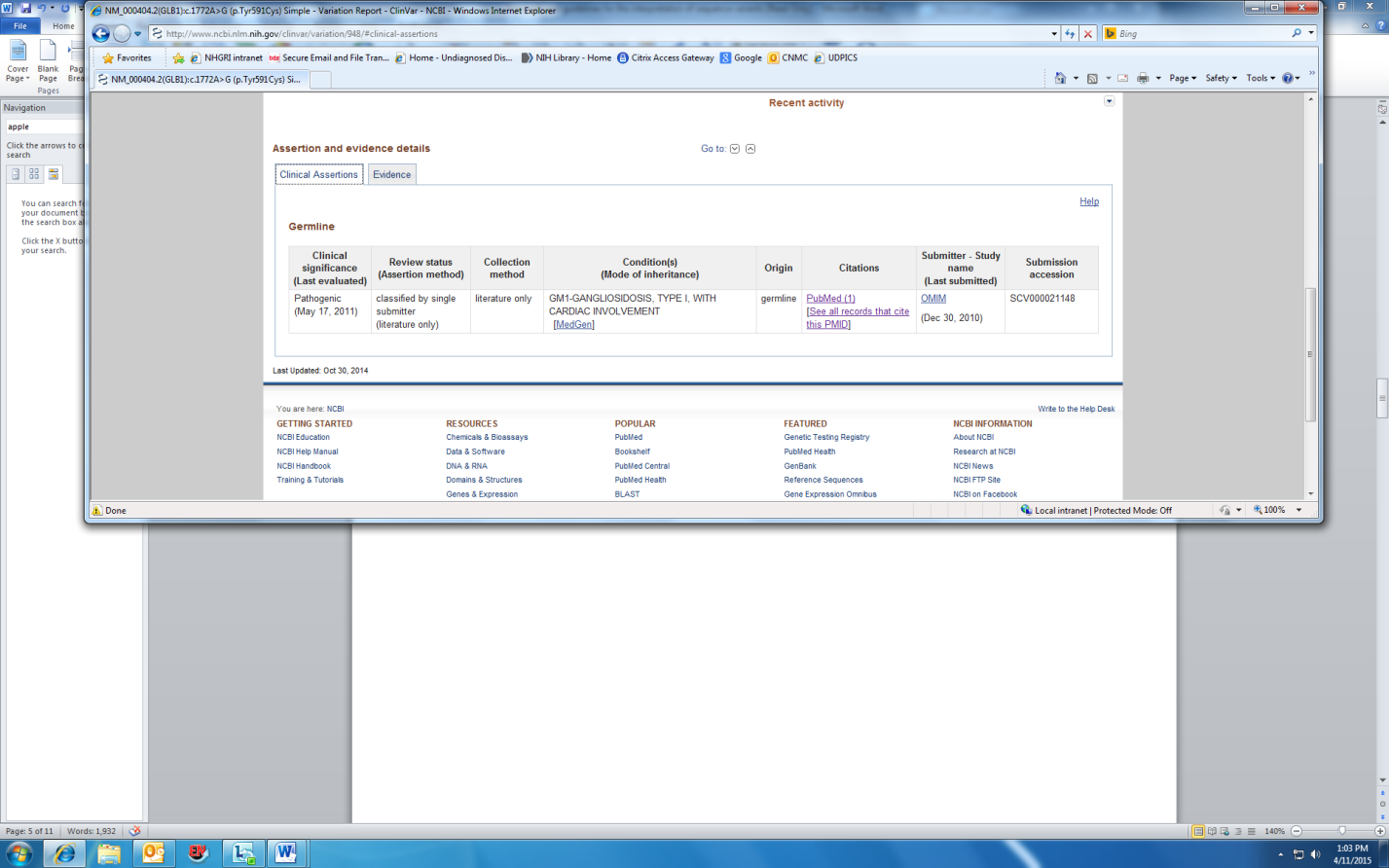
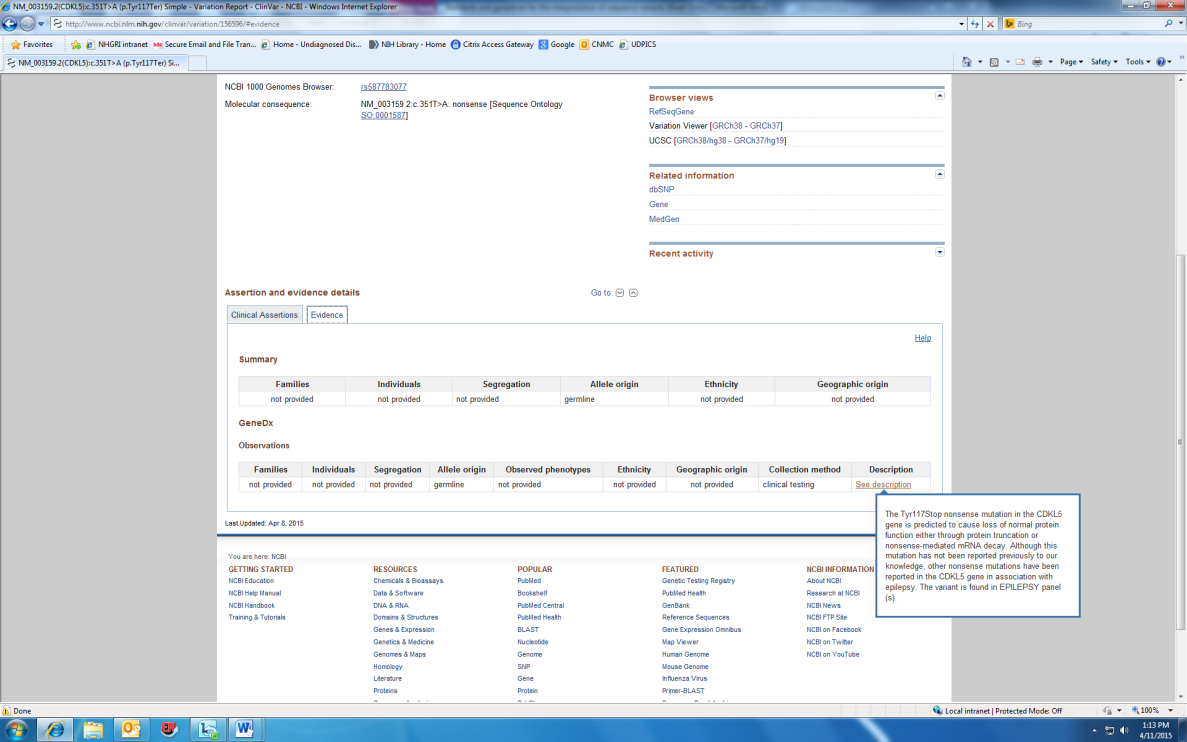
[*PubMed*](#pubmed)[*, ClinVar*](#clinvar)*, and* [*GeneReviews*](#GeneReview) *would be helpful in determining if new changes have been evaluated.*

BP7 A synonymous (silent) variant for which [splicing prediction algorithms](#splice) predict no impact to the splice consensus sequence nor the creation of a new splice site AND the nucleotide is not highly conserved

*Again, the* [*splice predictors*](#splice) *are fairly complicated to use and more than the first and second consensus site may be harder to predict. Some of the splice prediction programs are described, briefly. This is an area where a good mentor will likely be needed to feel confident in this decision making.*

**ClinVar**

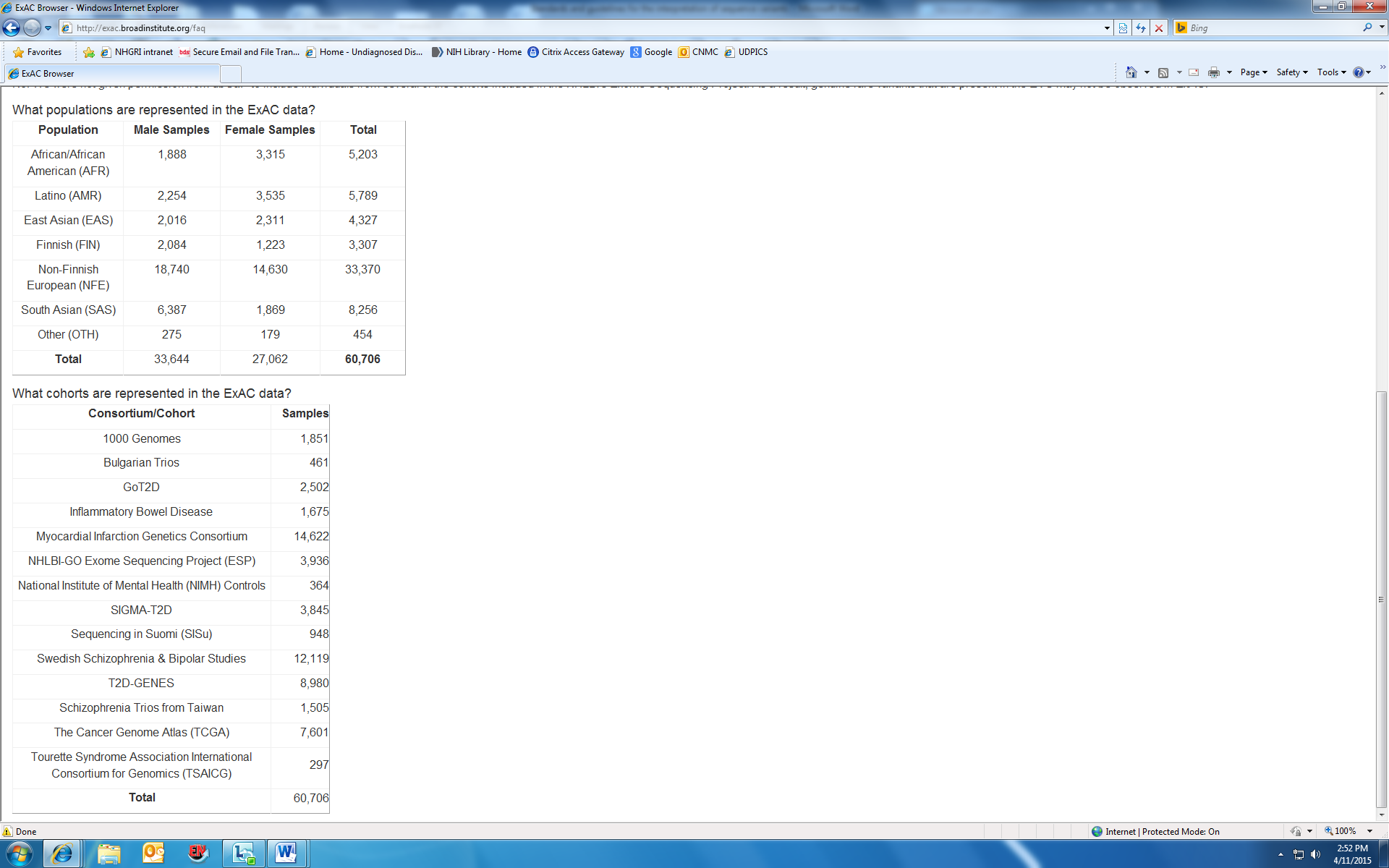
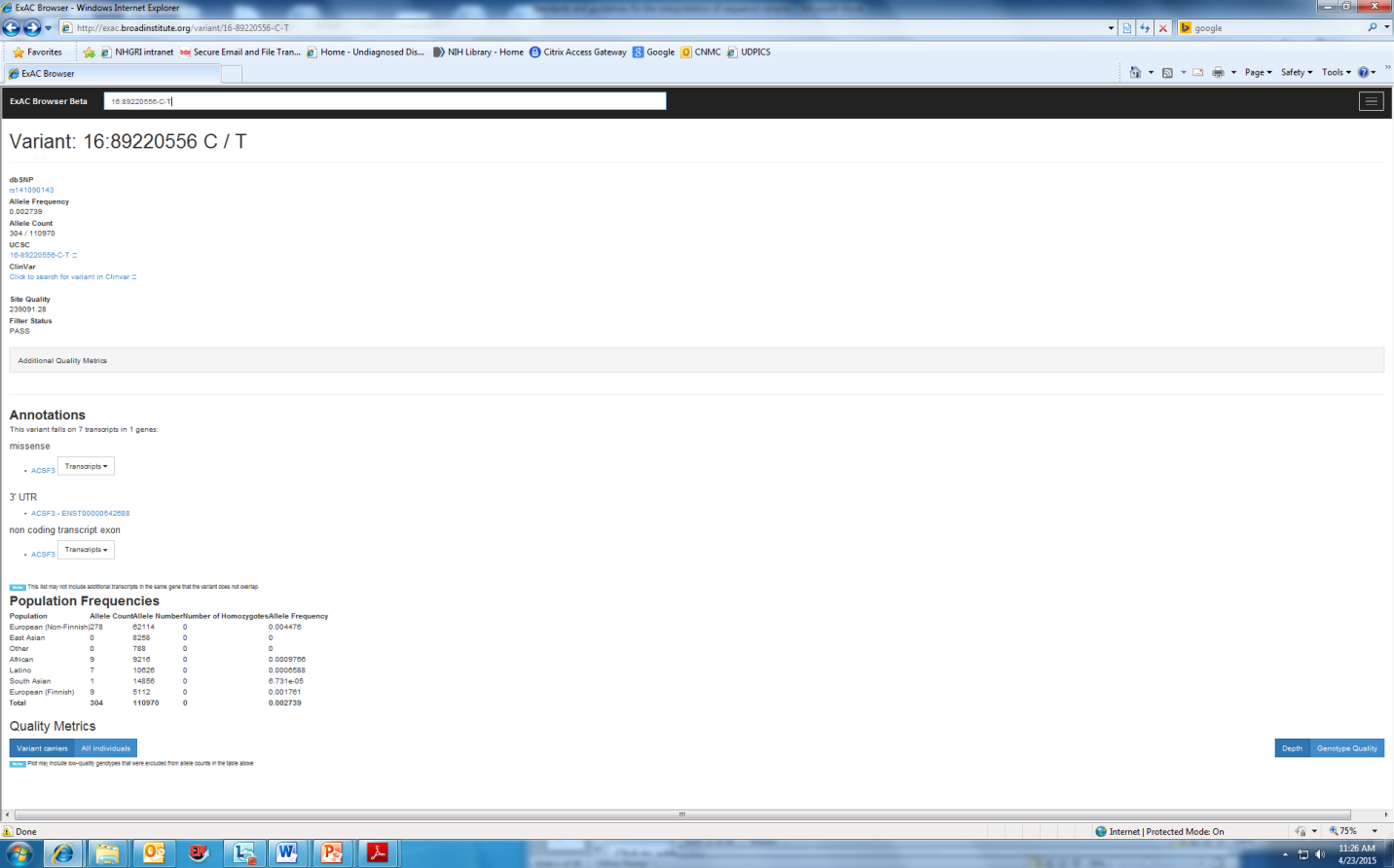


* 1.  ClinVar (<http://www.clinvar.com/>) is a public access site where companies and organizations that perform sequence analysis deposit their annotated genetic results. This means that changes are “classified” as pathogenic, likely pathogenic, variants of unknown significance, likely benign, or benign.
  2. THE BEST PART….no log in needed!
  3. ClinVar is growing every day, so repeat searching is very helpful with this resource
  4. ClinVar can be a bit less intuitive than HGMD for searching. First, you can use the advanced search (purple word below search box) and include both the gene name and the variant “name” (amino acid change in the example below.
  5. Once the list comes up, use the Display settings button below the search outcomes bar to choose “position” on far right. This will put the identified changes in the gene in numeric order from pter to qter. This might be in opposite order for your gene of interest, but at least it will be more easily searchable.
  6. Another option is the use the “find” function on your computer (apple F for mac, ctrl F for PC) and put in the number/term you are looking for.
  7. Once you find a position you want to read more about, click on the link. This will take you to a second page with articles, source of information, etc. at the bottom (SCROLL DOWN!)
  8. At the bottom of the variant page, be sure to toggle between the clinical assertions and evidence tabs.
  9. If you hover over the “see description” field, you can view the description of how that group made their decision. This is NOT required, so will not be available for all submissions. If there is more than one submitter, they could each have their own description. For example, GeneDX is submitting all of their descriptions for the Noonan gene panel (as of 4/11/2015), but Emory had not for the glb1 gene (as of 4/11/2015). So it is worth the time to check back if it is an important gene change for your work.

**ExAC**

The ExAC database is compiled from multiple sites (see below) and is exonic data from unaffected families. It is commonly parents of affected children. <http://exac.broadinstitute.org/>

Within the ExAC browser, type your genetic change (Note: you need to type it as chromosome number:location-c-t (\*see search box in picture)). The population frequencies include the allele count, and number of homozygous cases (this should worry you if it is a rare autosomal recessive disorder!). If you do find a homozygous change, consider sending the variant to the ExAC team to have them check the original data to see if it is clean and well-sequenced.

The ExAC database is compiled from multiple sites (see below) and is exonic data from unaffected families. It is commonly parents of affected children.

**Functional Domains**

Using your captured sequence (see reference sequence), you can then use multiple programs that predict the domain that your area of interest is in. Remember, if there is a HUMAN crystal structure, this is MUCH better than any predication model!

One easy to use example:

<http://www.uniprot.org/> (will need the NCBI number from your reference sequence). Very fast and will give you the general domains of the protein of interest.

Also, you can zoom into the active site/domains by clicking on the graphical view to see exactly which amino acid residues are responsible for the noted activity.



**GeneReviews**

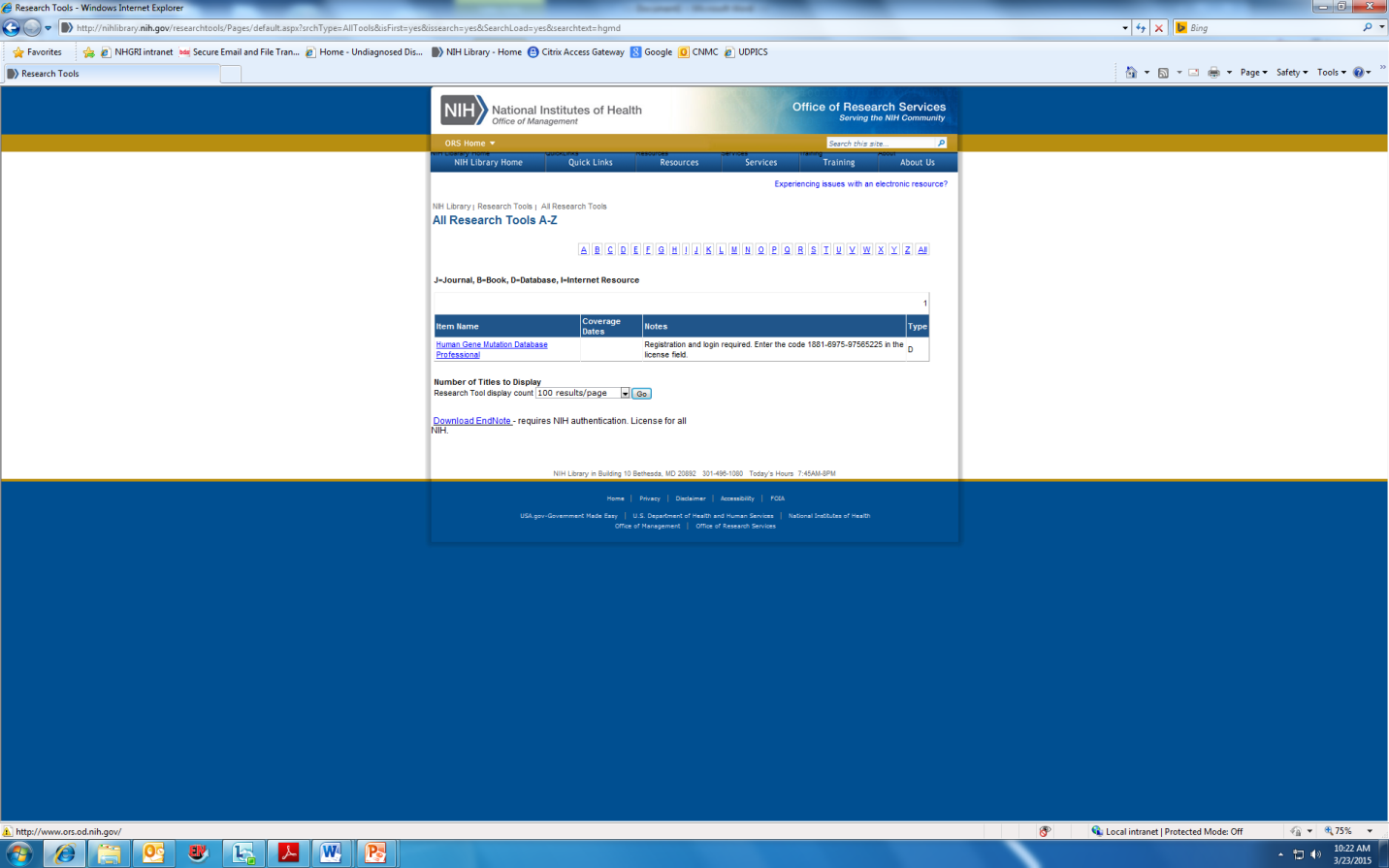
GeneReviews are expert opinion papers that are peer-reviewed to be a guide to the gene products and diseases. This is a primary place to check to see if a gene has a disease-associated, has a treatment protocol and/or recommendations, and a description of the normal mechanism of the gene product.

<http://www.ncbi.nlm.nih.gov/books/NBK1116/>

Note: if you try to go straight to GeneReviews on NIH campus, sometimes is comes up as blocked. Just google search GeneReview and link from there and it will go through.

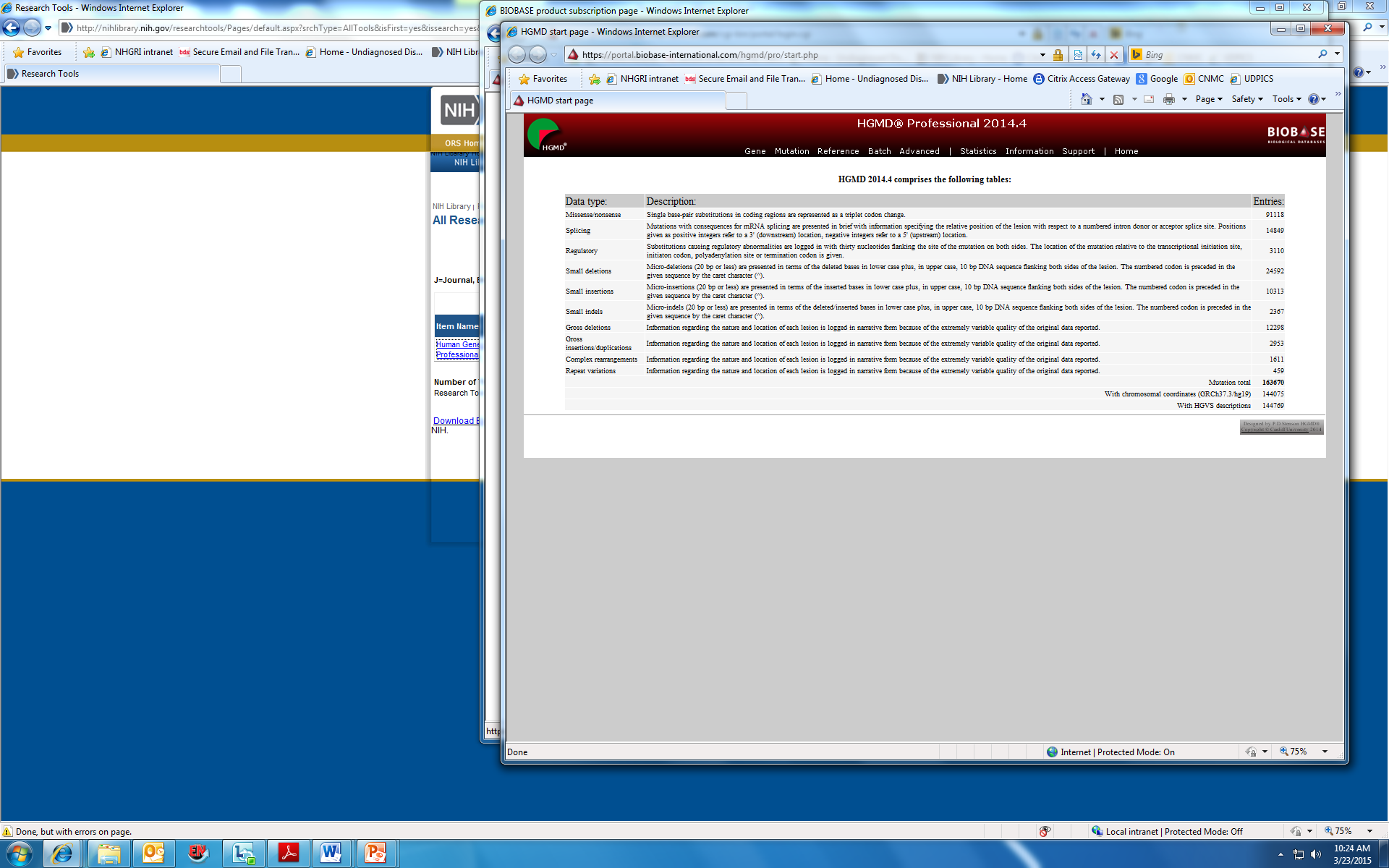
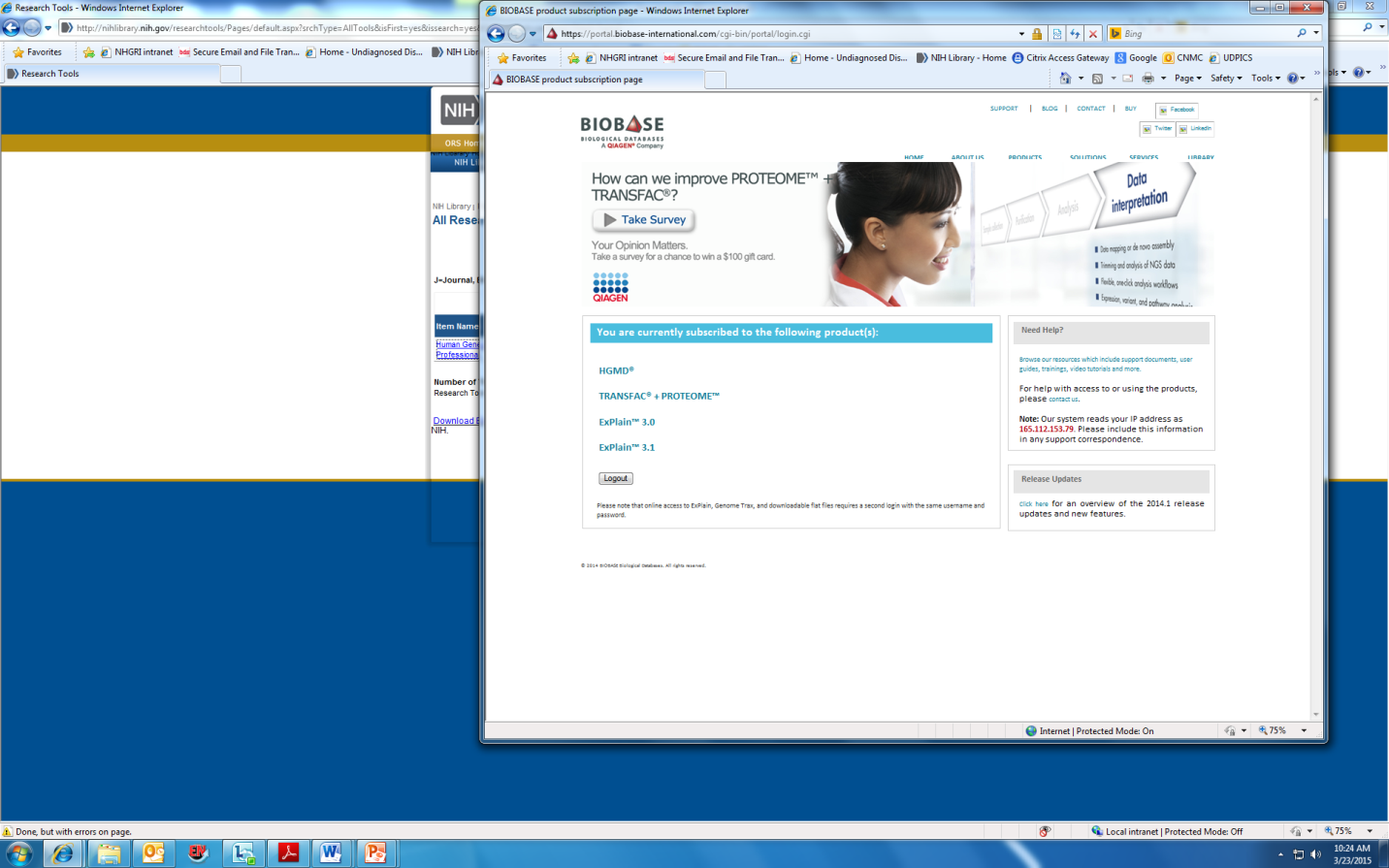
Caution: not all genes have a GeneReview. And if you complain too loudly, you might be recruited to write one that is missing!

**HGMD 101**

Go to NIH library website and search for HGMD for link and passcode.

Within the biobase webiste, create a log in. This can then beused both at NIH and offsite for review.

Next, it is often easiest to click on “gene” to look up your gene of interest. From there you will have a list of mutations types. If your exact mutation is not there, consider looking at changes in that region, or the same position.

If you find a change that you want more information on, click on the reference. Via the NIH library (if remote, go to NIH library website and use your username/password to do a remote log-in) most of the articles are immediately available.

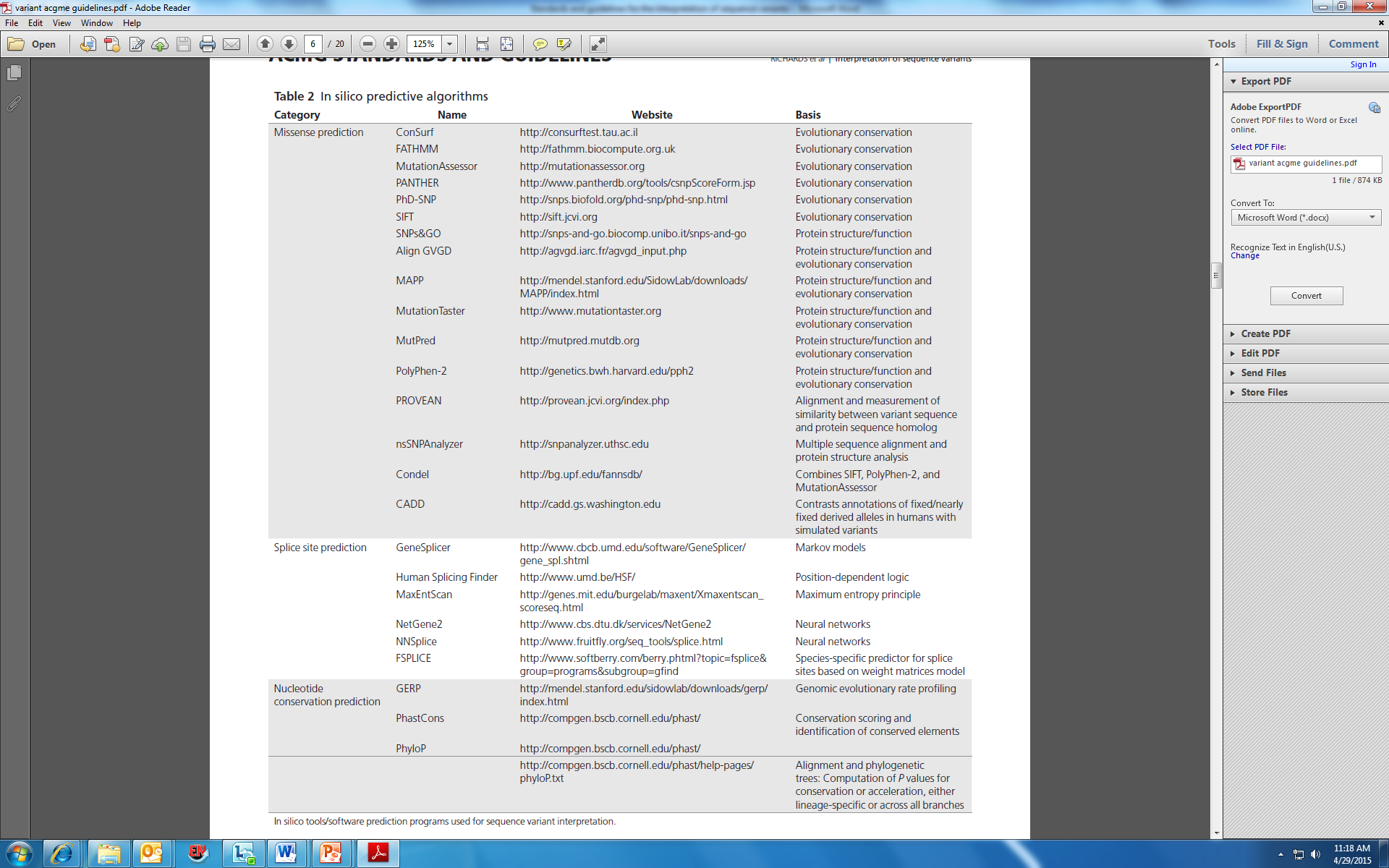
This is THE hardest part. Most geneticists would be worried if the DNA change has only been seen in one family or is of unclear significance attributing it to that patient. And it even gets trickier if you have compound heterozygosity. And JUST BECAUSE IT IS PUBLISHED DOES NOT MEAN IT IS REALLY DISEASE-CAUSING! Prove it to yourself!!!

**Mitochondrial Genome Changes**

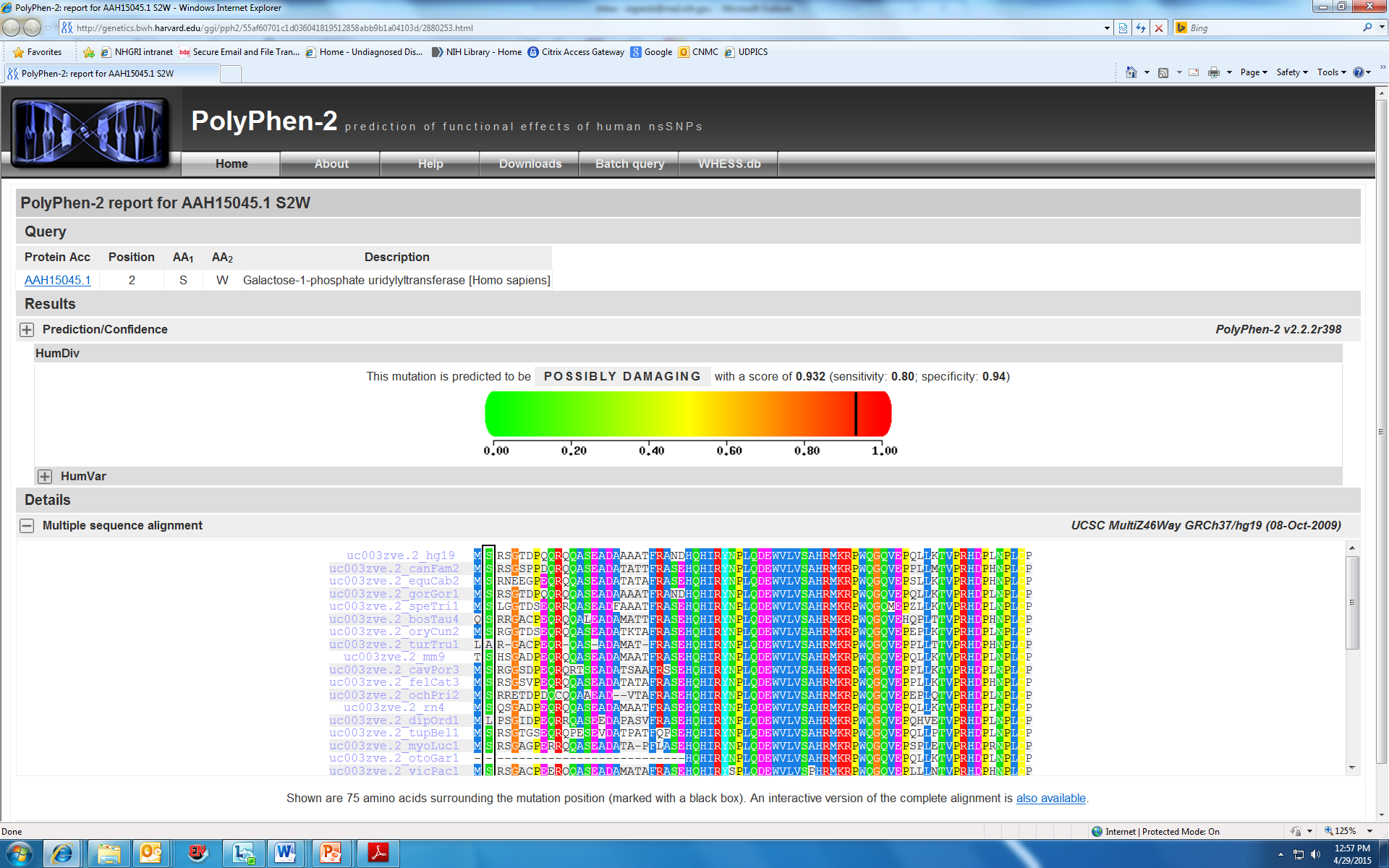
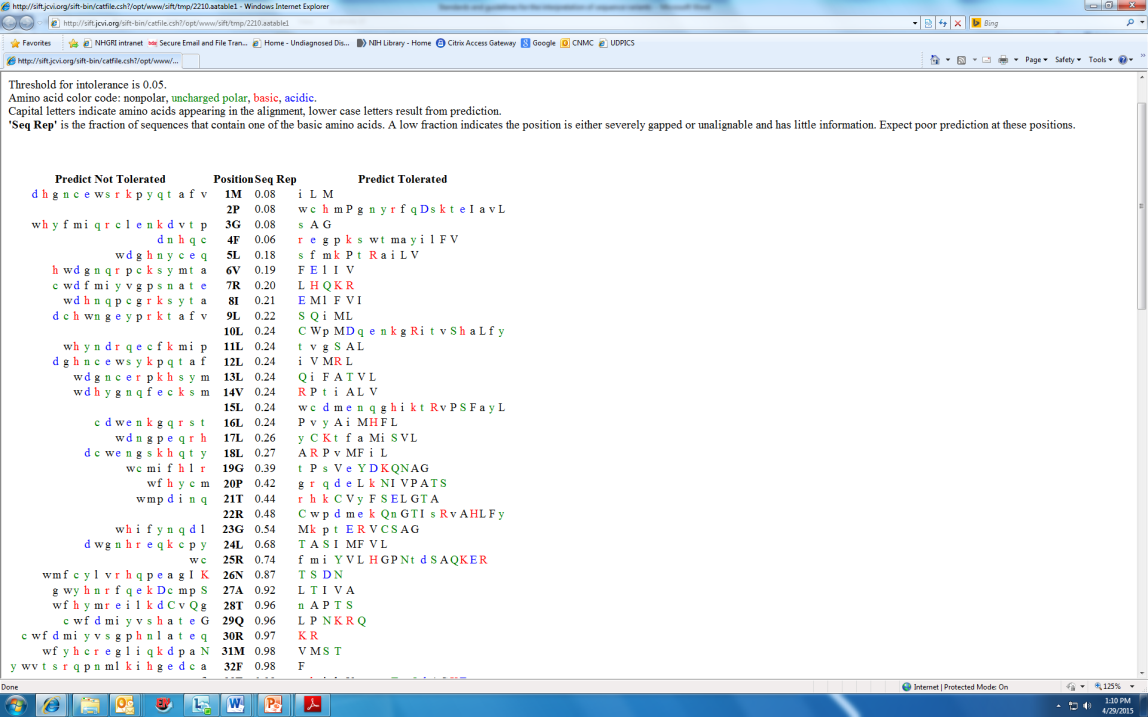
**Mito≠Genome**

**So, use the correct resources!**

1. Nomenclature. The mitochondrial nomenclature will be m. ###A>T and the p. numbersing system. There will not be a c.### given.
2. The currently used mitochondrial DNA sequence is NC\_012920gi251831106 (check to see if this is still the one in use at the time you are doing this.
3. Unless you have an understanding of heteroplasmy in the change of interest and how it affects the gene of interest, be very cautious in making decisions about mitochondrial changes.
4. There are 275 well characterized mitochondrial DNA variants that will often be the best/easiest place for you to search first in MitoMap (<http://mitomap.org/bin/view.pl/MITOMAP/WebHome>). MitoMap is considered the main source of information
5. A group at CHOP has begun an all-inclusive mitochondrial and genome affecting mitochdria resource for researchers. Currently it is a little bit cumbersome, but getting better every day (<https://mseqdr.org>). Requires registration/log-on but free.
6. Other helpful resources:
   1. Frequency information can also be found at <http://www.mtdb.igp.uu.se/>)
   2. secondary structures, sequences, and alignment of mitochondrial transfer RNAs (<http://mamittrna.u-strsbg.fr/>)
   3. mitochondrial haplogroups (<http://www.phylotree.org/>)
   4. miscellaneous information (<http://www.mtdnacommunity.org/default.aspx>)**In Silico Prediction Models**

Since the “rules” state multiple modalities should be evaluated, a few are shown here.

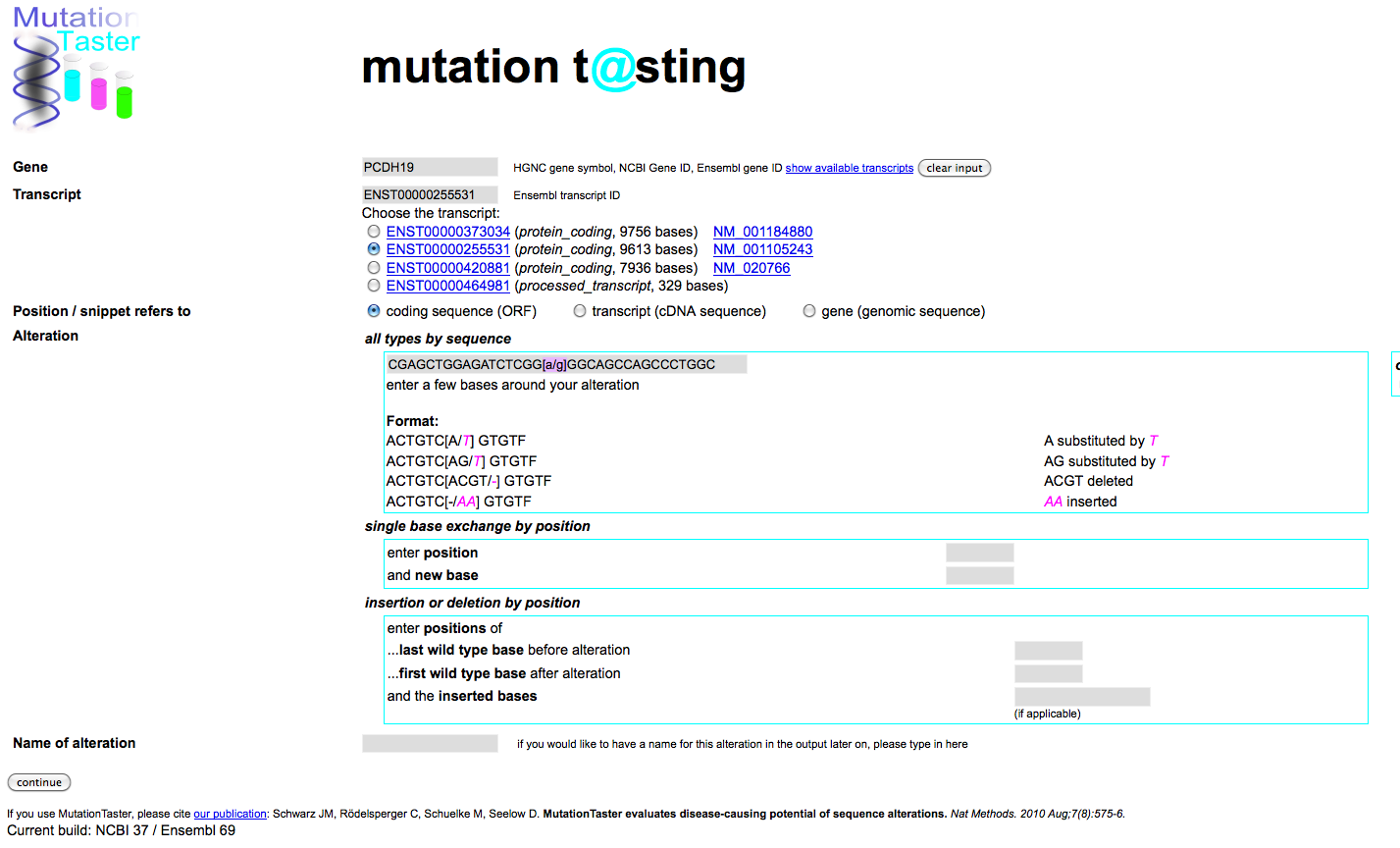
**PolyPhen-2**

1. uses both protein structure, function, and evolutionary conservation to predict pathogenicity. In the output, you’ll have an evolutionary conservation list that ma kes lovely pictures, too!
2. PolyPhen-2 directions: at the homepage (<http://genetics.bwh.harvard.edu/pph2/index.shtml>) put in the protein identifier (from reference sequence page). Then, you will choose the change (the top row is the normal amino acid, bottom is your amino acid change). Then hit submit. It may take up to a few hours to get your results. Just “refresh” the page until there is a “view” purple word that appears. Click on it and you’ll have your images.

**SIFT**

1. Only uses evolutionary conservation. Not that much easier than PolyPhen-2 but it is faster.
2. You will need the gi number (the numbers after the letters gi on your FASTA sequence (see how to get a reference sequence). Then you will also enter your amino acid change. Conversely, it also has a list of what it determines are OK and not OK changes (see picture below), which is nice if you’re looking for conservation over a range of amino acids.

**MUTATION TASTER**

* **Navigate to** [**http://www.mutationtaster.org/**](http://www.mutationtaster.org/) **and enter the gene and then select the proper transcript, and provide sequence context for accuracy.**
* **The probability value is the probability of the prediction, i.e. a value close to 1 indicates a high 'security' of the prediction**
* **This probability score is the number indicated in precomputes; however, the call of pathogenic or benign comes from the overall prediction.**
* **This information is combined to yield the internal “red” and “green” precompute**

**Pubmed Access**

<http://www.ncbi.nlm.nih.gov/pubmed>

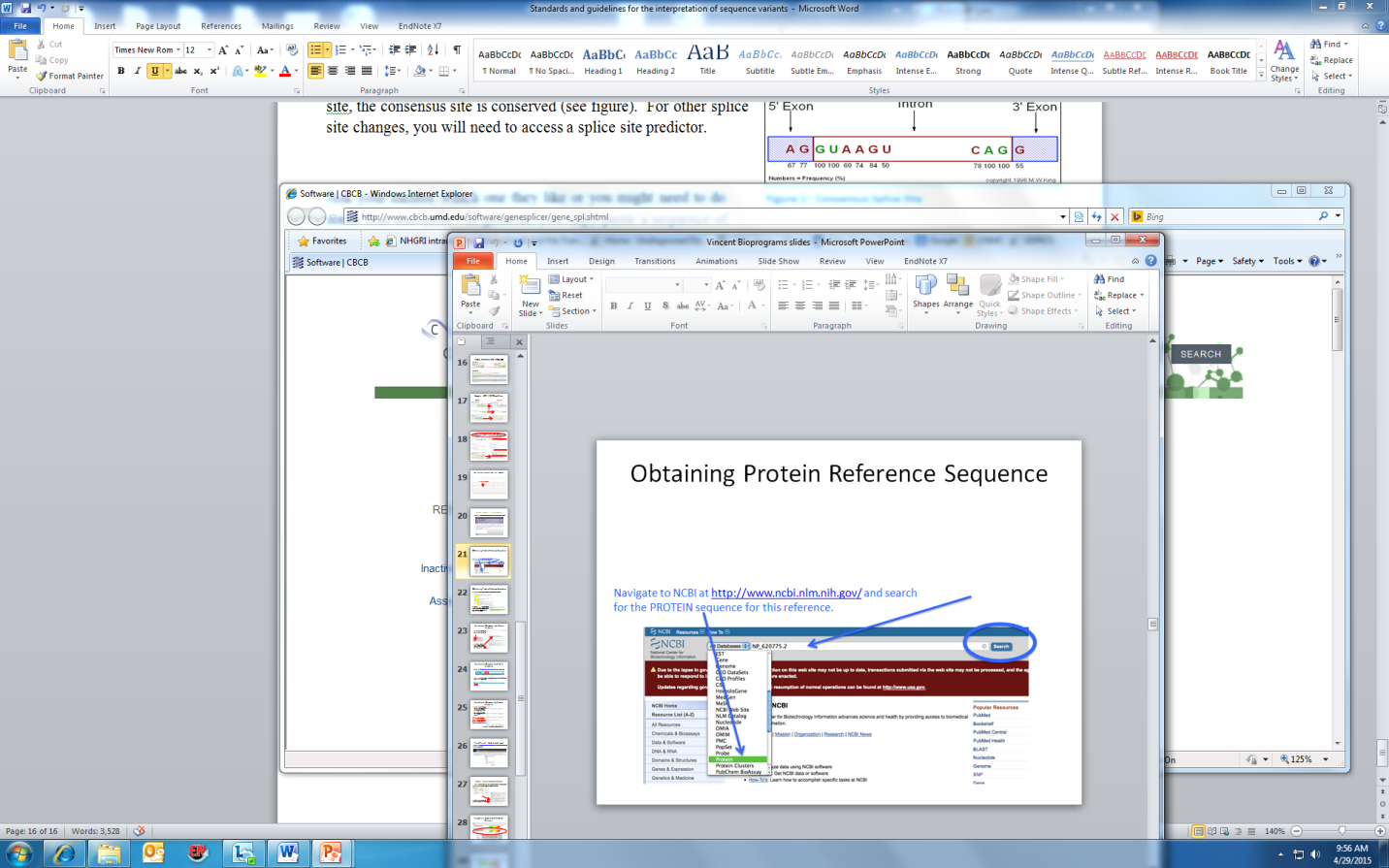
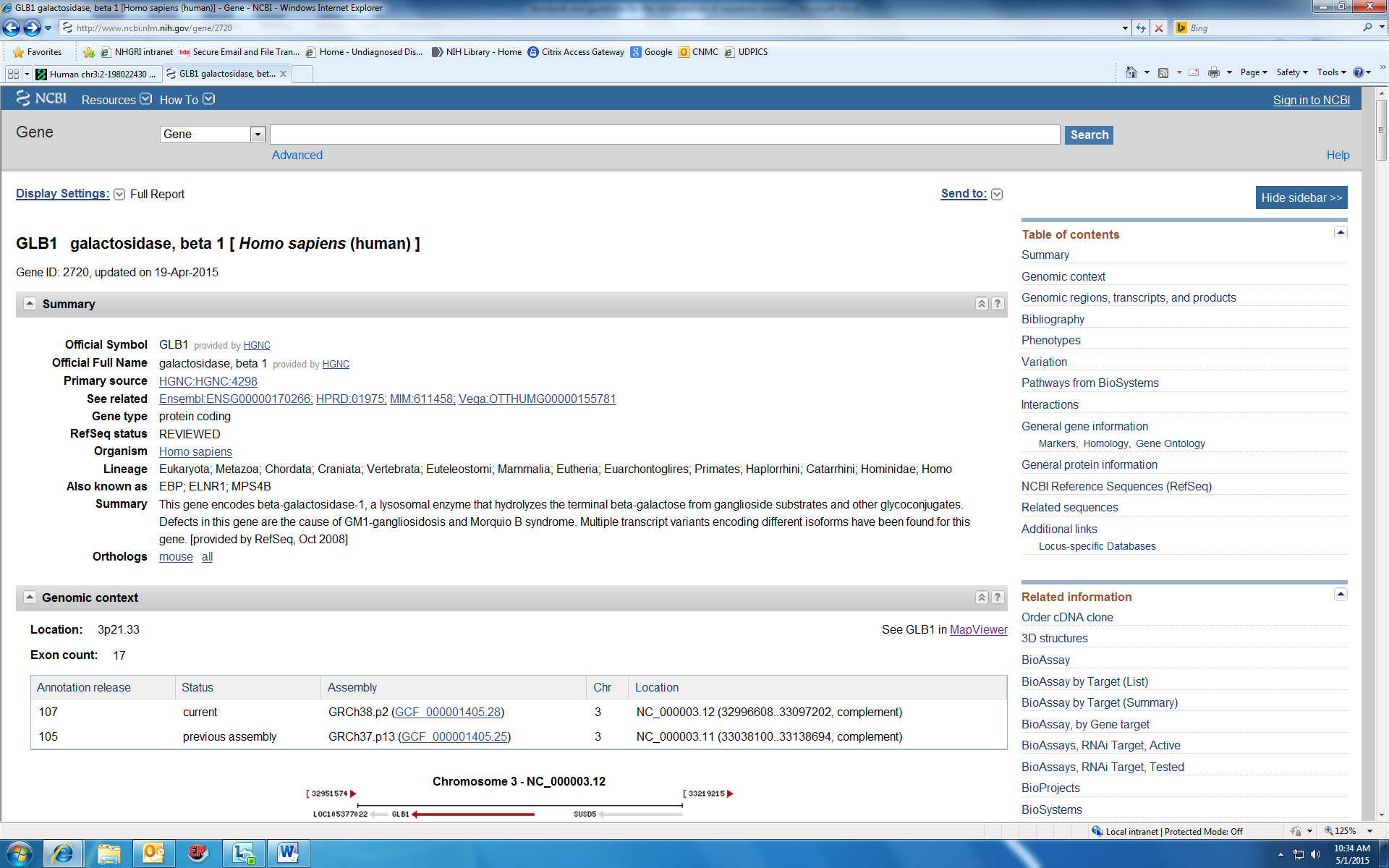
From the NIH library site, you can log in using the “remote login” button. Using your NIH Username and Login, a new pubmed window will appear. Then, you can search in pubmed as you would otherwise and the NIH library symbol will appear to the right on the abstract page for a paper if it is available.

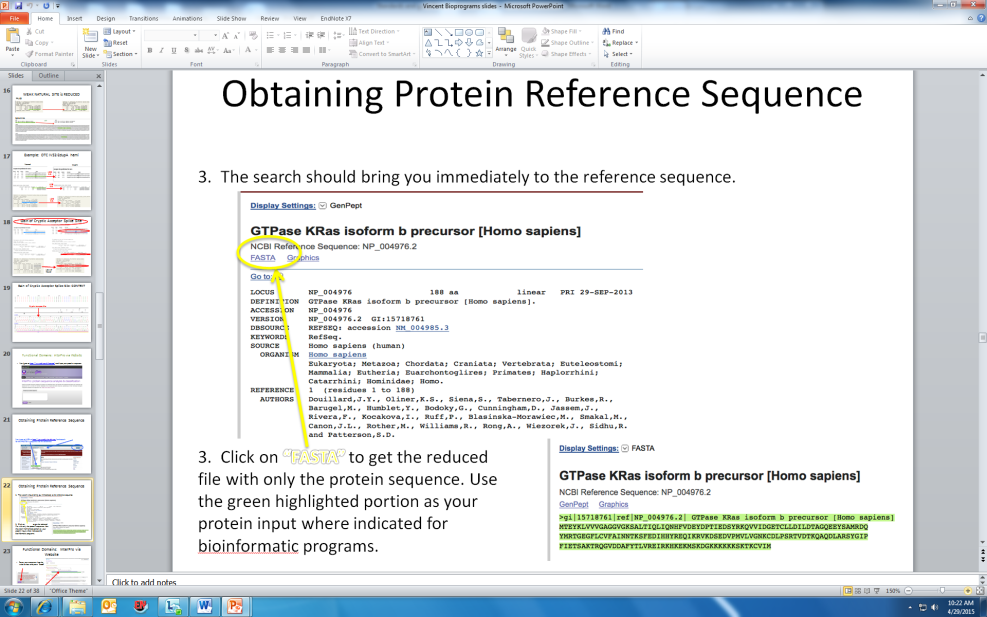
Papers not immediately available can be “ordered”. In general, an ordered article requires a PMID number.

To order an article, you will need the Personal Idenfitifer cose on the back of your PIV card or the front of your “physician” NIH card, if you have one.

**Reference Sequences**

**FOR DNA/CHROMOSOMAL LOCTION**

1. Navigate to NCBI at <http://www.ncbi.nlm.nih.gov/> and search for the Gene sequence for this reference (pulldown box to left of search box Figure 1).
2. If you are looking for the gene reference sequence so that you can use UCSC or another browser, then I use the same NCBI site and choose “gene” for the search criteria. From there, you can either copy/paste the location for UCSC, etc. or you can click on “map viewer” for a simplistic map of the chromosome region (figure 2).



**FOR PROTEIN SEQUENCES**

1. Navigate to NCBI at <http://www.ncbi.nlm.nih.gov/> and search for the PROTEIN sequence for this reference. Use the protein name and “human” in the search. You will still likely get multiple hits. By checking the genereview for the protein length you should be looking for it might help you to choose the correct sequence to use as your reference.
2. The search should bring you to the reference sequence.
3. Click on FASTA to get the reduced sequence you need to copy/paste into on-line resources

**Splice Site Analysis**

Splice site changes can be fairly difficult to call. If it is a +1 or 2 site, the consensus site is conserved (see figure). For other splice site changes, you will need to access a splice site predictor.

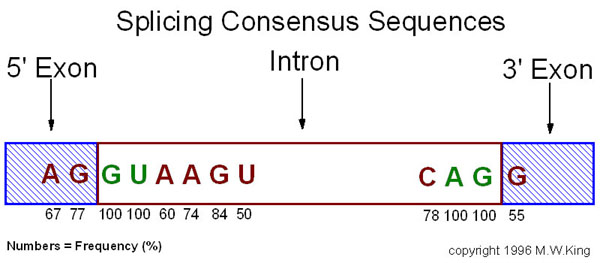
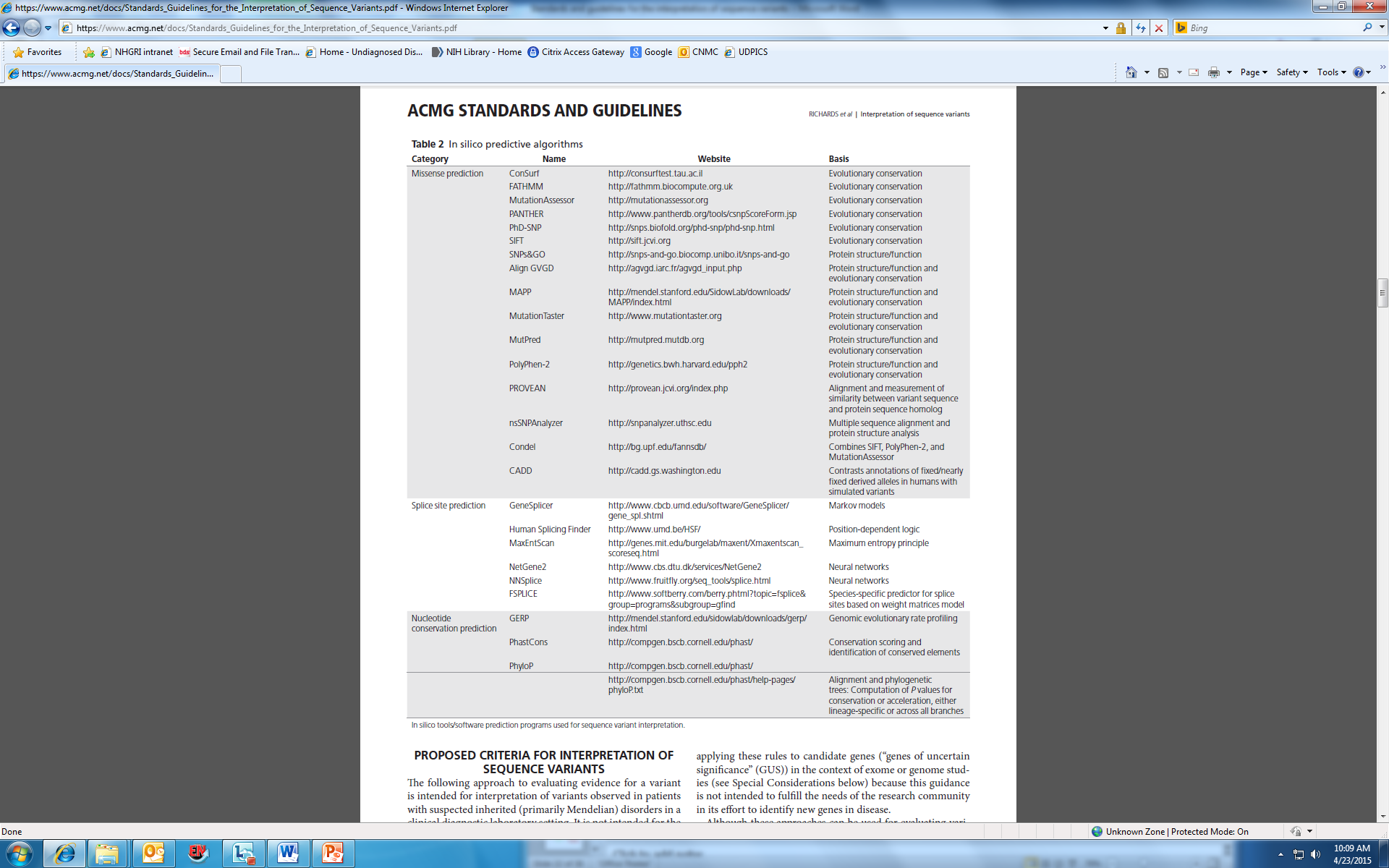


Figure 1: Consensus Splice Site

Ask your mentor which one they like or you might need to do some trial and error. In general, you copy/paste a sequence of the exon🡪intron🡪exon to find how the splice will be changed. You don’t need a log-in for many of these sites (horray!).

* BDGP/NNSplice Splice Site Predictor <http://www.fruitfly.org/seq_tools/splice.html>
* NetGene2 Splice Predictor <http://www.cbs.dtu.dk/services/NetGene2/>
* Softberry/FSPLICE Predictor [http://linux1.softberry.com/berry.phtml](http://linux1.softberry.com/berry.phtml?topic=fsplice&group=programs&subgroup=gfind))
* GeneSplicer (<http://www.cbcb.umd.edu/software/genesplicer/gene_spl.shtml>)
* Human Splicing Finder (<http://www.umd.be.HSF/>)
* MaxEntScan <http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html>)



**UCSC Gene Browser**

This is a publically available browser of gene location/alignment. It does have some mutations and deletions; however, HGMD and ClinVar has much better annotation.

Location: <http://genome.ucsc.edu/>

Click on Genome Browser

Then, in the search box, enter your gene’s location (see the section on reference sequences)