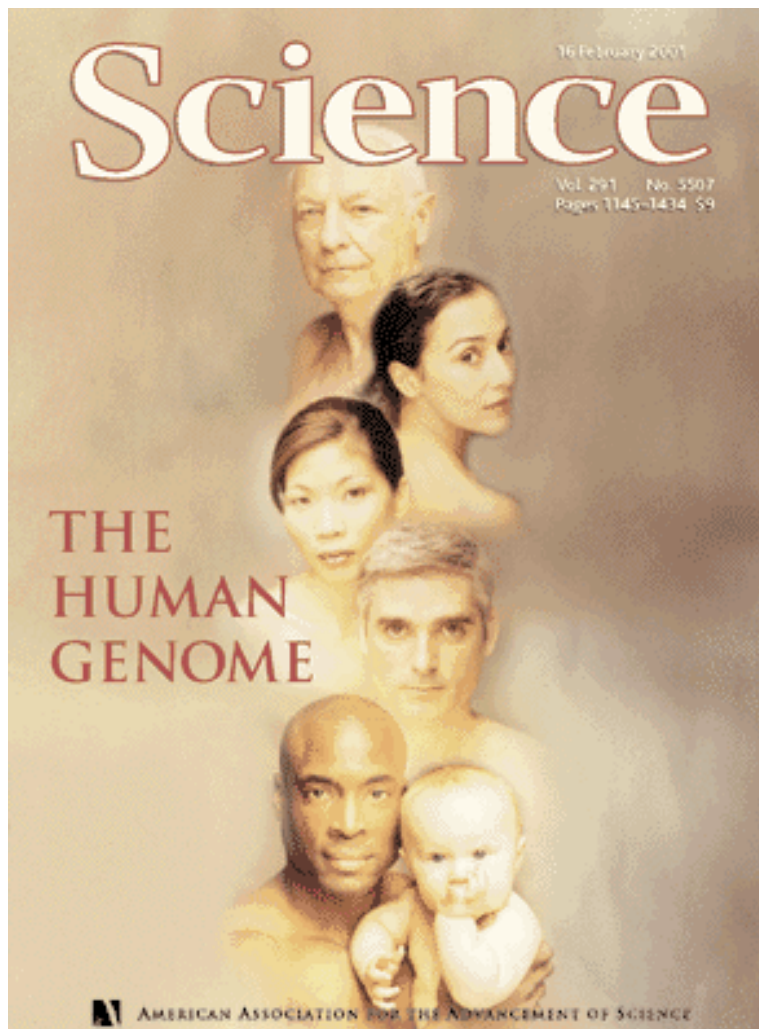


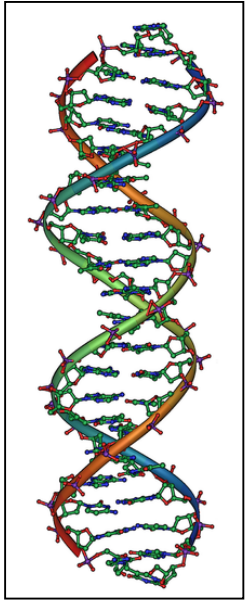
# Annotation Error in Public Databases

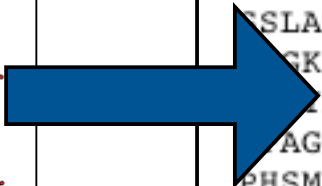
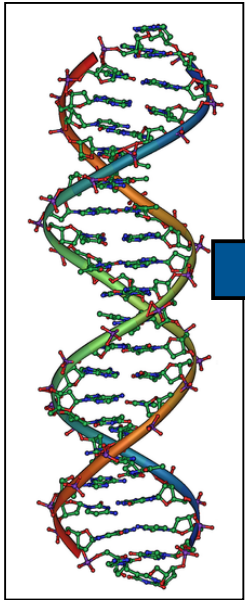
---

**ALEXANDRA SCHNOES**  
**UNIVERSITY OF CALIFORNIA, SAN FRANCISCO**  
**OCTOBER 25, 2010**



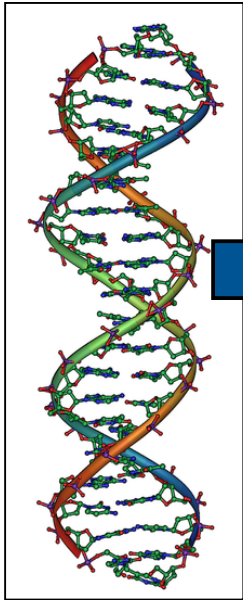
New genomes (and metagenomes) sequenced every day...





MKPRLETSQEFLEGRNI  
RPARMIFRNVPVDPQSI  
SSLATSLLNPDQMQLVI  
GKMLLGRRHVLPVSI  
HPLKMPGAVIPVSI  
AGTAIRFEPGDSKTI  
PHSMDRESYMRMFGATI  
ITNALIVDWTGIYVADI  
ITPDQVQEALASGVTTI  
QIRAGAAGLKLHEDWGI  
GCHADDTTSAVQEDNVI

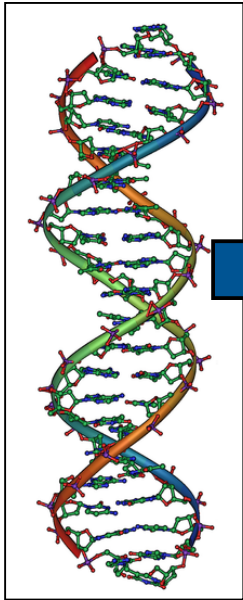




MKPRLETSQEFLEGRNI  
RPARMIFRNVPVPDQSI  
SSLATSLLNPDQMQLVI  
GKMMMLGRRHVLPSTI  
HPLKMPGAVI  
AGTAIRFEPGDSKTV  
PHSMDRESYMRMFGATV  
ITNALIVDWTGIYVADV  
ITPDQVQEALASGVTTI  
QIRAGAAGLKLHEDWGC  
GCHADDTTSAVQEDNVI

**Results: 1 to 20 of 782**

- [RecName: Full=Chloromuconate cycloisomerase; AltName: F](#)  
1. 370 aa protein  
P05404.4 GI:135651  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Protein](#)
- [RecName: Full=Chloromuconate cycloisomerase; AltName: F](#)  
2. 370 aa protein  
P27099.1 GI:135517  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Protein](#)
- [RecName: Full=Chloromuconate cycloisomerase; AltName: F](#)  
3. 372 aa protein

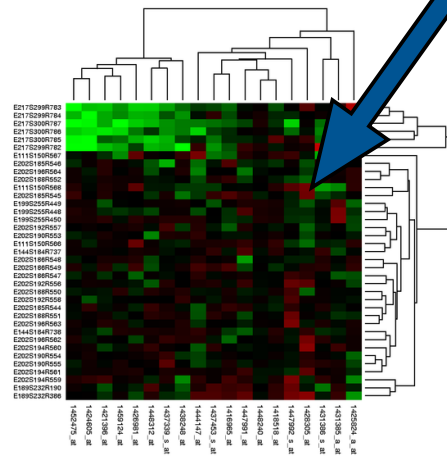


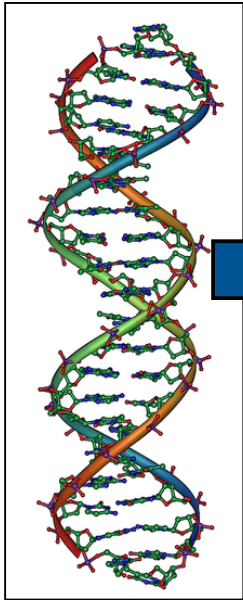
```

MKPRLETSQEFLEGRNI
RPARMIFRNVPVPDQS/
ESLATSLLNPDQMQLVI
GKMMMLGRRHVLPSTV
HPLKMPGAVI
AGTAIRFEPGDSKTV
PHSMDRESYMRMGATV
ITNALIVDWTGIYVADV
ITPDQVQEALASGVTTI
QIRAGAAGLKLHEDWGC
GCHADDTTSAVQRPNUV
  
```

**Results: 1 to 20 of 782**

- [RecName: Full=Chloromuconate cycloisomerase; AltName: F](#)
- 1. 370 aa protein  
P05404.4 GI:135651  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Protein](#)
- [RecName: Full=Chloromuconate cycloisomerase; AltName: F](#)
- 2. 370 aa protein  
P27099.1 GI:135517  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Protein](#)
- [RecName: Full=Chloromuconate cycloisomerase; AltName: F](#)
- 3. 370 aa protein



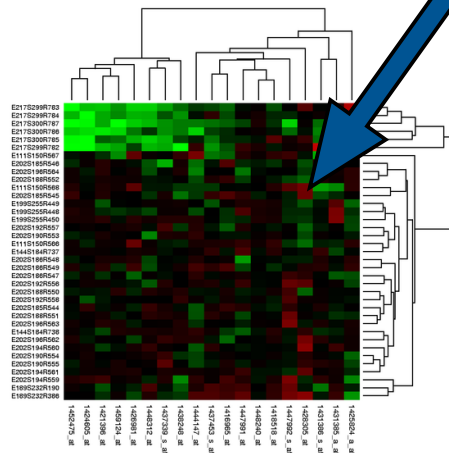


```

MKPRLETSQEFLEGRNI
RPARMI FRNVVDPQS/
ESLATSLLNPDQMLVI
GKMLLGRRHVLP SVI
HPLKMPGAVI
AGTAIRFEPGDSKTV
PHSMDRESYMRMGATV
ITNALIVDWTGIYVADV
ITPDQVQEALASGVTTI
QIRAGAAGLKLHEDWGC
GCHADDTTSAVQEDNVI
  
```

**Results: 1 to 20 of 782**

- [RecName: Full=Chloromuconate cycloisomerase; AltName: F](#)
- 1. 370 aa protein  
P05404.4 GI:135651  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Protein](#)
- [RecName: Full=Chloromuconate cycloisomerase; AltName: F](#)
- 2. 370 aa protein  
P27099.1 GI:135517  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Protein](#)
- [RecName: Full=Chloromuconate cycloisomerase; AltName: F](#)
- 3. 370 aa protein



**RESEARCH ARTICLE**

### Sequencing and Analysis of Neanderthal Genomic DNA

James P. Noonan,<sup>1,2</sup> Graham Coop,<sup>1</sup> Siddhartha Sarkar,<sup>1,2</sup> Dong Song,<sup>1</sup> Johannes Krause,<sup>1</sup> Ian Hogg,<sup>1</sup> Yang Chen,<sup>1</sup> Steven Pääbo,<sup>1</sup> Svante Pääbo<sup>1</sup>

**Abstract**  
Our knowledge of Neanderthals is based on a limited number of skeletal and artifact remains which we need better information about their biology, behavior, and evolutionary history. Here, we describe the first identification of the entire Neanderthal genome from a single high-throughput sequencing and analysis. Several lines of evidence indicate that the 45,216 base pair (bp) Neanderthal genome is identical to the 45,216 bp of Neanderthal origin, the divergence between the two genomes is approximately 0.1%. Several Neanderthal genes diverge from the human genome, including the Neanderthal-specific gene *FOXP2*, which is involved in language. We also identify a Neanderthal-specific gene, *STAT4*, which is involved in the immune system. We also identify a Neanderthal-specific gene, *STAT4*, which is involved in the immune system. We also identify a Neanderthal-specific gene, *STAT4*, which is involved in the immune system.

**Introduction**  
Neanderthals are the closest extinct relatives of modern humans (1). An early Neanderthal skeleton was discovered in 1908 at the site of Neander in the Neander Valley, Germany (2). The discovery of Neanderthals led to the development of a new branch of paleoanthropology, Neanderthal studies, which focuses on the biology, behavior, and evolutionary history of this extinct hominid species. Neanderthals are considered to be a subspecies of modern humans, *Homo neanderthalensis*, which diverged from the modern human lineage approximately 400,000 years ago (3). The discovery of Neanderthals led to the development of a new branch of paleoanthropology, Neanderthal studies, which focuses on the biology, behavior, and evolutionary history of this extinct hominid species.

**Methods**  
We used a high-throughput sequencing approach to identify and sequence Neanderthal DNA. We used a high-throughput sequencing approach to identify and sequence Neanderthal DNA. We used a high-throughput sequencing approach to identify and sequence Neanderthal DNA.

**Results**  
We identified the first Neanderthal genome from a single high-throughput sequencing and analysis. Several lines of evidence indicate that the 45,216 base pair (bp) Neanderthal genome is identical to the 45,216 bp of Neanderthal origin, the divergence between the two genomes is approximately 0.1%. Several Neanderthal genes diverge from the human genome, including the Neanderthal-specific gene *FOXP2*, which is involved in language. We also identify a Neanderthal-specific gene, *STAT4*, which is involved in the immune system. We also identify a Neanderthal-specific gene, *STAT4*, which is involved in the immune system.

**Discussion**  
Our findings provide the first complete Neanderthal genome, which is identical to the 45,216 bp of Neanderthal origin. This discovery provides a valuable resource for studying the biology, behavior, and evolutionary history of Neanderthals. We also identify several Neanderthal-specific genes, including *FOXP2* and *STAT4*, which are involved in language and the immune system, respectively. These findings provide a valuable resource for studying the biology, behavior, and evolutionary history of Neanderthals.

**Conclusion**  
We have identified the first Neanderthal genome from a single high-throughput sequencing and analysis. This discovery provides a valuable resource for studying the biology, behavior, and evolutionary history of Neanderthals. We also identify several Neanderthal-specific genes, including *FOXP2* and *STAT4*, which are involved in language and the immune system, respectively. These findings provide a valuable resource for studying the biology, behavior, and evolutionary history of Neanderthals.

**Supplementary Information**  
Supplementary Information is available for this article. Visit [http://www.sciencemag.org/suppl/](#) to view the full version of this article.

**References**  
1. Greenberg, J. H. R. *The Race that Brought Us* (Oxford University Press, 1994).  
2. Schönerer, A. *Neanderthal* (Springer, 2004).  
3. Harpending, D. C. *Neanderthal Demography* (Science, 1994).

**Fig. 1. Generation of ancient Neanderthal genomic DNA for direct selection and sequencing.**

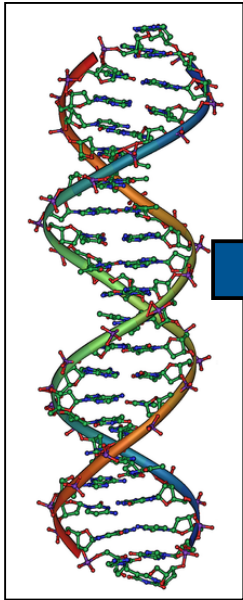
**Fig. 1. Generation of ancient Neanderthal genomic DNA for direct selection and sequencing.**

**Fig. 1. Generation of ancient Neanderthal genomic DNA for direct selection and sequencing.**









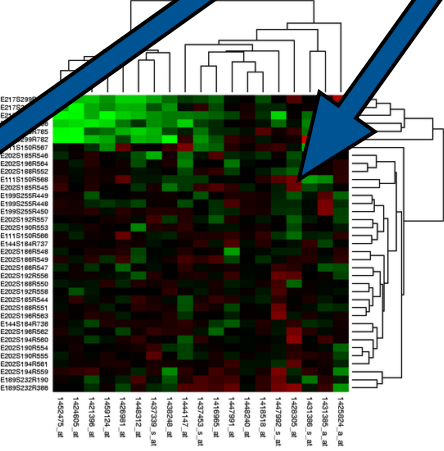
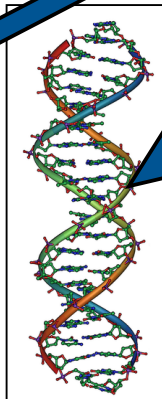
```

MKPRLETSQEFLEGRNI
RPARMI FRNVVVPDQS/
SLSLATSLLNPDQMLVI
GKMMMLGRRHVLPSVI
HPLKMPGAVI
AGTAIRFEPGDSKTV
PHSMDRESYMRMGATV
ITNALIVDWTGIYVADV
ITPDQVQEALASGVTTI
QIRAGAAGLKLHEDWGC
GCHADDTTSAVQPRNVI

```

**Results: 1 to 20 of 782**

- [RecName: Full=Chloromuconate cycloisomerase; AltName: F](#)  
 1. 370 aa protein  
 P05404.4 GI:135651  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Protein](#)
- [RecName: Full=Chloromuconate cycloisomerase; AltName: F](#)  
 2. 370 aa protein  
 P27099.1 GI:135517  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Protein](#)
- [RecName: Full=Chloromuconate cycloisomerase; AltName: F](#)  
 370 aa protein



**RESEARCH ARTICLE**

**Sequencing and Analysis of Neanderthal Genomic DNA**

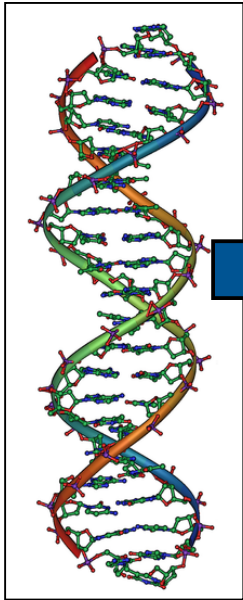
Jesse J. Hillier,<sup>1</sup> Graham Coop,<sup>1</sup> Stephen Huddleston,<sup>1</sup> Dong Shao,<sup>1</sup> Johannes Krause,<sup>1</sup> Ian Hogg,<sup>1</sup> Yang Chen,<sup>1</sup> Steven Pääbo,<sup>1</sup> Svante Pääbo,<sup>1</sup> Jonathan A. Hradek,<sup>1</sup> Tracy M. S. Black,<sup>1</sup>

Our knowledge of Neanderthals is based on a limited number of remains and artifacts from which we must make inferences about their biology, behavior, and evolutionary history. Here, we describe the identification of the nuclear genomes from a new perspective, based on the development of a Neanderthal reference library used for high-throughput sequencing and analysis. Several lines of evidence indicate that the 65210 base pair Neanderthal sequence is the first of its kind. The Neanderthal sequence is highly similar to the modern human genome, but it differs in several key regions. These include the Neanderthal-specific gene, the Neanderthal-specific gene, and the Neanderthal-specific gene. The Neanderthal-specific gene is located on the X chromosome and is highly conserved across all modern human populations. The Neanderthal-specific gene is located on the X chromosome and is highly conserved across all modern human populations. The Neanderthal-specific gene is located on the X chromosome and is highly conserved across all modern human populations.

**Fig. 1. Generation of ancient integrative library DNA for direct selection and sequencing.**

The diagram illustrates the process of generating an ancient integrative library DNA for direct selection and sequencing. It starts with a DNA sample, which is then fragmented and ligated with a specific adapter. This is followed by PCR amplification and sequencing. The resulting data is then analyzed to identify specific genes of interest.

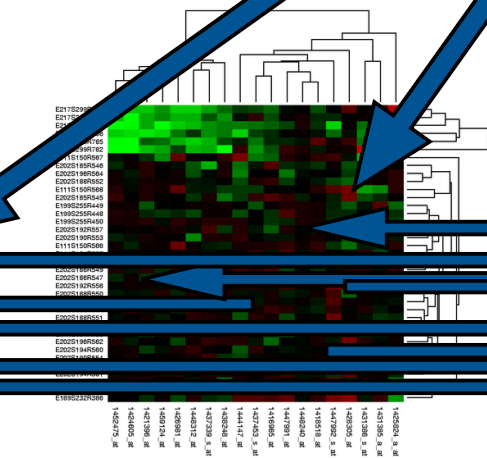
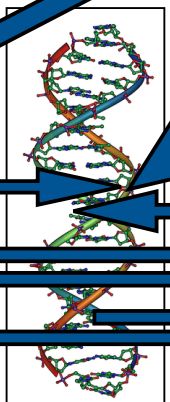
- [RecName: Full=Chlor](#)  
 1. 370 aa protein  
 P05404.4 GI:135651  
[GenPept](#) [FASTA](#) [Gra](#)
- [RecName: Full=Chlor](#)  
 2. 370 aa protein  
 P27099.1 GI:135517  
[GenPept](#) [FASTA](#) [Gra](#)
- [RecName: Full=Chlor](#)  
 3. 372 aa protein



MKPRLETSQEFLEGRNI  
 RPARMI FRNVVPDQS/  
 ESLATSL LNP DQMQLVI  
 CKMMLGRRHVLP SVI  
 HPLKMPGAVI  
 AGTAIRFEPGDSKT  
 PHSM DRESYMRMGAT  
 ITNALIVDWTGIYVAD  
 ITPDQVQEALASGVT  
 QIRAGAAGLKLHEDWGC  
 GCHADDTTSAVQRPNU

**Results: 1 to 20 of 782**

- [RecName: Full=Chloromuconate cycloisomerase; AltName: F](#)  
 1. 370 aa protein  
 P05404.4 GI:135651  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Protein](#)
- [RecName: Full=Chloromuconate cycloisomerase; AltName: F](#)  
 2. 370 aa protein  
 P27099.1 GI:135517  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Protein](#)
- [RecName: Full=Chloromuconate cycloisomerase; AltName: F](#)  
 370 aa protein



**RESEARCH ARTICLE**

**Sequencing and Analysis of Neanderthal Genomic DNA**

James R. Noonan,<sup>1,2</sup> Graham Coop,<sup>1</sup> Siddharth Nadkarni,<sup>1</sup> Doug Smith,<sup>1</sup> Matthew Krause,<sup>1</sup> Jay Altmann,<sup>1</sup> Yang Chen,<sup>1</sup> Steven Pääbo,<sup>1</sup> Svante Pääbo<sup>1\*</sup>

**Abstract**  
 Our knowledge of Neanderthal is based on a limited number of skeletal and artifact remains which we need better information about their biology, behavior, and evolutionary history. Here, we describe the characterization of the entire Neanderthal genome from a new perspective, based on the development of a Neanderthal reference library used for high-throughput sequencing and analysis. Several lines of evidence indicate that the 65210 base pair Neanderthal sequence is identical to the 28.3% of Neanderthal origin, the strongest evidence for the identification of Neanderthal origin. Several Neanderthal and chimpanzee genomes were sequenced to compare the Neanderthal genome to other hominid genomes. These results indicate that the Neanderthal genome is distinct from other hominid genomes and that the Neanderthal genome is closely related to the modern human genome. The Neanderthal genome sequence we obtained is of the same size as the human genome, with a total size of approximately 3.1 Gb. The Neanderthal genome is estimated to have diverged from the human genome approximately 380,000 years ago, before the emergence of anatomically modern humans. Our finding that the Neanderthal and human genomes are at least 99.7% identical led us to develop and successfully implement a targeted method for recovering specific ancient DNA sequences from non-specific libraries. This method applied to the Neanderthal genome sequence of the 40,000-year-old Vindija Neanderthal skeleton (Vindija 1) and the 40,000-year-old Mezmaiskaya Neanderthal skeleton (Mezmaiskaya 1) and was used to identify Neanderthal-specific variants. The Neanderthal genome sequence we obtained is of the same size as the human genome, with a total size of approximately 3.1 Gb. The Neanderthal genome is estimated to have diverged from the human genome approximately 380,000 years ago, before the emergence of anatomically modern humans. Our finding that the Neanderthal and human genomes are at least 99.7% identical led us to develop and successfully implement a targeted method for recovering specific ancient DNA sequences from non-specific libraries. This method applied to the Neanderthal genome sequence of the 40,000-year-old Vindija Neanderthal skeleton (Vindija 1) and the 40,000-year-old Mezmaiskaya Neanderthal skeleton (Mezmaiskaya 1) and was used to identify Neanderthal-specific variants.

**Introduction**  
 The Neanderthal genome sequence we obtained is of the same size as the human genome, with a total size of approximately 3.1 Gb. The Neanderthal genome is estimated to have diverged from the human genome approximately 380,000 years ago, before the emergence of anatomically modern humans. Our finding that the Neanderthal and human genomes are at least 99.7% identical led us to develop and successfully implement a targeted method for recovering specific ancient DNA sequences from non-specific libraries. This method applied to the Neanderthal genome sequence of the 40,000-year-old Vindija Neanderthal skeleton (Vindija 1) and the 40,000-year-old Mezmaiskaya Neanderthal skeleton (Mezmaiskaya 1) and was used to identify Neanderthal-specific variants.

**Methods**  
 We used a targeted method for recovering specific ancient DNA sequences from non-specific libraries. This method applied to the Neanderthal genome sequence of the 40,000-year-old Vindija Neanderthal skeleton (Vindija 1) and the 40,000-year-old Mezmaiskaya Neanderthal skeleton (Mezmaiskaya 1) and was used to identify Neanderthal-specific variants.

**Results**  
 The Neanderthal genome sequence we obtained is of the same size as the human genome, with a total size of approximately 3.1 Gb. The Neanderthal genome is estimated to have diverged from the human genome approximately 380,000 years ago, before the emergence of anatomically modern humans. Our finding that the Neanderthal and human genomes are at least 99.7% identical led us to develop and successfully implement a targeted method for recovering specific ancient DNA sequences from non-specific libraries. This method applied to the Neanderthal genome sequence of the 40,000-year-old Vindija Neanderthal skeleton (Vindija 1) and the 40,000-year-old Mezmaiskaya Neanderthal skeleton (Mezmaiskaya 1) and was used to identify Neanderthal-specific variants.

**Discussion**  
 Our findings indicate that the Neanderthal genome is distinct from other hominid genomes and that the Neanderthal genome is closely related to the modern human genome. The Neanderthal genome sequence we obtained is of the same size as the human genome, with a total size of approximately 3.1 Gb. The Neanderthal genome is estimated to have diverged from the human genome approximately 380,000 years ago, before the emergence of anatomically modern humans. Our finding that the Neanderthal and human genomes are at least 99.7% identical led us to develop and successfully implement a targeted method for recovering specific ancient DNA sequences from non-specific libraries. This method applied to the Neanderthal genome sequence of the 40,000-year-old Vindija Neanderthal skeleton (Vindija 1) and the 40,000-year-old Mezmaiskaya Neanderthal skeleton (Mezmaiskaya 1) and was used to identify Neanderthal-specific variants.

**Conclusions**  
 The Neanderthal genome sequence we obtained is of the same size as the human genome, with a total size of approximately 3.1 Gb. The Neanderthal genome is estimated to have diverged from the human genome approximately 380,000 years ago, before the emergence of anatomically modern humans. Our finding that the Neanderthal and human genomes are at least 99.7% identical led us to develop and successfully implement a targeted method for recovering specific ancient DNA sequences from non-specific libraries. This method applied to the Neanderthal genome sequence of the 40,000-year-old Vindija Neanderthal skeleton (Vindija 1) and the 40,000-year-old Mezmaiskaya Neanderthal skeleton (Mezmaiskaya 1) and was used to identify Neanderthal-specific variants.

**References**  
 1. ...

**Supplementary Information**  
 Supplementary Information is available for this article. Visit [www.nature.com/neanderthal](#) to view all supplementary information for this article.

**Additional Information**  
 Correspondence: Svante Pääbo (s.pääbo@ki.se)

**Supplementary Information**  
 Supplementary Information is available for this article. Visit [www.nature.com/neanderthal](#) to view all supplementary information for this article.

**Additional Information**  
 Correspondence: Svante Pääbo (s.pääbo@ki.se)

**Supplementary Information**  
 Supplementary Information is available for this article. Visit [www.nature.com/neanderthal](#) to view all supplementary information for this article.

**Additional Information**  
 Correspondence: Svante Pääbo (s.pääbo@ki.se)

[RecName: Full=Chlor](#)  
 1. 370 aa protein  
 P05404.4 GI:135651  
[GenPept](#) [FASTA](#) [Gra](#)

[RecName: Full=Chlor](#)  
 2. 370 aa protein  
 P27099.1 GI:135517  
[GenPept](#) [FASTA](#) [Gra](#)

[RecName: Full=Chlor](#)  
 3. 372 aa protein



# Computational Function Prediction Needed



**Total  
Sequences**

**Characterized  
Sequences**

What about the error that results from large scale function prediction?

Our focus: commonly used protein sequence databases

---

How prevalent is misannotation in common sequence databases?

What can we learn about these annotation errors and annotation in general?

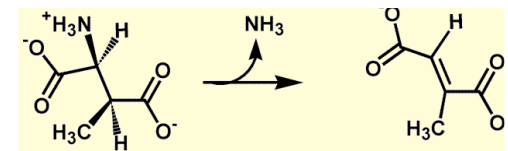
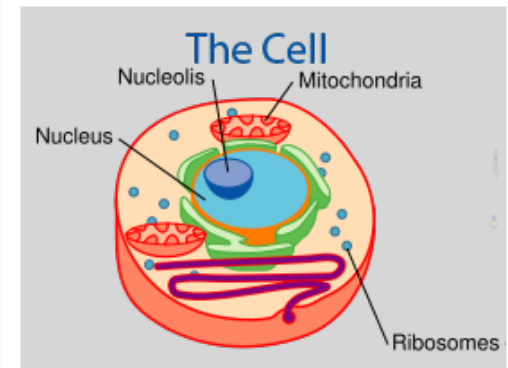
# What is 'function'?

## Many Possible Definitions

- Biological role
- Enzymatic activity
- Protein localization
- Protein interactions
- Protein expression
- Temporal activity
- Post-translational modifications
- Structural domain
- Sequence motif
- Structural motif
- Binding sites
- Functionally important residues
- Genomic context
- Metabolic pathway



Phenotype



Enzymatic Reaction

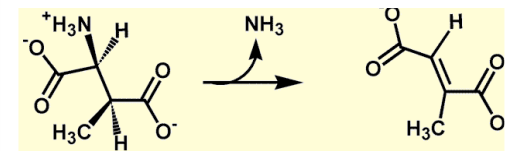
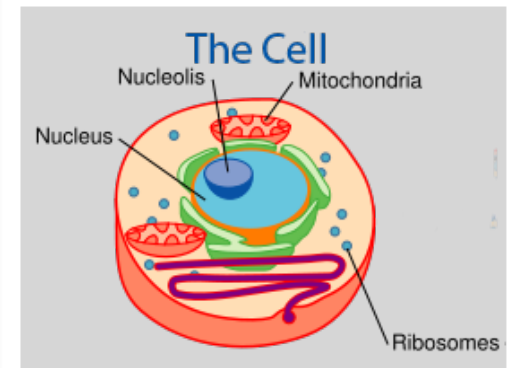
# What is 'function'?

## Many Possible Definitions

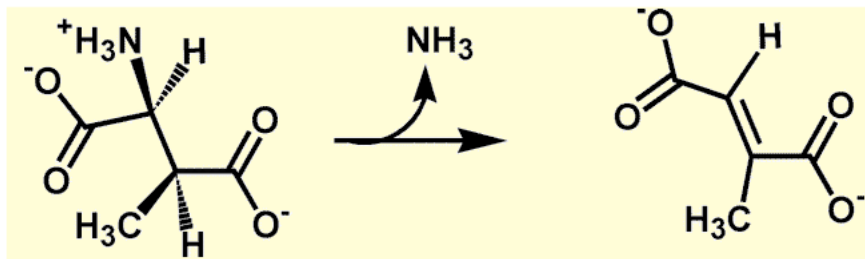
- Biological role
- Enzymatic activity
- Protein localization
- Protein interactions
- Protein expression
- Temporal activity
- Post-translational modifications
- Structural domain
- Sequence motif
- Structural motif
- Binding sites
- Functionally important residues
- Genomic context
- Metabolic pathway



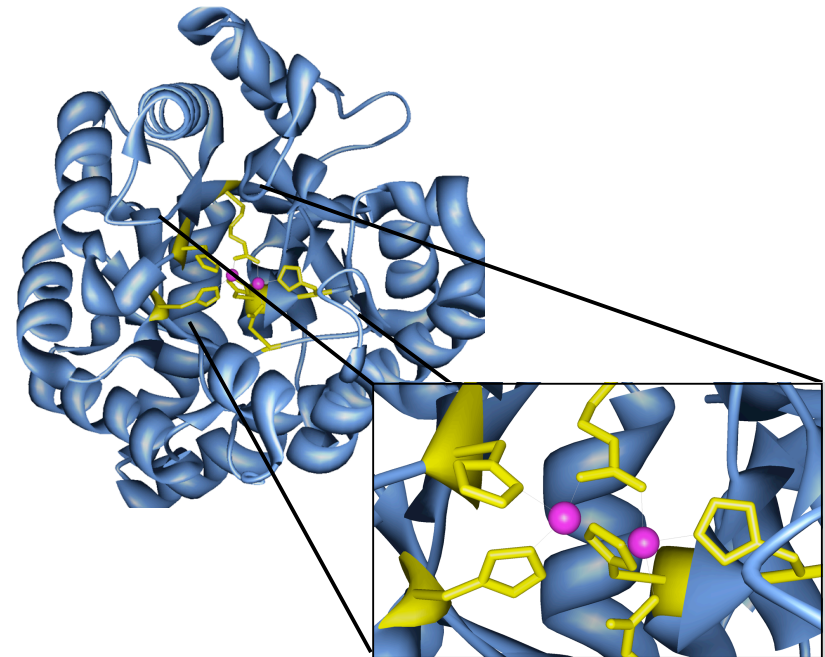
Phenotype



Enzymatic Reaction



- Concrete definition of function
  - Substrate
  - Product
  - Chemical conversion



- Function can be mapped to specific residues

Why use enzymes?

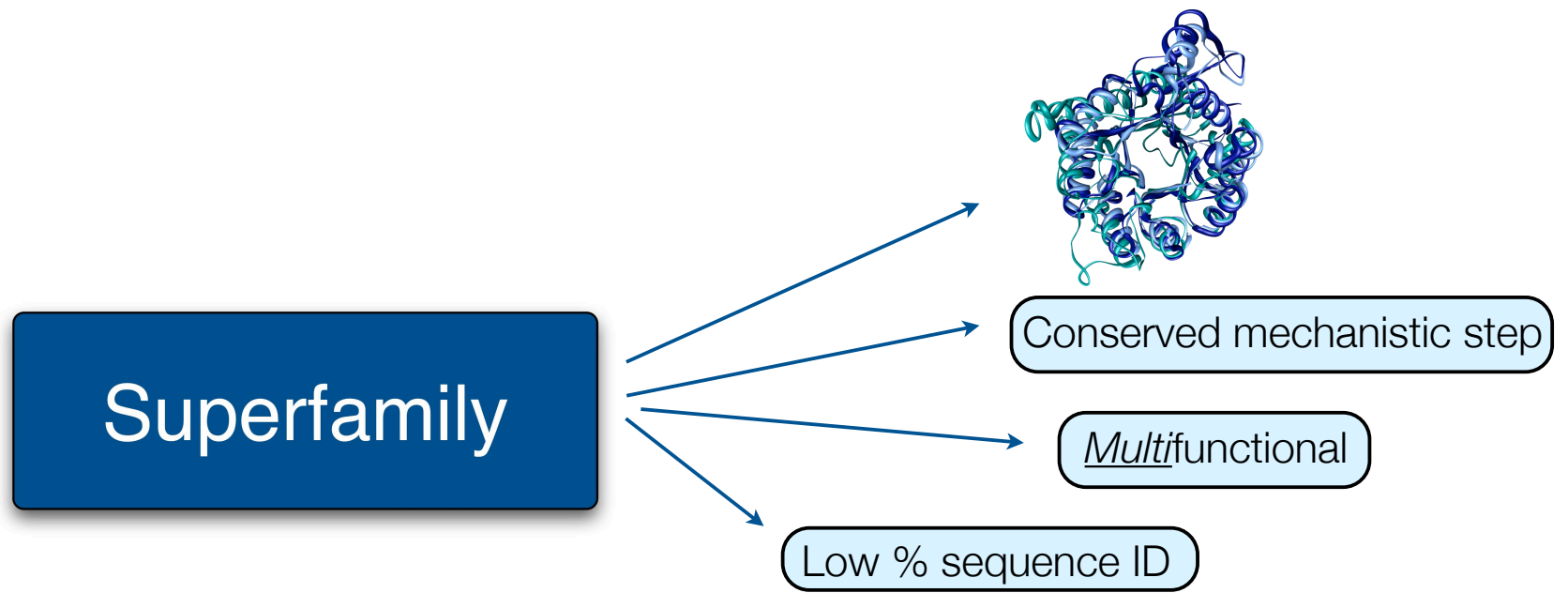
# Functionally Diverse Enzyme Superfamilies

---

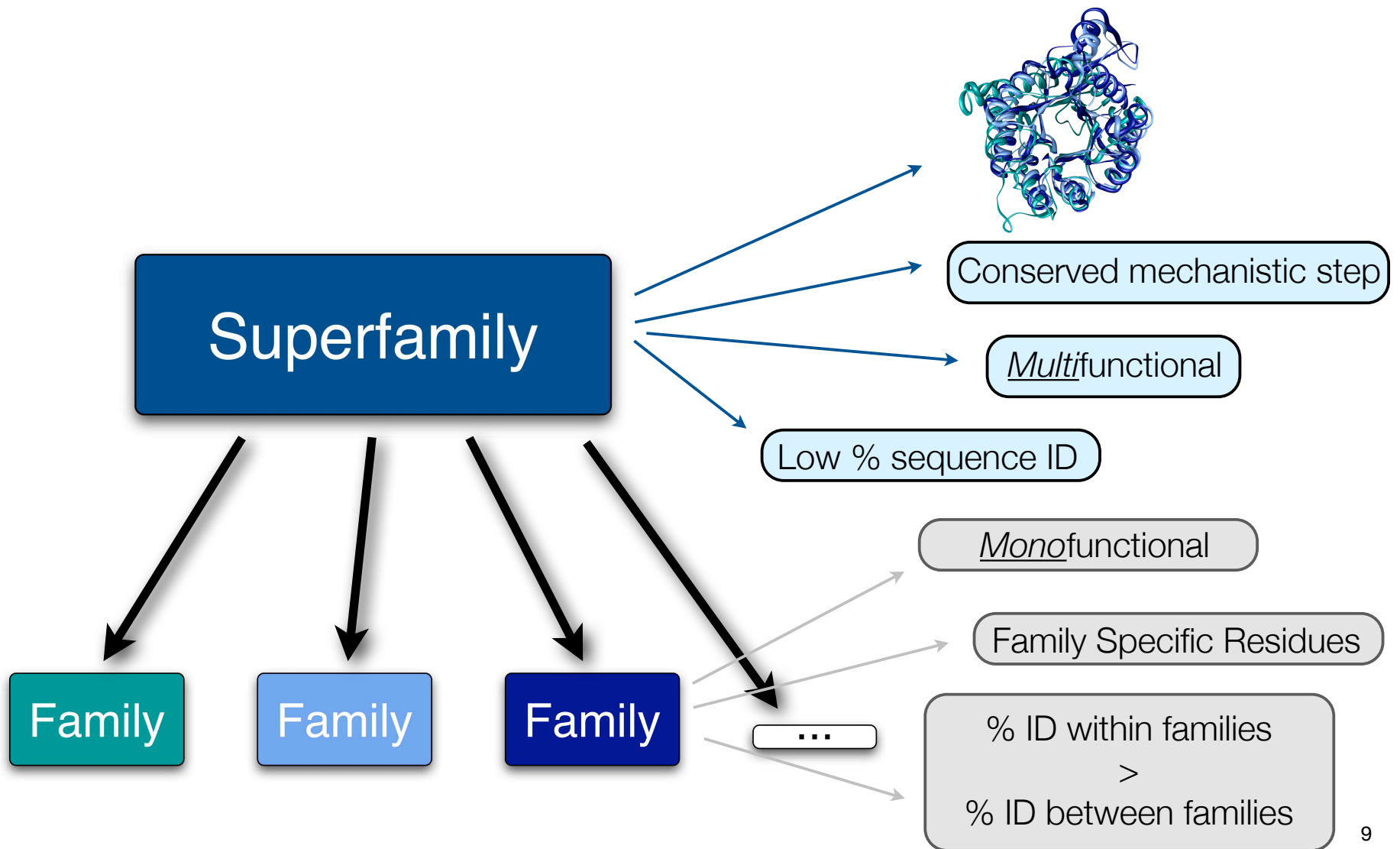


# Functionally Diverse Enzyme Superfamilies

---



# Functionally Diverse Enzyme Superfamilies



# What is needed for the misannotation analysis?

---

## **Gold Standard Sequence Set**

### **Requirements**

- Organized hierarchy & data
  - Superfamily definitions
  - Family definitions
  - Sequences
  - Sequence alignments
  - Statistical models
- Functions are experimentally characterized
- Understand functional mechanism
  - Structure
  - Active site
  - Functionally important residues
- Large set

# What is needed for the misannotation analysis?

---

## **Gold Standard Sequence Set**

### **Requirements**

- Organized hierarchy & data
  - Superfamily definitions
  - Family definitions
  - Sequences
  - Sequence alignments
  - Statistical models
- Functions are experimentally characterized
- Understand functional mechanism
  - Structure
  - Active site
  - Functionally important residues
- Large set

- 6 Superfamilies
- 5 Structural folds
- 37 Families
- 5/6 E.C. categories

Genome Biol. 2006;7(1):R8.

Structure-Function Linkage Database

UCSF University of California, San Francisco | About UCSF | UCSF Medical Center

SFLD RBVI

Browse by Superfamily Browse by Reaction Search by Enzyme Search by Reaction

Structure Function Linkage Database

Welcome to the Structure-Function Linkage Database

What is the SFLD?

- A database that links evolutionarily related sequences and structures from mechanistically diverse superfamilies of enzymes to their chemical reactions
- Correlates conserved active site residues with specific partial reactions that all members of a superfamily perform ([more details](#))

What makes the SFLD unique?

- The SFLD correlates conserved partial reactions associated with active site similarities in all members of a superfamily
- Provides the ability to search for related proteins by their partial chemical reactions

How can I use the SFLD? ([Instructions](#) & [Caveats](#))

- [Browse by superfamily, subgroup, and family](#)

Evidence Codes

Sequence Models (HMMs)

Superfamily

Family

Family

Family

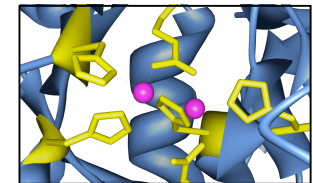
...

Hierarchically Organized

```

300      310      320      330
ELDGRGVD AELVAD EWCNTVEDVKFFTDNKAGHMVQIKTP
ELDGRGVD AELVAD EWCNTVEDVKFFTDNKAGHMVQIKTP
ALKEAEVKVEVVADEWCNTYEEIVEFVDAQADMVQIKTP
ELTRLGSGVKIVADEWCNTYQDIVDFDAASCHMVQIKTP
ELTRLGSGVKIVADEWCNTYQDIVDFDAASCHMVQIKTP
GLADAGVAVDIVADEWCDSRADVEAFVDAGAADVQVQKTP
ILDNRGSSARIVVDERCNTFEDIRLFAEAKATHLVQIKTP
  
```

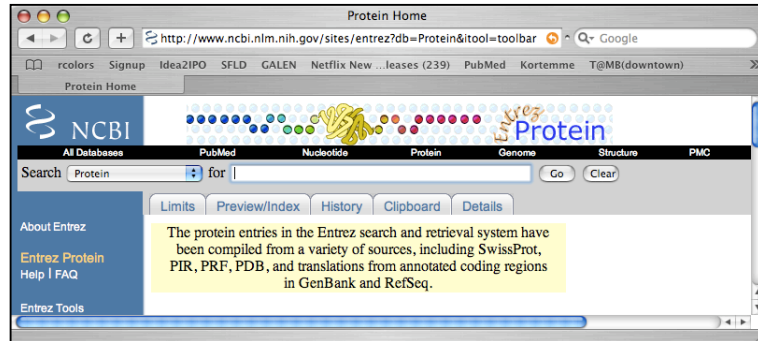
Hand-Curated Sequence Alignments



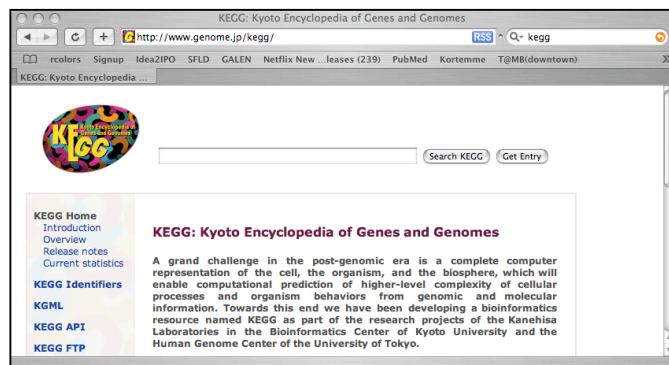
Functionally Important Residues

Gold Standard Sequence Set | [sfld.rbvi.ucsf.edu](http://sfld.rbvi.ucsf.edu)

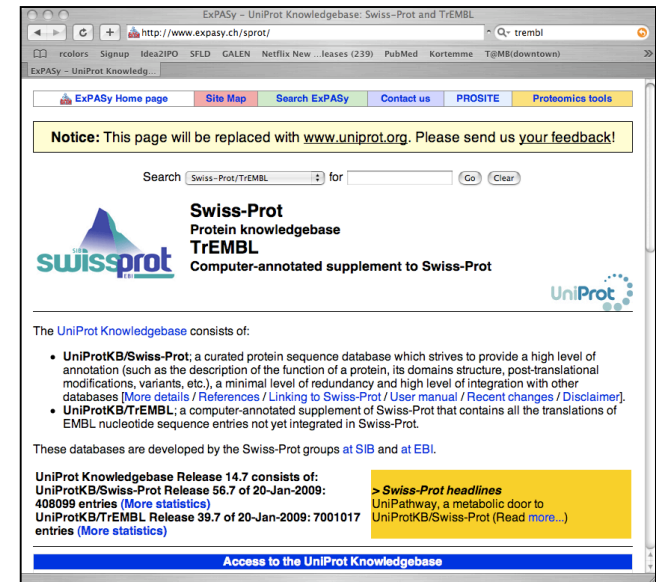
# Data Source: Commonly Used Sequence Databases



**NCBI**  
Automated  
Large



**KEGG**  
Automated



**TrEMBL**  
Automated  
Large

---

**Swiss-Prot**  
Curated  
Small

# Analysis Question

**Given:**

*A protein sequence annotated to a specific enzyme function*

**Is that annotation correct?**



# General Process

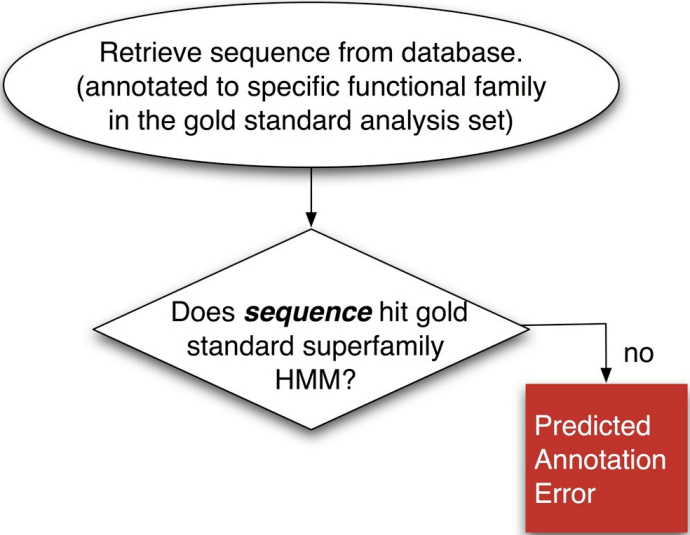
Gather all sequences labeled to a specific function

Using defined sequence metrics and functional information determine if each sequence appropriately maps to its labeled function

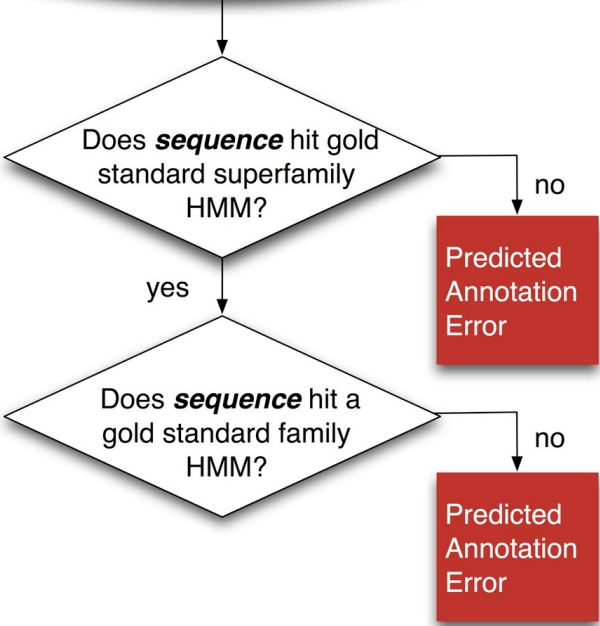
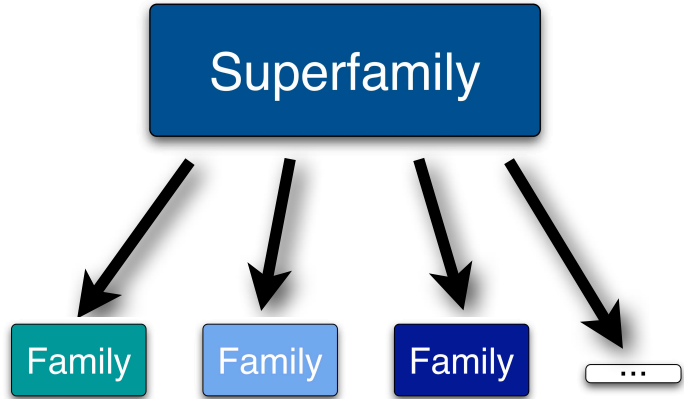
Output:  
% Misannotation  
in Gold Standard Set superfamilies & families

Retrieve sequence from database.  
(annotated to specific functional family  
in the gold standard analysis set)

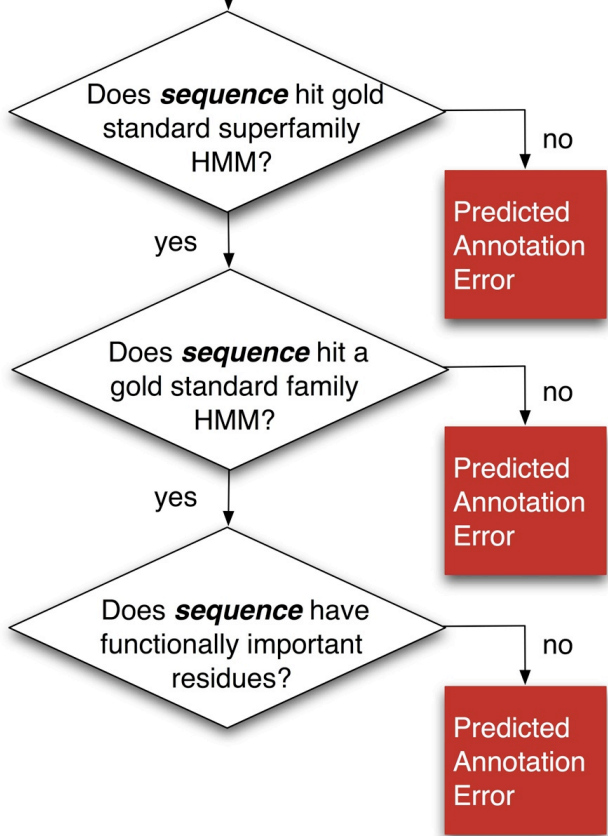
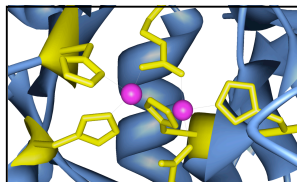
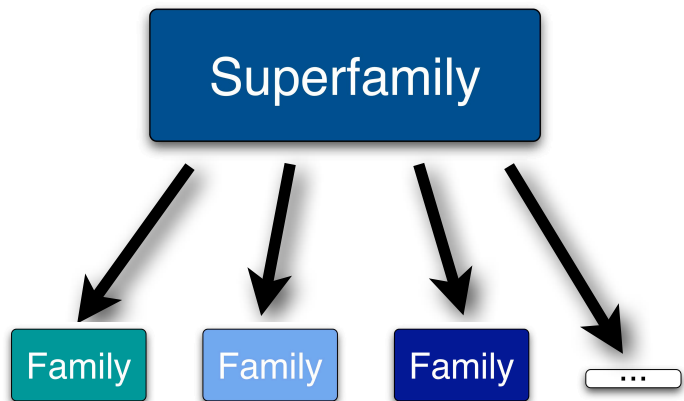
# Superfamily



Retrieve sequence from database.  
(annotated to specific functional family  
in the gold standard analysis set)

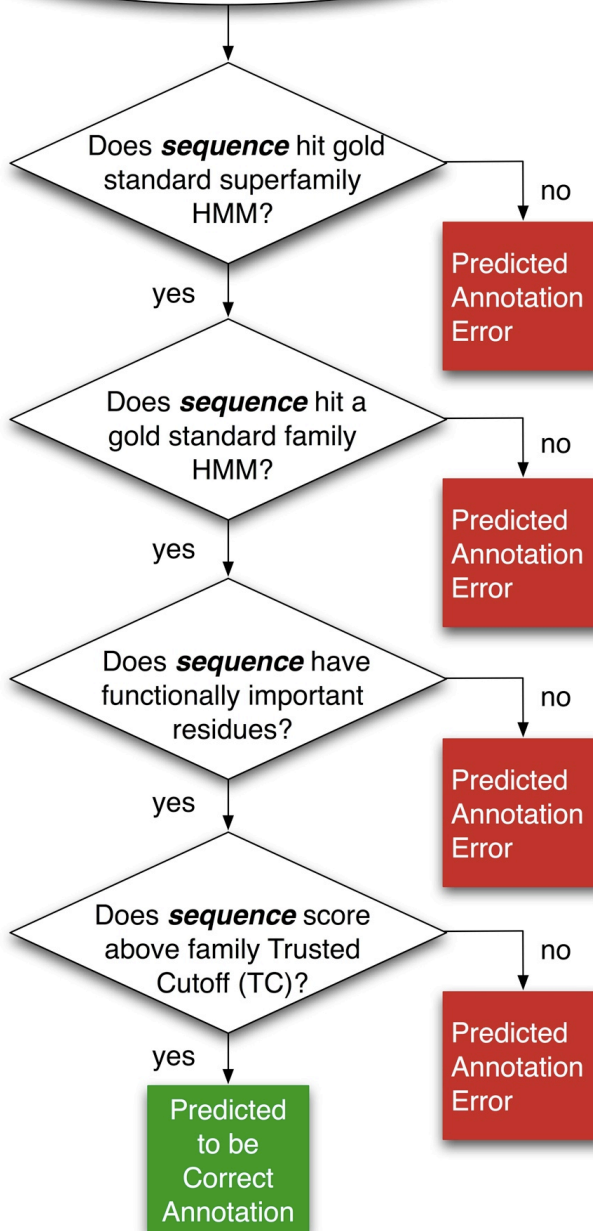
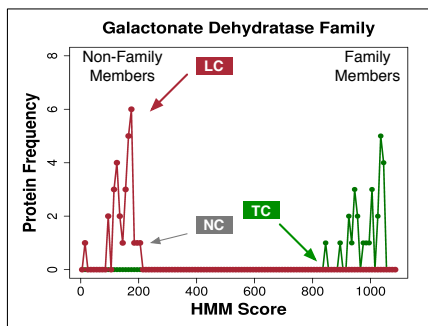
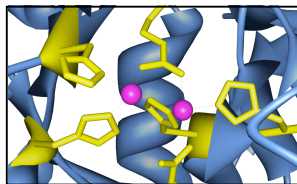
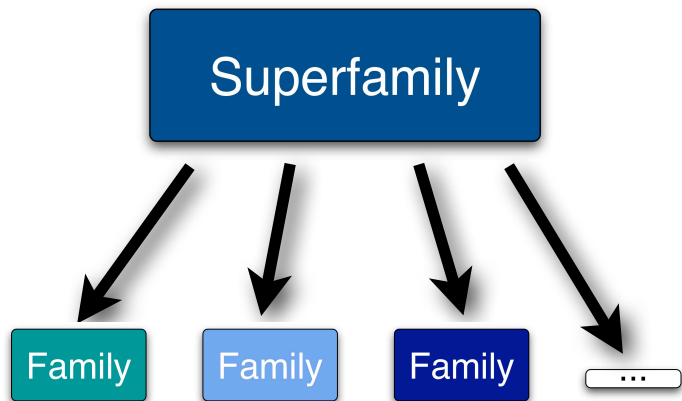


Retrieve sequence from database.  
(annotated to specific functional family  
in the gold standard analysis set)



300	310	320	330
ELDGRGVD	AELVAD	EWCN	TVEDVKFFTDNKAGHMVQIKTP
ELDGRGVD	AELVAD	EWCN	TVEDVKFFTDNKAGHMVQIKTP
ALKEAEVK	VEVVA	DEWC	NTYEEIVEFVDAQADMVQIKTP
ELTRLGSG	VKIVAD	DEWC	NTYQDIVDFTDAASCHMVQIKTP
ELTRLGSG	VKIVAD	DEWC	NTYQDIVDFTDAGSCHMVQIKTP
GLADAGVA	VDIVAD	DEWC	DSRADVEAFVDAGADVVQVQKTP
ILDNRGSS	ARIVVD	DERC	NTFEDIRLFAEAKATHLVQIKTP

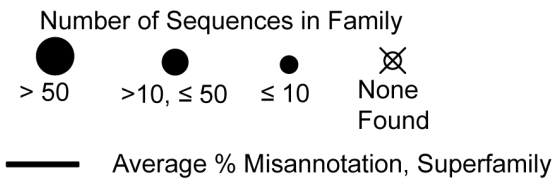
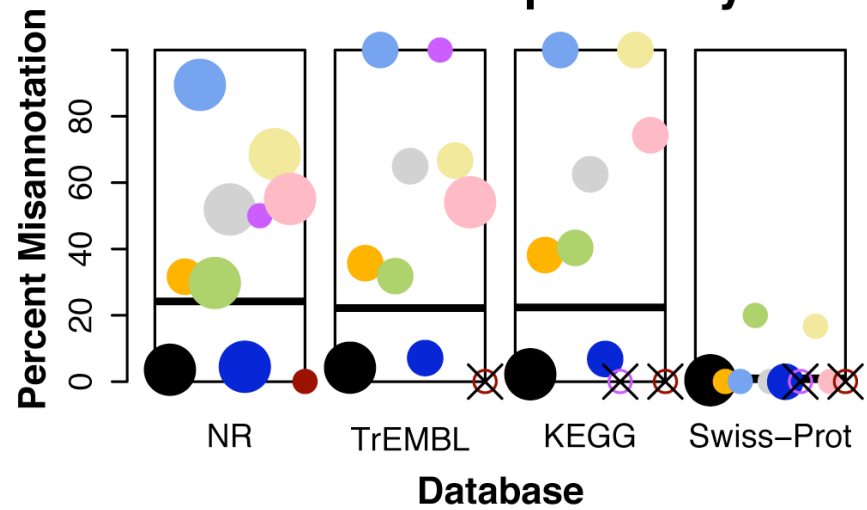
Retrieve sequence from database.  
(annotated to specific functional family  
in the gold standard analysis set)



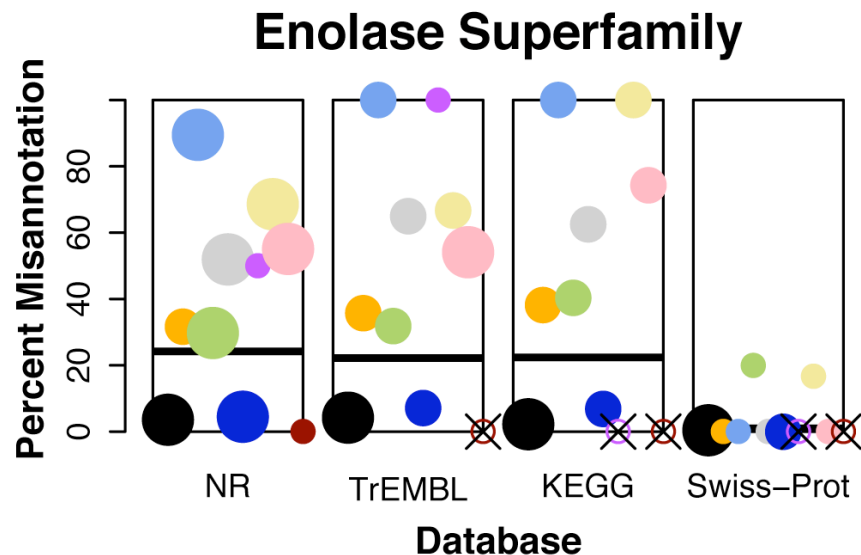
```

300      310      320      330
ELDGRGVD AELVAD EWCNTV EDVKFF TDNKAGH MVQIKTP
ELDGRGVD AELVAD EWCNTV EDVKFF TDNKAGH MVQIKTP
ALKEAEVKVEVVA DEWCNTY EEIVFV DAQAADM VQIKTP
ELTRLGSGVKIVAD EWCNTY QDIVDF TDAA SCHMVQIKTP
ELTRLGSGVKIVAD EWCNTY QDIVDF TDAGS CHMVQIKTP
GLADAGVAVDIVAD EWCDSR ADVEAFV DAGAADV VQVQKTP
ILLNRRGSSARIVVD ERCNTF EDIRLFAEAKATHLVQIKTP
  
```

## Enolase Superfamily



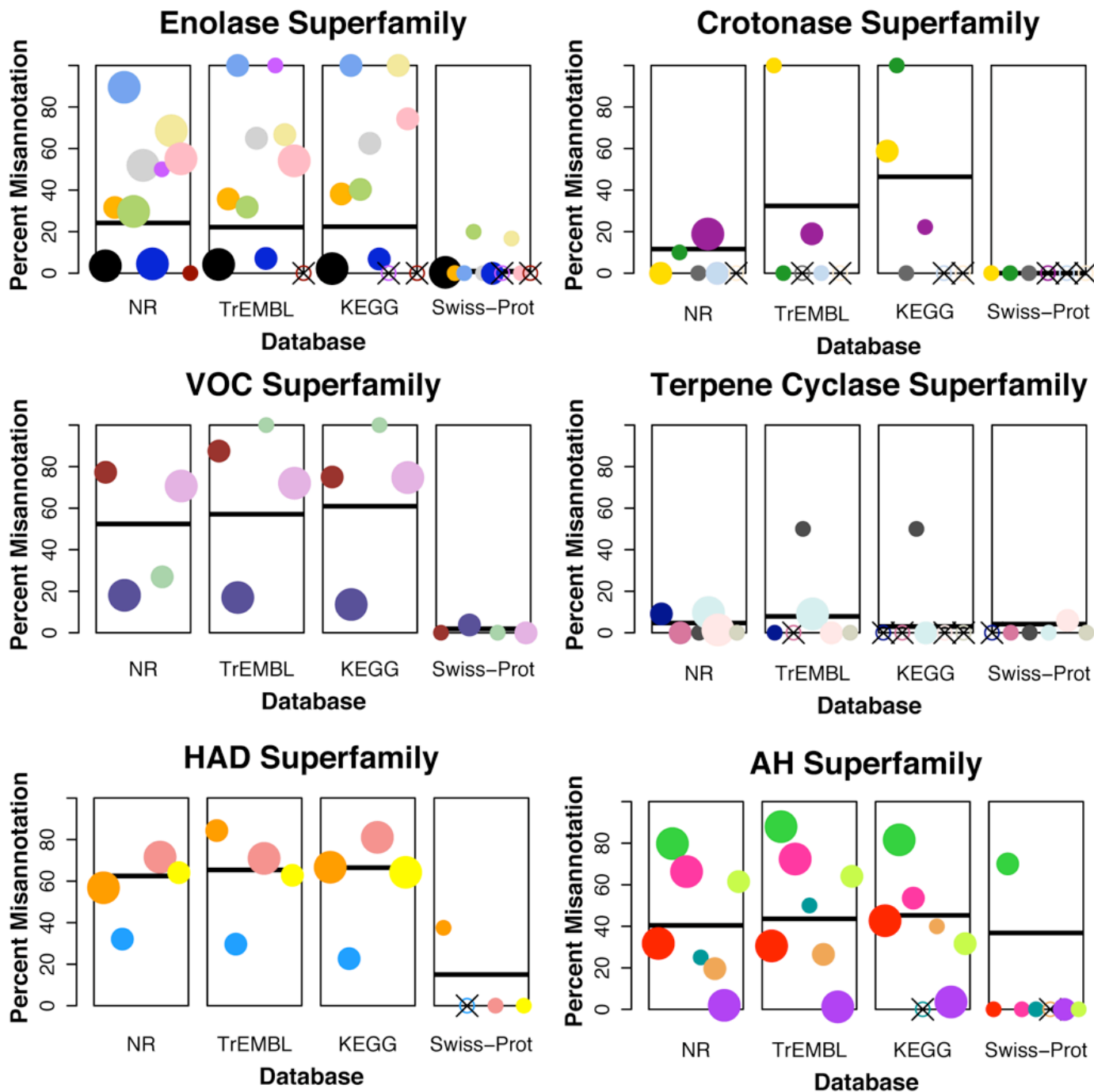




Superfamily	Family	E.C. No.	Family Color
Enolase	Enolase	4.2.1.11	●
	Galactonate dehydratase	4.2.1.6	●
	Mandelate racemase	5.1.2.2	●
	Glucarate dehydratase	4.2.1.40	●
	Methylaspartate ammonia-lyase	4.3.1.2	●
	<i>ortho</i> -succinyl benzoate synthase	4.2.1.113	●
	Dipeptide epimerase	—	●
	Chloromuconate cycloisomerase	5.5.1.7	●
	Muconate cycloisomerase	5.5.1.1	●
	L-fuconate dehydratase	4.2.1.68	●
Crotonase	Dodecenoyl-CoA delta-isomerase (mitochondrial)	5.3.3.8	●
	Delta(3,5)-delta(2,4)-dienoyl-CoA isomerase	—	●
	Methylmalonyl-CoA decarboxylase	4.1.1.41	●
	3-Hydroxyisobutyryl-CoA hydrolase	3.1.2.4	●
	4-Chlorobenzoate dehalogenase	3.8.1.7	●
	1,4-Dihydroxy-2-naphthoyl-CoA synthase	—	●
Vicinal Oxygen Chelate (VOC)	Methylmalonyl-CoA epimerase	5.1.99.1	●
	4-Hydroxyphenylpyruvate dioxygenase	1.13.11.27	●
	FosA	2.5.1.18	●
	Glyoxalase I	4.4.1.5	●
Terpene Cyclase	5-Epi-aristolochene synthase	—	●
	Bornyl diphosphate synthase	5.5.1.8	●
	Pentalenene synthase	4.2.3.7	●
	Squalene-hopene synthase	5.4.99.17	●
	Trichodiene synthase	4.2.3.6	●
	Aristolochene synthase	4.2.3.9	●
Haloacid Dehalogenase (HAD)	Deoxy-D-mannose-octulosonate 8-phosphate phosphatase	3.1.3.45	●
	Phosphonoacetaldehyde hydrolase	3.11.1.1	●
	2-Haloacid dehalogenase	3.8.1.2	●
	Beta-phosphoglucomutase	5.4.2.6	●
Amidohydrolase (AH)	Cytosine deaminase	3.5.4.1	●
	Adenosine deaminase	3.5.4.4	●
	N-acyl-D-amino-acid deacylase	3.5.1.81	●
	L-hydantoinase	3.5.2.2	●
	D-hydantoinase	3.5.2.2	●
	Urease	3.5.1.5	●
Isoaspartyl dipeptidase	—	●	

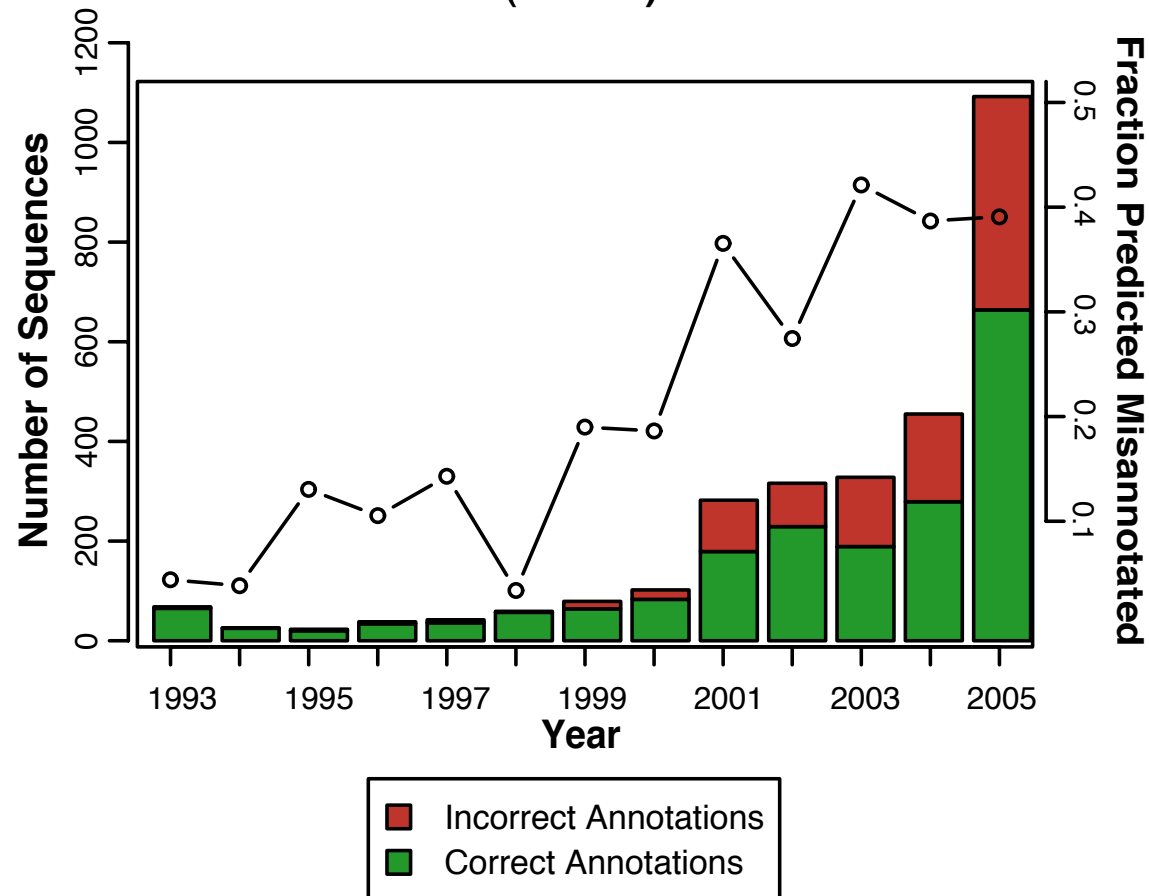
**Variable  
percent  
misannotation**

**Manually  
curated Swiss-  
Prot is most  
accurate**

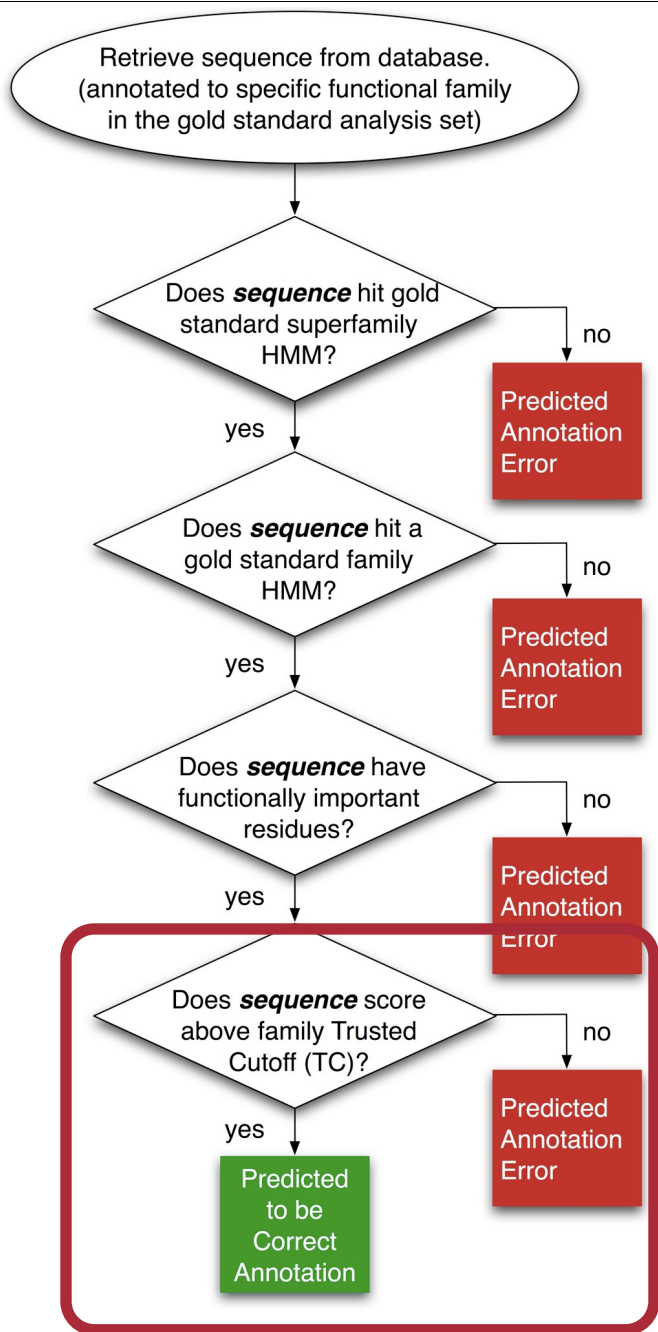


# Misannotation Problem is Getting Worse

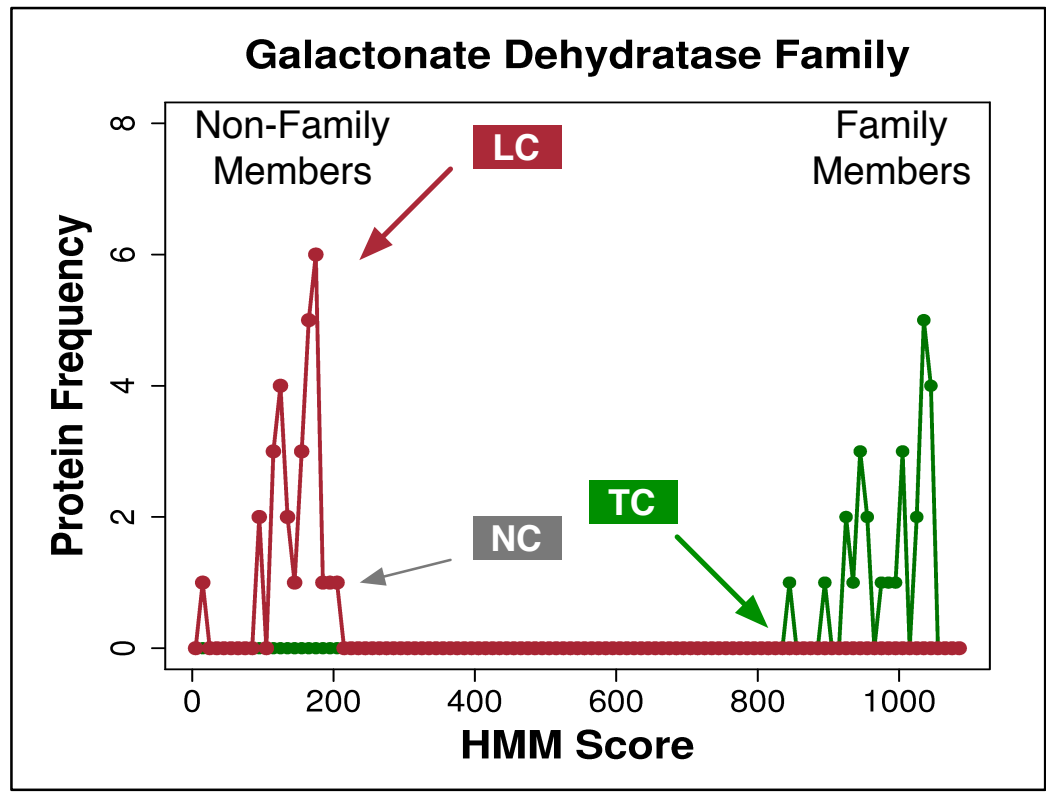
Sequences Deposited by Year and the Fraction Predicted to be Misannotated (NR DB)



What are the characteristics of these misannotations?

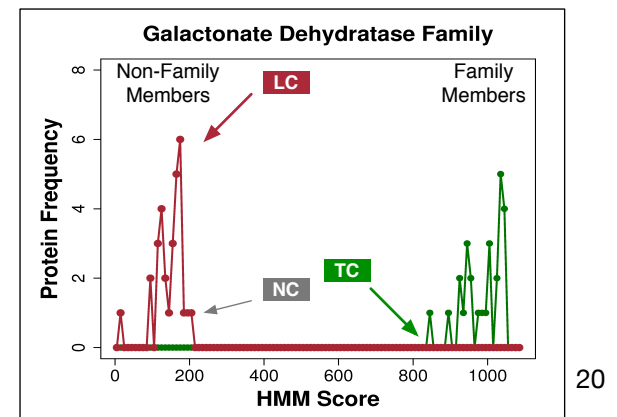
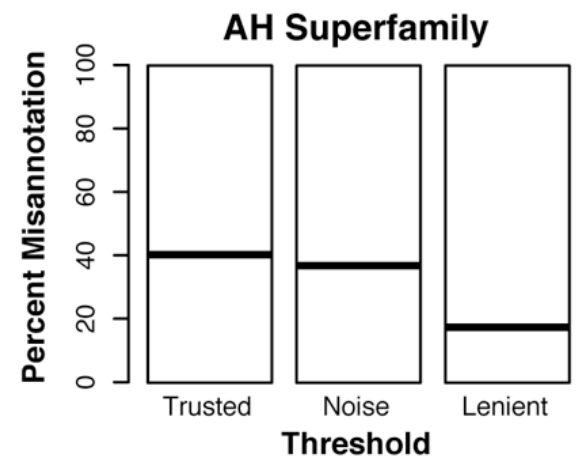
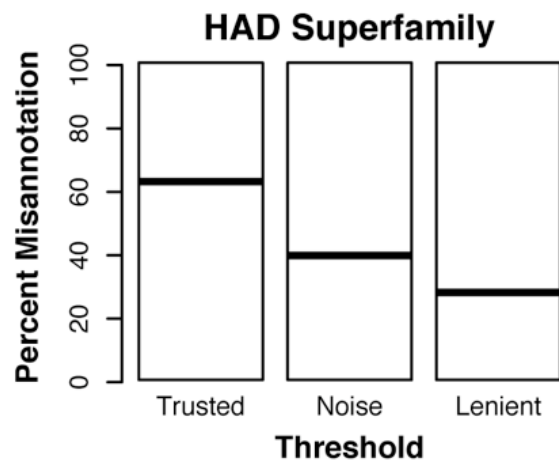
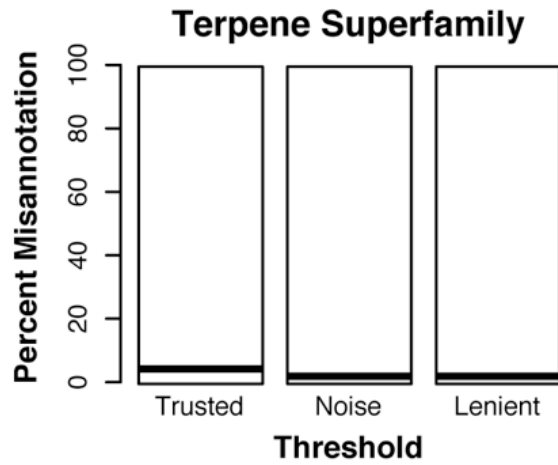
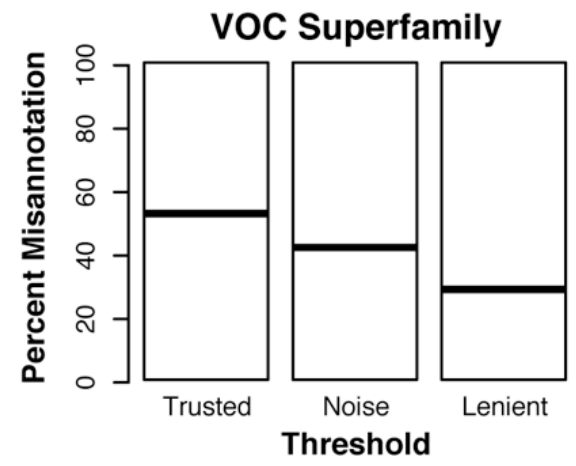
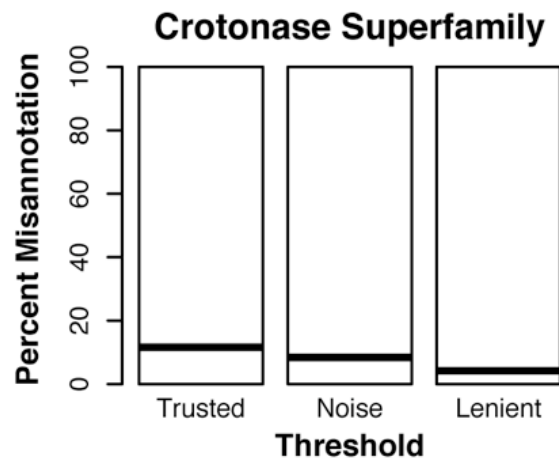
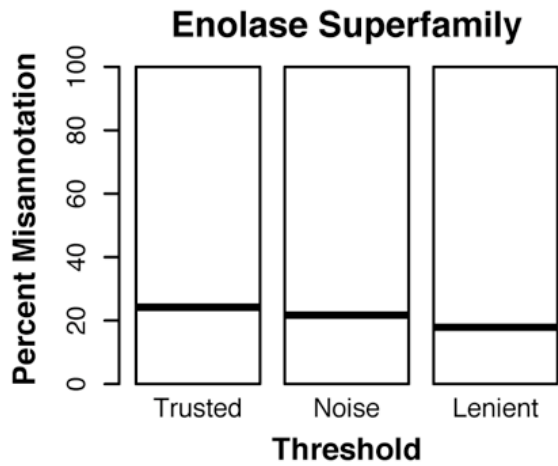


Sensitivity to threshold change



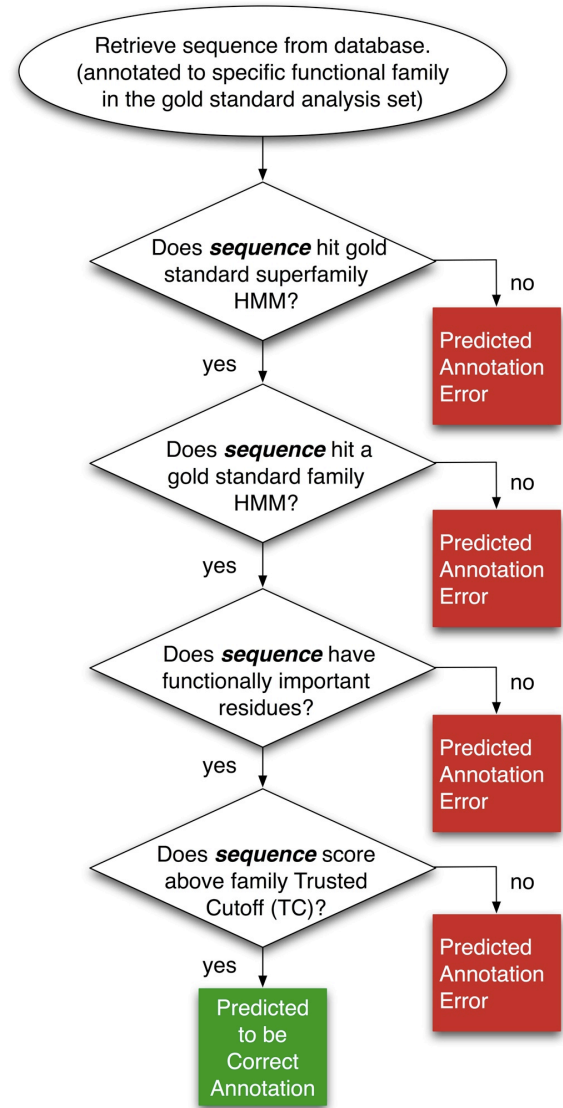
- TC** — Trusted Cutoff
- NC** — Noise Cutoff
- LC** — Lenient Cutoff

Sensitivity to threshold change

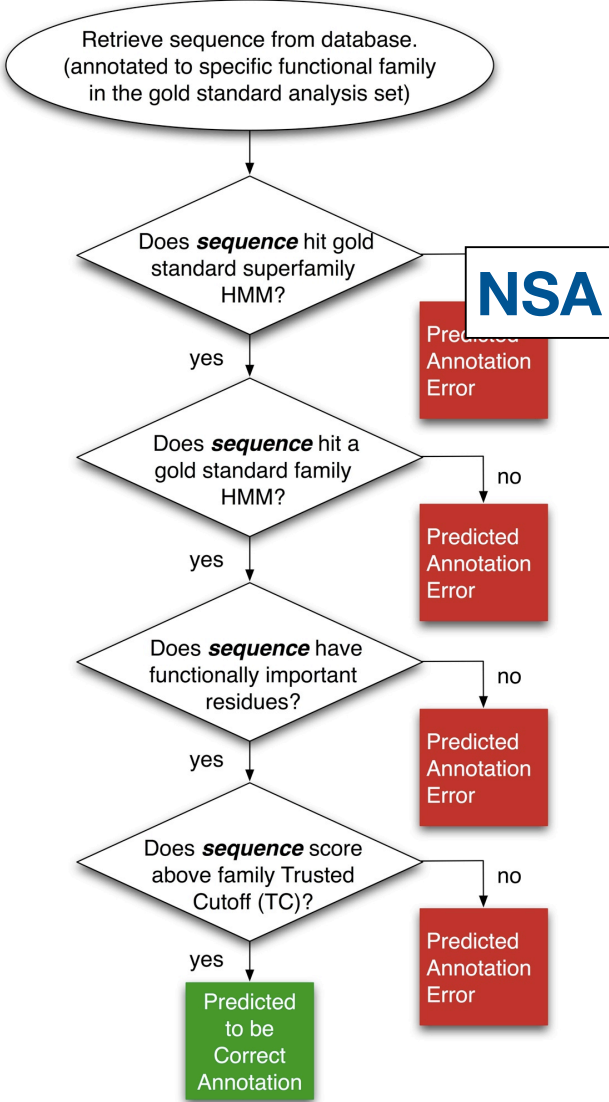


Sensitivity to threshold change

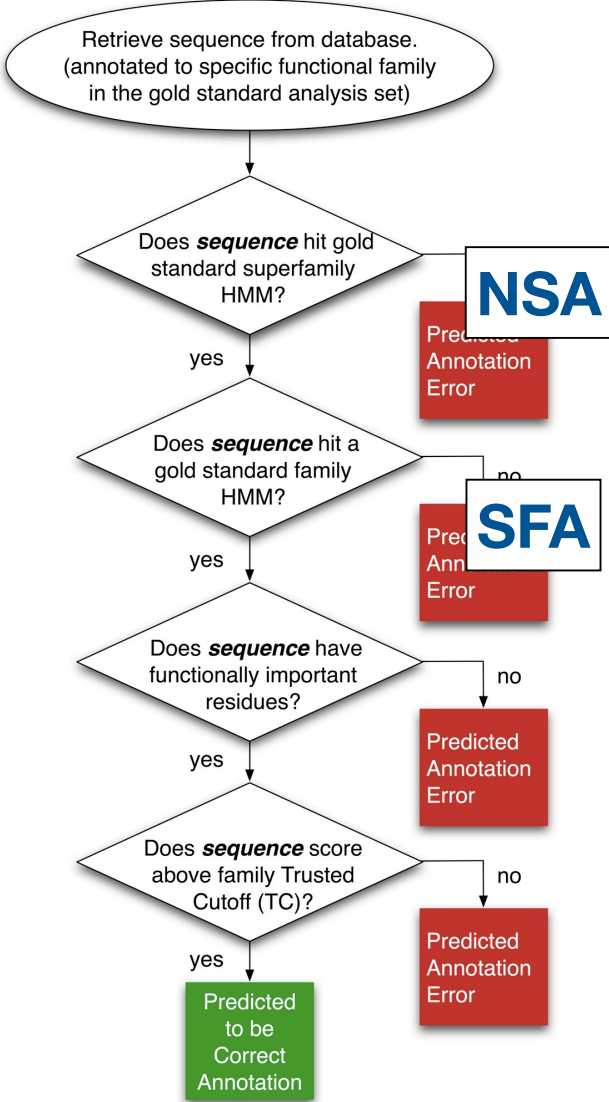




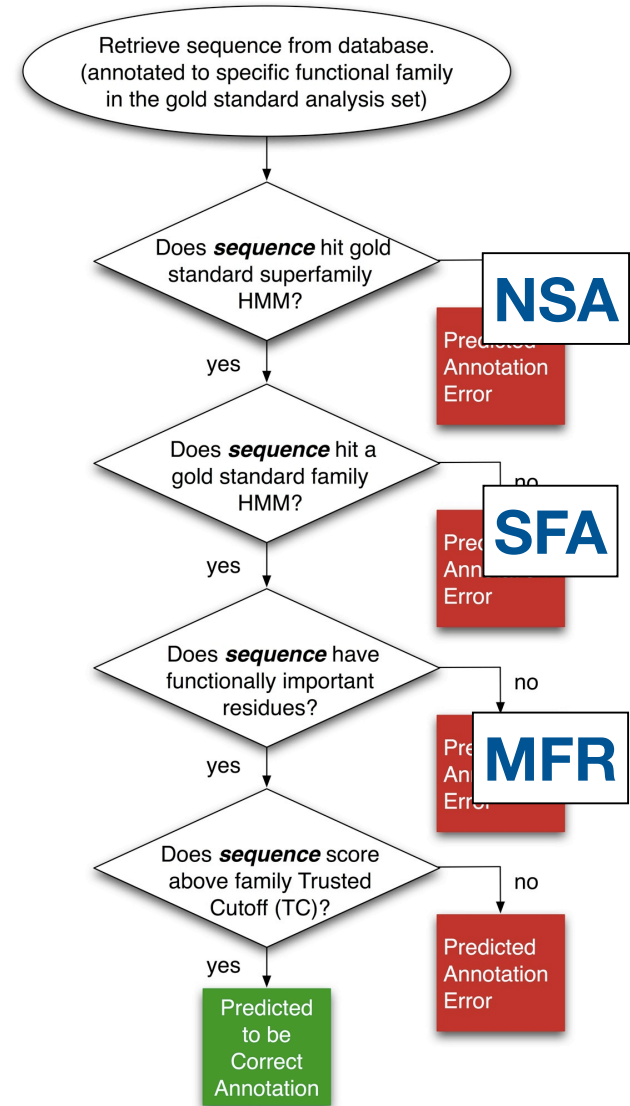
**NSA** — No Superfamily Association



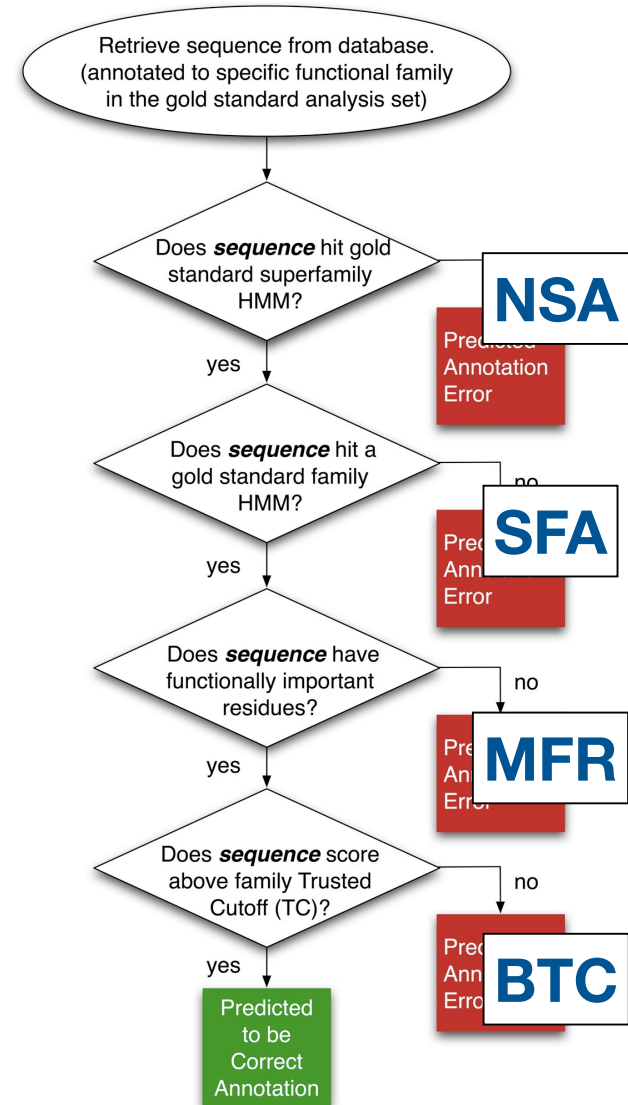
**NSA** — No Superfamily Association  
**SFA** — Superfamily Association Only



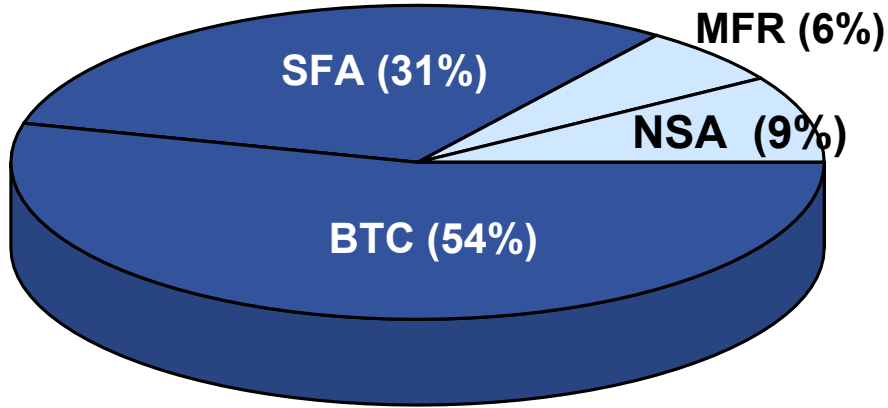
- NSA** — No Superfamily Association
- SFA** — Superfamily Association Only
- MFR** — Missing Functionally Important Residues



- NSA** — No Superfamily Association
- SFA** — Superfamily Association Only
- MFR** — Missing Functionally Important Residues
- BTC** — Below Trusted Cutoff

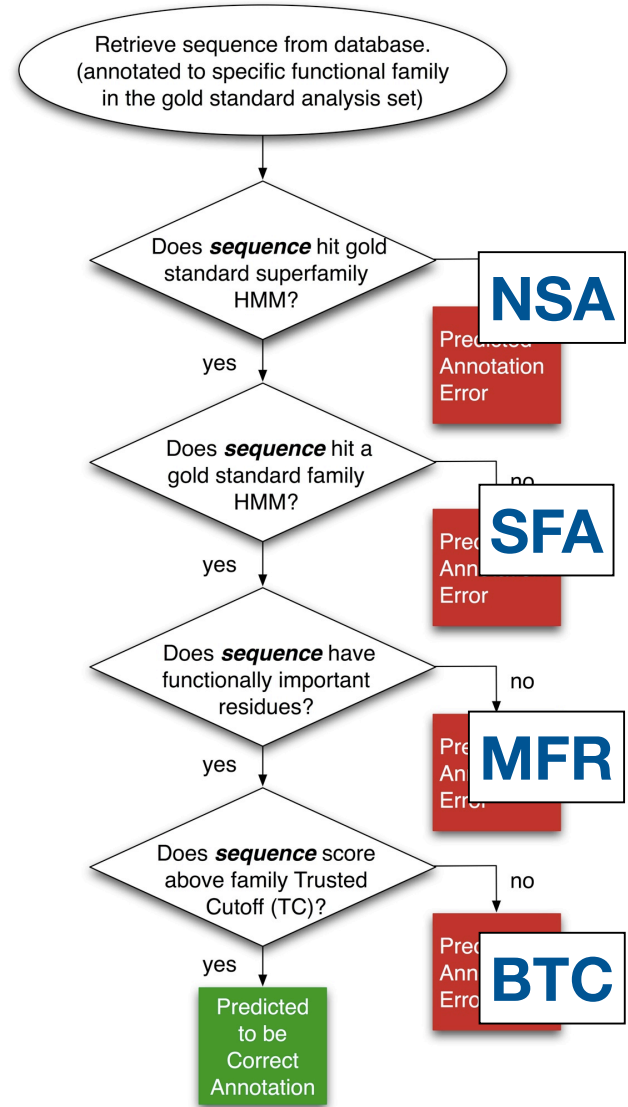


# Types of Misannotation

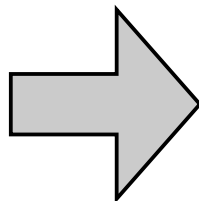


- Misannotations due to overprediction
- Misannotations not due to overprediction

- NSA** — No Superfamily Association
- SFA** — Superfamily Association Only
- MFR** — Missing Functionally Important Residues
- BTC** — Below Trusted Cutoff



Biggest Problem



**Predicting function without sufficient evidence**

# Dipeptide Epimerase

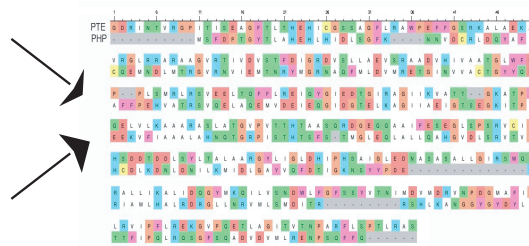
>gj13786715|pdb|1HZY|A Chaperone-Independent Resolution Structure Of The Zinc-Containing Phosphotriesterase From *Pseudomonas* *Dominata*  
GDRINTVRGPITISEAGFTL...  
FDIGRDVSLLAIEVSRAD...  
GKATPFQELVLKAAARAS...  
AARGYLIGLDHHPHSAIGL...  
MDVMDRVNPDGMAFIPL...



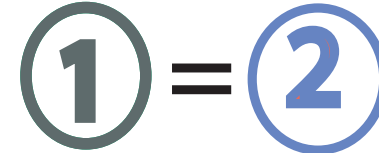
>gj1176259|sp|P45548|P... ECOM Phosphotriesterase homology protein  
MSFDPTGYTLAHEHLHID...  
RETGINVACTGYYQDA...  
FIAAALAHNQTGRPISTH...  
IGKNSYYPDEKRIAMLH...  
VMLRENPSQFFQ



Unknown Function



# Dipeptide Epimerase



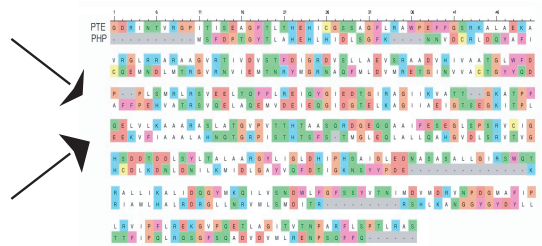
Dipeptide Epimerase

Error Propagation

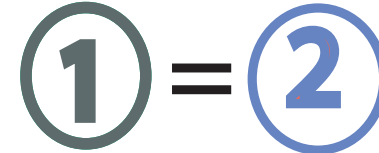
# Dipeptide Epimerase

>gj13786715|pdb|1HZY|A Chaperone Resolution Structure Of The Zinc-Containing Phosphotriesterase From *Pseudomonas* *Dominata*  
GDRINTVRGPIITISEAGFTL...  
FDIGRDVSLLAIEVSRAAD...  
GKATPFQELVLKAAARAS...  
AARGYLIGLDHHPHSAIGL...  
MDVMDRVNPDGMAFIPL...

>gj1176259|sp|P45548|P... ECOM Phosphotriesterase homology protein  
MSFDPTGYTLAHEHLHID...  
RETGINVACTGYYQDAI...  
FIAAALAHNQTGRPISTH...  
IGKNSYYPDEKRIAMLH...  
VMLRENPSQFFQ



# Dipeptide Epimerase



Dipeptide Epimerase

Unknown Function



Error Propagation



**BLAST sequence similarity network**

- E-value  $1 \times 10^{-30}$  or lower
- Distance between nodes reflects level of sequence similarity

— Sequence similarity

○ Correct annotation

△ Incorrect annotation

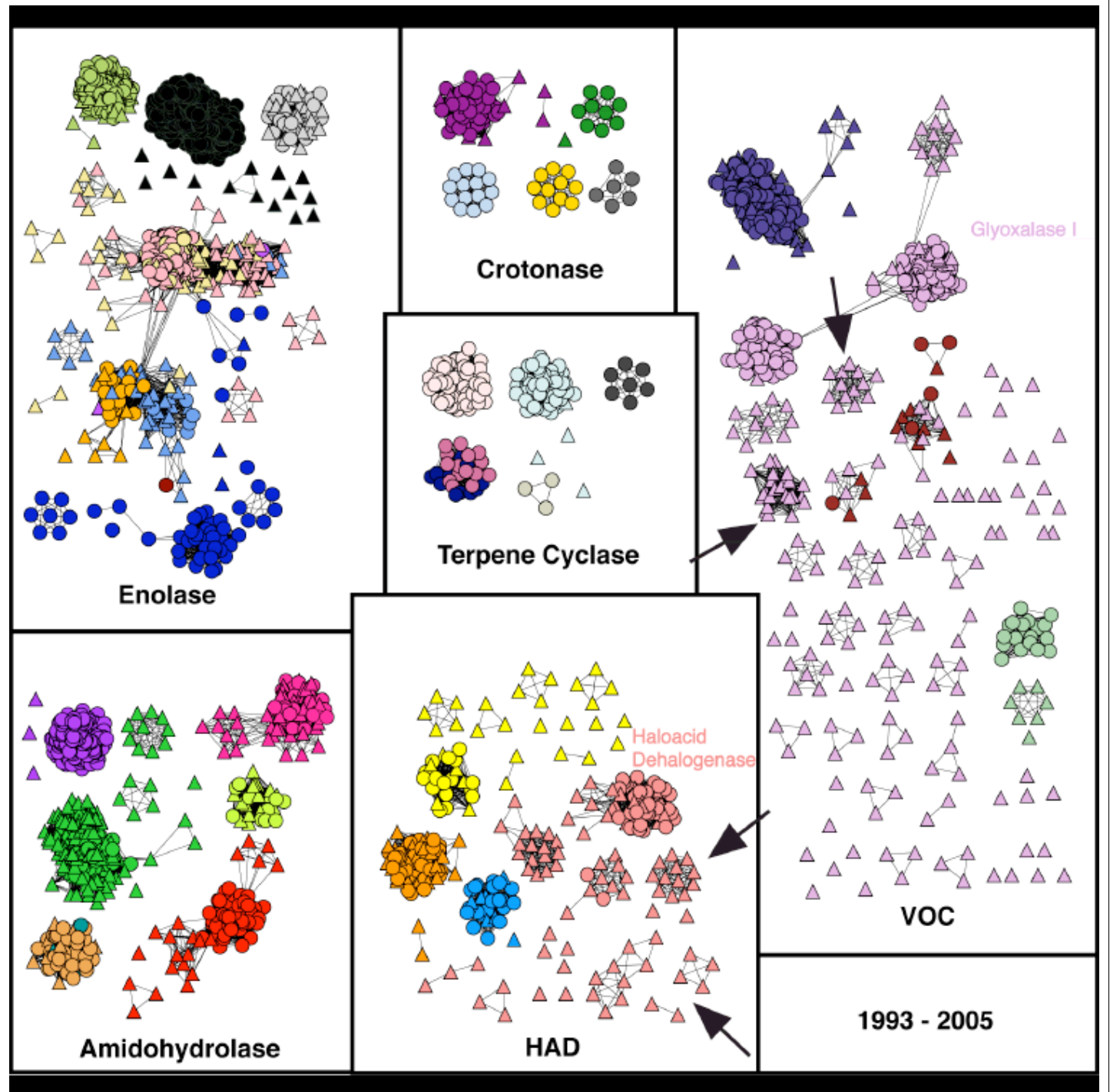
**BLAST sequence similarity network**

- E-value  $1 \times 10^{-30}$  or lower
- Distance between nodes reflects level of sequence similarity

— Sequence similarity

● Correct annotation

▲ Incorrect annotation



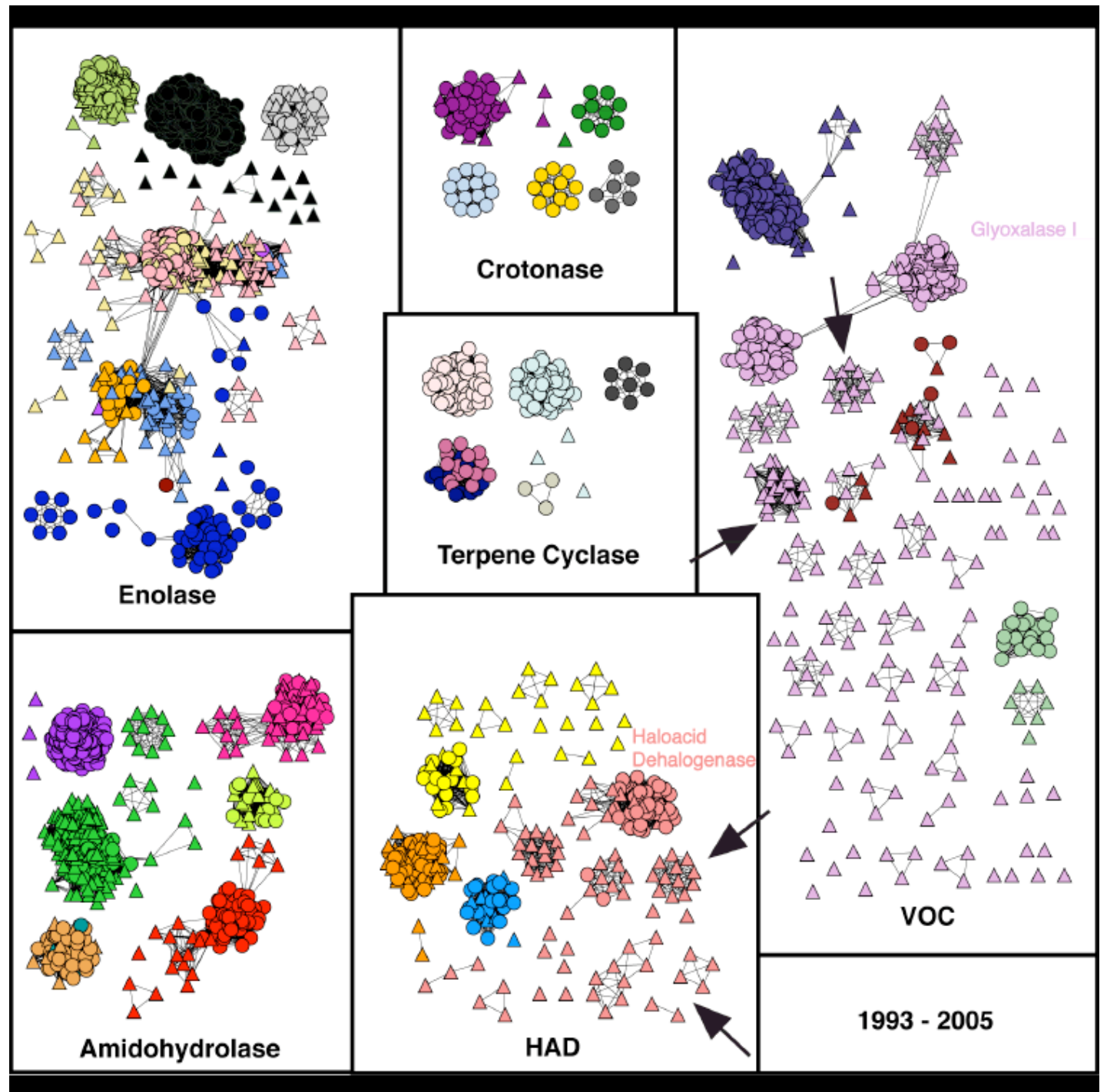
# Misannotations

- Cluster with each other
- Indication of error propagation

## BLAST sequence similarity network

- E-value  $1 \times 10^{-30}$  or lower
- Distance between nodes reflects level of sequence similarity

- Sequence similarity
- Correct annotation
- △ Incorrect annotation



## In Conclusion...

---

- Misannotation is a serious problem
  - Automated databases
  - Across multiple folds, functions and superfamilies
  - Hard to predict misannotation *a priori*
  - Manual curation delivers the highest quality
- Misannotation problem is getting worse
- Overprediction is a common problem
- Error propagation appears to be a common source of misannotation

# Acknowledgements



**Patricia Babbitt & lab**  
**Shoshana Brown**

**Igor Dodevski**  
**University of Zürich**

**Tanja Kortemme & Lab**  
**Colin Smith**

**Jim Wells Lab**  
**Emily Crawford**

**\$\$ Howard Hughes Pre-Doctoral Fellowship**  
**NIH & NSF**

**PLoS Comput Biol. 2009 Dec;5(12):e1000605.**