U.S. FOOD & DRUG ADMINISTRATION

Creating new and updated codon usage tables in HIVE using species-specific genomic and tissue-specific transcriptomic information

Luis Santana-Quintero, PhD
Office of Biostatistics and Pharmacovigilance
Center for Biologics Evaluation and Research

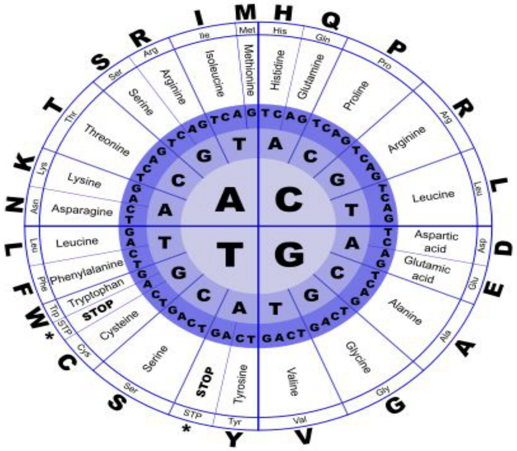
Disclaimer: *The information in this presentation represents the opinions of the speaker and does not necessarily represent FDA's position or policy.*

Outline

- Background
  - Codon, Codon Pairs, dinucleotides and junction dinucleotides
- Introduction
- Methods
  - Codon Usage Tables (CUT)
  - Transcriptomic Weighted Usage
- HIVE Platform
- Results and Examples
- Conclusions
- Acknowledgements

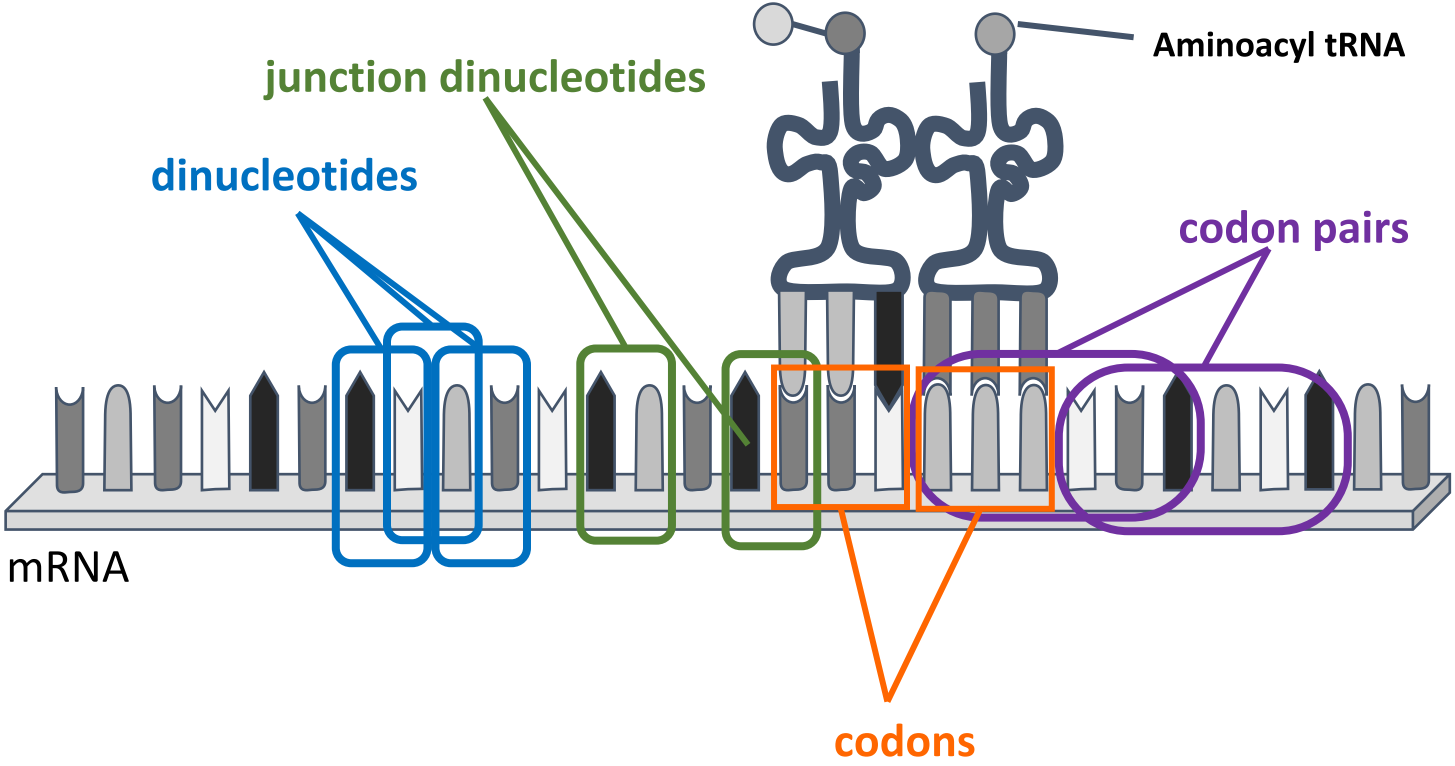
Background

- A codon is a DNA (or RNA) sequence of three nucleotides that encodes a particular **amino acid** or signaling the start/stop of protein synthesis.
- Genetic code is made up of codons and there are **64 different codons**.



- The codon usage tables are linked to a taxonomic reference, and they allow comparative analysis of the **codon usage frequencies**.

Background: Codons, Codon Pairs and Dinucleotides



Modified from Pearson Prentice Hall copyrighted material

Introduction

- We have created updated usage tables with the sequence information from GenBank and RefSeq. Including information for the non-random distribution of occurrences in genes within a given species:
  - Codon
  - Codon pair
  - Dinucleotide
  - Junction Dinucleotide
  - GC content

In total, 288 million coding sequences (35 million from GenBank, 253 million from RefSeq) were included in the database, resulting in the creation of over 855,000 codon usage tables.

- Accounting for differential gene expression profiles in various human tissues, we have also created usage tables for normal **human tissues** and for **human primary tumors**.

Methods

We created a publicly accessible repository of comprehensive, regularly updated Codon Usage Tables, HIVE-CUTs.

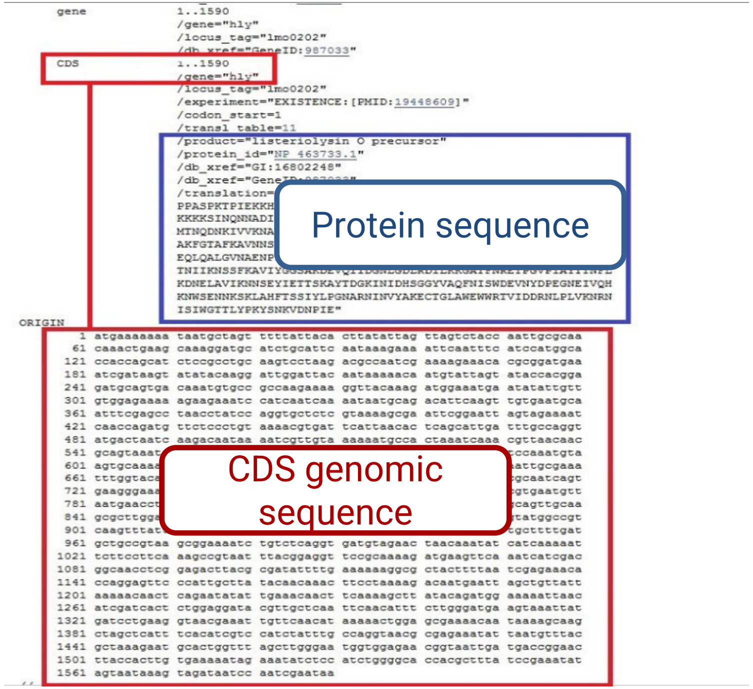
- CoCoPUTs – Codon and Codon Pair Usage Tables at species level (Genbank & RefSeq)
- TissueCoCoPUTs – 52 Human tissue-specific from GTEx Portal
- CancerCoCoPUTs – Tumor-specific contain transcriptome-derived data from 32 primary cancer types from TCGA

Athey *et. al.*, *BMC Bioinformatics*, 2017
Alexaki *et. al,* Journal of Molecular Biology, 2019
Holcomb *et. al*, *Infection, Genetics and Evolution*, 2019
Kames *et. al.*, *Journal of Molecular Biology*, 2019
Meyer *et. al.*, *Genome Medicine*, 2021

Codon Usage Table

Input: Listeria Genbank file (.gbff file)

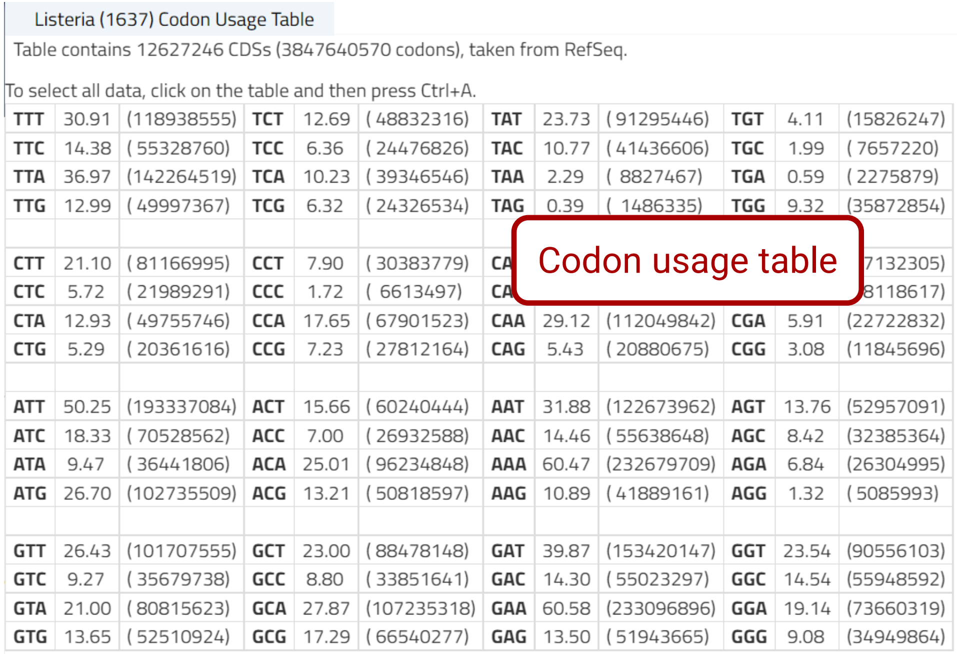
Output: Listeria raw count and proportion of each codon that appears on the sequence



```
     gene            1..1590
                     /gene="hly"
                     /locus_tag="lmo0202"
                     /db_xref="GeneID:987033"
     CDS             1..1590
                     /gene="hly"
                     /locus_tag="lmo0202"
                     /experiment="EXISTENCE:[PMID:19448609]"
                     /codon_start=1
                     /transl_table=11
                     /product="listeriolysin O precursor"
                     /protein_id="NP_463733.1"
                     /db_xref="GI:16802248"
                     /db_xref="Gene
                     /translation=
                     PPASPKTPIEKKH
                     KKKKSINQNNADI
                     MTNQDNKIVVKNA
                     AKFGTAFKAVNNS
                     EQLQALGVNAENP
                     TNIIKNSSFKAVIYGGSAKDEVQIIDGNLGDLRDILKKGATFNRETPGVPIAYTTNFL
                     KDNELAVIKNNSEYIETTSKAYTDGKINIDHSGGYVAQFNISWDEVNYDPEGNEIVQH
                     KNWSENNKSKLAHFTSSIYLPGNARNINVYAKECTGLAWEWWRTVIDDRNLPLVKNRN
                     ISIWGTTLYPKYSNKVDNPIE"
ORIGIN
        1 atgaaaaaaa taatgctagt ttttattaca cttatattag ttagtctacc aattgcgcaa
       61 caaactgaag caaaggatgc atctgcattc aataaagaaa attcaatttc atccatggca
      121 ccaccagcat ctccgcctgc aagtcctaag acgccaatcg aaaagaaaca cgcggatgaa
      181 atcgataagt atatacaagg attggattac aataaaaaca atgtattagt ataccacgga
      241 gatgcagtga caaatgtgcc gccaagaaaa ggttacaaag atggaaatga atatattgtt
      301 gtggagaaaa agaagaaatc catcaatcaa aataatgcag acattcaagt tgtgaatgca
      361 atttcgagcc taacctatcc aggtgctctc gtaaaagcga attcggaatt agtagaaaat
      421 caaccagatg ttctccctgt aaaacgtgat tcattaacac tcagcattga tttgccaggt
      481 atgactaatc aagacaataa aatcgttgta aaaaatgcca ctaaatcaaa cgttaacaac
      541 gcagtaaat                                              ccaaatgta
      601 agtgcaaaa                                              attgcgaaa
      661 tttggtaca                                              gcaatcagt
      721 gaagggaaa                                              gtgaatgtt
      781 aatgaacct                                              cagttgcaa
      841 gcgcttgga                                              tatggccgt
      901 caagtttat                                              tgcttttgat
      961 gctgccgtaa gcggaaaatc tgtctcaggt gatgtagaac taacaaatat catcaaaaat
     1021 tcttccttca aagccgtaat ttacggaggt tccgcaaaag atgaagttca aatcatcgac
     1081 ggcaacctcg gagacttacg cgatattttg aaaaaaggcg ctacttttaa tcgagaaaca
     1141 ccaggagttc ccattgctta tacaacaaac ttcctaaaag acaatgaatt agctgttatt
     1201 aaaaacaact cagaatatat tgaaacaact tcaaaagctt atacagatgg aaaaattaac
     1261 atcgatcact ctggaggata cgttgctcaa ttcaacattt cttgggatga agtaaattat
     1321 gatcctgaag gtaacgaaat tgttcaacat aaaaactgga gcgaaaacaa taaaagcaag
     1381 ctagctcatt tcacatcgtc catctatttg ccaggtaacg cgagaaatat taatgtttac
     1441 gctaaagaat gcactggttt agcttgggaa tggtggagaa cggtaattga tgaccggaac
     1501 ttaccacttg tgaaaaatag aaatatctcc atctggggca ccacgcttta tccgaaatat
     1561 agtaataaag tagataatcc aatcgaataa
```

Protein sequence

CDS genomic sequence



Listeria (1637) Codon Usage Table

Table contains 12627246 CDSs (3847640570 codons), taken from RefSeq.

To select all data, click on the table and then press Ctrl+A.

TTT	30.91	(118938555)	TCT	12.69	(48832316)	TAT	23.73	(91295446)	TGT	4.11	(15826247)
TTC	14.38	(55328760)	TCC	6.36	(24476826)	TAC	10.77	(41436606)	TGC	1.99	(7657220)
TTA	36.97	(142264519)	TCA	10.23	(39346546)	TAA	2.29	(8827467)	TGA	0.59	(2275879)
TTG	12.99	(49997367)	TCG	6.32	(24326534)	TAG	0.39	(1486335)	TGG	9.32	(35872854)
CTT	21.10	(81166995)	CCT	7.90	(30383779)	CA					7132305)
CTC	5.72	(21989291)	CCC	1.72	(6613497)	CA					8118617)
CTA	12.93	(49755746)	CCA	17.65	(67901523)	CAA	29.12	(112049842)	CGA	5.91	(22722832)
CTG	5.29	(20361616)	CCG	7.23	(27812164)	CAG	5.43	(20880675)	CGG	3.08	(11845696)
ATT	50.25	(193337084)	ACT	15.66	(60240444)	AAT	31.88	(122673962)	AGT	13.76	(52957091)
ATC	18.33	(70528562)	ACC	7.00	(26932588)	AAC	14.46	(55638648)	AGC	8.42	(32385364)
ATA	9.47	(36441806)	ACA	25.01	(96234848)	AAA	60.47	(232679709)	AGA	6.84	(26304995)
ATG	26.70	(102735509)	ACG	13.21	(50818597)	AAG	10.89	(41889161)	AGG	1.32	(5085993)
GTT	26.43	(101707555)	GCT	23.00	(88478148)	GAT	39.87	(153420147)	GGT	23.54	(90556103)
GTC	9.27	(35679738)	GCC	8.80	(33851641)	GAC	14.30	(55023297)	GGC	14.54	(55948592)
GTA	21.00	(80815623)	GCA	27.87	(107235318)	GAA	60.58	(233096896)	GGA	19.14	(73660319)
GTG	13.65	(52510924)	GCG	17.29	(66540277)	GAG	13.50	(51943665)	GGG	9.08	(34949864)

Codon usage table

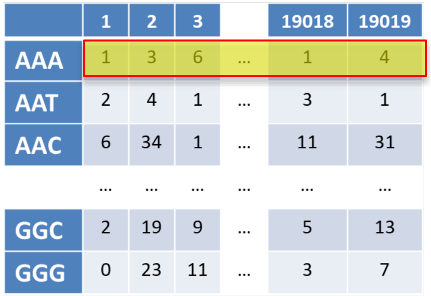
Cancer CoCoPUTs

RNA-seq files from NCI containing counts from 32 primary cancer types from TCGA

- Primary tumor
- Solid Tissue Normal

https://cancergenome.nih.gov/

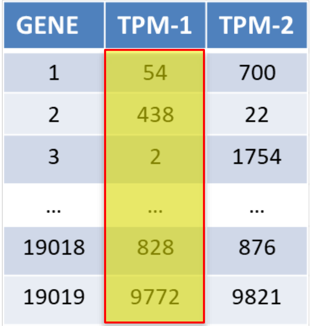
The Cancer Genome Atlas



	1	2	3		19018	19019
AAA	1	3	6	...	1	4
AAT	2	4	1	...	3	1
AAC	6	34	1	...	11	31
	...	...	...	...	...	...
GGC	2	19	9	...	5	13
GGG	0	23	11	...	3	7

Gene level usage (codon and codon pair)

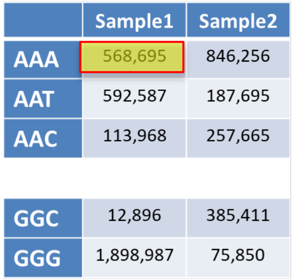
×



GENE	TPM-1	TPM-2
1	54	700
2	438	22
3	2	1754
...	...	...
19018	828	876
19019	9772	9821

Primary transcript quantifications (TPM for each gene)

=



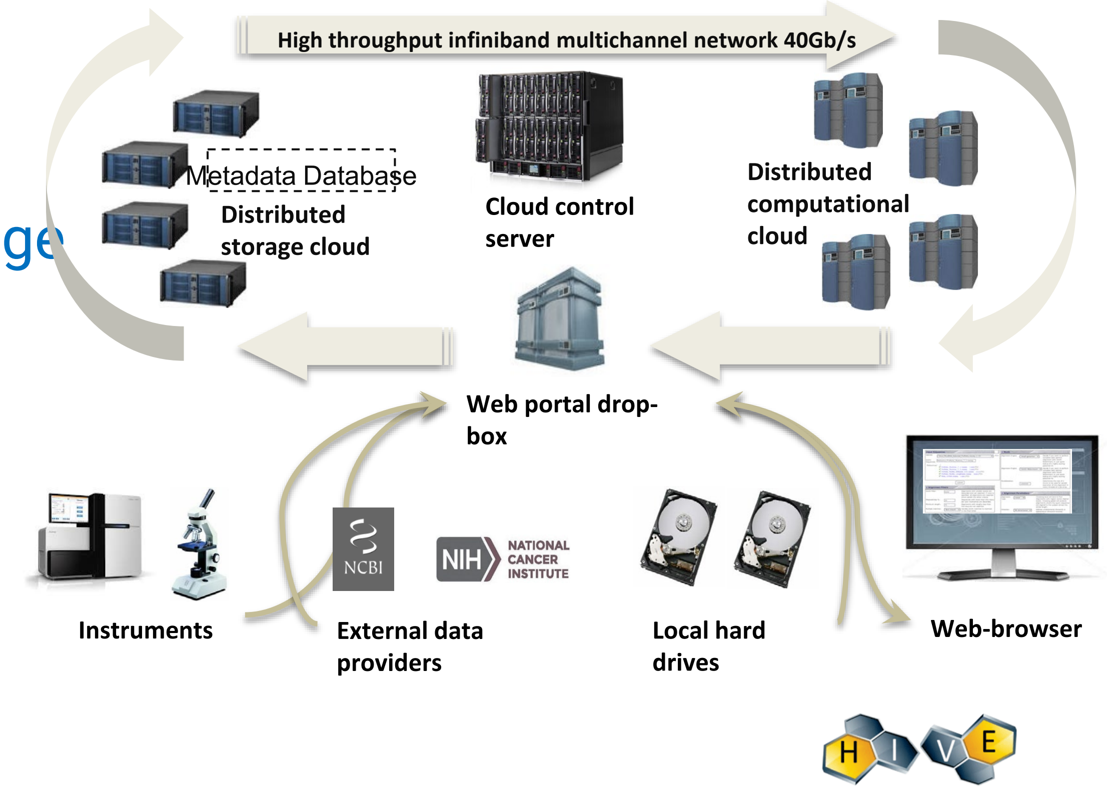
	Sample1	Sample2
AAA	568,695	846,256
AAT	592,587	187,695
AAC	113,968	257,665
GGC	12,896	385,411
GGG	1,898,987	75,850

Transcriptomic Weighted Usage

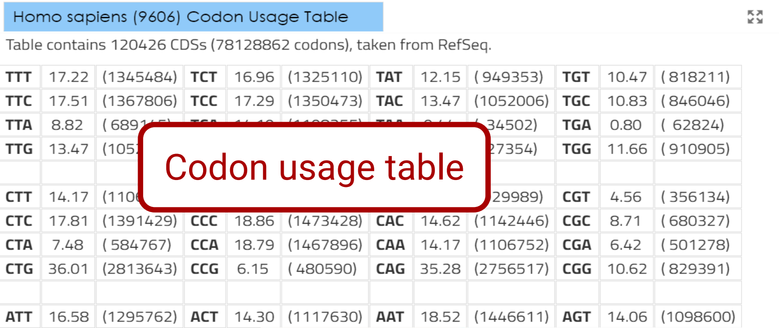
HIVE Platform

- A **cloud-based environment** that comprises both a storage library of data and a powerful computing capacity.

- Can **consume**, **digest**, **analyze**, **manage**, and **share all** this data.


High throughput infiniband multichannel network 40Gb/s
Metadata Database
Distributed storage cloud
Cloud control server
Distributed computational cloud
Web portal drop-box
Instruments
NCBI
NIH NATIONAL CANCER INSTITUTE
External data providers
Local hard drives
Web-browser
HIVE

Results: CoCoPUTs

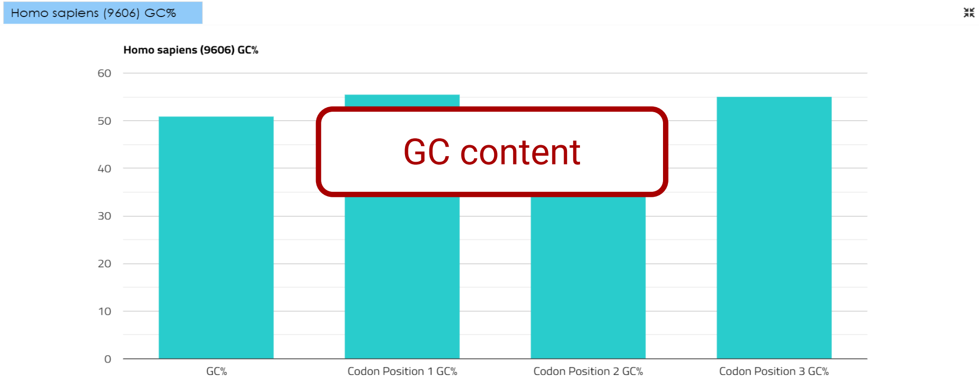


Homo sapiens (9606) Codon Usage Table

Table contains 120426 CDSs (78128862 codons), taken from RefSeq.

TTT	17.22	(1345484)	TCT	16.96	(1325110)	TAT	12.15	(949353)	TGT	10.47	(818211)
TTC	17.51	(1367806)	TCC	17.29	(1350473)	TAC	13.47	(1052006)	TGC	10.83	(846046)
TTA	8.82	(6891	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	34502)	TGA	0.80	(62824)
TTG	13.47	(105	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	27354)	TGG	11.66	(910905)
CTT	14.17	(110	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	29989)	CGT	4.56	(356134)
CTC	17.81	(1391429)	CCC	18.86	(1473428)	CAC	14.62	(1142446)	CGC	8.71	(680327)
CTA	7.48	(584767)	CCA	18.79	(1467896)	CAA	14.17	(1106752)	CGA	6.42	(501278)
CTG	36.01	(2813643)	CCG	6.15	(480590)	CAG	35.28	(2756517)	CGG	10.62	(829391)
ATT	16.58	(1295762)	ACT	14.30	(1117630)	AAT	18.52	(1446611)	AGT	14.06	(1098600)
ATC	18.68	(1459807)	ACC	17.7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
ATA	8.16	(637556)	ACA	16.5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
ATG	21.48	(1677888)	ACG	5.5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
GTT	11.80	(922234)	GCT	18.8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
GTC	13.44	(1050009)	GCC	25.6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
GTA	7.71	(602084)	GCA	17.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
GTG	25.75	(2011761)	GCG	5.9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

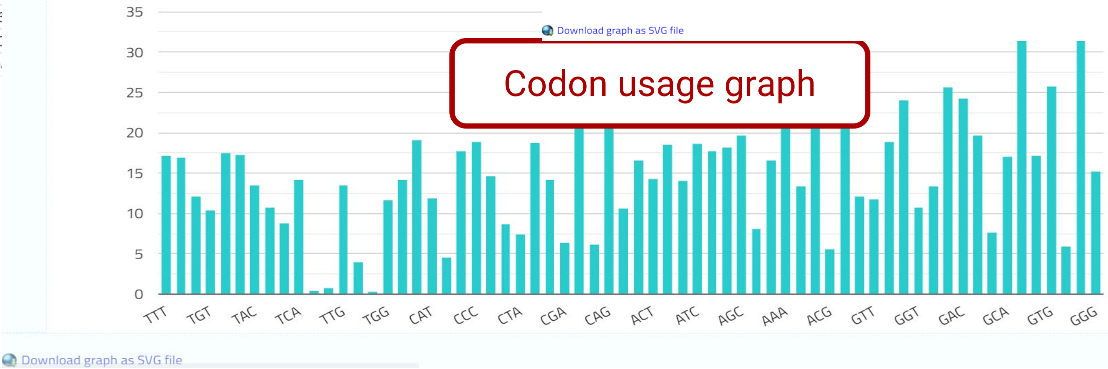
Codon usage table



Homo sapiens (9606) GC%

Download graph as SVG file

GC content



Codon Frequencies ✕

Download graph as SVG file

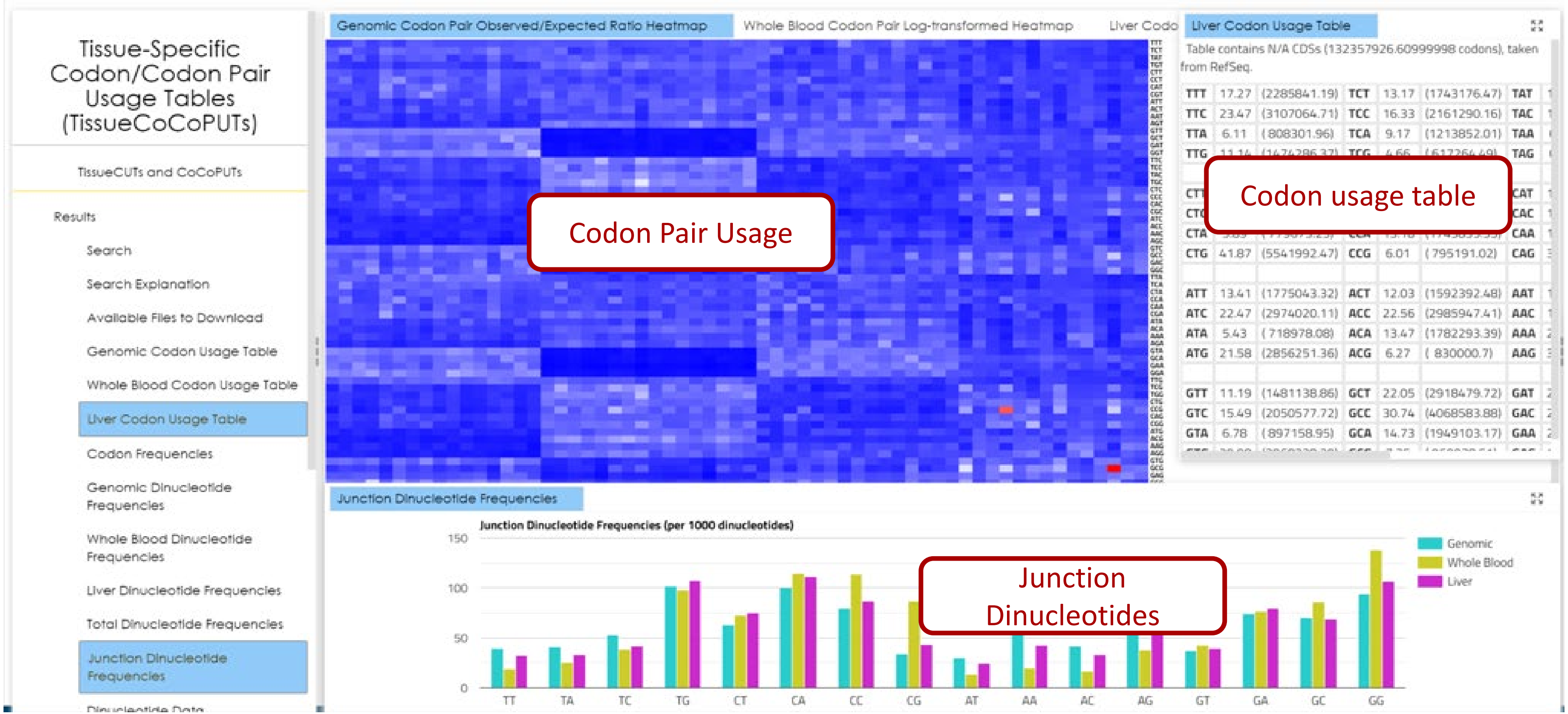
Codon usage graph

Codon Pair Usage

Dinucleotides

Junction Dinucleotides

Results: TissueCoCoPUTs



Tissue-Specific Codon/Codon Pair Usage Tables (TissueCoCoPUTs)

- TissueCUTs and CoCoPUTs
- Results
  - Search
  - Search Explanation
  - Available Files to Download
  - Genomic Codon Usage Table
  - Whole Blood Codon Usage Table
  - Liver Codon Usage Table
  - Codon Frequencies
  - Genomic Dinucleotide Frequencies
  - Whole Blood Dinucleotide Frequencies
  - Liver Dinucleotide Frequencies
  - Total Dinucleotide Frequencies
  - Junction Dinucleotide Frequencies
  - Dinucleotide Data

Genomic Codon Pair Observed/Expected Ratio Heatmap | Whole Blood Codon Pair Log-transformed Heatmap | Liver Codo

Codon Pair Usage

Liver Codon Usage Table

Table contains N/A CDSs (132357926.60999998 codons), taken from RefSeq.

TTT	17.27	(2285841.19)	TCT	13.17	(1743176.47)	TAT	1
TTC	23.47	(3107064.71)	TCC	16.33	(2161290.16)	TAC	1
TTA	6.11	(808301.96)	TCA	9.17	(1213852.01)	TAA	(
TTG	11.14	(1474286.37)	TCG	4.66	(617264.49)	TAG	(
CTT	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	CAT	1
CTC	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	CAC	1
CTA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	CAA	1
CTG	41.87	(5541992.47)	CCG	6.01	(795191.02)	CAG	3
ATT	13.41	(1775043.32)	ACT	12.03	(1592392.48)	AAT	1
ATC	22.47	(2974020.11)	ACC	22.56	(2985947.41)	AAC	1
ATA	5.43	(718978.08)	ACA	13.47	(1782293.39)	AAA	2
ATG	21.58	(2856251.36)	ACG	6.27	(830000.7)	AAG	3
GTT	11.19	(1481138.86)	GCT	22.05	(2918479.72)	GAT	2
GTC	15.49	(2050577.72)	GCC	30.74	(4068583.88)	GAC	2
GTA	6.78	(897158.95)	GCA	14.73	(1949103.17)	GAA	2

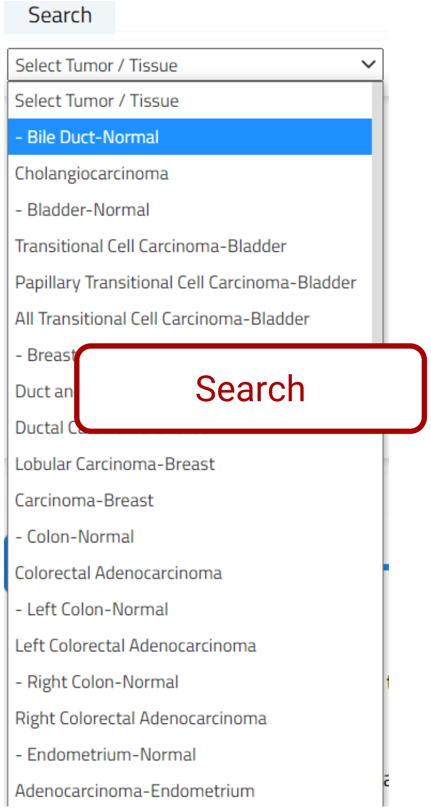
Codon usage table

Junction Dinucleotide Frequencies

Junction Dinucleotide Frequencies (per 1000 dinucleotides)

Junction Dinucleotides

Results: CancerCoCoPUTs

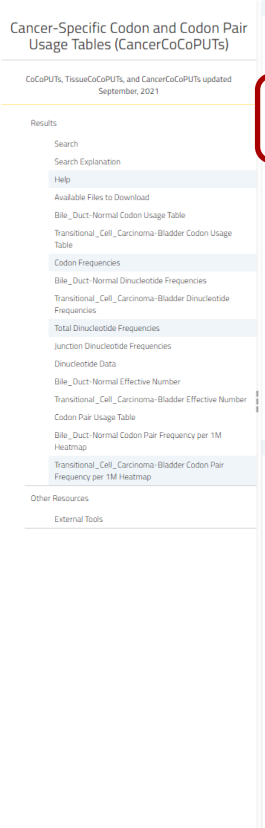


Search

Select Tumor / Tissue

- Select Tumor / Tissue
- - Bile Duct-Normal
- Cholangiocarcinoma
- - Bladder-Normal
- Transitional Cell Carcinoma-Bladder
- Papillary Transitional Cell Carcinoma-Bladder
- All Transitional Cell Carcinoma-Bladder
- - Breast
- Duct an
- Ductal C
- Lobular Carcinoma-Breast
- Carcinoma-Breast
- - Colon-Normal
- Colorectal Adenocarcinoma
- - Left Colon-Normal
- Left Colorectal Adenocarcinoma
- - Right Colon-Normal
- Right Colorectal Adenocarcinoma
- - Endometrium-Normal
- Adenocarcinoma-Endometrium

Search



Cancer-Specific Codon and Codon Pair Usage Tables (CancerCoCoPUTs)

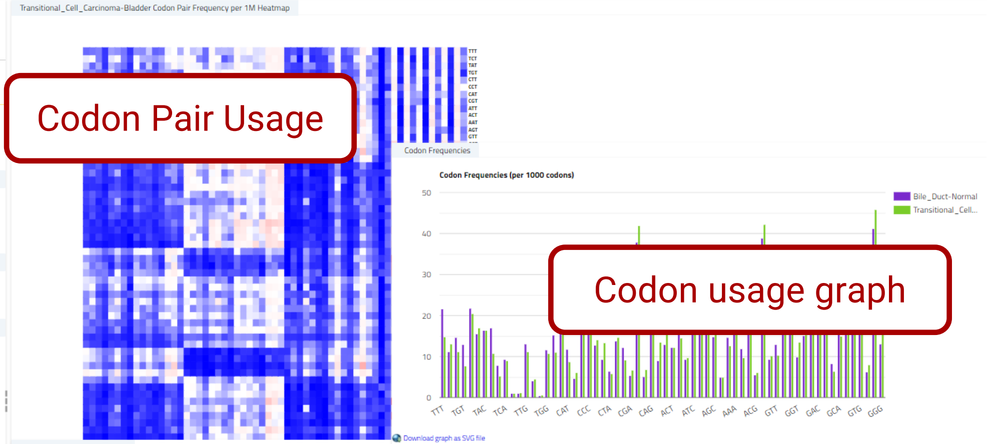
CoCoPUTs, TissueCoCoPUTs, and CancerCoCoPUTs updated September, 2021

Results

- Search
- Search Explanation
- Help
- Available Files to Download
- Bile_Duct-Normal Codon Usage Table
- Transitional_Cell_Carcinoma-Bladder Codon Usage Table
- Codon Frequencies
- Bile_Duct-Normal Dinucleotide Frequencies
- Transitional_Cell_Carcinoma-Bladder Dinucleotide Frequencies
- Total Dinucleotide Frequencies
- Junction Dinucleotide Frequencies
- Dinucleotide Data
- Bile_Duct-Normal Effective Number
- Transitional_Cell_Carcinoma-Bladder Effective Number
- Codon Pair Usage Table
- Bile_Duct-Normal Codon Pair Frequency per 1M Heatmap
- Transitional_Cell_Carcinoma-Bladder Codon Pair Frequency per 1M Heatmap

Other Resources

- External Tools



Transitional_Cell_Carcinoma-Bladder Codon Pair Frequency per 1M Heatmap

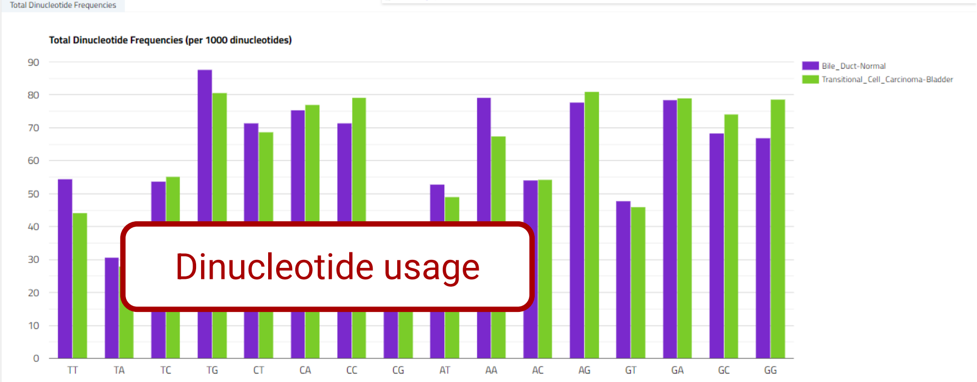
Codon Pair Usage

Codon Frequencies

Codon Frequencies (per 1000 codons)

Download graph as SVG file

Codon usage graph



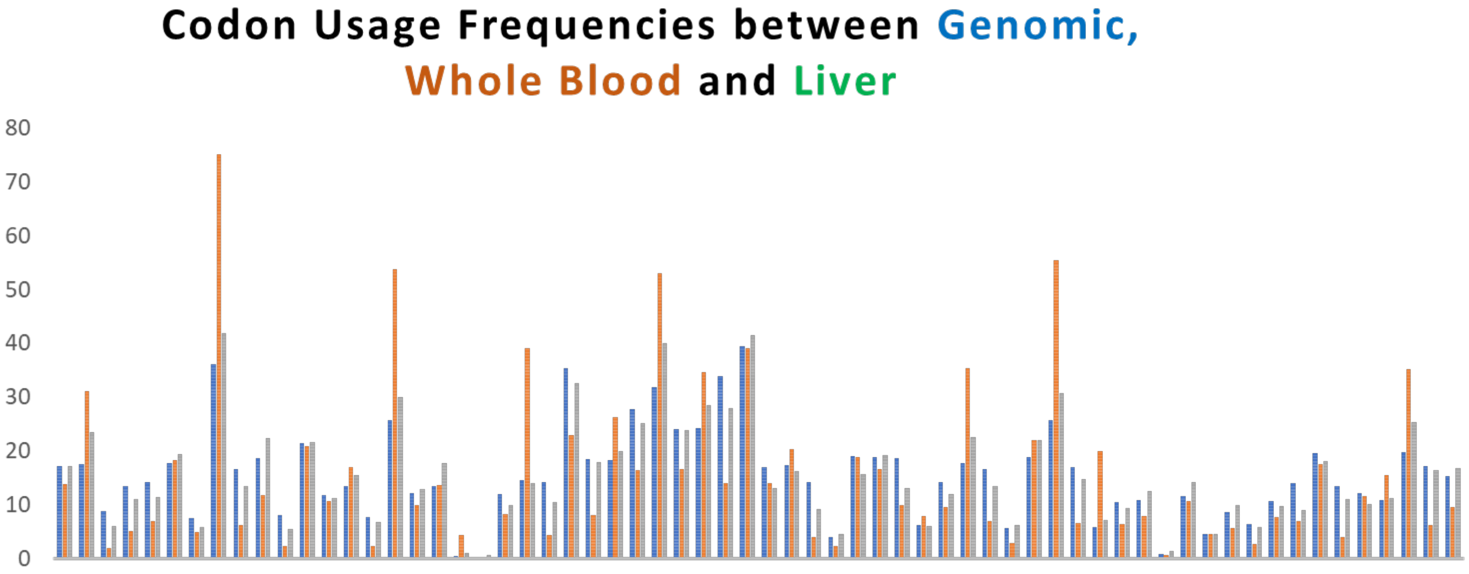
Total Dinucleotide Frequencies

Total Dinucleotide Frequencies (per 1000 dinucleotides)

Dinucleotide usage

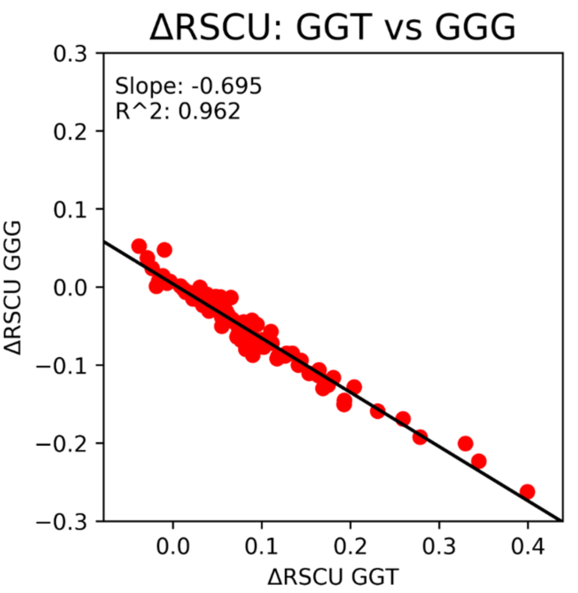
Example: Tissue-Specific Codon Usage

- Calculated from tissue-specific transcriptome (GTEx Portal and Protein Atlas) and human codon usage per CDS
- Results show large variability between tissues

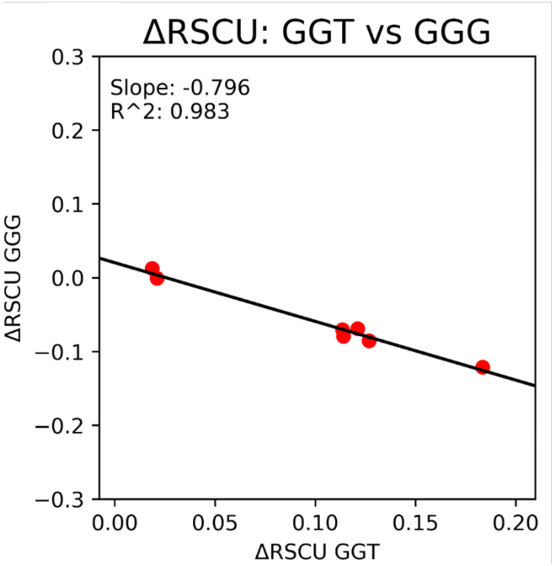


Example: Strong Negative Relationship Between GGT and GGG across Breast Cancer Subtypes

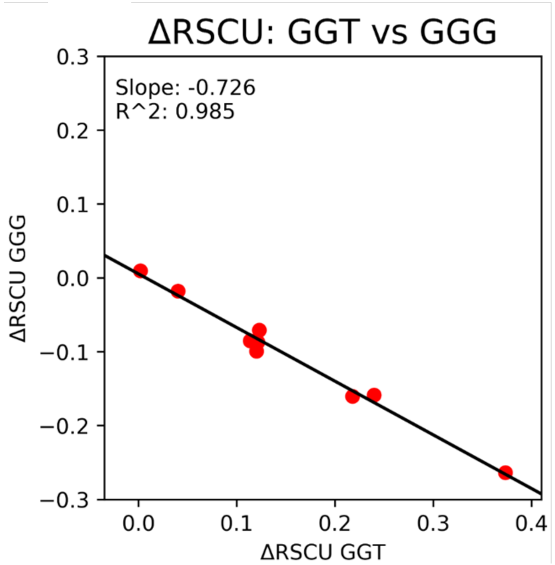
Ductal Carcinoma



Lobular Carcinoma



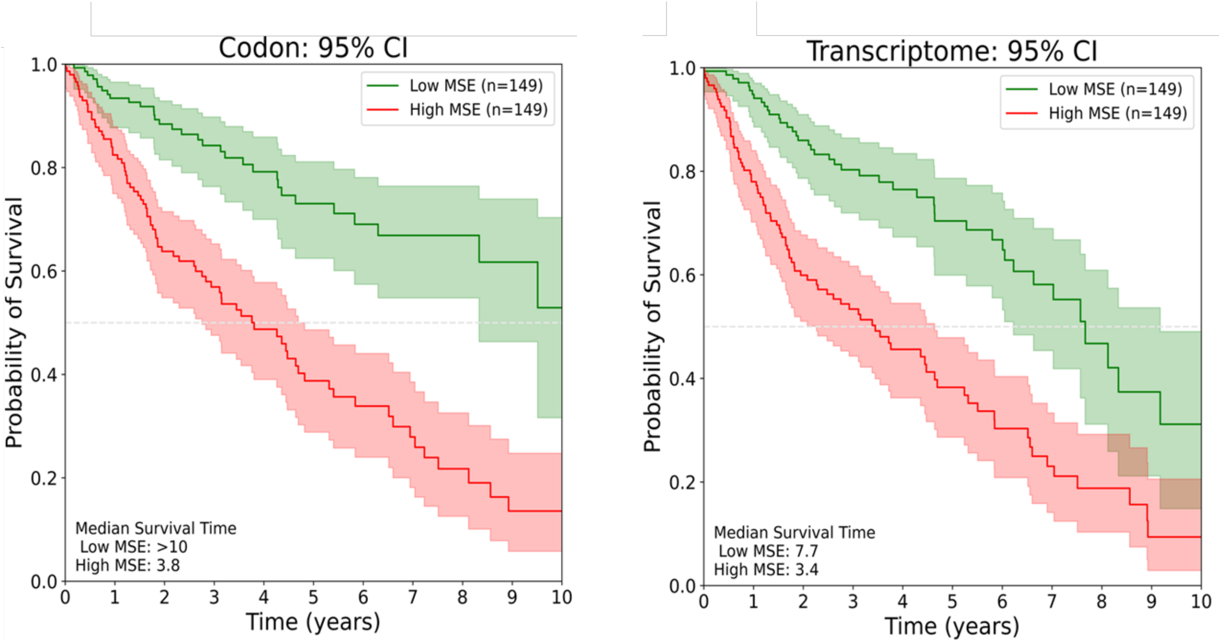
Duct and Lobular Carcinoma



Meyer D., Kames J. *et. al.*, *Genome Medicine*, 2021

Example: Change in Codon Usage Associated with Patient Survival Across All Cancers

MSE (Mean Squared Error):
Degree of Change



Meyer D., Kames J. *et. al.*, *Genome Medicine*, 2021

Conclusions

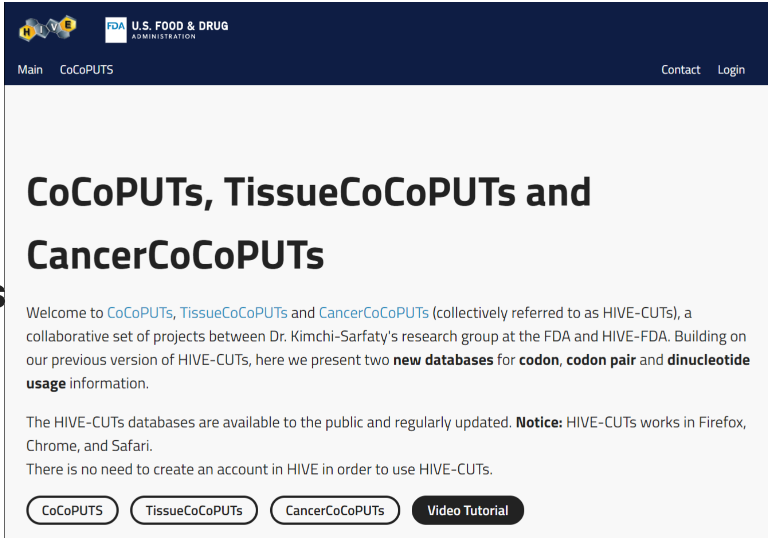
Codon Usage Tables provide the frequency of occurrence of codons and is essential in many biological studies and applications, e.g.

- Drug Development
- Gene Therapy
- Vaccine Development
- Implications of synonymous variants
- Recombinant Therapeutic protein
- Personalized Cancer Medicine
- Virus de-optimization design
- Evolutionary and Translational studies

Availability

https://dnahive.fda.gov/dna.cgi?cmd=cuts_main

1. Clickable heatmap displaying codon pair information
2. Text-based codon usage table
3. Graphs plotting the codon frequencies per 1000 codons
4. Graphs showing total dinucleotide and junction dinucleotide frequencies
5. Graphs showing the GC% content
6. Effective Number of Codons (ENC) and Codon Pairs (ENCP), metrics measuring codon and codon pair usage bias
7. Taxonomy tree that displays each query and traces the route back to their last shared classification (only available for species related queries)



CoCoPUTs, TissueCoCoPUTs and CancerCoCoPUTs

Welcome to CoCoPUTs, TissueCoCoPUTs and CancerCoCoPUTs (collectively referred to as HIVE-CUTs), a collaborative set of projects between Dr. Kimchi-Sarfaty's research group at the FDA and HIVE-FDA. Building on our previous version of HIVE-CUTs, here we present two **new databases** for **codon**, **codon pair** and **dinucleotide usage** information.

The HIVE-CUTs databases are available to the public and regularly updated. **Notice:** HIVE-CUTs works in Firefox, Chrome, and Safari.
There is no need to create an account in HIVE in order to use HIVE-CUTs.

Acknowledgements

Kimchi-Sarfaty Group (FDA)
Douglas Meyer
Jacob Kames
David Holcomb
Katerina Alexaki
John Athey
Ryan Hunt
Nobuko Katagiri
Upendra Katneni

HIVE (FDA)
Luis Santana-Quintero
Konstantinos Karagiannis
Anton Golikov
Sean Smith
Marianna Faradhzeva

Harvard University
Juan Ibla

University of Connecticut
Haim Bar

Cleveland State University
Anton Komar

NCBI
Michael DiCuccio

Thank you!

Disclaimer: *The information in this presentation represents the opinions of the speaker and does not necessarily represent FDA’s position or policy.*

FDA U.S. FOOD & DRUG ADMINISTRATION