Using Hypothes.is to annotate articles for ClinGen Gene Curation
Standard Operating Procedure version 2

Background

Hypothes.is is a web-based tool that allows users to create, share, and search annotations of web-based articles. For the purposes of Clingen Gene Curation, annotated articles may include primary research papers, reviews, and web pages (e.g. OMIM, GeneReviews). The goal of annotating articles for the ClinGen gene curation process is to allow multiple users to be able to search for, and easily find, relevant information within an article. For example, a curator could annotate an article in order to share relevant information with other members of a gene curation committee.

Required components

- Prior to using Hypothes.is to tag articles for ClinGen gene curation, users should be familiar with the ClinGen gene curation framework including:
    - Gene Clinical Validity Curation Process SOP.
    - Strande and Riggs et al, 2017, Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. Am J Hum Genet. 100:895-906 (PMID: 28552198).
    - Guidelines of the ClinGen Lumping and Splitting Working Group.
- Users will also require:
    - Internet access
    - A Hypothes.is account. To sign up for a Hypothes.is account, go to https://web.hypothes.is/
        - Click "get started" in the top right to sign up.
        - Follow directions on the Hypothes.is website to install Chrome extension.
- Basic knowledge on how to use Hypothes.is and further details on obtaining an account are available in a presentation on Hypothes.is, given to the ClinGen Biocurator WG by Maryanne Martone, PhD, at **http://tinyurl.com/y9gkjyht**
  Curators must view this presentation prior to annotating articles for the ClinGen gene curation process in order to have general instructions on how to use Hypothes.is.

List of websites that may be needed during annotation:

- PubMed: https://www.ncbi.nlm.nih.gov/pubmed
- GeneReviews: https://www.ncbi.nlm.nih.gov/books/NBK1116/
- OMIM: http://omim.org/
- ClinVar: https://www.ncbi.nlm.nih.gov/clinvar/
- ExAC: http://exac.broadinstitute.org/
- gnomAD: http://gnomad.broadinstitute.org/
- HGNC: http://www.genenames.org/
- HPO: http://compbio.charite.de/hpoweb/showterm?id=HP:0000118
- MonDO: https://www.ebi.ac.uk/ols/ontologies/mondo

Annotating articles

The goal of annotating articles for the ClinGen gene curation process is to allow multiple users to be able to search for, and easily find, relevant information within an article. In order to do so, a set of tags has been developed. These tags were chosen to match specific fields in the Gene Curation Interface (GCI). NOTE: When searching an annotated article, Hypothesis cannot search for tags with multiple words. **Therefore, camel case is used for tagging articles e.g. the tag for Protein Interaction is ProteinInteraction**.

Not all information in the article may be relevant to the current curation. However, feel free to highlight and/ or add tags that could be helpful to others in the future.

1. Tag the title of the article with:
    - ClinGen
    - General
    - The HGNC name of the gene
    - **In cases where a gene is involved in more than one disease entity, add the tag for the disease acronym.**
        - For example, the gene *RET* is associated with multiple endocrine neoplasia type 2 A and B, so you can tag with two possible tags here: (1) RET and MEN2A, and (2) RET and MEN2B

2. Sub-tag the following information within the annotation
    - The PMID of the article
    - The gene name (using HGNC nomenclature)
    - The name of the disease being curated.
    - The MonDO ID for the disease.
        i. Ideally, the MonDO ID should match the MonDO ID that will be entered into the Gene Curation Interface (GCI). However, as many human genes are associated with more than one disorder, it is understood that the MonDO ID may not yet have been decided upon as the curator begins to collect and assess the literature. In that case, the appropriate MonDO ID can be added once the disease entity has been approved by the Gene Curation Expert Panel (GCEP).

3. Annotate the article.
    - Annotate and tag pertinent areas of text according to the types of evidence collected for assessing gene:disease validity using the ClinGen Gene Curation framework. A list of tags to be used is given in Table 1. These tags were chosen to match fields in the Gene Curation Interface and must be used in order to maintain consistency for different users of the same tagged article. Primary tags describe major categories of information to be curated. Secondary tags used within the annotation, sub-categorize the evidence described by the primary tag e.g. a variant identified in an individual would be give the primary tag “Individual” and the secondary tag “variant” links to the HGVS variant nomenclature, and either the ClinVar or Clingen Allele Registry ID.
    - ***For Case level annotation, a curator may find it helpful to add all the information for each case report in one annotation. In doing so you can add all the tags that are listed. This can allow for***

easy access to the information in one annotation to transfer information to the GCI for curation (refer to helpful notes #2).

- If there are multiple genes/proteins included in one article, include the specific gene name tag on each annotation to ensure that the correct evidence is linked to the specific gene.
- **IMPORTANT: After adding tags, make sure to click "post to" and choose the appropriate group. Otherwise, the annotations will not be saved.**

Table 1. Tags to be used in annotation

Category	Primary tag	Secondary tag	Additional information to be added to annotation
General	ClinGen General	HGNC symbol of gene of interest, i.e. BRCA1, ATM, etc.	**PMID:** add PMID or article **Gene:** add HGNC name **Disease:** lay term of disease **MonDO:** label with MonDO ID for the disease entity being curated if known, or the GCEP approved disease entity nomenclature. ***Note:*** if no MonDO ID, label with the MIM phenotype number in OMIM. **InheritancePattern:** note inheritance pattern being curated. **Prevalence:** only if known, not required **Penetrance:** only if known, not required.
	Contradictory Evidence		**ContradictoryEvidence:** This is to highlight that this paper contains contradictory evidence. Annotate and highlight the contradictory information. Add appropriate notes to capture the contradictory evidence.
Lumping and splitting	LumpingSplitting **Note:** the article constitutes the Assertion criteria.		**DiseaseMechanism:** Annotate notes pertaining to difference in disease mechanism **PhenotypicVariability:** Annotate with notes assessing the phenotype, whether intra- or inter-familial variability or consistency observed. **InheritancePattern:** Annotate with any noted changes in inheritance pattern from the one being curated for.

Case	Group	GroupPhenotypes	**GroupName:** Annotate with group name, either listed in publication or one that is generated based on the available information. **GroupPresentingHPOs:** Annotate with the HPOs that all members of the group tested have in common. **GroupNotHPOs:** Annotate with the HPOs that were noted **not** to manifest within all members of the groups tested.
		GroupDemographics	**List as known:** **NumberMales:** as mentioned **NumberFemales:** as mentioned **CountryOrigin:** as mentioned **Ethnicity:** as mentioned **Race:** as mentioned **AgeRange:** as mentioned **TotalNumber**: total number of individuals in group **FamilyInfo:** # individuals with family information. **NoFamily:** # of individuals without family **VariantCount:** # individuals with variant of interest **NonVariant:** # individuals with no variation in the gene of interest. **OtherVariant:** # individuals with variant found in another gene. And list the other genes.
		GroupGenotypingMethod	**GroupPreviousTesting:** note if occurred, describe results, and note is genome wide analysis methods were used. **GroupMethod1:** as mentioned **GroupMethod2:** as mentioned **GroupGenotypingMethod:** as mentioned
		GroupAddInfo	**GroupAddInfo:** Annotate any other pertinent information mentioned about the group.
	Family	FamilyInfo	**FamilyLabel:** Annotate with the family name or number, if applicable.

			Family#PresentingHPOs: Annotate with the HPOs that all members of the family have in common. **FamilyNotHPOs:** Annotate with the HPOs that were noted **not** to manifest within all members of the family tested. **CountryOrigin:** as mentioned **Ethnicity:** as mentioned **Race:** as mentioned **Family#PreviousTesting:** Annotate with a description of method and whether genome-wide analysis methods were used. **Family#Method1:** as mentioned **Family#Method2:** as mentioned **Family#GenotypingMethod:** Describe method as needed, and any additional pertinent information about the family method. **Affected:** as mentioned **Unaffected:** as mentioned. Required for autosomal recessive inheritance. **Segregations:** Annotate the number of segregations reported, and whether any were inconsistent. Also note if consanguinity was mentioned and the figure containing the pedigree for the family in question. **PublishedLOD:** record if mentioned in the article **eLOD:** record if you have estimated the LOD per the ClinGen segregation protocols in the SOP. Check that the GCI calculation matches your eLOD. **AddSegregationInfo:** Annotate with other pertinent information as needed. **ProbandLabel:** record the proband label for the individual in the family you plan to score. **Variant:** Annotate with the HGNC nomenclature for the variant **ClinVar:** if known

			CAID: If no associated ClinVar ID, register the variant with the ClinGen Allele Registry (CA) and copy or create a CA ID. **gnomAD:** Annotate with the maximum allele frequency for the variant in question if found in gnomAD, and add the link to the specific variant page. **FamilyAddInfo:** Annotate with pertinent information as needed. **FamilyPMIDs:** Annotate with the PMIDs that correspond with previous publications about the family of interest.
	Individual	CaseInfo	**Case#:** Annotate with the Individual label, sex, age, ethnicity if known. **CasePresentingHPOs:** Annotate with the HPOs presenting specifically in the proband of interest. **CaseHPOFreeText:** Annotate with phenotyping information that does not have an appropriate HPO number, or requires explanation. **CaseNotHPOs:** Annotate with the HPOs that were noted **not** to manifest within the proband of interest. **CaseNotHPOFreeText:** Annotate with phenotyping information that appears normal or unaltered in the patient, yet does not have an appropriate HPO number, or requires explanation. **CasePreviousTesting:** Annotate with a description of method and whether genome-wide analysis methods were used. **CaseMethod1:** as mentioned **CaseMethod2:** as mentioned **CaseGenotypingMethod:** Describe method as needed, and any additional pertinent information about the methods used on the proband.

			Variant: Annotate with the HGVS nomenclature for the variant. ***Note:*** *For autosomal recessive inheritance of compound mutations, you will need to separate variant entries and two separate ClinVar IDs or CA IDs.* **ClinVar:** if known **CAID:** If no associated ClinVar ID, register the variant with the ClinGen Allele Registry (CA) and copy or create a CA ID. **gnomAD:** Annotate with the maximum allele frequency for the variant in question if found in gnomAD, and add the link to the specific variant page. **VariantEvidence:** Annotate with information that supports the pathogenicity of the variant. Include PMIDs if the evidence was presented in another paper. ***This is useful for non-LOF variants and required for awarding more than 0.1 pts.*** **CaseAddInfo:** Annotate with any additional information about the proband that is pertinent to the curation. **CasePMIDs:** Annotate with PMID numbers of article that have previously reported the proband of interest.
Case-control studies	Case-control	Controls	**CaseControlLabel:** Annotate with a label to describe the data. **CaseCohortLabel:** Annotate with a label to describe the case cohort. **ControlCohortLabel:** Annotate with a label to describe the control cohort. **CaseControlPhenotype:** Annotate with the HPOs presenting within all of the case control cohort. **CaseControlNOTPhenotype:** Annotate with the HPOs numbers noted as **NOT** being present in the case cohort.

			CaseDemographics: Annotate with the demographics if known, including number of males and females, country of origin, ethnicity, race, and age range. **ControlDemographics:** Annotate with the demographics if known, including number of males and females, country of origin, ethnicity, race, and age range. **CaseMethod:** Annotate with information describing previous testing and genotyping methods used. **ControlMethod:** Annotate with information describing previous testing and genotyping methods used. **CasePower:** Annotate with the number of cases with the variants in the gene in question, and include the number of all cases genotyped. **ControlPower:** Annotate with the number of Controls that had a variant in the gene in question, and include the number of all cases genotyped. **CaseAddInfo:** Annotate with any additional information about the case cohort that is pertinent to the curation, including PMIDS were evidence may have previously been reported. **ControlAddInfo:** Annotate with any additional information about the control cohort that is pertinent to the curation, including PMIDS were evidence may have previously been reported. **CaseControlEvaluation:** Annotate with the study type and detection method used. **CaseControlStats:** Annotate with the test statistics, value, p-value, and confidence interval if known. **CaseControlBias:** Annotate with information regarding the

			matching of demographic information, matching for ancestry, evaluation of diseaseoutcome, and difference in variables between cases and controls. **CaseControlComments:** Annotate with any additional pertinent information about the study.

Table 1. Tags to be used in annotation (continued)

Category	**Primary tag**	**Secondary tag**	**Additional information to be added to annotation**
Experimental	Function	BiochemicalFunction	**Selection A- Gene with same function implicated in disease** **ExperimentName:** per curator's discretion **GOterm:** Add appropriate GO term(s) for the gene sharing the same function. **FunctionalEvidence:** Describe the evidence. **HGNC:** Annotate with the HGNC nomenclature of the gene with a shared function. **GeneEvidence:** Annotate with the evidence for the same function between the genes. **SharedGeneImplication:** Annotate with notes on how the shared gene was implicated in the disease. **Selection B- Gene function is consistent with phenotype.** **GOterm:** Add appropriate GO term(s) for the gene sharing the same function. **FunctionalEvidence:** Describe the evidence. **FunctionalHPO:** Annotate with HPO terms of the phenotype consistent with function.

			PhenotypeEvidence: Annotate with an explanation of how the phenotype is consistent with disease. Note where the evidence was shown (figure, table, etc.).
		ProteinInteraction	**ExperimentName:** per curator's discretion **InteractingGene:** Provide HGNC symbol for the interacting gene. *Note: this gene must be implicated in the same disease as the gene being curated.* **InteractionType:** Annotate with the interaction type, i.e. physical, genetic, etc. **InteractionMethod:** Annotate with the method used to identify the interaction, i.e. yeast-2-hybrid (Y2H), coimmunoprecipitation, affinity chromatography, etc. **InteractionEvidence:** Annotate with a description of the evidence supporting the interaction. Note where the evidence can be found (figure, etc).
		Expression	**Selection A- Gene is expressed in tissues relevant to disease:** **ExperimentName:** per curator's discretion **TissueExpression:** Annotate with tissue name and the UBERON number for the tissue that gene expression was observed in. **ExpressionEvidence:** Annotate with the evidence that supports the expression. Note where the evidence is found in the article.

			Selection B- Expression altered in patients: **ExperimentName:** per curator's discretion **TissueExpression:** Annotate with tissue name and the UBERON number for the tissue that gene expression was observed in. **ExpressionEvidence:** Annotate with the evidence that supports the expression. Note where the evidence is found in the article.
	FunctionalAlteration	PatientCells	**ExperimentName:** per curator's discretion **PatientCell:** Annotate with the name and Cell Ontology number for the cell line used. **GeneFunction:** Annotate with a GO term describing the normal function of the gene product and the GO number. **GeneAlteration:** Describe the gene alteration. **FunctionalEvidence:** Annotate with a description of the evidence supporting an altered function. Note where evidence is provided in the article. **FAVariants:** If a variant for the gene of interest was used, annotate with the HGNC nomenclature to describe the variant. *Also annotate with the ClinVar ID or CA ID.*
		NonPatientCells	**ExperimentName:** per curator's discretion **NonPatientCell:** Annotate with the name and either the EFO or Cell Ontology number for the cell line used. **GeneFunction:** Annotate with GO term describing the normal function of the gene product and the GO number.

			GeneAlteration: Describe the gene alteration methods used. **FunctionalEvidence:** Annotate with a description of the evidence supporting an altered function. Note where evidence is provided in the article. **FAVariants:** If a variant for the gene of interest was used, annotate with the HGNC nomenclature to describe the variant. *Also annotate with the ClinVar ID or CA ID.*
	ModelSystems	ModelOrganism	**ExperimentName:** per curator's discretion **NonHumanModel:** Annotate the species of the model system. **GeneAlteration:** Annotate with a description of the genetic alteration of the model. **ModelPhenotype:** Annotate with HPOs terms that are present in the non-human model system. **HumanPhenotype:** Annotate with HPOs terms that present in the human disease that are similar to the ones presentin in the mouse. **ModelEvidence:** Annotate with an explanation of how the model system phenotype recapitulates the human phenotype. Note where the evidence is found in the article. **ModelVariant:** Annotate with the HGNC nomenclature of the variant(s), as well as the *ClinVar ID or CA ID for that variant.*
		CellCultureModel	**ExperimentName:** per curator's discretion

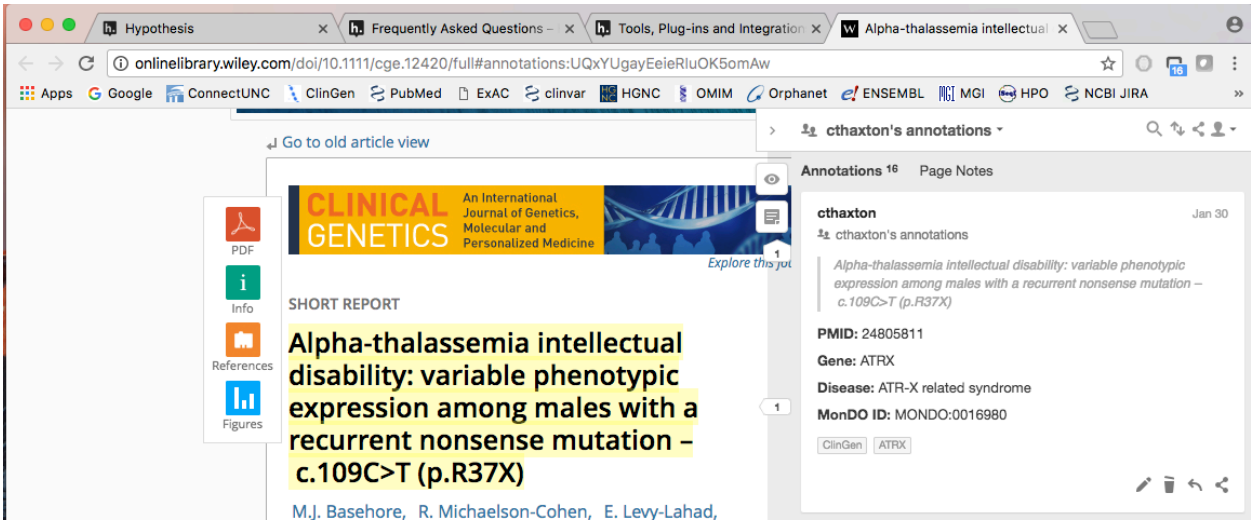
			CellLine: Annotate with the name of the cell line used. **CellEFO:** Provide the EFO or Cell Ontology number. **GeneAlteration:** Describe the genetic alteration of the cell line. **CellPhenotype:** Annotate with the phenotype observed and provide appropriate HPOs terms. **CellEvidence:** Annotate with a description of how the cell phenotype correlates with the observed human phenotypes. Note where the evidence is found in the article. **CellVariant:** Annotate with the HGNC nomenclature of the variant(s), as well as the *ClinVar ID or CA ID for that variant.*
	Rescue	RescueHuman	**ExperimentName:** per curator's discretion **RescueProband:** Annotate with proband name and/or description. **GeneAlteration:** Describe the genetic alteration in the proband. **RescueVariant:** Annotate with the HGNC nomenclature of the variant(s), as well as the *ClinVar ID or CA ID for that variant.* **RescuedPhenotype:** Annotate with the appropriate HPOs terms of the phenotype preent in the proband that is being targeted for rescue. **RescueMethod:** Describe the method used to rescue the phenotype in a human. **RescueEvidence:** Annotate with an explanation of the rescue of phenotype. Note

			where evidence can be found in the article.
		RescueModelOrganism	**ExperimentName:** per curator's discretion **RescueModel:** Annotate the species of the model system. **GeneAlteration:** Annotate with a description of the genetic alteration of the model. **RescueVariant:** Annotate with the HGNC nomenclature of the variant(s), as well as the *ClinVar ID or CA ID for that variant.* **RescuePhenotype:** Annotate with HPOs terms that are present in the non-human model system. **RescueMethod:** Annotate with a description of the method used to rescue the phenotype. **RescueEvidence:** Annotate with an explanation of how the model system phenotype is rescued. Note where the evidence is found in the article. **ModelVariant:** Annotate with the HGNC nomenclature of the variant(s), as well as the *ClinVar ID or CA ID for that variant.*
		RescueCellCulture	**ExperimentName:** per curator's discretion **RescueCellLine:** Annotate with the name of the cell line used. **RescueCellEFO:** Provide the EFO or Cell Ontology number. **GeneAlteration:** Describe the genetic alteration present in the cell line. **RescueCellPhenotype:** Annotate with the phenotype rescued and provide appropriate HPOs terms.

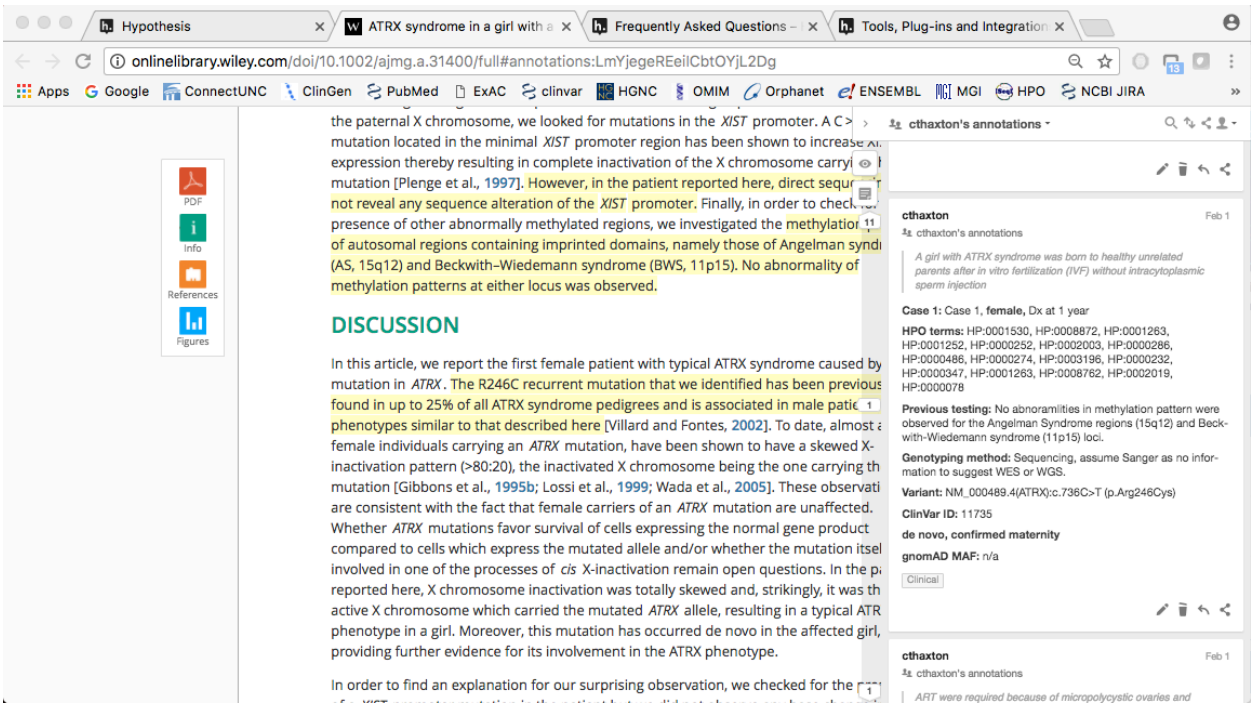
			RescueMethod: Annotate with a description of the method used to rescue the phenotype. **RescueEvidence:** Annotate with a description of how the cell phenotype is rescued. Note where the evidence is found in the article. **RescueCellVariant:** Annotate with the HGNC nomenclature of the variant(s), as well as the *ClinVar ID or CA ID for that variant.*
		RescuePatientCells	**ExperimentName:** per curator's discretion **RescuePatientCell:** Annotate with the name and Cell Ontology number for the cell line used. **GeneAlteration:** Describe the gene alteration present in the patient cells. **RescuePhenotype:** Annotate with the phenotype rescued and provide appropriate HPOs terms. **RescueMethod:** Annotate with a description of the method used to rescue the phenotype. **RescueEvidence:** Annotate with a description of how the cell phenotype is rescued. Note where the evidence is found in the article. **RescueVariants:** If a variant for the gene of interest was used, annotate with the HGNC nomenclature to describe the variant. *Also annotate with the ClinVar ID or CA ID.*

Helpful notes and sample images of Hypothes.is annotations:

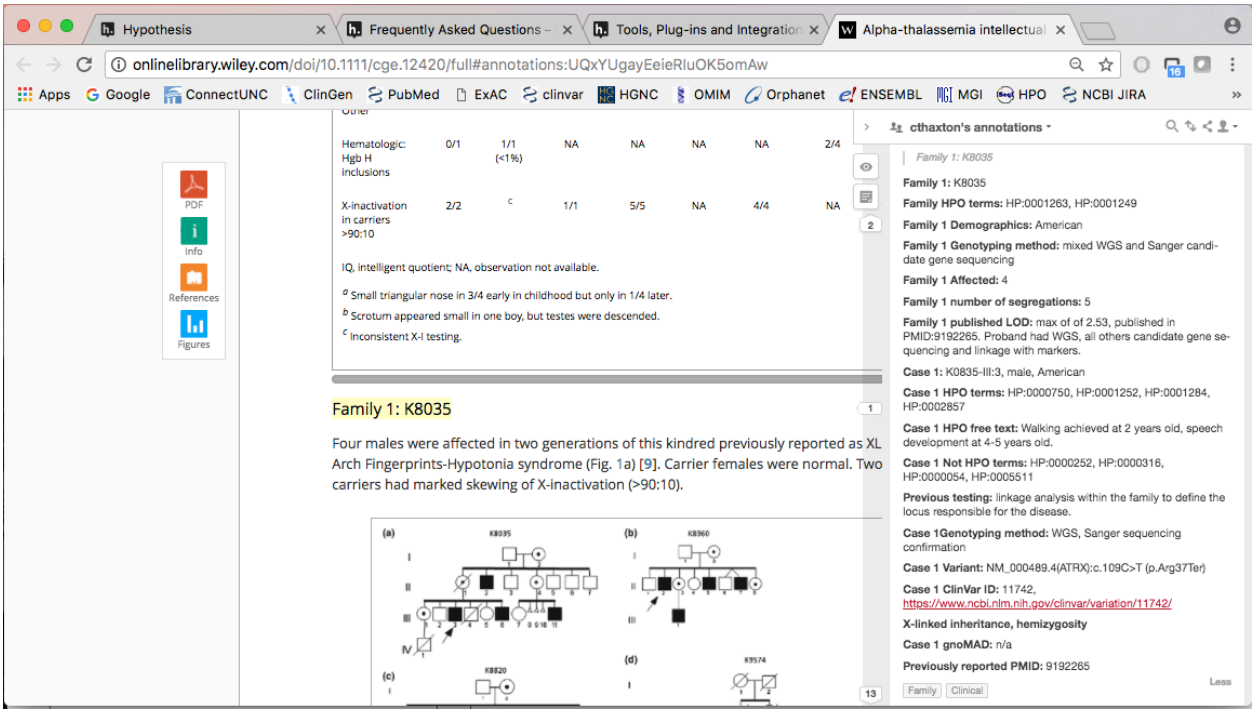
1. Highlight title of article and annotate with PMID, Gene name, Disease name, and MonDO ID for the disease entity.



2. For Case level data, annotate with all the pertinent information as outlined in the SOP in one annotation per case:



3. For Case level data, that has family data (i.e. segregation, etc), and case level proband data, you can combine the two sets of information into one annotation to help with ease of data transfer into the GCI:



4. In cases where multiple PMIDs may reflect the same data it is useful to annotate and highlight/ bold these additional PMIDS for entry into the GCI. A good example is mouse models, in which the base model may have been studied multiple times in the literature. MGI (Mouse Genome Informatics) catalogs almost all genetic strains of mice based on the gene alteration and has compiled lists of references.