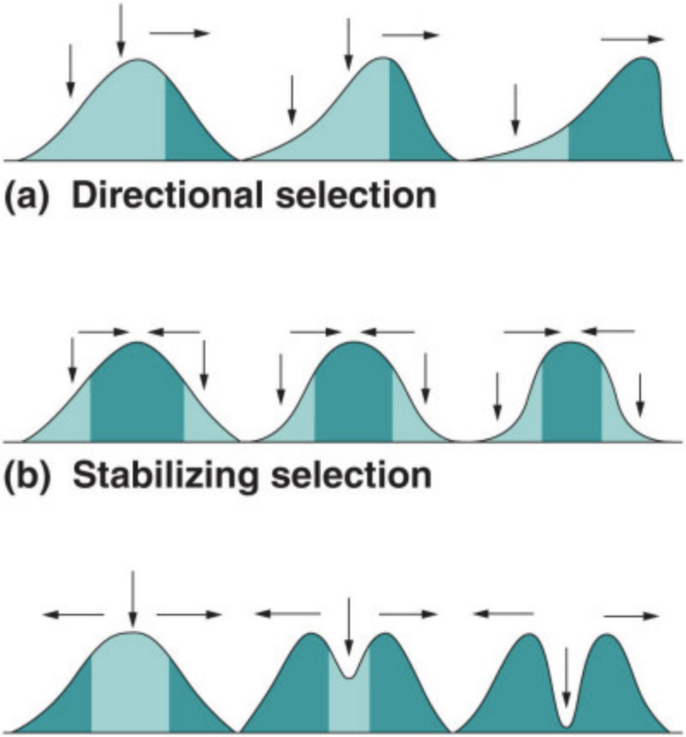
Evolution & BLAST

Lecture 4
Sept 7, 2016

Announcements

- AWS???
  - “AWS Educate Application Approved”
  - https://aws.amazon.com/education/awseducate/contact-us/
- Reading posted later today

SELECTION



(a) Directional selection

(b) Stabilizing selection

(c) Disruptive selection

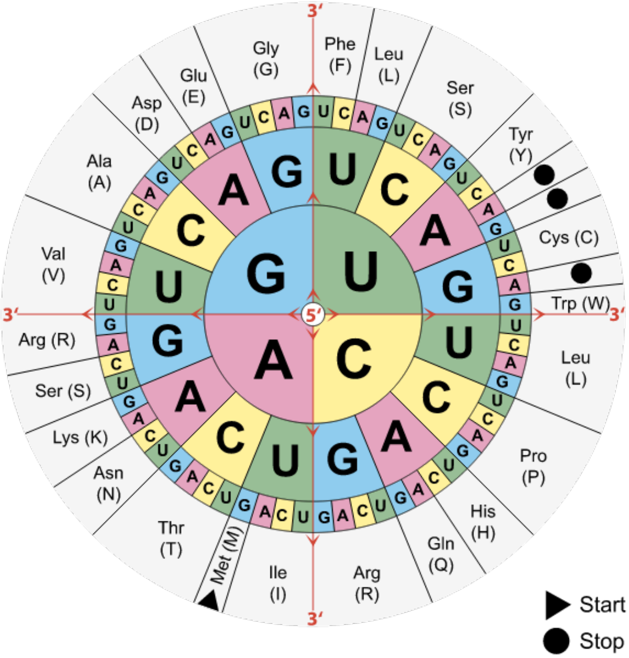
SELECTION

- Synonymous change

- Non-synonymous

Selection

5′-AUGCAGGCAUGA-3′

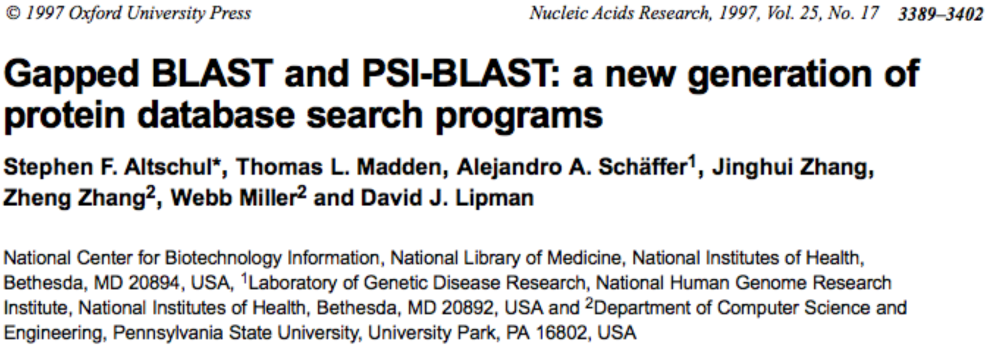


SELECTION

- dN/dS

BLAST

(Basic Local Alignment and Search Tool)



© 1997 Oxford University Press *Nucleic Acids Research, 1997, Vol. 25, No. 17* ***3389–3402***

Gapped BLAST and PSI-BLAST: a new generation of protein database search programs

Stephen F. Altschul*, Thomas L. Madden, Alejandro A. Schäffer[1], Jinghui Zhang, Zheng Zhang[2], Webb Miller[2] and David J. Lipman

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA, [1]Laboratory of Genetic Disease Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA and [2]Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802, USA

BLAST

- Query

- Database

BLAST databases

What type of BLAST??

- What type of query sequence do you have?
  - Nucleotide?

BLASTn

BLASTx

tBLASTx

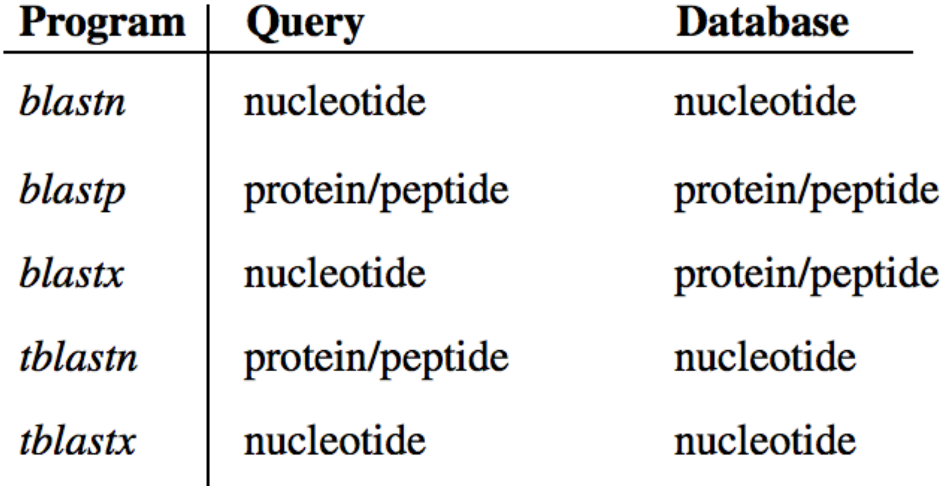
What type of BLAST??

- What type of query sequence do you have?
  - Protein?

BLASTp

tBLASTn

What type of BLAST??



Program	Query	Database
blastn	nucleotide	nucleotide
blastp	protein/peptide	protein/peptide
blastx	nucleotide	protein/peptide
tblastn	protein/peptide	nucleotide
tblastx	nucleotide	nucleotide

BLAST

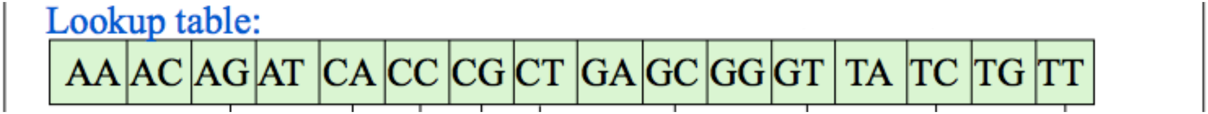
Steps in BLAST

BLAST

1. Build Lookup table

Preprocess: Build a *lookup table* of size $|\Sigma|^w$ for all *w*-length words in D

Σ={A,C,G,T}
$w = 2$
→ 4^2 (=16) entries in lookup table



Lookup table:

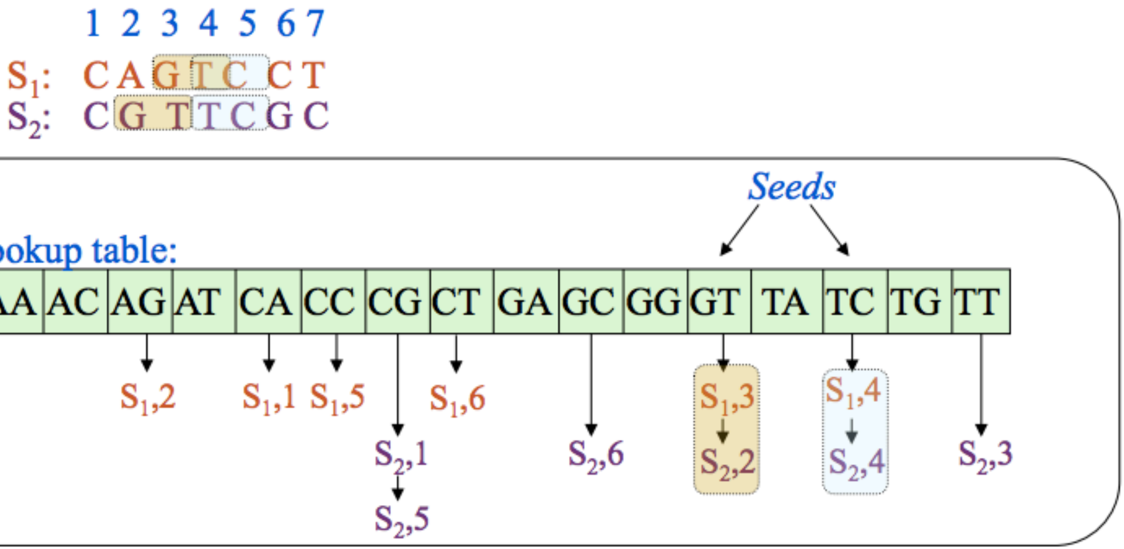
AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT

BLAST

Word size related to sensitivity of BLAST

BLAST

2. Filter low complexity and identify seeds



1 2 3 4 5 6 7

S_1: C A G T C C T

S_2: C G T T C G C

Seeds

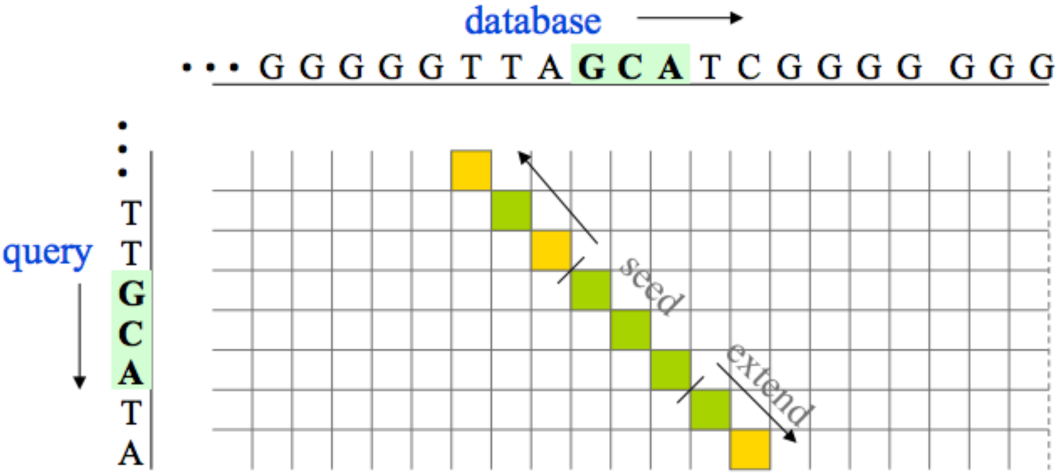
Lookup table:

AA AC AG AT CA CC CG CT GA GC GG GT TA TC TG TT

S_1,2 S_1,1 S_1,5 S_2,1 S_2,5 S_1,6 S_2,6 S_1,3 S_2,2 S_1,4 S_2,4 S_2,3

BLAST

3. Bidirectional extension – (Smith Waterman algorithm)



(Big detour through local alignment)

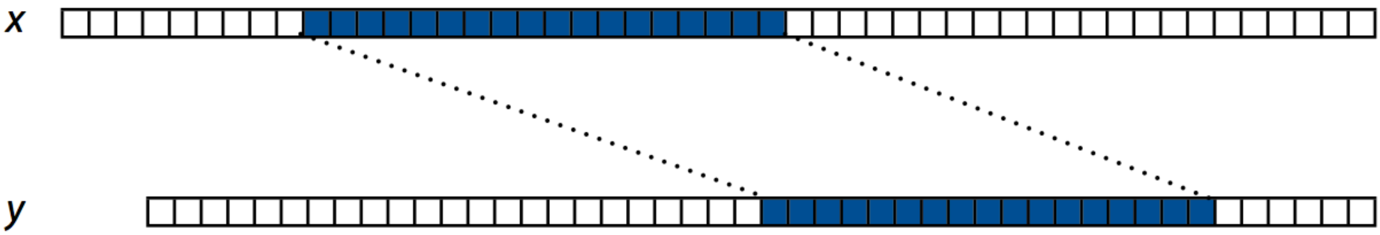
Smith Waterman local alignment.

The **Smith–Waterman algorithm** performs local sequence alignment; that is, for determining similar regions between two strings or nucleotide or protein sequences. Instead of looking at the total sequence, the Smith–Waterman algorithm compares segments of all possible lengths and optimizes the similarity measure.

(https://en.wikipedia.org/wiki/Smith%E2%80%93Waterman_algorithm)

Smith Waterman

Given strings *x* and *y*, what is the optimal global alignment value of a *substring* of *x* to a *substring* of *y*. This is *local alignment*.



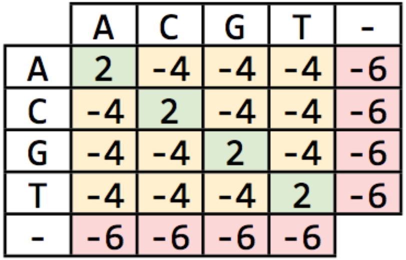
Smith Waterman

Let $V[0, j] = 0$, and let $V[i, 0] = 0$

$$\text{Otherwise, let } V[i,j] = \max \begin{cases} V[i-1, j] + s(x[i-1], -) \\ V[i, j-1] + s(-, y[j-1]) \\ V[i-1, j-1] + s(x[i-1], y[j-1]) \\ 0 \end{cases}$$

$s(a, b)$ assigns a score to a particular match, gap, or replacement

$s(a, b)$



	A	C	G	T	-
A	2	-4	-4	-4	-6
C	-4	2	-4	-4	-6
G	-4	-4	2	-4	-6
T	-4	-4	-4	2	-6
-	-6	-6	-6	-6	

Local alignment: Smith-Waterman

Let $V[0,j]=0$, and let $V[i,0]=0$

$$\text{Otherwise, let } V[i,j] = \max \begin{cases} V[i-1,j] + s(x[i-1],-) \\ V[i,j-1] + s(-,y[j-1]) \\ V[i-1,j-1] + s(x[i-1],y[j-1]) \\ 0 \end{cases}$$

$s(a,b)$ assigns a score to a particular match, gap, or replacement

Y

			T	A	T	A	T	G	C	G	G	C	G	T	T	T
		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	G	0														
	G	0														
	T	0														
	A	0														
	T	0														
	G	0														
X	C	0														
	T	0														
	G	0														
	G	0														
	C	0														
	G	0														
	C	0														
	T	0														
	A	0														

$$s(a,b)$$

	A	C	G	T	-
A	2	-4	-4	-4	-6
C	-4	2	-4	-4	-6
G	-4	-4	2	-4	-6
T	-4	-4	-4	2	-6
-	-6	-6	-6	-6	

Smith Waterman

$$V[i,j] = \max \begin{cases} V[i-1,j] + s(x[i-1], -) \\ V[i,j-1] + s(-, y[j-1]) \\ V[i-1,j-1] + s(x[i-1], y[j-1]) \\ 0 \end{cases}$$

	ϵ	T	A	T	A	T	G	C	G	G	C	G	T	T	T
ϵ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	2	0	2	2	0	2	0	0	0
G	0	0	0	0	0	0	2	0	2	4	0	2	0	0	0
T	0	2	0	2	0	2	0	0	0	0	0	0	4	2	2
A	0	0	4	0	?										
T	0														
G	0														
C	0														
T	0														
G	0														
G	0														
C	0														
G	0														
C	0														
T	0														
A	0														

$s(a, b)$

	A	C	G	T	-
A	2	-4	-4	-4	-6
C	-4	2	-4	-4	-6
G	-4	-4	2	-4	-6
T	-4	-4	-4	2	-6
-	-6	-6	-6	-6	

Smith Waterman

$$V[i,j] = \max \begin{cases} V[i-1,j] + s(x[i-1], -) \\ V[i,j-1] + s(-, y[j-1]) \\ V[i-1,j-1] + s(x[i-1], y[j-1]) \\ 0 \end{cases}$$

	ϵ	T	A	T	A	T	G	C	G	G	C	G	T	T	T
ϵ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	2	0	2	2	0	2	0	0	0
G	0	0	0	0	0	0	2	0	2	4	0	2	0	0	0
T	0	2	0	2	0	2	0	0	0	0	0	0	4	2	2
A	0	0	4	0	4	0	0	0	0	0	0	0	0	0	0
T	0	2	0	6	0	6	0	0	0	0	0	0	2	2	2
G	0	0	0	0	2	0	8	2	2	2	0	2	0	0	0
C	0	0	0	0	0	0	2	10	4	0	4	0	0	0	0
T	0	2	0	2	0	2	0	4	6	0	0	0	2	2	2
G	0	0	0	0	0	0	4	0	6	8	2	2	0	0	0
G	0	0	0	0	0	0	2	0	2	8	4	4	0	0	0
C	0	0	0	0	0	0	0	4	0	2	10	4	0	0	0
G	0	0	0	0	0	0	2	0	6	2	4	12	6	0	0
C	0	0	0	0	0	0	0	4	0	2	4	6	8	2	0
T	0	2	0	2	0	2	0	0	0	0	0	0	8	10	4
A	0	0	4	0	4	0	0	0	0	0	0	0	2	4	6

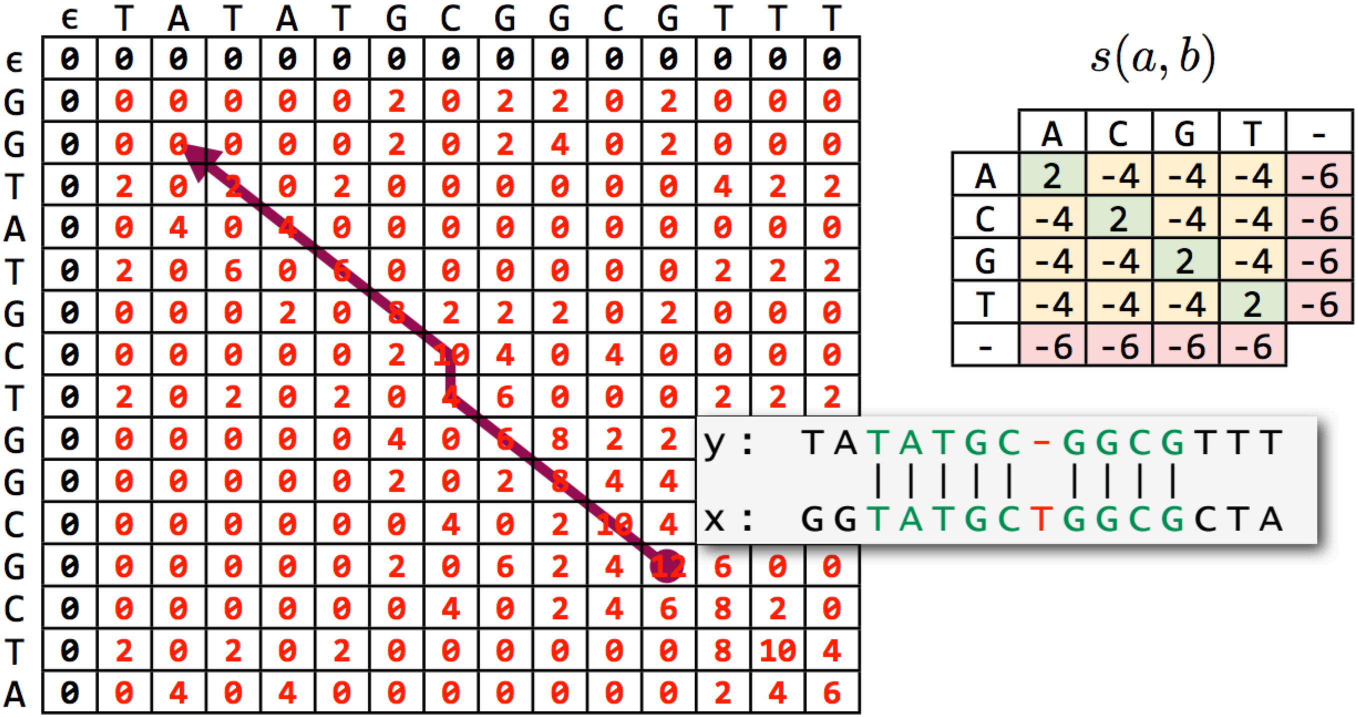
$s(a, b)$

	A	C	G	T	-
A	2	-4	-4	-4	-6
C	-4	2	-4	-4	-6
G	-4	-4	2	-4	-6
T	-4	-4	-4	2	-6
-	-6	-6	-6	-6	

0's in essence allow peaks of similarity to rise above "background" of 0s

Smith Waterman

Backtrace: (a) start from *maximal* cell in the matrix, (b) stop backtrace when we reach a cell with score = 0



	ϵ	T	A	T	A	T	G	C	G	G	C	G	T	T	T
ϵ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	2	0	2	2	0	2	0	0	0
G	0	0	0	0	0	0	2	0	2	4	0	2	0	0	0
T	0	2	0	2	0	2	0	0	0	0	0	0	4	2	2
A	0	0	4	0	4	0	0	0	0	0	0	0	0	0	0
T	0	2	0	6	0	6	0	0	0	0	0	0	2	2	2
G	0	0	0	0	2	0	8	2	2	2	0	2	0	0	0
C	0	0	0	0	0	0	2	10	4	0	4	0	0	0	0
T	0	2	0	2	0	2	0	4	6	0	0	0	2	2	2
G	0	0	0	0	0	0	4	0	6	8	2	2			
G	0	0	0	0	0	0	2	0	2	8	4	4			
C	0	0	0	0	0	0	0	4	0	2	10	4			
G	0	0	0	0	0	0	2	0	6	2	4	12	6	0	0
C	0	0	0	0	0	0	0	4	0	2	4	6	8	2	0
T	0	2	0	2	0	2	0	0	0	0	0	0	8	10	4
A	0	0	4	0	4	0	0	0	0	0	0	0	2	4	6

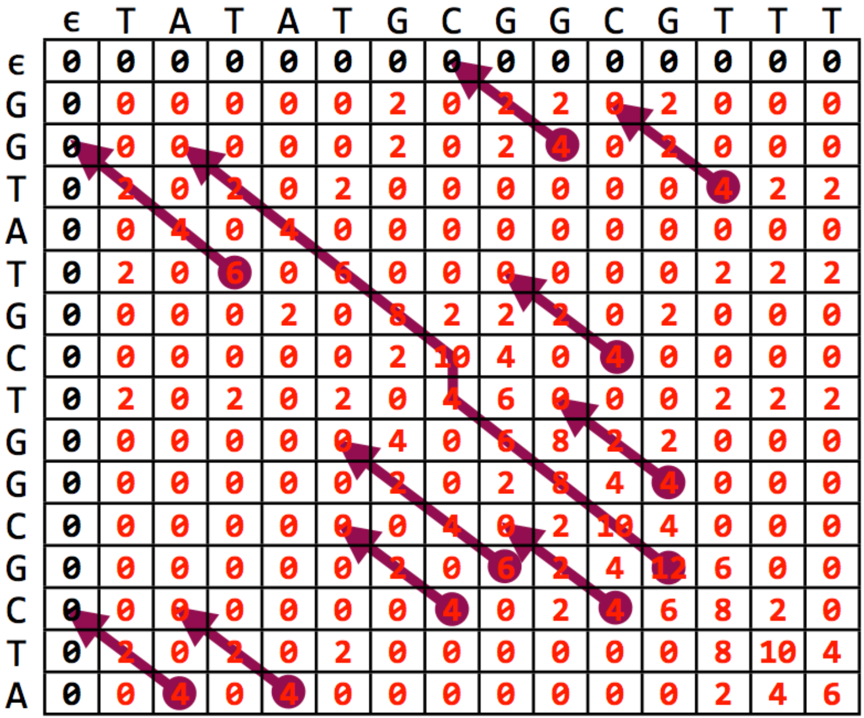
```
y:  T A T A T G C - G G C G T T T
        | | | | |   | | | |
x:  G G T A T G C T G G C G C T A
```

$s(a,b)$

	A	C	G	T	-
A	2	-4	-4	-4	-6
C	-4	2	-4	-4	-6
G	-4	-4	2	-4	-6
T	-4	-4	-4	2	-6
-	-6	-6	-6	-6	

Smith Waterman

We might be interested in the *best* local alignment, or in many *good-enough* local alignments



	ϵ	T	A	T	A	T	G	C	G	G	C	G	T	T	T
ϵ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	2	0	2	2	0	2	0	0	0
G	0	0	0	0	0	0	2	0	2	4	0	2	0	0	0
T	0	2	0	2	0	2	0	0	0	0	0	0	4	2	2
A	0	0	4	0	4	0	0	0	0	0	0	0	0	0	0
T	0	2	0	6	0	6	0	0	0	0	0	0	2	2	2
G	0	0	0	0	2	0	8	2	2	2	0	2	0	0	0
C	0	0	0	0	0	0	2	10	4	0	4	0	0	0	0
T	0	2	0	2	0	2	0	4	6	0	0	0	2	2	2
G	0	0	0	0	0	0	4	0	6	8	2	2	0	0	0
G	0	0	0	0	0	0	2	0	2	8	4	4	0	0	0
C	0	0	0	0	0	0	0	4	0	2	10	4	0	0	0
G	0	0	0	0	0	0	2	0	6	2	4	12	6	0	0
C	0	0	0	0	0	0	0	4	0	2	4	6	8	2	0
T	0	2	0	2	0	2	0	0	0	0	0	0	8	10	4
A	0	0	4	0	4	0	0	0	0	0	0	0	2	4	6

Reducing *good-enough* threshold risks allowing lots of tiny alignments that aren't very relevant

BLAST

4. Rank and report

BLAST

Stats

$$E = Kmne^{-\lambda S}$$

BLAST

Stats

$$\text{p} = 1 - e^{-E}$$

BLAST

Is my p-value significant?

	H_o true	H_o false
Reject H_o	**Type 1 error (false pos)**	**Correct!**
Accept H_o	Correct!	Type 2 error (false neg)

BLAST null: There is no match between query and database entry

BLAST

Multiple testing correction

http://www.r-fiddle.org/#/

Finding Data

Read data

- http://www.ebi.ac.uk/ena
- http://www.ncbi.nlm.nih.gov/sra
- http://metagenomics.anl.gov/?page=MetagenomeSelect

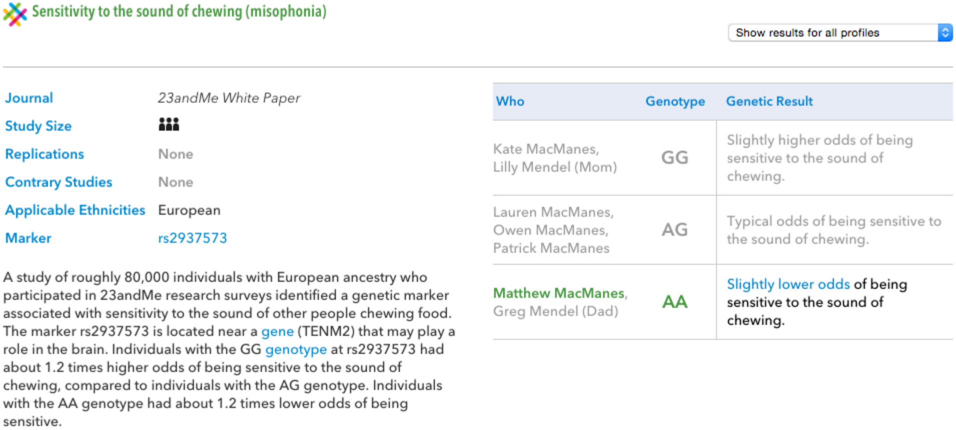
Assembly (and other) Data

- http://useast.ensembl.org/info/data/ftp/index.html
- http://www.ncbi.nlm.nih.gov/genome/
- http://datadryad.org/
- http://figshare.com/

Finding Data

Human Stuff

- http://www.ncbi.nlm.nih.gov/clinvar/
- http://www.ncbi.nlm.nih.gov/omim
- http://snpedia.com/index.php/SNPedia



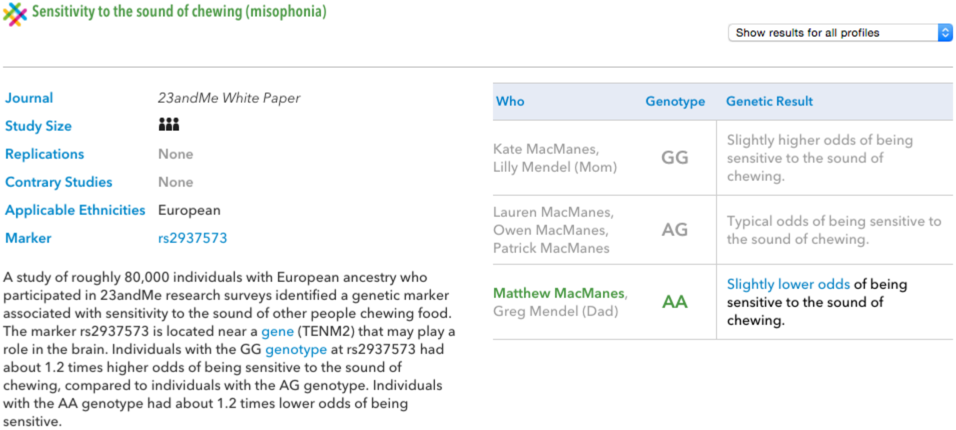
Sensitivity to the sound of chewing (misophonia)

Show results for all profiles

Journal	*23andMe White Paper*
Study Size	👤👤👤
Replications	None
Contrary Studies	None
Applicable Ethnicities	European
Marker	rs2937573

A study of roughly 80,000 individuals with European ancestry who participated in 23andMe research surveys identified a genetic marker associated with sensitivity to the sound of other people chewing food. The marker rs2937573 is located near a gene (TENM2) that may play a role in the brain. Individuals with the GG genotype at rs2937573 had about 1.2 times higher odds of being sensitive to the sound of chewing, compared to individuals with the AG genotype. Individuals with the AA genotype had about 1.2 times lower odds of being sensitive.

Who	Genotype	Genetic Result
Kate MacManes, Lilly Mendel (Mom)	GG	Slightly higher odds of being sensitive to the sound of chewing.
Lauren MacManes, Owen MacManes, Patrick MacManes	AG	Typical odds of being sensitive to the sound of chewing.
Matthew MacManes, Greg Mendel (Dad)	AA	Slightly lower odds of being sensitive to the sound of chewing.



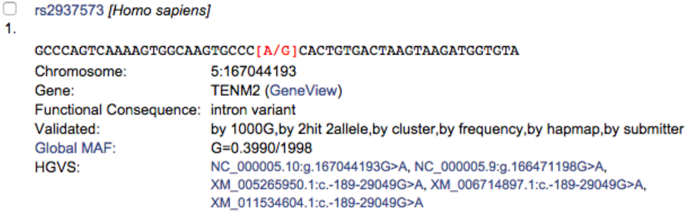
Sensitivity to the sound of chewing (misophonia)

Show results for all profiles

Journal	*23andMe White Paper*
Study Size	●●●
Replications	None
Contrary Studies	None
Applicable Ethnicities	European
Marker	rs2937573

A study of roughly 80,000 individuals with European ancestry who participated in 23andMe research surveys identified a genetic marker associated with sensitivity to the sound of other people chewing food. The marker rs2937573 is located near a gene (TENM2) that may play a role in the brain. Individuals with the GG genotype at rs2937573 had about 1.2 times higher odds of being sensitive to the sound of chewing, compared to individuals with the AG genotype. Individuals with the AA genotype had about 1.2 times lower odds of being sensitive.

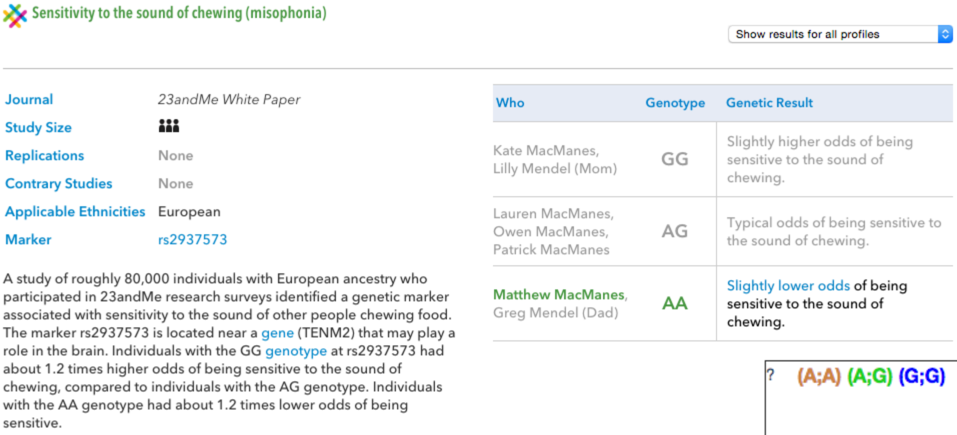
Who	Genotype	Genetic Result
Kate MacManes, Lilly Mendel (Mom)	GG	Slightly higher odds of being sensitive to the sound of chewing.
Lauren MacManes, Owen MacManes, Patrick MacManes	AG	Typical odds of being sensitive to the sound of chewing.
Matthew MacManes, Greg Mendel (Dad)	AA	Slightly lower odds of being sensitive to the sound of chewing.



□ 1. rs2937573 *[Homo sapiens]*

```
GCCCAGTCAAAAGTGGCAAGTGCCC[A/G]CACTGTGACTAAGTAAGATGGTGTA
```

Chromosome:	5:167044193
Gene:	TENM2 (GeneView)
Functional Consequence:	intron variant
Validated:	by 1000G,by 2hit 2allele,by cluster,by frequency,by hapmap,by submitter
Global MAF:	G=0.3990/1998
HGVS:	NC_000005.10:g.167044193G>A, NC_000005.9:g.166471198G>A, XM_005265950.1:c.-189-29049G>A, XM_006714897.1:c.-189-29049G>A, XM_011534604.1:c.-189-29049G>A



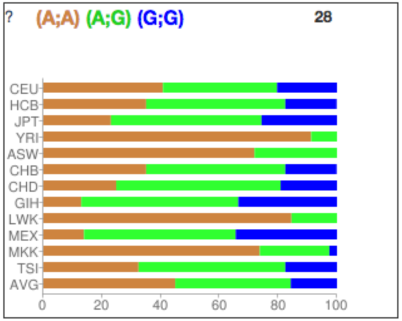
Sensitivity to the sound of chewing (misophonia)

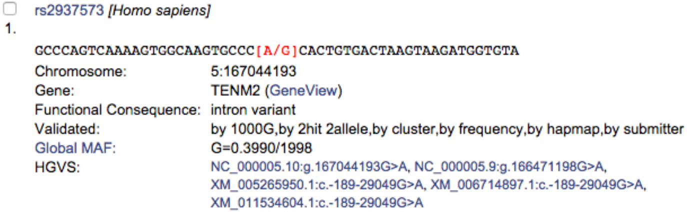
Show results for all profiles

Journal	*23andMe White Paper*
Study Size	●●●
Replications	None
Contrary Studies	None
Applicable Ethnicities	European
Marker	rs2937573

A study of roughly 80,000 individuals with European ancestry who participated in 23andMe research surveys identified a genetic marker associated with sensitivity to the sound of other people chewing food. The marker rs2937573 is located near a gene (TENM2) that may play a role in the brain. Individuals with the GG genotype at rs2937573 had about 1.2 times higher odds of being sensitive to the sound of chewing, compared to individuals with the AG genotype. Individuals with the AA genotype had about 1.2 times lower odds of being sensitive.

Who	Genotype	Genetic Result
Kate MacManes, Lilly Mendel (Mom)	GG	Slightly higher odds of being sensitive to the sound of chewing.
Lauren MacManes, Owen MacManes, Patrick MacManes	AG	Typical odds of being sensitive to the sound of chewing.
Matthew MacManes, Greg Mendel (Dad)	AA	Slightly lower odds of being sensitive to the sound of chewing.

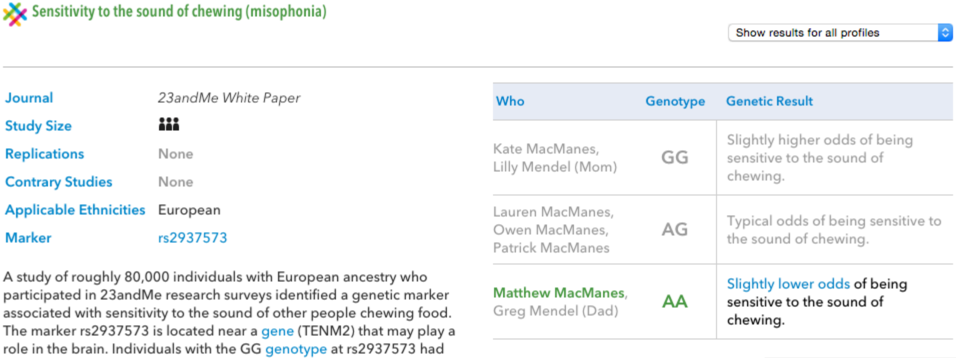




□ 1. rs2937573 *[Homo sapiens]*

```
GCCCAGTCAAAAGTGGCAAGTGCCC[A/G]CACTGTGACTAAGTAAGATGGTGTA
```

Chromosome:	5:167044193
Gene:	TENM2 (GeneView)
Functional Consequence:	intron variant
Validated:	by 1000G,by 2hit 2allele,by cluster,by frequency,by hapmap,by submitter
Global MAF:	G=0.3990/1998
HGVS:	NC_000005.10:g.167044193G>A, NC_000005.9:g.166471198G>A, XM_005265950.1:c.-189-29049G>A, XM_006714897.1:c.-189-29049G>A, XM_011534604.1:c.-189-29049G>A

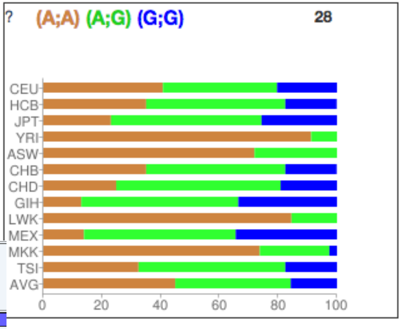


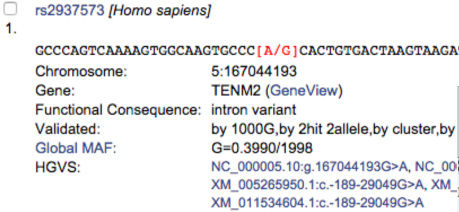
Sensitivity to the sound of chewing (misophonia)

Journal	*23andMe White Paper*
Study Size	👥
Replications	None
Contrary Studies	None
Applicable Ethnicities	European
Marker	rs2937573

A study of roughly 80,000 individuals with European ancestry who participated in 23andMe research surveys identified a genetic marker associated with sensitivity to the sound of other people chewing food. The marker rs2937573 is located near a gene (TENM2) that may play a role in the brain. Individuals with the GG genotype at rs2937573 had about 1.2 times higher odds of being sensitive to the sound of chewing, compared to individuals with the AG genotype. Individuals with the AA genotype had about 1.2 times lower odds of being sensitive.

Who	Genotype	Genetic Result
Kate MacManes, Lilly Mendel (Mom)	GG	Slightly higher odds of being sensitive to the sound of chewing.
Lauren MacManes, Owen MacManes, Patrick MacManes	AG	Typical odds of being sensitive to the sound of chewing.
Matthew MacManes, Greg Mendel (Dad)	AA	Slightly lower odds of being sensitive to the sound of chewing.



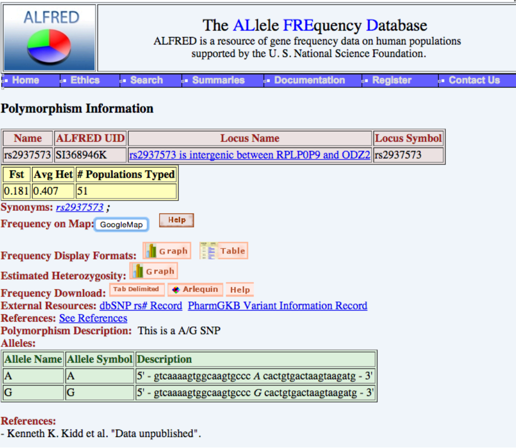


☐ rs2937573 *[Homo sapiens]*

1.

GCCCAGTCAAAAGTGGCAAGTGCCC[A/G]CACTGTGACTAAGTAAGATGGTGTA

Chromosome:	5:167044193
Gene:	TENM2 (GeneView)
Functional Consequence:	intron variant
Validated:	by 1000G,by 2hit 2allele,by cluster,by
Global MAF:	G=0.3990/1998
HGVS:	NC_000005.10:g.167044193G>A, NC_00 XM_005265950.1:c.-189-29049G>A, XM_ XM_011534604.1:c.-189-29049G>A



The ALlele FREquency Database

ALFRED is a resource of gene frequency data on human populations supported by the U. S. National Science Foundation.

Polymorphism Information

Name	ALFRED UID	Locus Name	Locus Symbol
rs2937573	SI368946K	rs2937573 is intergenic between RPLP0P9 and ODZ2	rs2937573

Fst	Avg Het	# Populations Typed
0.181	0.407	51

Synonyms: *rs2937573* ;

Frequency on Map: GoogleMap Help

Frequency Display Formats: Graph Table

Estimated Heterozygosity: Graph

Frequency Download: Tab Delimited Arlequin Help

External Resources: dbSNP rs# Record PharmGKB Variant Information Record

References: See References

Polymorphism Description: This is a A/G SNP

Alleles:

Allele Name	Allele Symbol	Description
A	A	5' - gtcaaaagtggcaagtgccc *A* cactgtgactaagtaagatg - 3'
G	G	5' - gtcaaaagtggcaagtgccc *G* cactgtgactaagtaagatg - 3'

References:

- Kenneth K. Kidd et al. "Data unpublished".