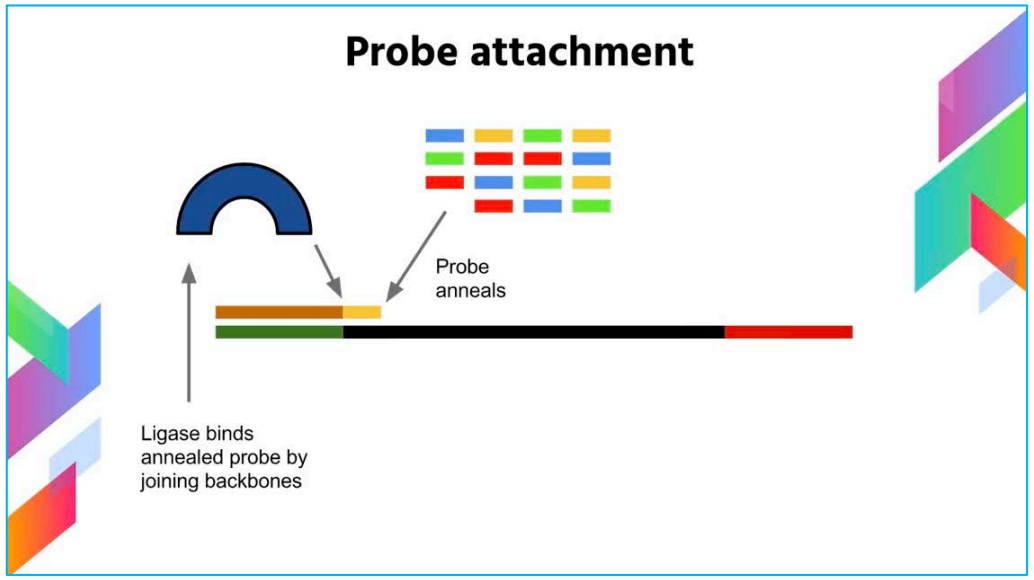
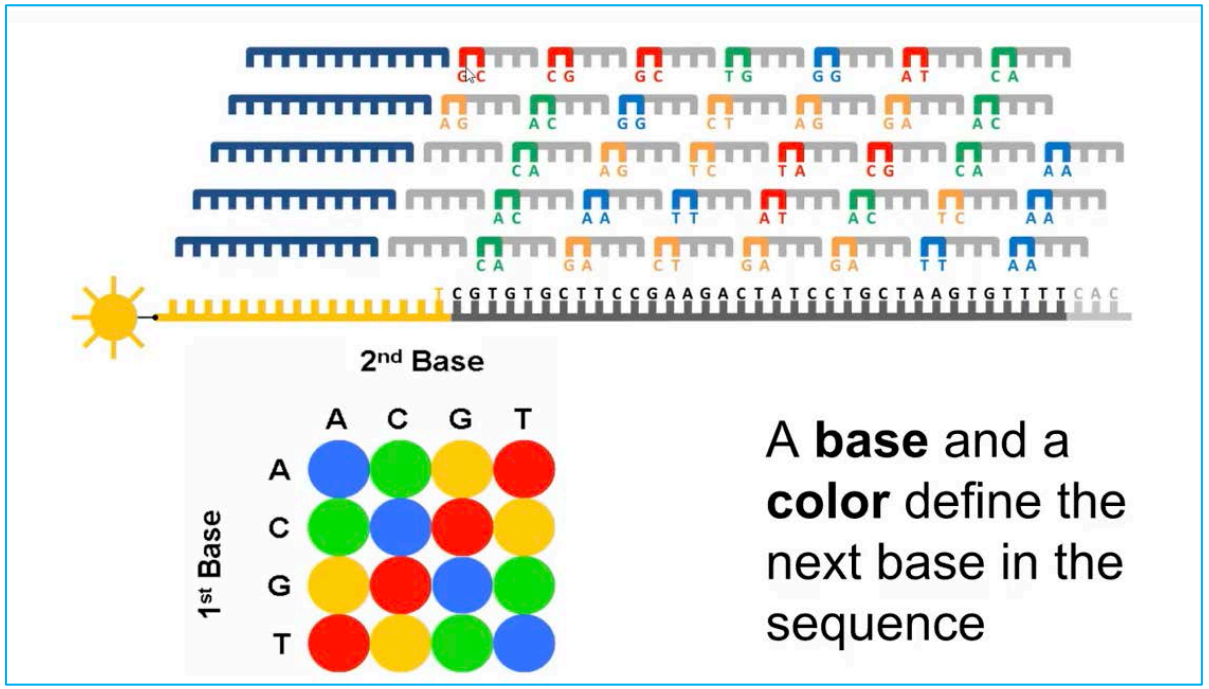
Sequencing technologies:

ABI & others - Sanger sequencing by capillary electrophoresis
Illumina - Next generation Sequencing
ABI - SOLiD sequencing

SOLiD Sequencing Strategy



https://youtu.be/YLT-DUeaLms



Sequencing Technologies

Short presentations on Thursday – 1 student per group

GROUPS	Assay Comparison
A & D	Illumina vs SOLiD
B & E	Illumina vs pyrosequencing
C & F	Illumina vs Pacific Biosciences

https://youtu.be/1zw2RE-PavQ

https://youtu.be/_OSRKtT_9vw

When Whales Walked

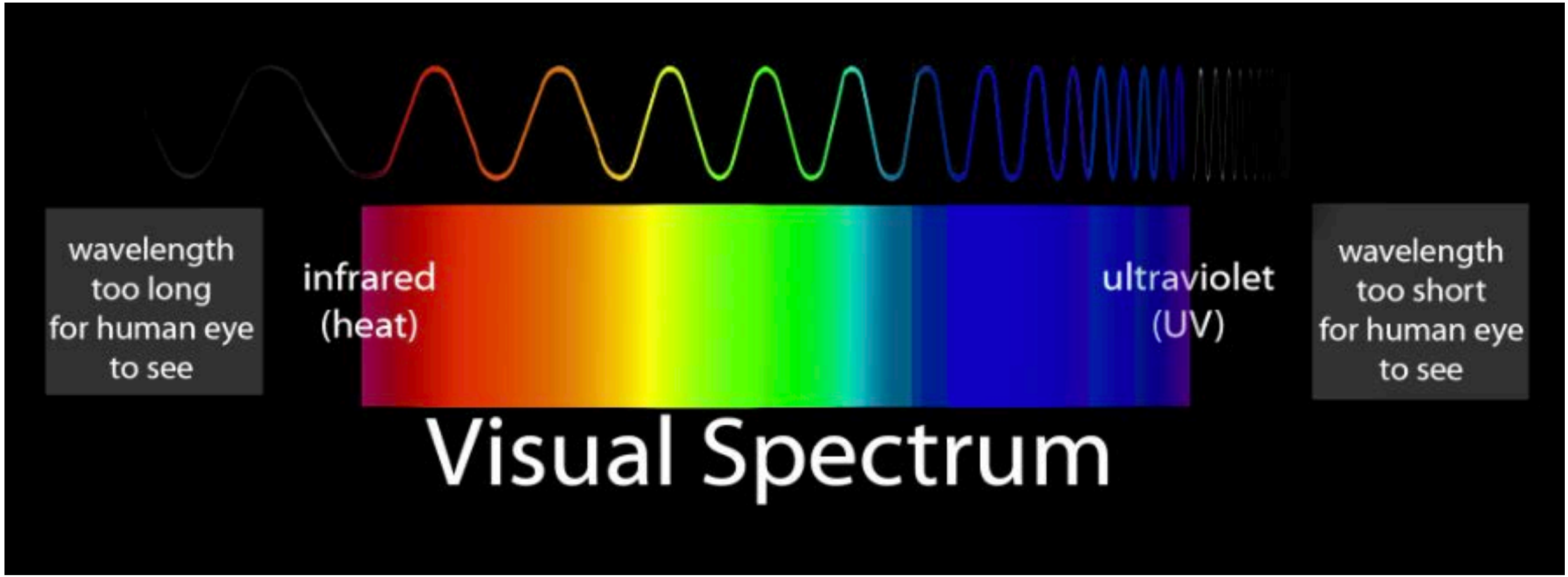
PBS Eons

Subscribe 704K

1,253,999 views

YouTube Search

The Origins of Human Color Vision – HHMI BioInteractive Video


wavelength too long for human eye to see
infrared (heat)
ultraviolet (UV)
wavelength too short for human eye to see
Visual Spectrum

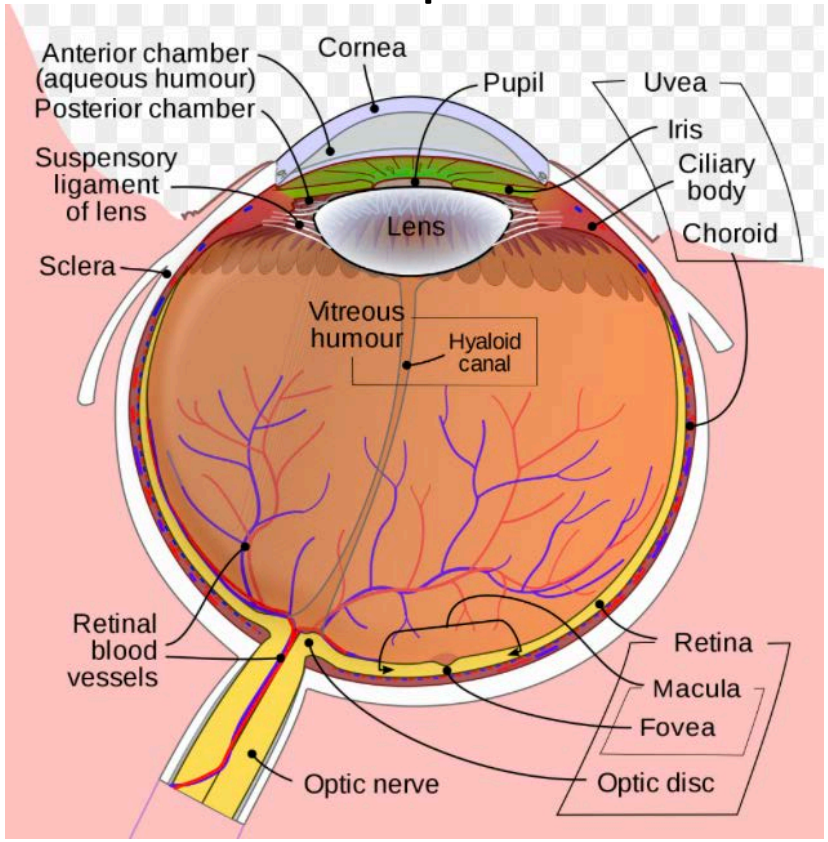
Primates have evolved a unique ability for three-dimensional color vision (trichromacy) from the two-dimensional color vision (dichromacy) present in the majority of other mammals. This was accomplished via allelic differentiation (e.g. most New World monkeys) or gene duplication (e.g. Old World primates) of the middle to long-wavelength sensitive (M/LWS, or red–green) opsin gene.

Allelic differentiation of the M/LWS opsins results in extensive color vision variability in New World monkeys where trichromats and dichromats are found in the same breeding population.

The visual system of primates:

- anthropoids (simians) [catarrhines (humans, apes and Old World monkeys)
- platyrrhines (New World monkeys)

At least two different spectral classes of cone photoreceptors are necessary in the retina to perceive differences of wavelength compositions (i.e. colors).



RH1 (rhodopsin or rod opsin for dim-light vision) and
four cone opsins:

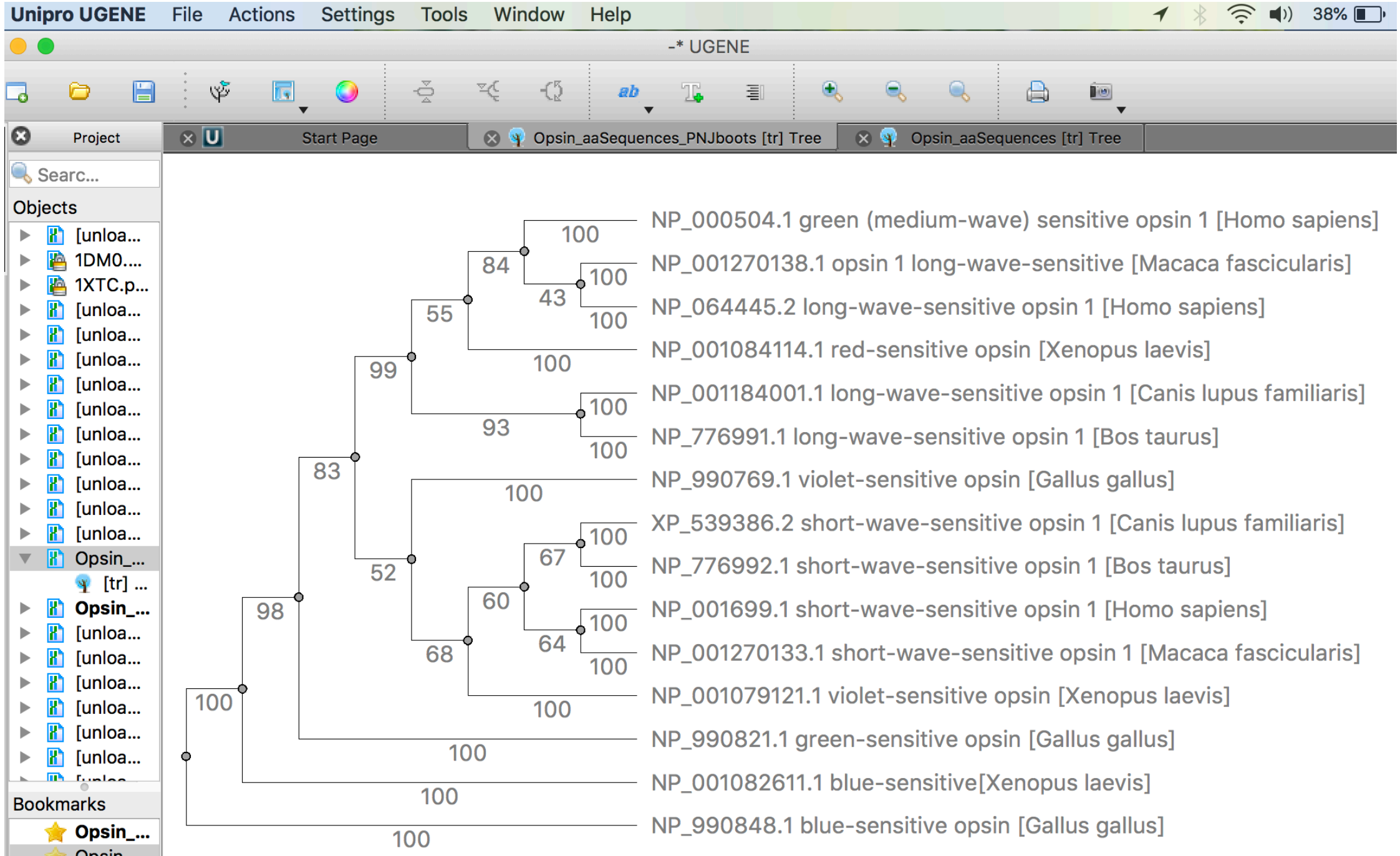
- RH2 (rhodopsin-like, or green),
- SWS1 (short wavelength-sensitive type 1, or ultraviolet-blue),
- SWS2 (short wavelength-sensitive type 2, or blue)
- M/LWS (middle to long wavelength-sensitive, or red–green)

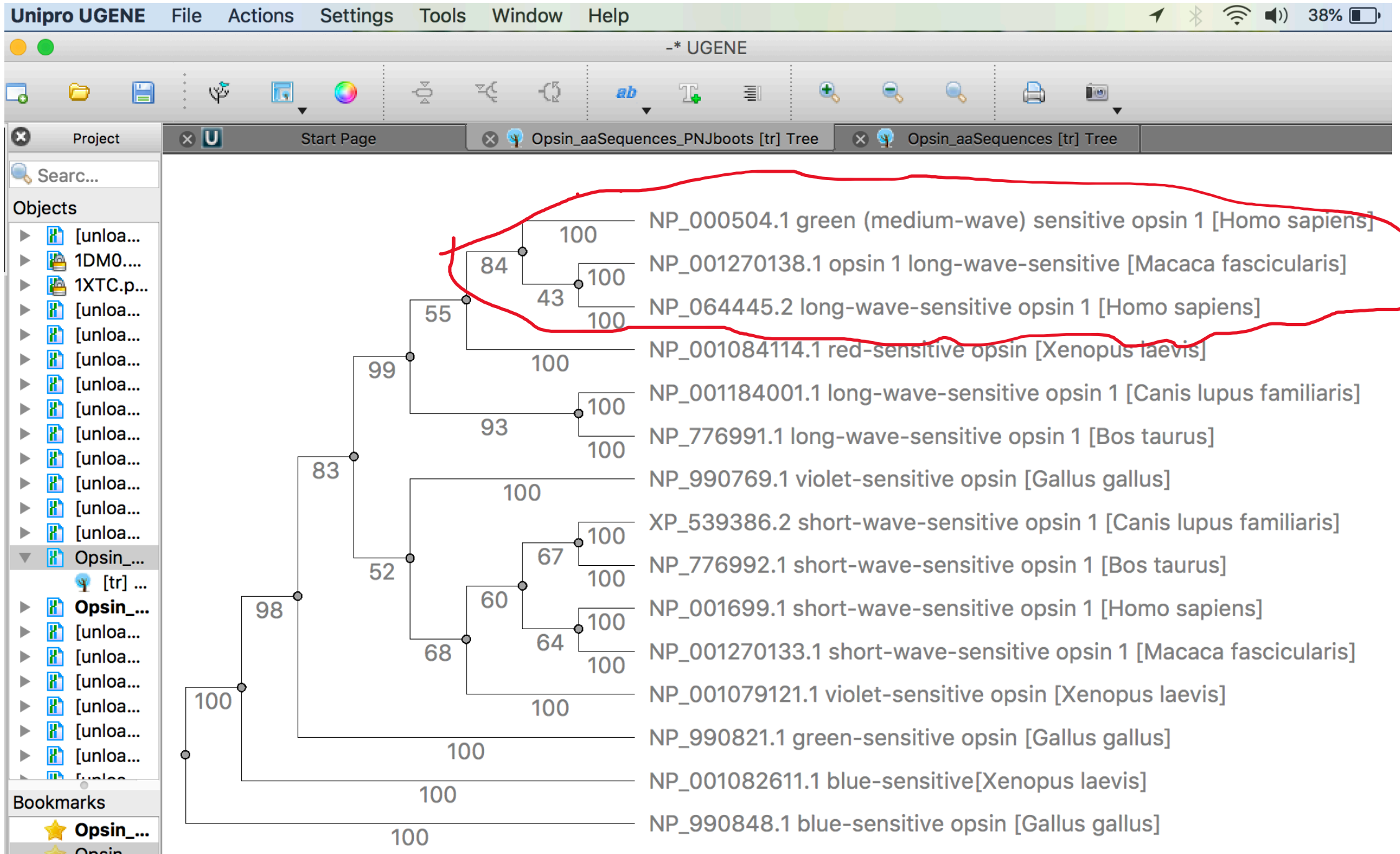
Thus, early vertebrates could already have had four-dimensional color vision (tetrachromacy).
Placental mammals maintain only two types of cone visual opsins, SWS1 and M/LWS, in addition to the RH1 rod opsin, and are hence dichromatic in color vision

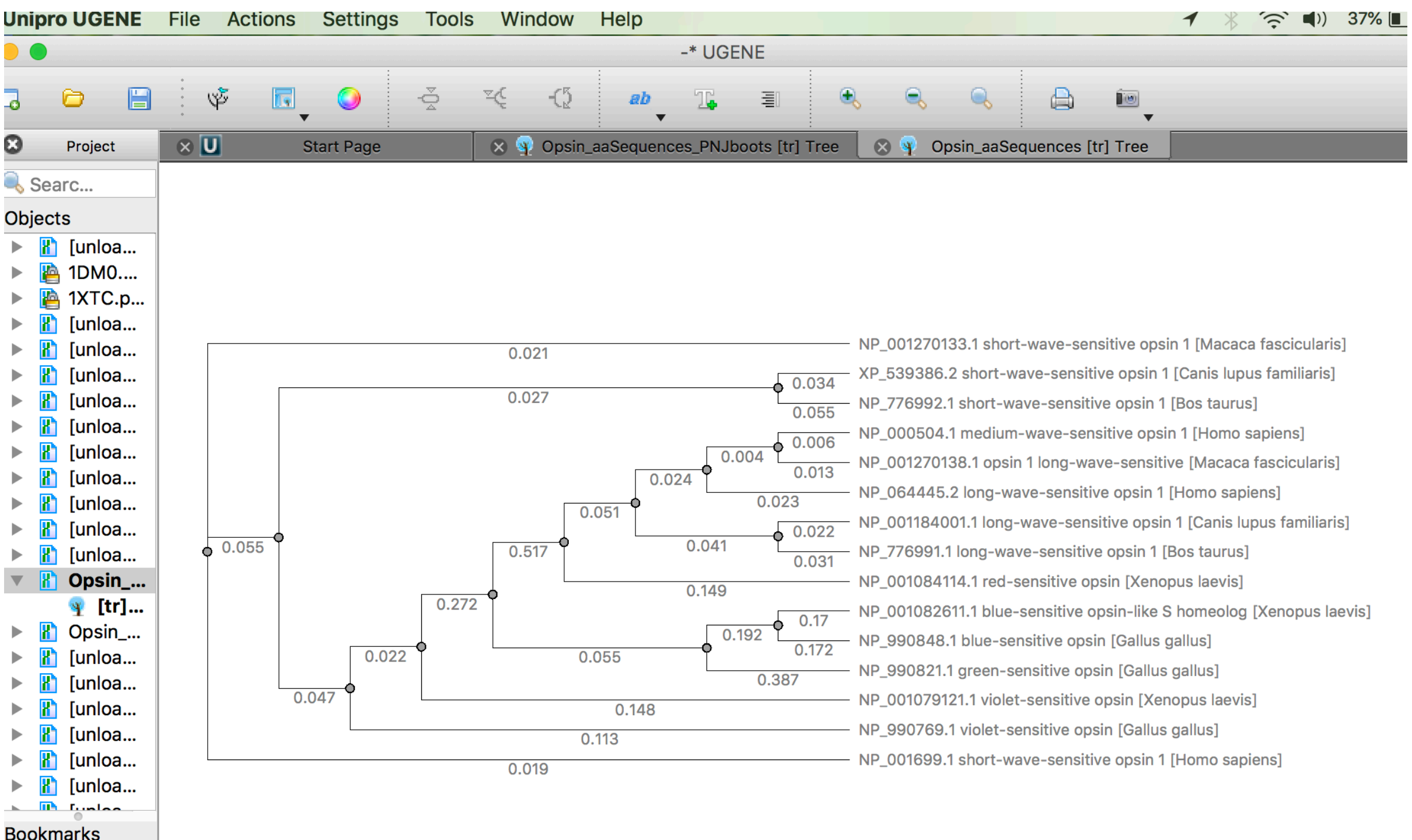
Midget ganglion pathway- a visual-signal-processor that creates fine spatial resolution

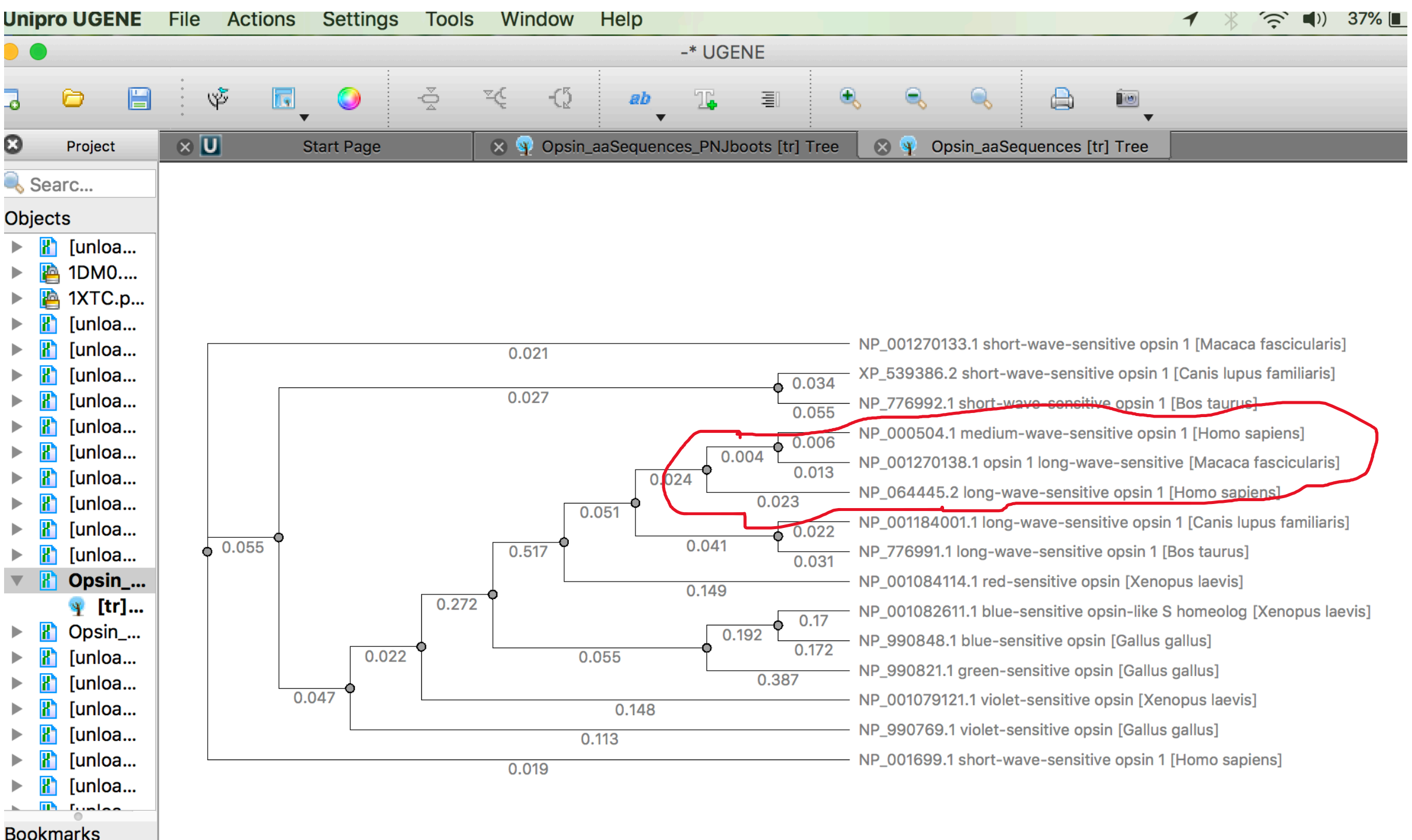
Among mammals, only primates are equipped with the midget ganglion pathway in the retina, which enables finer resolution of the spectral signals from two spectral classes of L/M cone cells. The midget pathway receives input from only one to approximately five L/M cone cells in the center of a midget ganglion receptive field and compares it to inputs from surrounding L/M cones, allowing high spatial acuity.

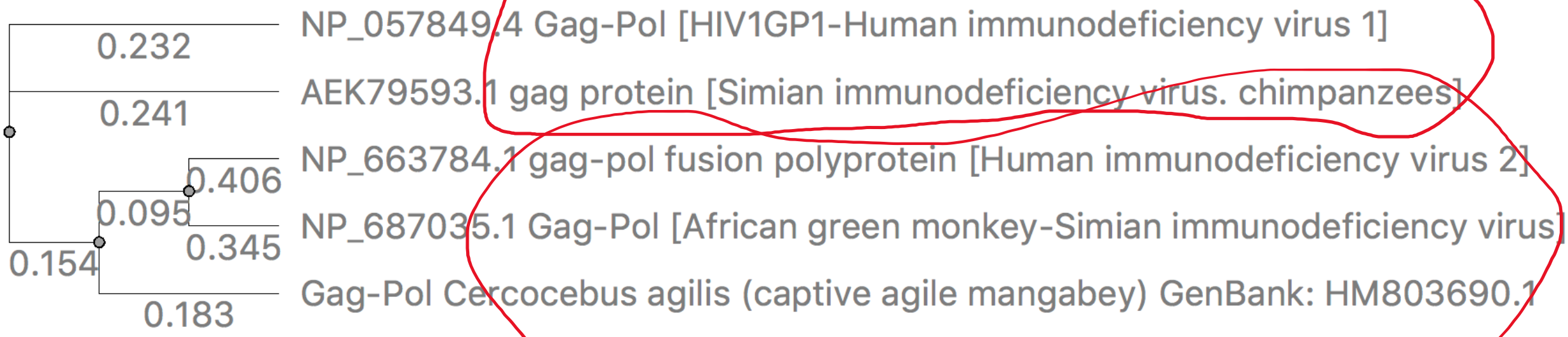
Phylogenetics Assignments

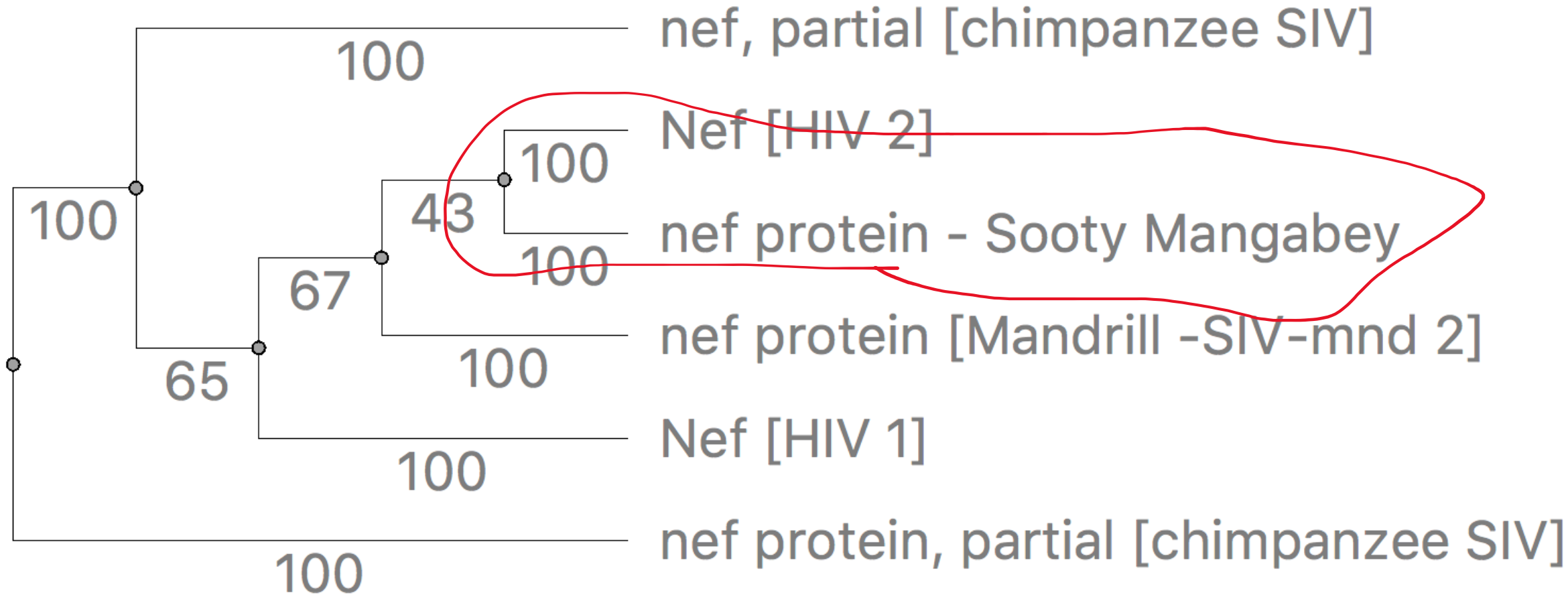

Unipro UGENE File Actions Settings Tools Window Help
38%
-* UGENE
Project
Start Page
Opsin_aaSequences_PNJboots [tr] Tree
Opsin_aaSequences [tr] Tree
Searc...
Objects
[unloa...
1DM0....
1XTC.p...
Opsin_...
[tr] ...
Bookmarks
Opsin_...
Opsin_...
100
84
100
43
100
55
100
99
100
93
100
83
100
52
100
67
100
60
100
64
100
68
100
98
100
100
100
100
NP_000504.1 green (medium-wave) sensitive opsin 1 [Homo sapiens]
NP_001270138.1 opsin 1 long-wave-sensitive [Macaca fascicularis]
NP_064445.2 long-wave-sensitive opsin 1 [Homo sapiens]
NP_001084114.1 red-sensitive opsin [Xenopus laevis]
NP_001184001.1 long-wave-sensitive opsin 1 [Canis lupus familiaris]
NP_776991.1 long-wave-sensitive opsin 1 [Bos taurus]
NP_990769.1 violet-sensitive opsin [Gallus gallus]
XP_539386.2 short-wave-sensitive opsin 1 [Canis lupus familiaris]
NP_776992.1 short-wave-sensitive opsin 1 [Bos taurus]
NP_001699.1 short-wave-sensitive opsin 1 [Homo sapiens]
NP_001270133.1 short-wave-sensitive opsin 1 [Macaca fascicularis]
NP_001079121.1 violet-sensitive opsin [Xenopus laevis]
NP_990821.1 green-sensitive opsin [Gallus gallus]
NP_001082611.1 blue-sensitive[Xenopus laevis]
NP_990848.1 blue-sensitive opsin [Gallus gallus]

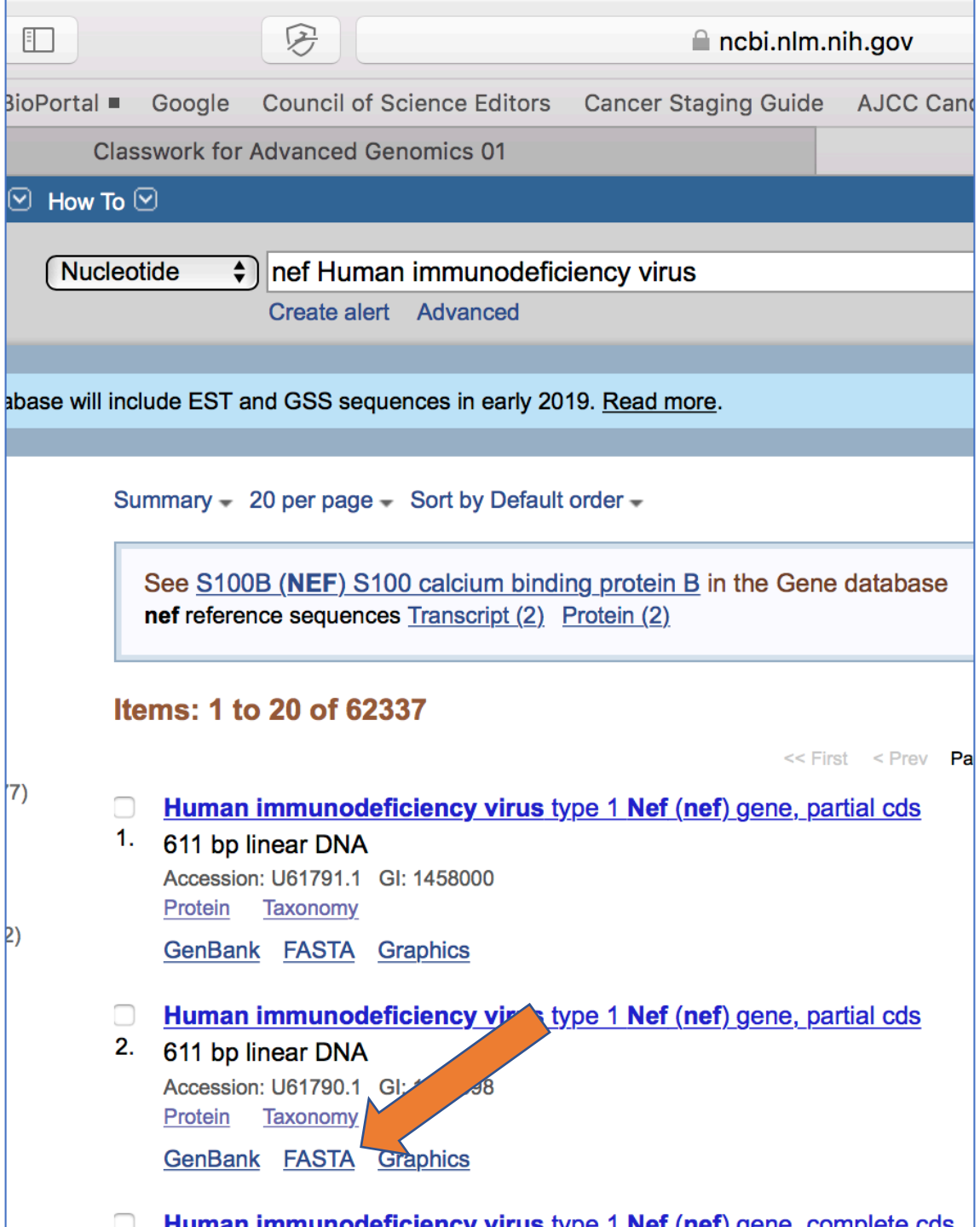

Unipro UGENE File Actions Settings Tools Window Help
38%
-* UGENE
Project
Start Page
Opsin_aaSequences_PNJboots [tr] Tree
Opsin_aaSequences [tr] Tree
Searc...
Objects
Bookmarks
NP_000504.1 green (medium-wave) sensitive opsin 1 [Homo sapiens]
NP_001270138.1 opsin 1 long-wave-sensitive [Macaca fascicularis]
NP_064445.2 long-wave-sensitive opsin 1 [Homo sapiens]
NP_001084114.1 red-sensitive opsin [Xenopus laevis]
NP_001184001.1 long-wave-sensitive opsin 1 [Canis lupus familiaris]
NP_776991.1 long-wave-sensitive opsin 1 [Bos taurus]
NP_990769.1 violet-sensitive opsin [Gallus gallus]
XP_539386.2 short-wave-sensitive opsin 1 [Canis lupus familiaris]
NP_776992.1 short-wave-sensitive opsin 1 [Bos taurus]
NP_001699.1 short-wave-sensitive opsin 1 [Homo sapiens]
NP_001270133.1 short-wave-sensitive opsin 1 [Macaca fascicularis]
NP_001079121.1 violet-sensitive opsin [Xenopus laevis]
NP_990821.1 green-sensitive opsin [Gallus gallus]
NP_001082611.1 blue-sensitive[Xenopus laevis]
NP_990848.1 blue-sensitive opsin [Gallus gallus]


Unipro UGENE File Actions Settings Tools Window Help
-* UGENE
Project
Start Page
Opsin_aaSequences_PNJboots [tr] Tree
Opsin_aaSequences [tr] Tree
Searc...
Objects
Bookmarks
0.021
0.027
0.034
0.055
0.004
0.006
0.013
0.024
0.023
0.051
0.041
0.022
0.031
0.055
0.517
0.149
0.272
0.192
0.17
0.172
0.055
0.387
0.022
0.047
0.148
0.113
0.019
NP_001270133.1 short-wave-sensitive opsin 1 [Macaca fascicularis]
XP_539386.2 short-wave-sensitive opsin 1 [Canis lupus familiaris]
NP_776992.1 short-wave-sensitive opsin 1 [Bos taurus]
NP_000504.1 medium-wave-sensitive opsin 1 [Homo sapiens]
NP_001270138.1 opsin 1 long-wave-sensitive [Macaca fascicularis]
NP_064445.2 long-wave-sensitive opsin 1 [Homo sapiens]
NP_001184001.1 long-wave-sensitive opsin 1 [Canis lupus familiaris]
NP_776991.1 long-wave-sensitive opsin 1 [Bos taurus]
NP_001084114.1 red-sensitive opsin [Xenopus laevis]
NP_001082611.1 blue-sensitive opsin-like S homeolog [Xenopus laevis]
NP_990848.1 blue-sensitive opsin [Gallus gallus]
NP_990821.1 green-sensitive opsin [Gallus gallus]
NP_001079121.1 violet-sensitive opsin [Xenopus laevis]
NP_990769.1 violet-sensitive opsin [Gallus gallus]
NP_001699.1 short-wave-sensitive opsin 1 [Homo sapiens]


Unipro UGENE File Actions Settings Tools Window Help
-* UGENE
Project
Start Page
Opsin_aaSequences_PNJboots [tr] Tree
Opsin_aaSequences [tr] Tree
Searc...
Objects
Bookmarks
0.021
0.027
0.034
0.055
0.006
0.004
0.013
0.024
0.023
0.051
0.022
0.041
0.031
0.055
0.517
0.149
0.272
0.17
0.192
0.172
0.022
0.055
0.387
0.047
0.148
0.113
0.019
NP_001270133.1 short-wave-sensitive opsin 1 [Macaca fascicularis]
XP_539386.2 short-wave-sensitive opsin 1 [Canis lupus familiaris]
NP_776992.1 short-wave-sensitive opsin 1 [Bos taurus]
NP_000504.1 medium-wave-sensitive opsin 1 [Homo sapiens]
NP_001270138.1 opsin 1 long-wave-sensitive [Macaca fascicularis]
NP_064445.2 long-wave-sensitive opsin 1 [Homo sapiens]
NP_001184001.1 long-wave-sensitive opsin 1 [Canis lupus familiaris]
NP_776991.1 long-wave-sensitive opsin 1 [Bos taurus]
NP_001084114.1 red-sensitive opsin [Xenopus laevis]
NP_001082611.1 blue-sensitive opsin-like S homeolog [Xenopus laevis]
NP_990848.1 blue-sensitive opsin [Gallus gallus]
NP_990821.1 green-sensitive opsin [Gallus gallus]
NP_001079121.1 violet-sensitive opsin [Xenopus laevis]
NP_990769.1 violet-sensitive opsin [Gallus gallus]
NP_001699.1 short-wave-sensitive opsin 1 [Homo sapiens]


0.232
0.241
0.406
0.095
0.345
0.154
0.183
NP_057849.4 Gag-Pol [HIV1GP1-Human immunodeficiency virus 1]
AEK79593.1 gag protein [Simian immunodeficiency virus. chimpanzees]
NP_663784.1 gag-pol fusion polyprotein [Human immunodeficiency virus 2]
NP_687035.1 Gag-Pol [African green monkey-Simian immunodeficiency virus]
Gag-Pol Cercocebus agilis (captive agile mangabey) GenBank: HM803690.1

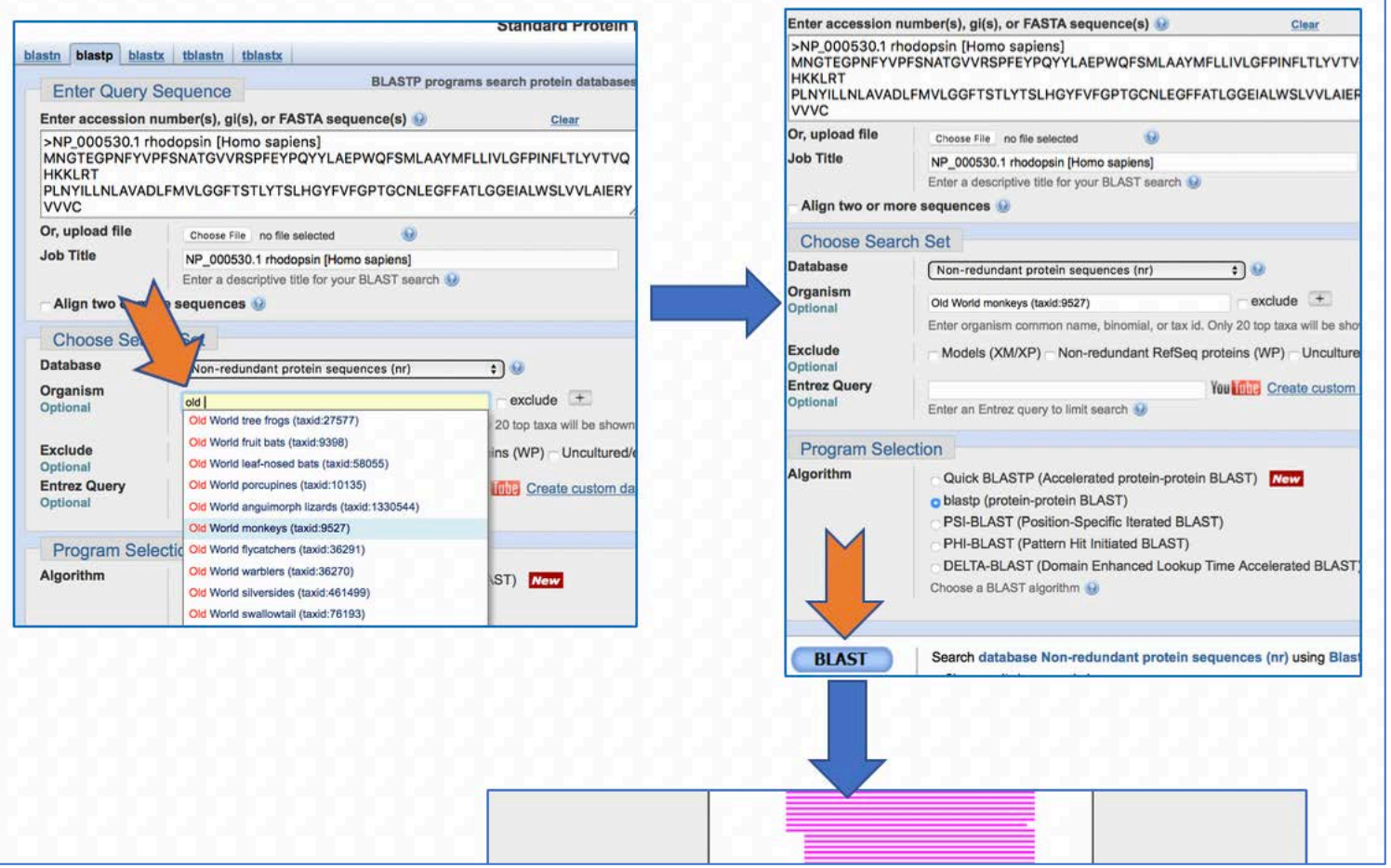

nef, partial [chimpanzee SIV]
Nef [HIV 2]
nef protein - Sooty Mangabey
nef protein [Mandrill -SIV-mnd 2]
Nef [HIV 1]
nef protein, partial [chimpanzee SIV]
100
100
43
100
100
67
65
100
100
100

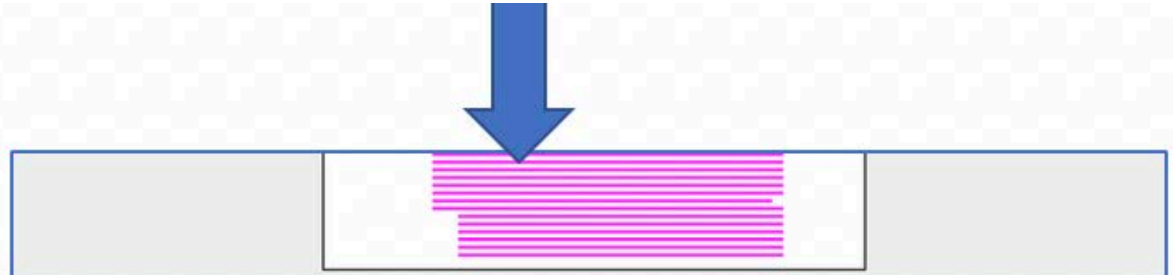

ncbi.nlm.nih.gov
BioPortal
Google
Council of Science Editors
Cancer Staging Guide
AJCC Canc
Classwork for Advanced Genomics 01
How To
Nucleotide
nef Human immunodeficiency virus
Create alert
Advanced
abase will include EST and GSS sequences in early 2019. Read more.
Summary
20 per page
Sort by Default order
See S100B (NEF) S100 calcium binding protein B in the Gene database
nef reference sequences Transcript (2) Protein (2)
Items: 1 to 20 of 62337
<< First
< Prev
Pa
Human immunodeficiency virus type 1 Nef (nef) gene, partial cds
1.
611 bp linear DNA
Accession: U61791.1 GI: 1458000
Protein
Taxonomy
GenBank
FASTA
Graphics
Human immunodeficiency vir type 1 Nef (nef) gene, partial cds
2.
611 bp linear DNA
Accession: U61790.1 GI:
Protein
Taxonomy
GenBank
FASTA
Graphics


ncbi.nlm.nih.gov
PanCancerAtlas cBioPortal
Google
Council of Science Editors
Cancer Staging Guide
AJCC
Classwork for Advanced Genomics 01
Humar
NCBI
Resources
How To
Nucleotide
Nucleotide
Advanced
The Nucleotide database will include EST and GSS sequences in early 2019. Read more.
FASTA
Human immunodeficiency virus type 1 Nef (nef) gene, partial c
GenBank: U61791.1
GenBank
Graphics
>U61791.1 Human immunodeficiency virus type 1 Nef (nef) gene, partial cds
ATGGGTGGCAAGTGGTCAAAATGTAGTGTGGTTGGGTGGCCTACTGTAAGGGAAAGAATGAGACGAGCGG
AGCCAGCAGCAGATGGGGTGGGAGCAGTATCTCGAGACCTGGAAAAACATGGAGCAATCACAAGTAGCAA
TACAGCAGCTAACAATGCTGATTGTGCCTGGCTAGAAGCACAAGAGGAGGAGGAGGTGGGTTTTCCAGTC
AGACCTCAGGTACCTTTAAGACCTATGACTTACAAGGGAGCTTTAGATCTTAGCCACTTTTTAAAAGAAA
AGGGGGGACTGGAAGGGCTAATTTACTCCCGAAAAAGACAAGAGATCCTTGATCTGTGGGTCTACCACAC
ACAAGGCTACTTCCCTGATTGGCAGAACTACACACCAGGGCCAGGGGTCAGATATCCACTGACCCTTGGA
TGGTGCTTCAAGCTAGTACCAGTTGATCCAGAGAAGGTAGAAGAGGCCAGTGAAGGAGAGAACAACAGCT
TGTTACACCCTACGAGCCTGCATGGGATGGAGGACCCGGAGAGAGAAGTGTTAGAGTGGAGGTTTGACAG
CCGCCTAGCATTTCATCACATGGCCCGAGAGCTGCATACAAGGACTGCTGA

>NP_000530.1 rhodopsin [Homo sapiens]
MNGTEGPNFYVPFSNATGVVRSPFEYPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLYVTVQHKKLRT
PLNYILLNLAVADLFMVLGGFTSTLYTSLHGYFVFGPTGCNLEGFFATLGGEIALWSLVVLAIERYVVVC
KPMSNFRFGENHAIMGVAFTWVMALACAAPPLAGWSRYIPEGLQCSCGIDYYTLKPEVNNESFVIYMFVV
HFTIPMIIIFFCYGQLVFTVKEAAAQQQESATTQKAEKEVTRMVIIMVIAFLICWVPYASVAFYIFTHQG
SNFGPIFMTIPAFFAKSAAIYNPVIYIMMNKQFRNCMLTTICCGKNPLGDDEASATVSKTETSQVAPA

Go to https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins





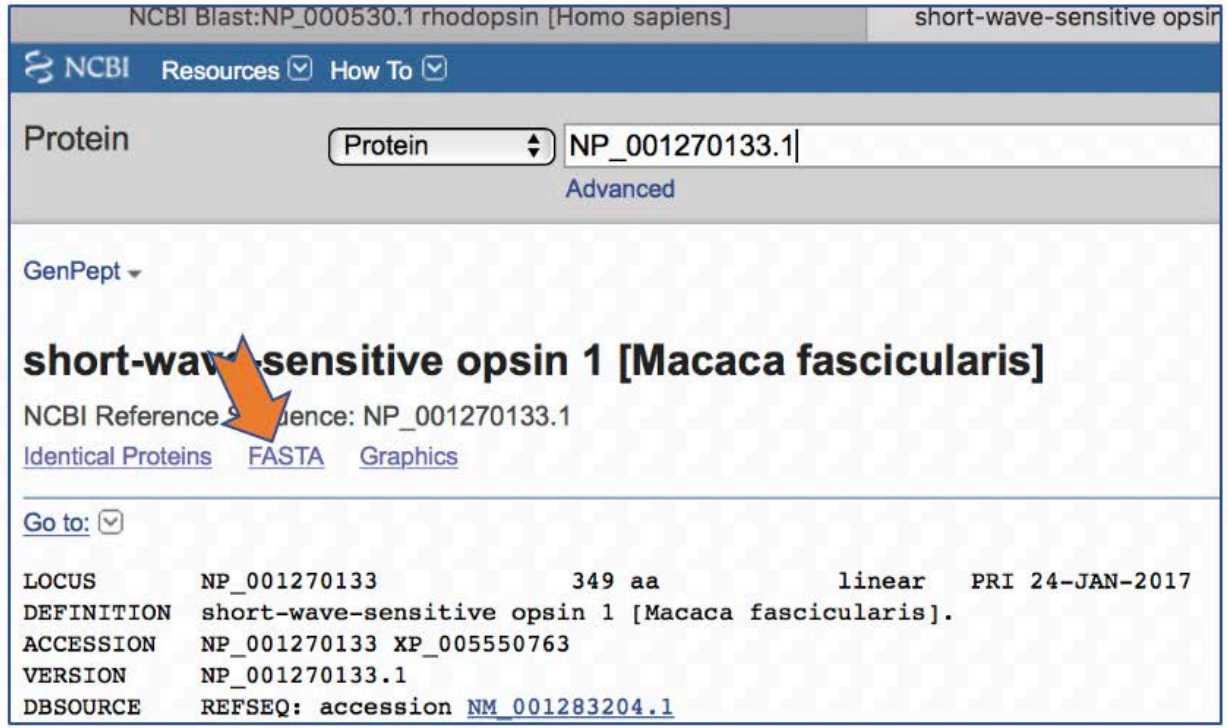
scriptions

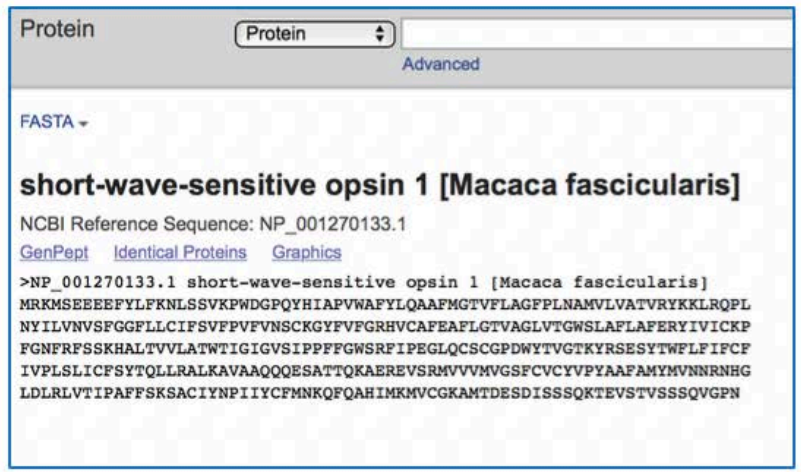
Sequences producing significant alignments:

Select: All None Selected:2

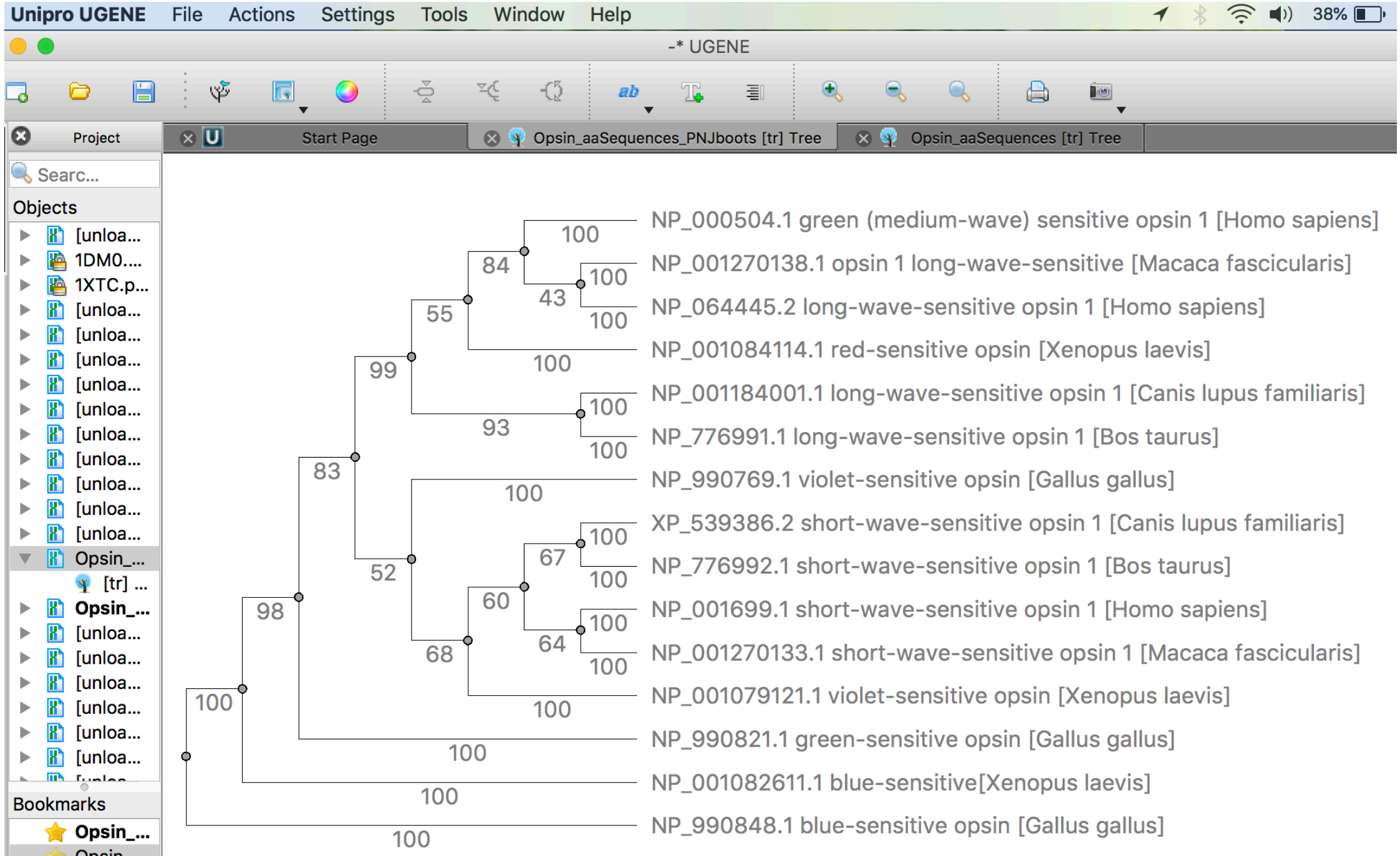
Alignments Download GenPept Graphics Distance tree of results Multiple alignment

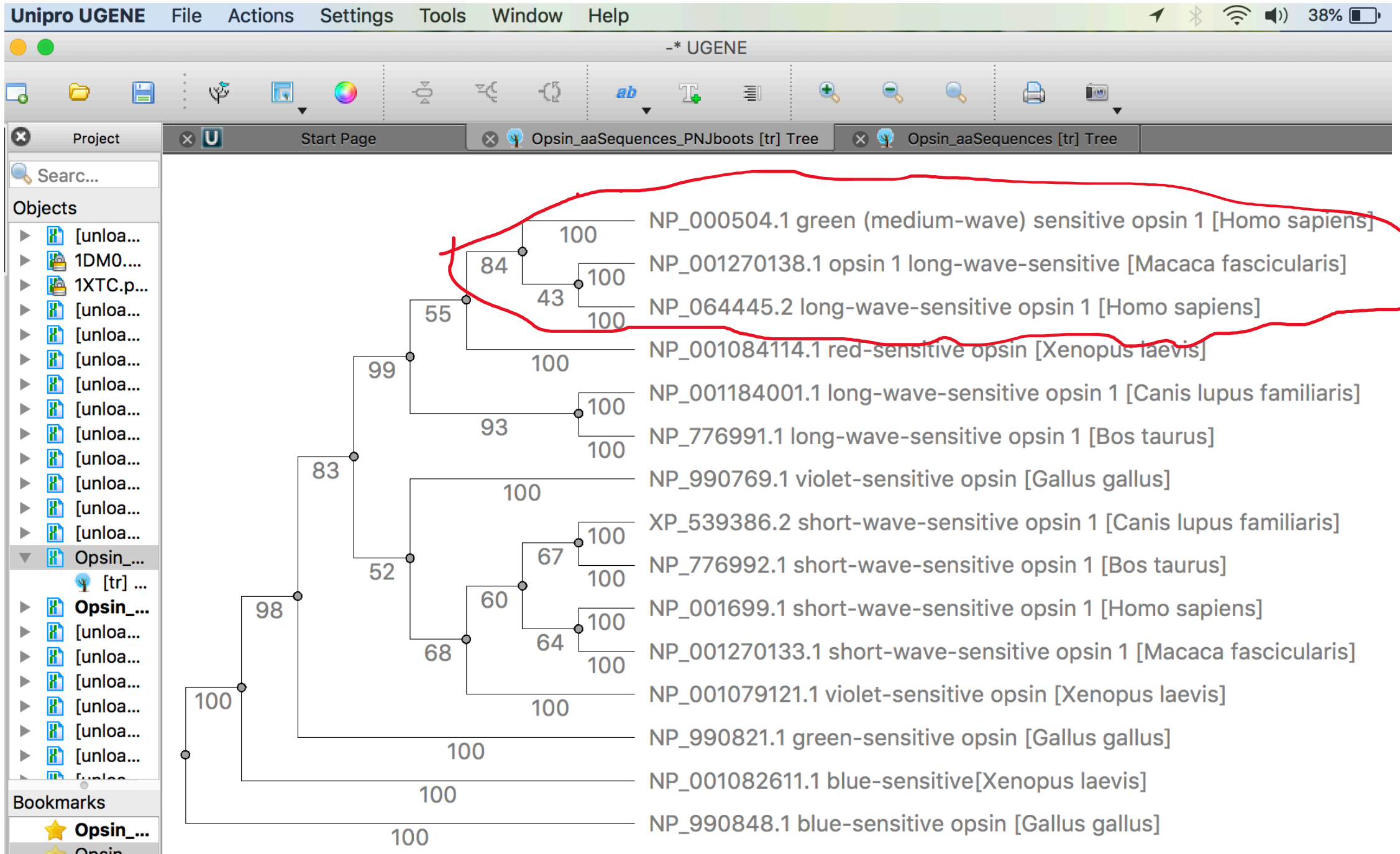
	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	PREDICTED: rhodopsin [Rhinopithecus roxellana]	710	710	100%	0.0	99%	XP_010379807.1
☐	PREDICTED: rhodopsin [Chlorocebus sabaeus]	709	709	100%	0.0	99%	XP_007983569.1
☐	PREDICTED: rhodopsin [Macaca mulatta]	707	707	100%	0.0	98%	XP_001094250.1
☐	rhodopsin [Macaca fascicularis]	705	705	100%	0.0	98%	NP_001270289.1
☐	short-wave-sensitive opsin 1 [Piliocolobus tephrosceles]	315	315	98%	5e-105	45%	XP_023079306.1
☐	PREDICTED: short-wave-sensitive opsin 1 [Rhinopithecus roxellana]	315	315	98%	7e-105	45%	XP_010377042.1
☐	PREDICTED: short-wave-sensitive opsin 1 [Chlorocebus sabaeus]	314	314	98%	1e-104	45%	XP_007981042.1
☐	PREDICTED: short-wave-sensitive opsin 1 [Mandrillus leucophaeus]	312	312	98%	7e-104	45%	XP_011851295.1
☐	PREDICTED: short-wave-sensitive opsin 1 [Macaca mulatta]	311	311	98%	1e-103	45%	XP_001091869.1
☑	short-wave-sensitive opsin 1 [Macaca fascicularis]	311	311	98%	2e-103	45%	NP_001270133.1
☐	short-wave-sensitive opsin 1 [Papio anubis]	310	310	98%	6e-103	45%	XP_003896610.1
☐	short-wave-sensitive opsin 1 isoform X1 [Macaca nemestrina]	295	295	89%	1e-97	46%	XP_024646465.1
☐	green opsin [Papio anubis]	286	286	97%	2e-93	43%	NP_001162268.1
☐	PREDICTED: long-wave-sensitive opsin 1 [Chlorocebus sabaeus]	286	286	97%	4e-93	43%	XP_007991323.1
☐	PREDICTED: long-wave-sensitive opsin 1 [Mandrillus leucophaeus]	286	286	97%	4e-93	43%	XP_011857677.1
☐	LW/MW hybrid opsin [Macaca fascicularis]	285	285	97%	5e-93	44%	AAD41526.1
☐	long-wave-sensitive opsin 1-like [Theropithecus gelada]	285	285	97%	1e-92	43%	XP_025228546.1
☐	PREDICTED: medium-wave-sensitive opsin 1 [Cercocebus atys]	284	284	97%	1e-92	44%	XP_011943704.1
☑	opsin 1 (cone pigments), long-wave-sensitive [Macaca fascicularis]	283	283	97%	2e-92	44%	NP_001270138.1
☐	green opsin (predicted) [Papio anubis]	283	283	97%	2e-92	44%	ABX10983.1

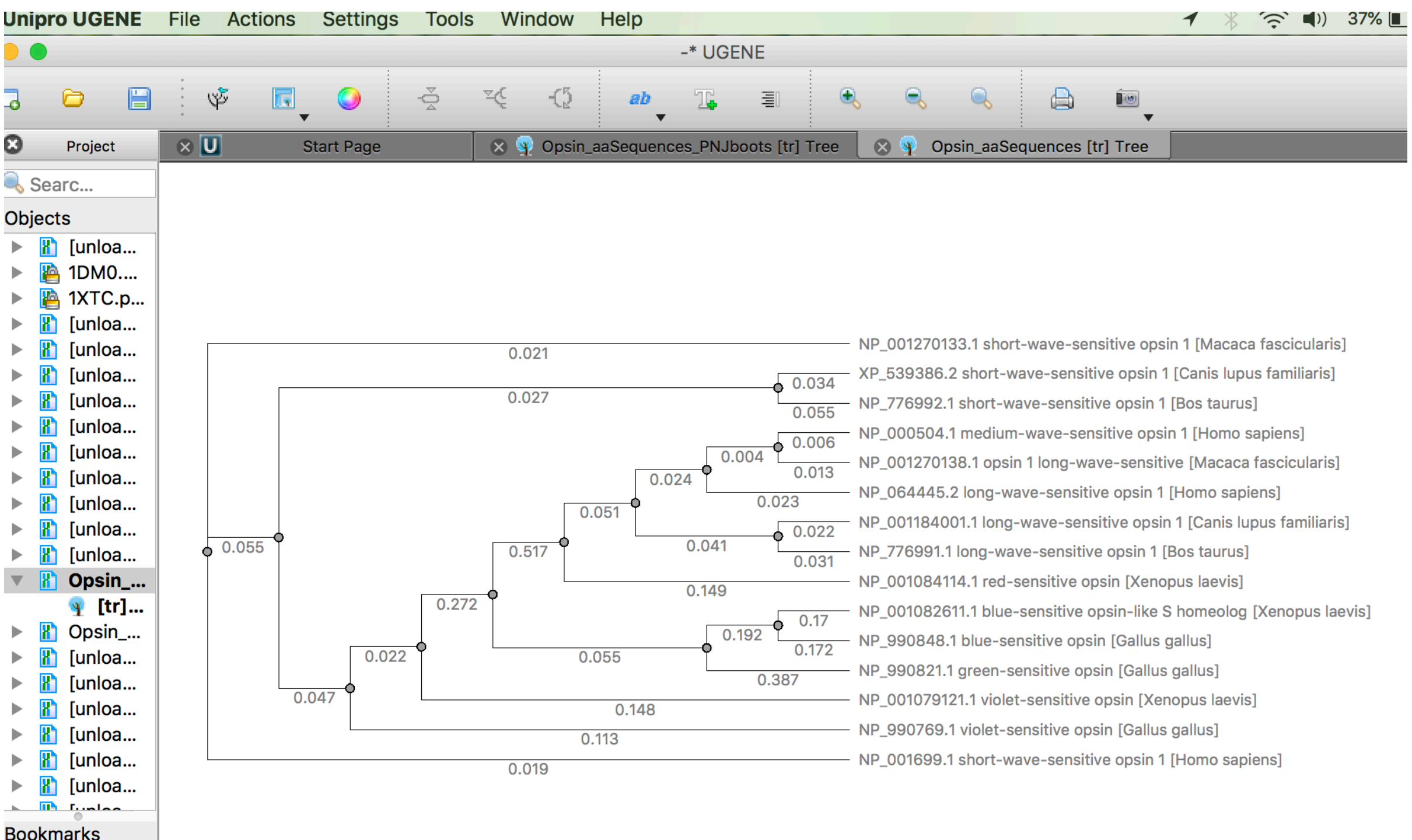

NCBI Blast:NP_000530.1 rhodopsin [Homo sapiens]
short-wave-sensitive opsir
NCBI
Resources
How To
Protein
Protein
NP_001270133.1
Advanced
GenPept
short-wave-sensitive opsin 1 [Macaca fascicularis]
NCBI Reference Sequence: NP_001270133.1
Identical Proteins
FASTA
Graphics
Go to:
LOCUS NP_001270133 349 aa linear PRI 24-JAN-2017
DEFINITION short-wave-sensitive opsin 1 [Macaca fascicularis].
ACCESSION NP_001270133 XP_005550763
VERSION NP_001270133.1
DBSOURCE REFSEQ: accession NM_001283204.1

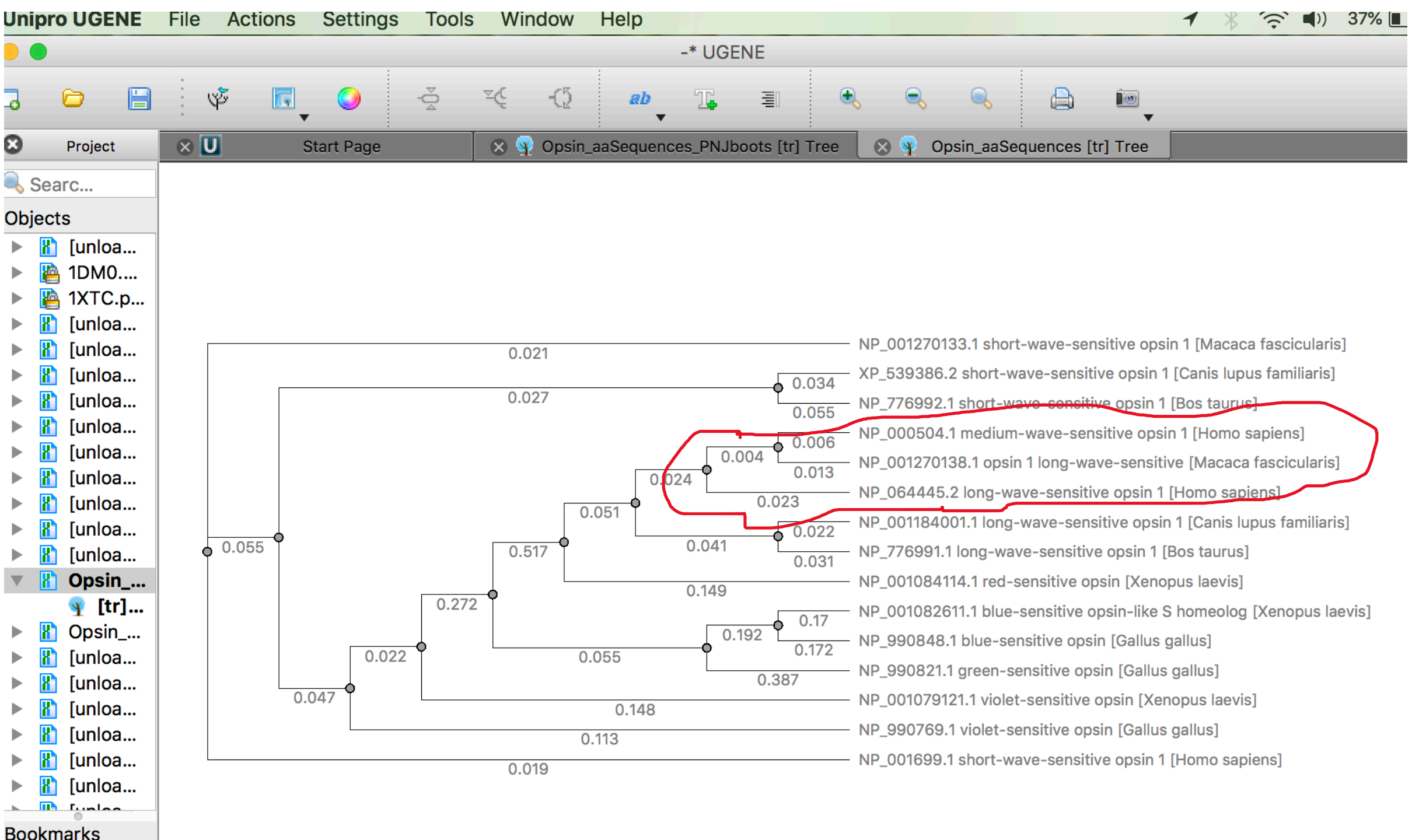

Protein
Protein
Advanced
FASTA
short-wave-sensitive opsin 1 [Macaca fascicularis]
NCBI Reference Sequence: NP_001270133.1
GenPept
Identical Proteins
Graphics
>NP_001270133.1 short-wave-sensitive opsin 1 [Macaca fascicularis]
MRKMSEEEEFYLFKNLSSVKPWDGPQYHIAPVWAFYLQAAFMGTVFLAGFPLNAMVLVATVRYKKLRQPL
NYILVNVSFGGFLLCIFSVFPVFVNSCKGYFVFGRHVCAFEAFLGTVAGLVTGWSLAFLAFERYIVICKP
FGNFRFSSKHALTVVLATWTIGIGVSIPPFFGWSRFIPEGLQCSCGPDWYTVGTKYRSESYTWFLFIFCF
IVPLSLICFSYTQLLRALKAVAAQQQESATTQKAEREVSRMVVVMVGSFCVCYVPYAAFAMYMVNNRNHG
LDLRLVTIPAFFSKSACIYNPIIYCFMNKQFQAHIMKMVCGKAMTDESDISSSQKTEVSTVSSSQVGPN

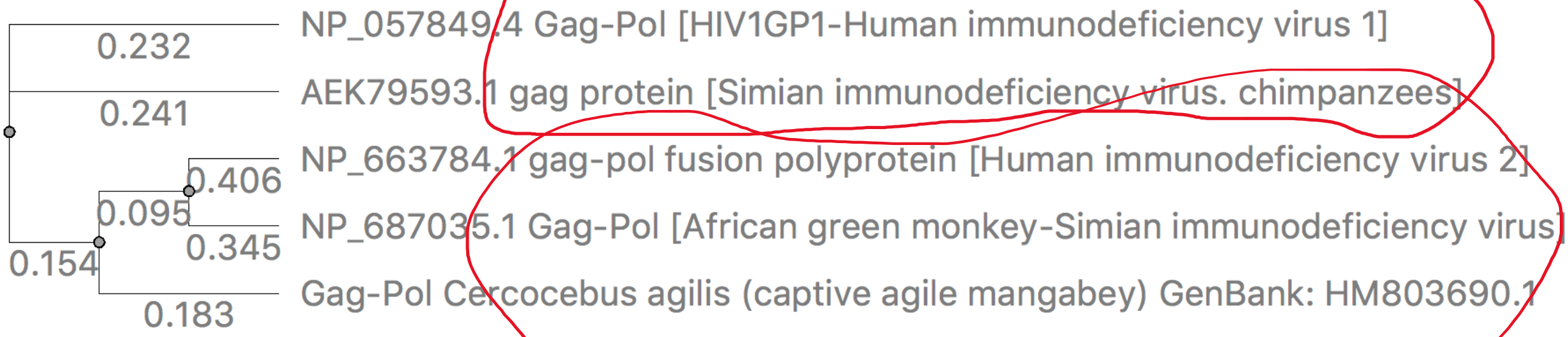
Phylogenetics Assignments

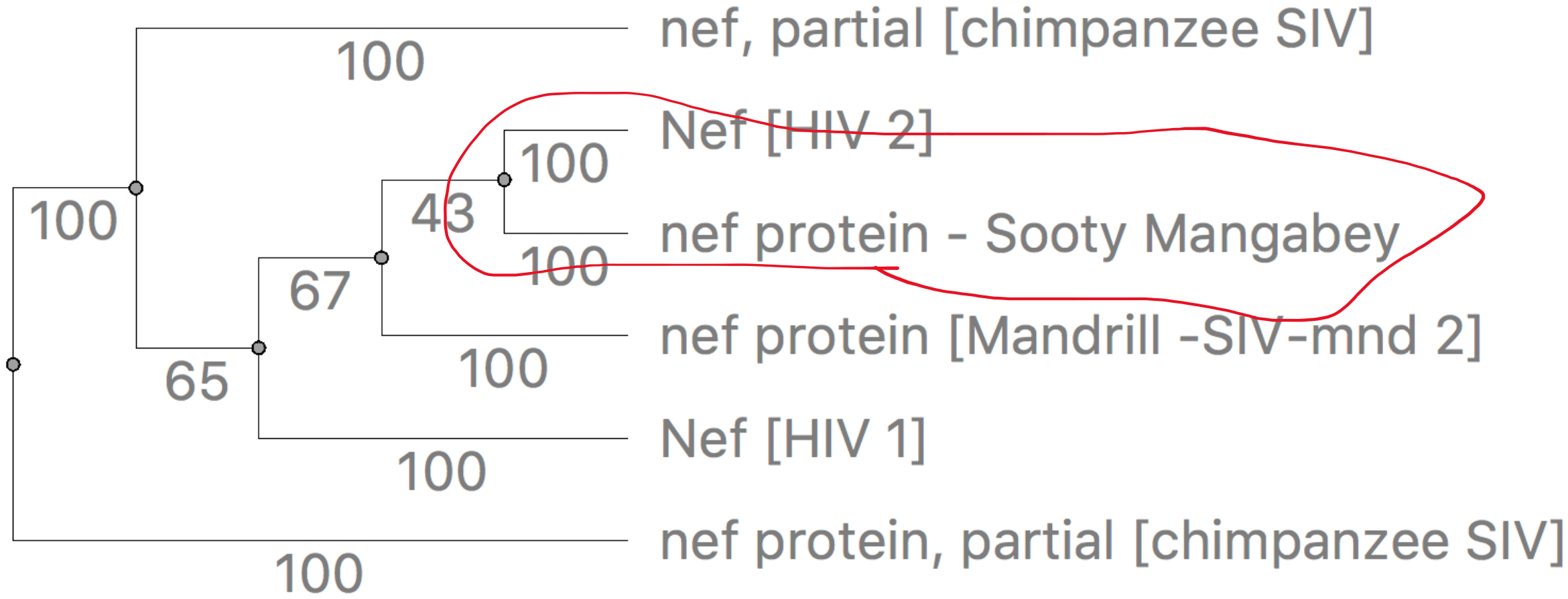

Unipro UGENE File Actions Settings Tools Window Help
38%
-* UGENE
Project
Start Page
Opsin_aaSequences_PNJboots [tr] Tree
Opsin_aaSequences [tr] Tree
Searc...
Objects
[unloa...
1DM0....
1XTC.p...
Opsin_...
[tr] ...
Bookmarks
Opsin_...
Opsin_...
100
84
100
43
100
55
100
99
100
93
100
83
100
52
100
67
100
60
100
64
100
68
100
98
100
100
100
100
NP_000504.1 green (medium-wave) sensitive opsin 1 [Homo sapiens]
NP_001270138.1 opsin 1 long-wave-sensitive [Macaca fascicularis]
NP_064445.2 long-wave-sensitive opsin 1 [Homo sapiens]
NP_001084114.1 red-sensitive opsin [Xenopus laevis]
NP_001184001.1 long-wave-sensitive opsin 1 [Canis lupus familiaris]
NP_776991.1 long-wave-sensitive opsin 1 [Bos taurus]
NP_990769.1 violet-sensitive opsin [Gallus gallus]
XP_539386.2 short-wave-sensitive opsin 1 [Canis lupus familiaris]
NP_776992.1 short-wave-sensitive opsin 1 [Bos taurus]
NP_001699.1 short-wave-sensitive opsin 1 [Homo sapiens]
NP_001270133.1 short-wave-sensitive opsin 1 [Macaca fascicularis]
NP_001079121.1 violet-sensitive opsin [Xenopus laevis]
NP_990821.1 green-sensitive opsin [Gallus gallus]
NP_001082611.1 blue-sensitive[Xenopus laevis]
NP_990848.1 blue-sensitive opsin [Gallus gallus]


Unipro UGENE File Actions Settings Tools Window Help
38%
-* UGENE
Project
Start Page
Opsin_aaSequences_PNJboots [tr] Tree
Opsin_aaSequences [tr] Tree
Searc...
Objects
Bookmarks
NP_000504.1 green (medium-wave) sensitive opsin 1 [Homo sapiens]
NP_001270138.1 opsin 1 long-wave-sensitive [Macaca fascicularis]
NP_064445.2 long-wave-sensitive opsin 1 [Homo sapiens]
NP_001084114.1 red-sensitive opsin [Xenopus laevis]
NP_001184001.1 long-wave-sensitive opsin 1 [Canis lupus familiaris]
NP_776991.1 long-wave-sensitive opsin 1 [Bos taurus]
NP_990769.1 violet-sensitive opsin [Gallus gallus]
XP_539386.2 short-wave-sensitive opsin 1 [Canis lupus familiaris]
NP_776992.1 short-wave-sensitive opsin 1 [Bos taurus]
NP_001699.1 short-wave-sensitive opsin 1 [Homo sapiens]
NP_001270133.1 short-wave-sensitive opsin 1 [Macaca fascicularis]
NP_001079121.1 violet-sensitive opsin [Xenopus laevis]
NP_990821.1 green-sensitive opsin [Gallus gallus]
NP_001082611.1 blue-sensitive[Xenopus laevis]
NP_990848.1 blue-sensitive opsin [Gallus gallus]


Unipro UGENE File Actions Settings Tools Window Help
-* UGENE
Project
Start Page
Opsin_aaSequences_PNJboots [tr] Tree
Opsin_aaSequences [tr] Tree
Searc...
Objects
Bookmarks
0.021
0.027
0.034
0.055
0.004
0.006
0.013
0.024
0.023
0.051
0.041
0.022
0.031
0.055
0.517
0.149
0.272
0.192
0.17
0.172
0.055
0.387
0.022
0.047
0.148
0.113
0.019
NP_001270133.1 short-wave-sensitive opsin 1 [Macaca fascicularis]
XP_539386.2 short-wave-sensitive opsin 1 [Canis lupus familiaris]
NP_776992.1 short-wave-sensitive opsin 1 [Bos taurus]
NP_000504.1 medium-wave-sensitive opsin 1 [Homo sapiens]
NP_001270138.1 opsin 1 long-wave-sensitive [Macaca fascicularis]
NP_064445.2 long-wave-sensitive opsin 1 [Homo sapiens]
NP_001184001.1 long-wave-sensitive opsin 1 [Canis lupus familiaris]
NP_776991.1 long-wave-sensitive opsin 1 [Bos taurus]
NP_001084114.1 red-sensitive opsin [Xenopus laevis]
NP_001082611.1 blue-sensitive opsin-like S homeolog [Xenopus laevis]
NP_990848.1 blue-sensitive opsin [Gallus gallus]
NP_990821.1 green-sensitive opsin [Gallus gallus]
NP_001079121.1 violet-sensitive opsin [Xenopus laevis]
NP_990769.1 violet-sensitive opsin [Gallus gallus]
NP_001699.1 short-wave-sensitive opsin 1 [Homo sapiens]


Unipro UGENE File Actions Settings Tools Window Help
-* UGENE
Project
Start Page
Opsin_aaSequences_PNJboots [tr] Tree
Opsin_aaSequences [tr] Tree
Searc...
Objects
Bookmarks
0.021
0.027
0.034
0.055
0.006
0.004
0.013
0.024
0.023
0.051
0.022
0.041
0.031
0.055
0.517
0.149
0.272
0.17
0.192
0.172
0.022
0.055
0.387
0.047
0.148
0.113
0.019
NP_001270133.1 short-wave-sensitive opsin 1 [Macaca fascicularis]
XP_539386.2 short-wave-sensitive opsin 1 [Canis lupus familiaris]
NP_776992.1 short-wave-sensitive opsin 1 [Bos taurus]
NP_000504.1 medium-wave-sensitive opsin 1 [Homo sapiens]
NP_001270138.1 opsin 1 long-wave-sensitive [Macaca fascicularis]
NP_064445.2 long-wave-sensitive opsin 1 [Homo sapiens]
NP_001184001.1 long-wave-sensitive opsin 1 [Canis lupus familiaris]
NP_776991.1 long-wave-sensitive opsin 1 [Bos taurus]
NP_001084114.1 red-sensitive opsin [Xenopus laevis]
NP_001082611.1 blue-sensitive opsin-like S homeolog [Xenopus laevis]
NP_990848.1 blue-sensitive opsin [Gallus gallus]
NP_990821.1 green-sensitive opsin [Gallus gallus]
NP_001079121.1 violet-sensitive opsin [Xenopus laevis]
NP_990769.1 violet-sensitive opsin [Gallus gallus]
NP_001699.1 short-wave-sensitive opsin 1 [Homo sapiens]


0.232
0.241
0.406
0.095
0.345
0.154
0.183
NP_057849.4 Gag-Pol [HIV1GP1-Human immunodeficiency virus 1]
AEK79593.1 gag protein [Simian immunodeficiency virus. chimpanzees]
NP_663784.1 gag-pol fusion polyprotein [Human immunodeficiency virus 2]
NP_687035.1 Gag-Pol [African green monkey-Simian immunodeficiency virus]
Gag-Pol Cercocebus agilis (captive agile mangabey) GenBank: HM803690.1

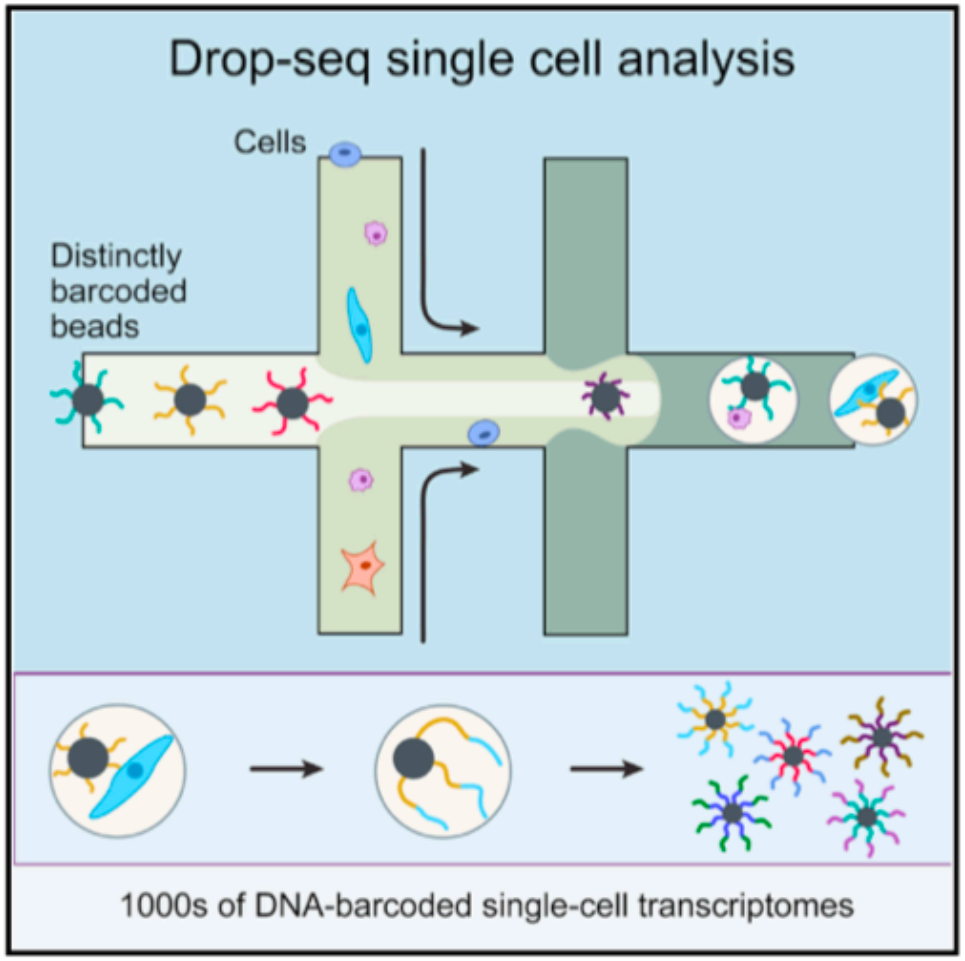

nef, partial [chimpanzee SIV]
Nef [HIV 2]
nef protein - Sooty Mangabey
nef protein [Mandrill -SIV-mnd 2]
Nef [HIV 1]
nef protein, partial [chimpanzee SIV]
100
100
43
100
100
67
65
100
100
100

Resource

Cell

Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets

Graphical Abstract



Authors

Evan Z. Macosko, Anindita Basu, ..., Aviv Regev, Steven A. McCarroll

Correspondence

emacosko@genetics.med.harvard.edu (E.Z.M.), mccarroll@genetics.med.harvard.edu (S.A.M.)

In Brief

Capturing single cells along with sets of uniquely barcoded primer beads together in tiny droplets enables large-scale, highly parallel single-cell transcriptomics. Applying this analysis to cells in mouse retinal tissue revealed transcriptionally distinct cell populations along with molecular markers of each type.

RESULTS

Drop-seq consists of the following steps (Figure 1A): (1) prepare a single-cell suspension from a tissue; (2) co-encapsulate each cell with a distinctly barcoded microparticle (bead) in a nanoliter-scale droplet; (3) lyse cells after they have been isolated in droplets; (4) capture a cell's mRNAs on its companion microparticle, forming STAMPs (single-cell transcriptomes attached to microparticles); (5) reverse-transcribe, amplify, and sequence thousands of STAMPs in one reaction; and (6) use the STAMP barcodes to infer each transcript's cell of origin.

A Split-Pool Synthesis Approach to Generate Large Numbers of Distinctly Barcoded Beads

To deliver large numbers of distinctly barcoded primer molecules into individual droplets, we use microparticles (beads). We synthesized oligonucleotide primers directly on beads (from 5′ to 3′, yielding free 3′ ends available for enzymatic priming). Each oligonucleotide is composed of four parts (Figure 1B): (1) a constant sequence (identical on all primers and beads) for use as a priming site for downstream PCR and sequencing; (2) a "cell barcode" (identical across all the primers on the surface of any one bead, but different from the cell barcodes on other beads); (3) a Unique Molecular Identifier (UMI) (different on each primer, to identify PCR duplicates) (Kivioja et al., 2012); and (4) an oligo-dT sequence for capturing polyadenylated mRNAs and priming reverse transcription.

To efficiently generate massive numbers of beads, each with a distinct barcode, we developed a "split-and-pool" DNA synthesis strategy (Figure 1C). A pool of millions of microparticles is divided into four equally sized groups; a different DNA base (A, G, C, or T) is then added to each. All microparticles are then re-pooled, mixed, and re-split at random into another four groups, and then a different DNA base (A, G, C, or T) is added to each of the four new groups. After 12 cycles of split-and-pool DNA synthesis, the primers on any given microparticle possess the same one of 4^{12} = 16,777,216 possible 12-bp barcodes, but different microparticles have different sequences (Figure 1C). The entire microparticle pool then undergoes eight rounds of degenerate oligonucleotide synthesis to generate the UMI on each oligo (Figure 1D); finally, an oligo-dT sequence (T30) is synthesized on the 3′ end of all oligos on all beads.

To confirm that we could distinguish RNAs based on attached barcodes, we reverse-transcribed a pool of synthetic RNAs onto 11 microparticles and sequenced the resulting cDNAs (Figure S1A and Supplemental Experimental Procedures); 11 microparticle barcodes each constituted 3.5%–14% of the resulting sequencing reads, whereas the next-most-abundant 12-mer constituted only 0.06% (Figure S1A). These results suggested that the microparticle-of-origin for most cDNAs can be recognized by sequencing. We also found that each bead contained more than 10^8 barcoded primer sites and that the sequence complexity of the barcodes approached theoretical limits (Figures S1B and S1C, Supplemental Experimental Procedures).

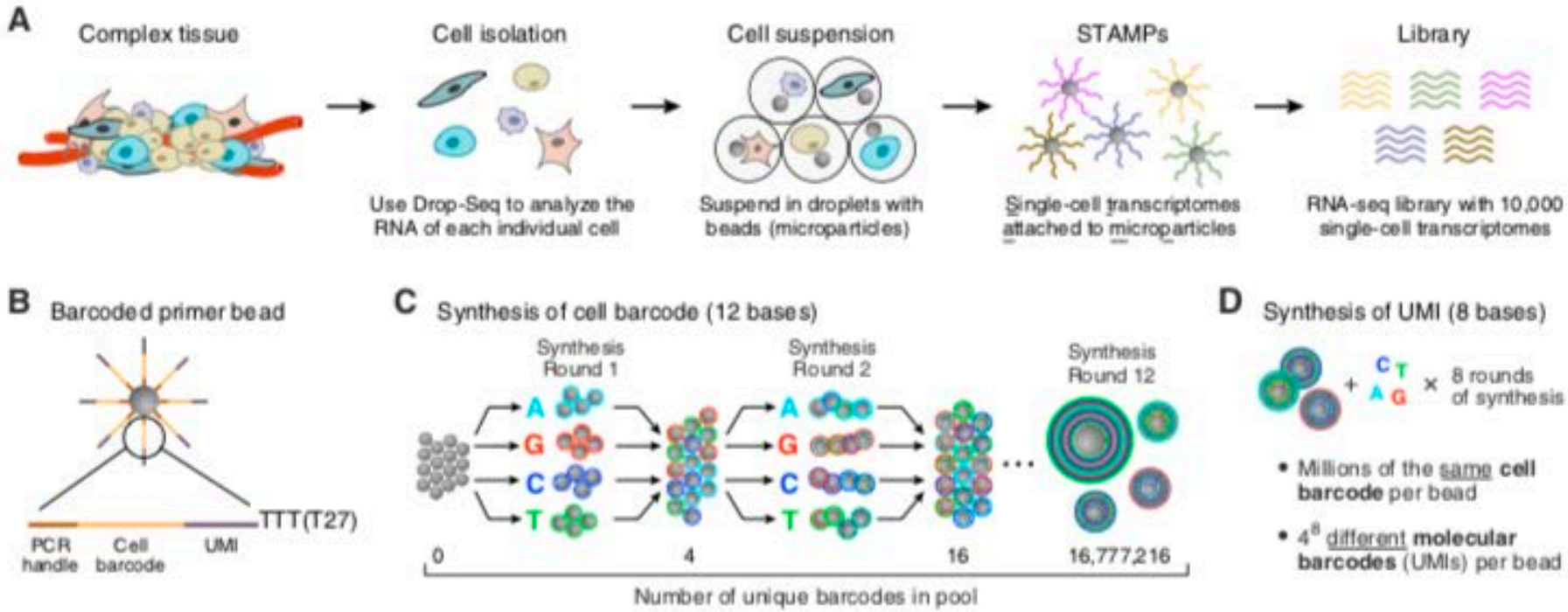


Figure 1. Molecular Barcoding of Cellular Transcriptomes in Droplets

(A) Drop-Seq barcoding schematic. A complex tissue is dissociated into individual cells, which are then encapsulated in droplets together with microparticles (gray circles) that deliver barcoded primers. Each cell is lysed within a droplet; its mRNAs bind to the primers on its companion microparticle. The mRNAs are reverse-transcribed into cDNAs, generating a set of beads called "single-cell transcriptomes attached to microparticles" (STAMPs). The barcoded STAMPs can then be amplified in pools for high-throughput mRNA-seq to analyze any desired number of individual cells.

(B) Sequence of primers on the microparticle. The primers on all beads contain a common sequence ("PCR handle") to enable PCR amplification after STAMP formation. Each microparticle contains more than 10^8 individual primers that share the same "cell barcode" (C) but have different unique molecular identifiers (UMIs), enabling mRNA transcripts to be digitally counted (D). A 30-bp oligo dT sequence is present at the end of all primer sequences for capture of mRNAs.

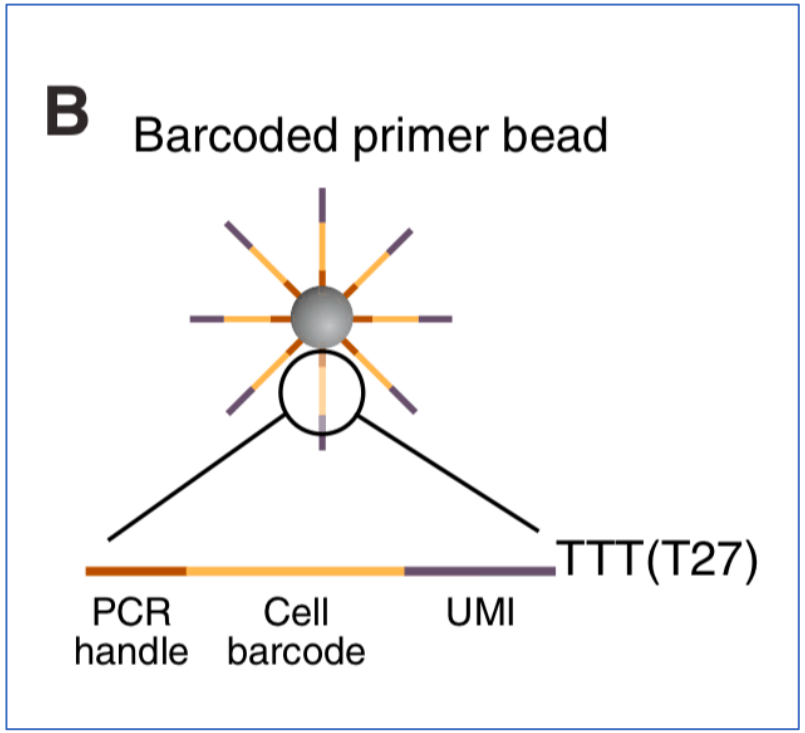
(C) Split-and-pool synthesis of the cell barcode. To generate the cell barcode, the pool of microparticles is repeatedly split into four equally sized oligonucleotide synthesis reactions, to which one of the four DNA bases is added, and then pooled together after each cycle, in a total of 12 split-pool cycles. The barcode synthesized on any individual bead reflects that bead's unique path through the series of synthesis reactions. The result is a pool of microparticles, each possessing one of 4^{12} (16,777,216) possible sequences on its entire complement of primers (see also Figure S1).

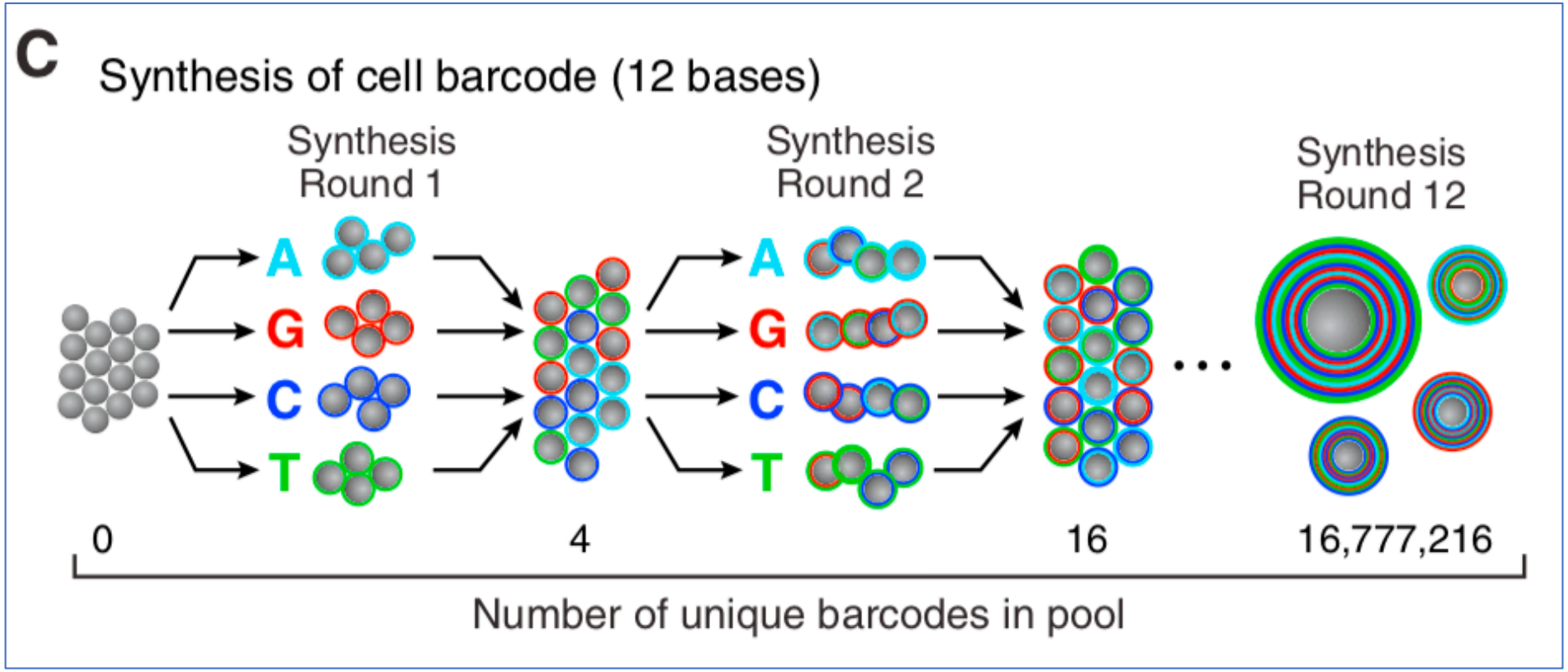
(D) Synthesis of a unique molecular identifier (UMI). Following the completion of the "split-and-pool" synthesis cycles, all microparticles are together subjected to eight rounds of degenerate synthesis with all four DNA bases available during each cycle, such that each individual primer receives one of 4^8 (65,536) possible sequences (UMIs).

Figure 1. Molecular Barcoding of Cellular Transcriptomes in Droplets

(A) Drop-Seq barcoding schematic. A complex tissue is dissociated into individual cells, which are then encapsulated in droplets together with microparticles (gray circles) that deliver barcoded primers. Each cell is lysed within a droplet; its mRNAs bind to the primers on its companion microparticle. The mRNAs are reverse-transcribed into cDNAs, generating a set of beads called "single-cell transcriptomes attached to microparticles" (STAMPs). The barcoded STAMPs can then be amplified in pools for high-throughput mRNA-seq to analyze any desired number of individual cells.

(B) Sequence of primers on the microparticle. The primers on all beads contain a common sequence ("PCR handle") to enable PCR amplification after STAMP formation. Each microparticle contains more than 10^8 individual primers that share the same "cell barcode" (C) but have different unique molecular identifiers (UMIs), enabling mRNA transcripts to be digitally counted (D). A 30-bp oligo dT sequence is present at the end of all primer sequences for capture of mRNAs.

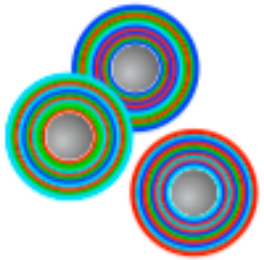




# possible bases	possible combination of 4 bases in a 12-base sequence		
4	16,777,216	unique cell-barcodes	

Primers on each bead can only **access one particular nucleotide**…and then all beads are **pooled together after each cycle**, in a total of 12 split-pool cycles.

D Synthesis of UMI (8 bases)



\+ C T A G × 8 rounds of synthesis

- Millions of the same **cell barcode** per bead
- 4^8 different **molecular barcodes** (UMIs) per bead

# possible bases	possible combination of 4 bases in an 8-base sequence	
4	65,536	unique molecular identifier

Every primer on each bead has **access to all 4 nucleotides** so each primer can incorporate one at random during each of 8 cycles……. each individual primer receives one of 48 (65,536) possible sequences (UMIs).

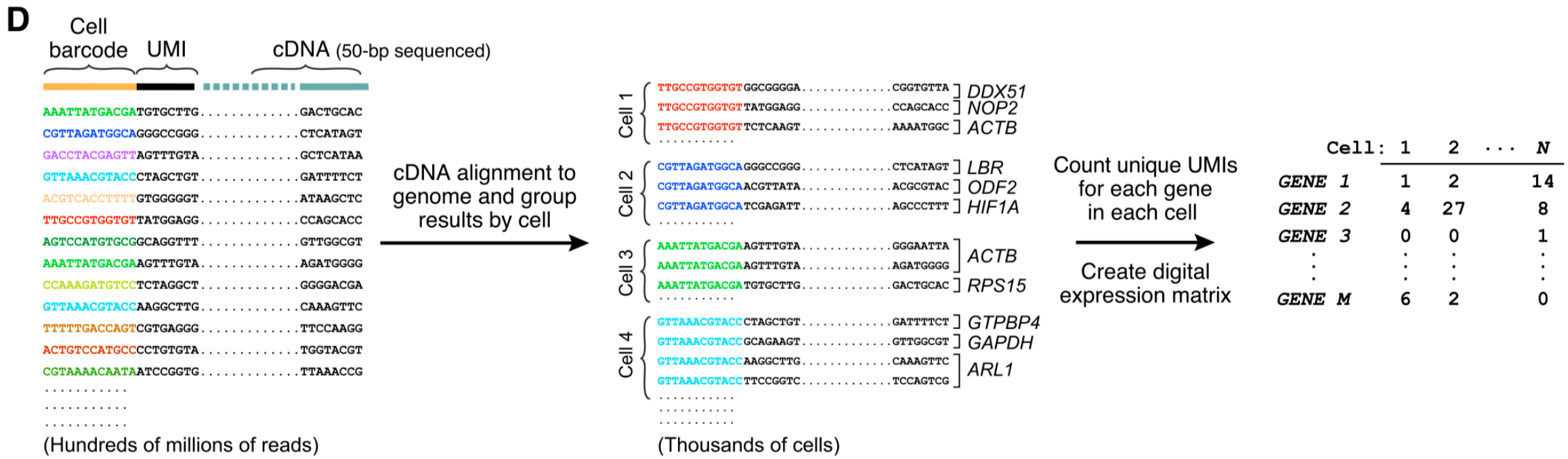


Figure 2. Extraction and Processing of Single-Cell Transcriptomes by Drop-Seq

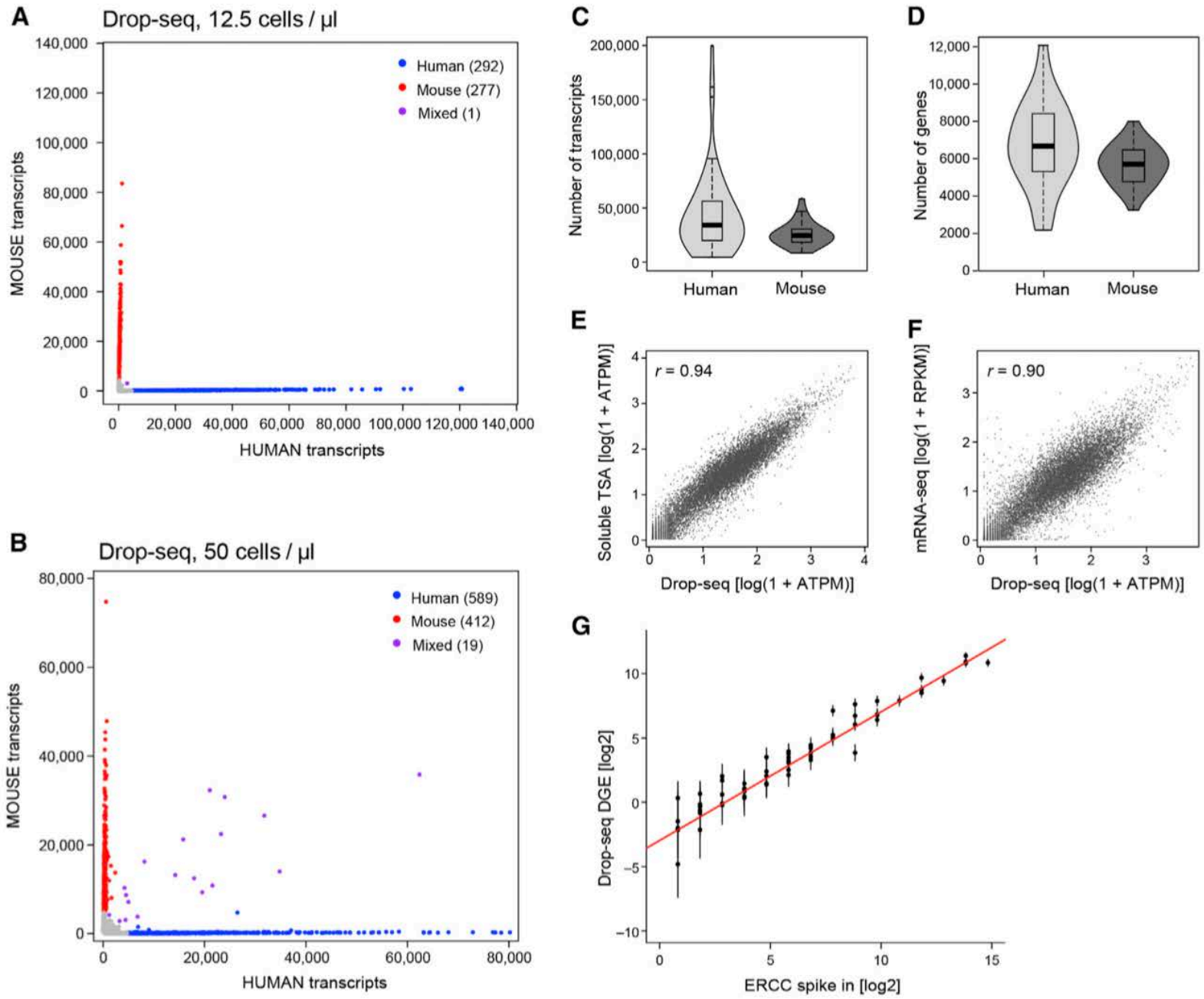
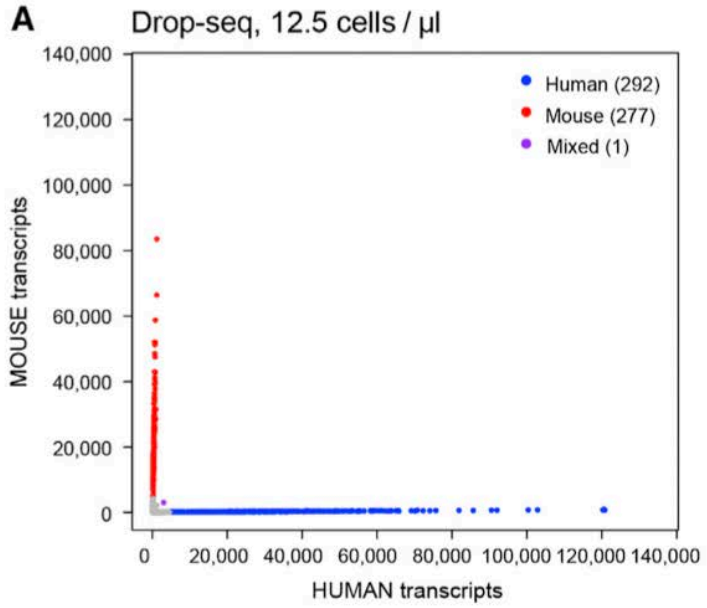
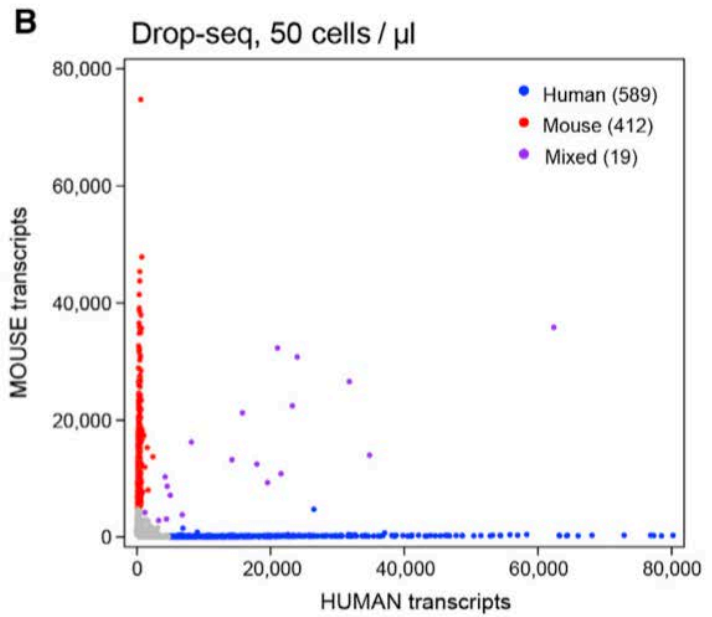


Figure 3. Critical Evaluation of Drop-Seq Using Species-Mixing Experiments





The Single-Cell Accuracy and Sensitivity of Drop-Seq Libraries

To measure the accuracy with which Drop-seq remembers the cell-of-origin of each mRNA, we analyzed mixtures of cultured human (HEK) and mouse (3T3) cells, scoring the numbers of human and mouse transcripts that associated with each cell barcode (Figures 3A, 3B, and S3A). We found that the individual STAMPs created by Drop-seq were highly organism-specific (Figures 3A and 3B), indicating high single-cell integrity of the libraries. At saturating levels of sequence coverage, we detected an average of 44,295 mRNA transcripts from 6,722 genes in HEK cells and 26,044 transcripts from 5,663 genes in 3T3 cells (Figures 3C and 3D).

Visualizing Data using t-SNE

Laurens van der Maaten LVDMAATEN@GMAIL.COM

- In this paper, we describe a way of *converting a high-dimensional data set into a matrix of pair- wise similarities* and we introduce a new technique, called “t-SNE”, for visualizing the resulting similarity data. t-SNE is capable of capturing much of the local structure of the high-dimensional data very well, while also revealing global structure such as the presence of clusters at several scales.

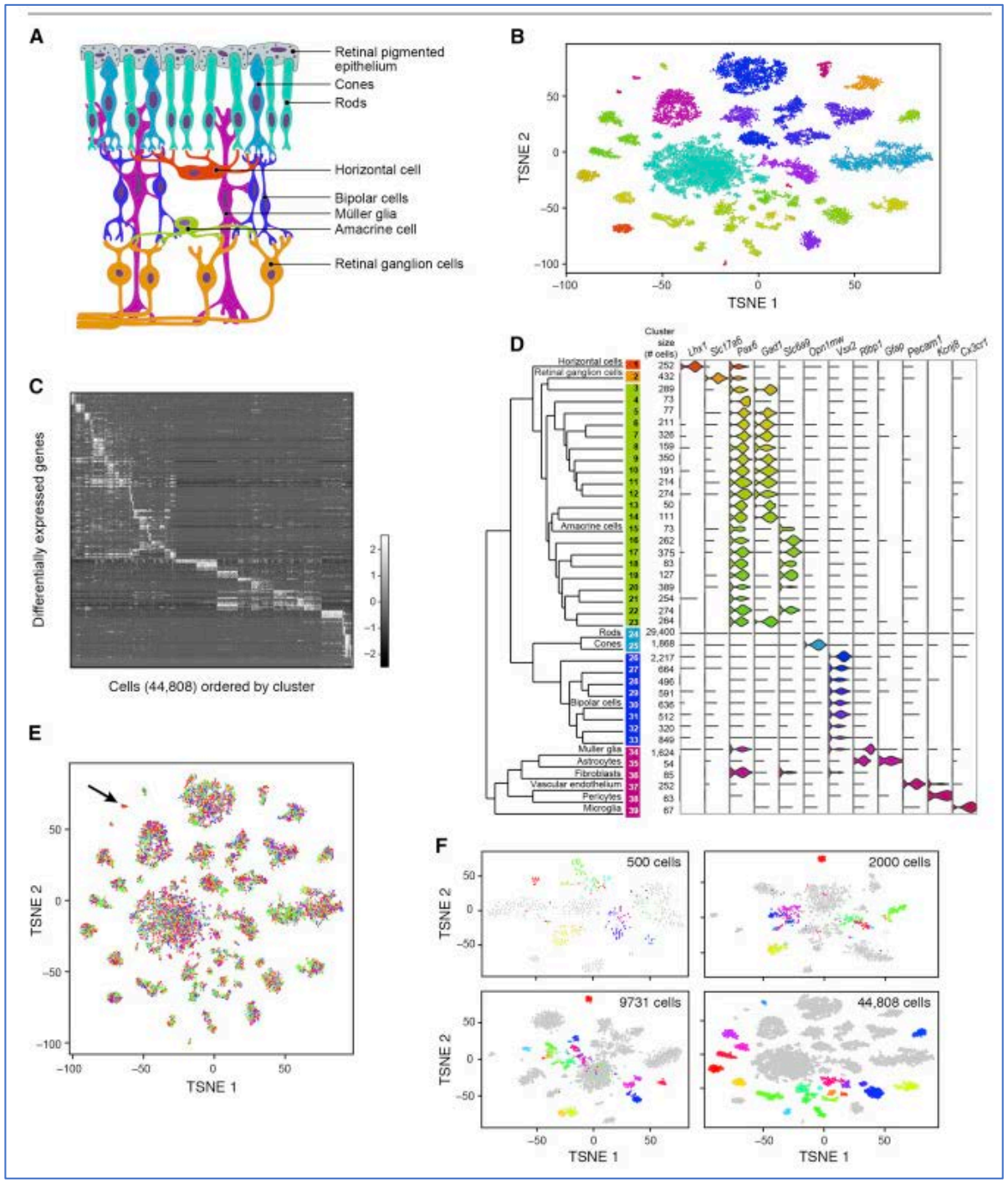
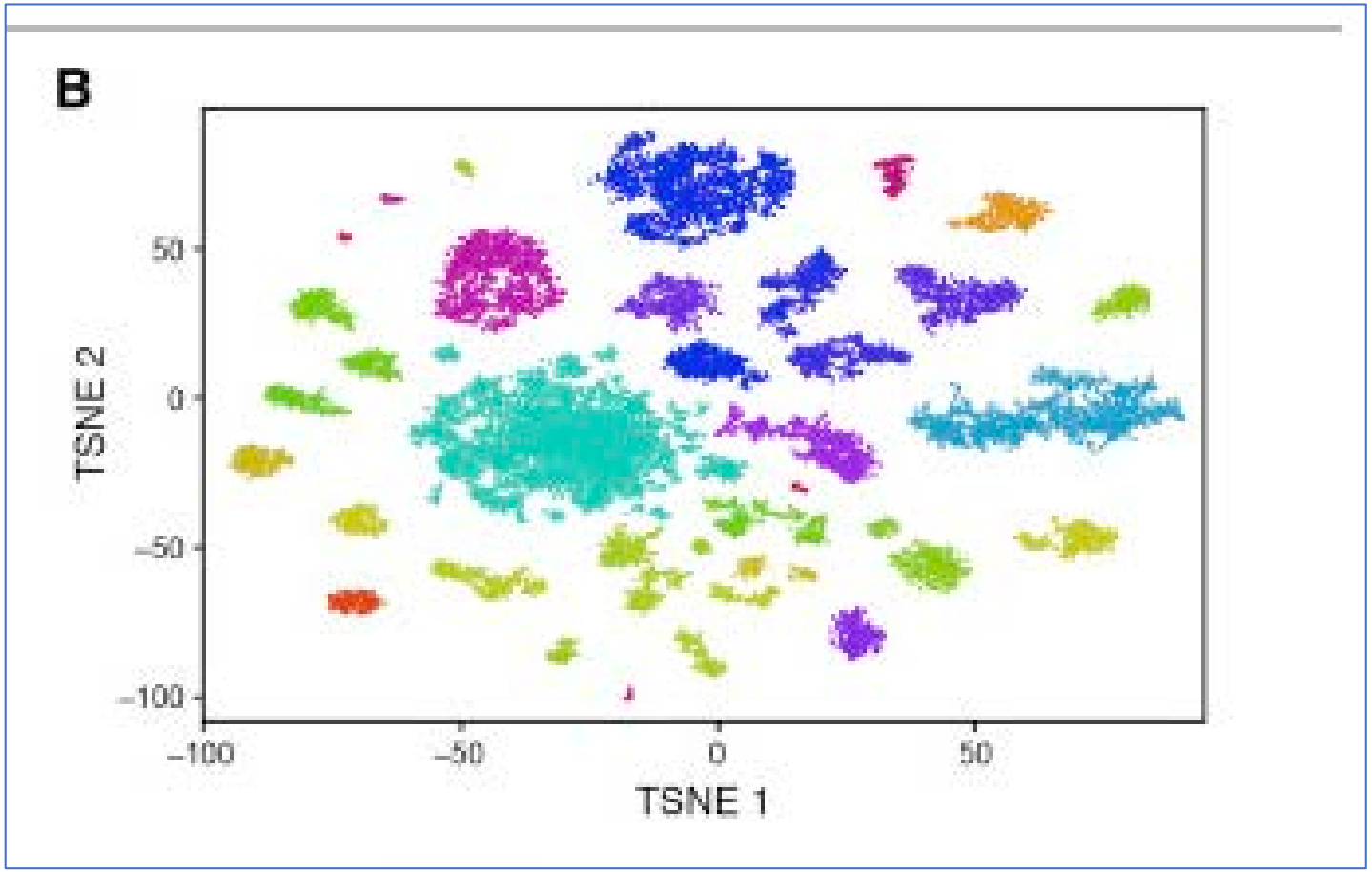


Figure 5. *Ab Initio* Reconstruction of Retinal Cell Types from 44,808 Single-Cell Transcription Profiles Prepared by Drop-Seq

(B) *Clustering of 44,808 Drop-seq single-cell expression profiles into 39 retinal cell populations*. The plot shows a two-dimensional representation (tSNE) of global gene expression relationships among 44,808 cells; clusters are colored by cell class, according to Figure 5A.

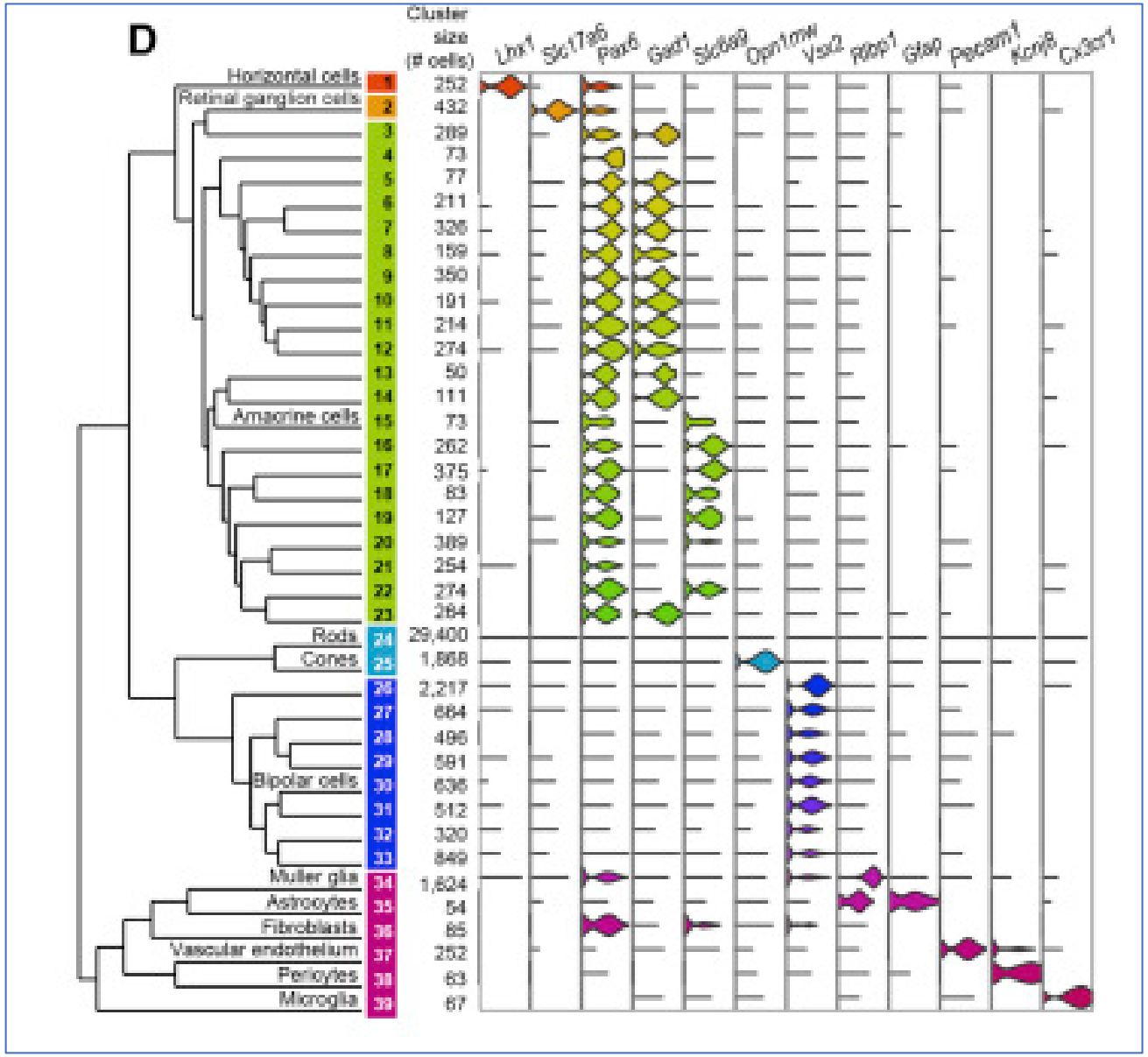
(D) *Gene expression similarity relationships among 39 inferred cell populations*. Average expression across all detected genes was calculated for each of 39 cell clusters, and the relative (Euclidean) distances between gene-expression patterns for the 39 clusters are represented by a dendrogram. The branches of the dendrogram were annotated by examining the differential expression of known markers for retina cell classes and types. Twelve examples are shown at right, using violin plots to represent the distribution of expression within the clusters. Violin plots for additional genes are in Figure S6A.

Figure 5. *Ab Initio* Reconstruction of Retinal Cell Types from 44,808 Single-Cell Transcription Profiles Prepared by Drop-Seq



(B) *Clustering of 44,808 Drop-seq single-cell expression profiles into 39 retinal cell populations*. The plot shows a two-dimensional representation (tSNE) of global gene expression relationships among 44,808 cells; clusters are colored by cell class, according to Figure 5A.

Figure 5. *Ab Initio* Reconstruction of Retinal Cell Types from 44,808 Single-Cell Transcription Profiles Prepared by Drop-Seq



(D) *Gene expression similarity relationships among 39 inferred cell populations*. Average expression across all detected genes was calculated for each of 39 cell clusters, and the relative (Euclidean) distances between gene-expression patterns for the 39 clusters are represented by a dendrogram. The branches of the dendrogram were annotated by examining the differential expression of known markers for retina cell classes and types.

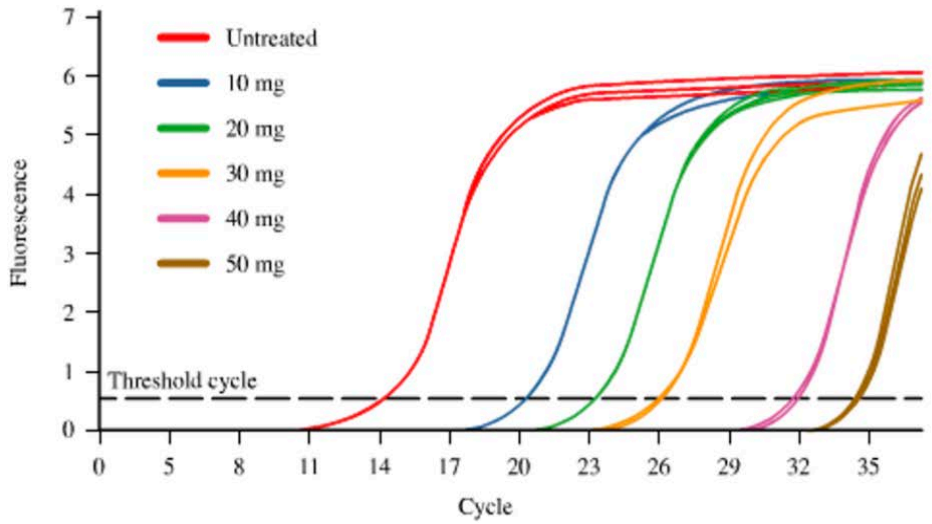
Advantages of ddPCR for Target Detection & Quantification

ddPCR offers greater sensitivity & precision than qPCR for quantifying nucleic acids:

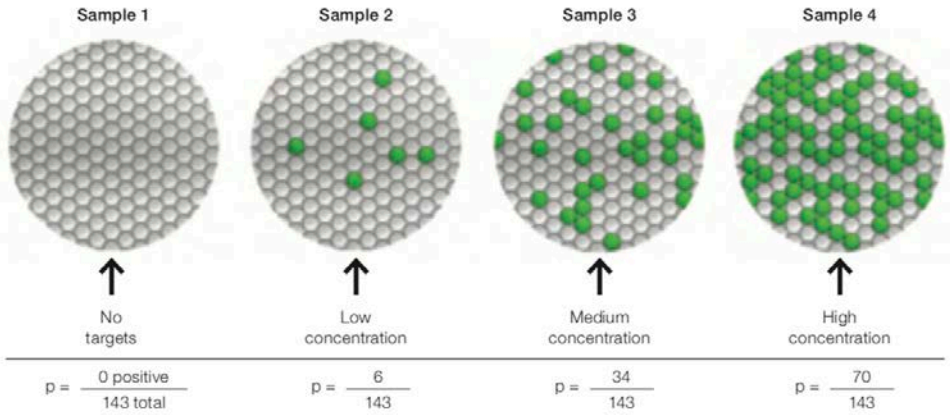
- ddPCR partitions target molecules in droplets where PCR-enrichment & fluorescent labeling boost detection sensitivity and precision compared to qPCR.
- Target is detected directly based on endpoint signal amplitude of target + droplets rather than measured indirectly against a standard dilution curve, as in qPCR.
- Absolute quantification of target molecules enables LoD down to 1 target / 10,000 to 100,000+ background molecules.
- Because target molecules are partitioned in 1ul droplets, error due to PCR efficiency & template bias is eliminated.
- Target molecule partitioning generates 1,000's of data-points from a single 20ul Rx thereby amplifying statistical power of samples that are in limited supply.

Conventional qPCR versus ddPCR

Conventional qPCR quantifies target indirectly by plotting signal intensities against a standard dilution curve and extrapolating the original concentration:



ddPCR partitions target molecules into oil droplets for PCR-based-labeling & direct detection & counting. The ratio of target +/– droplets is used to calculate [target] based on an (*approximately*) binomial distribution analysis:



Statistical Considerations for Different Applications

- **Absolute quantification**

  The recommended upper limit of detection (5 targets/droplet) is dictated by the requirement for a sufficient number of empty (target negative) droplets.

  ddPCR provides sensitivity to detect 10% mRNA differential expression with 95% confidence.

  For intermediate to high [target], *partitioning statistics* compensate for expected occurrences of >1 target copies per droplet:

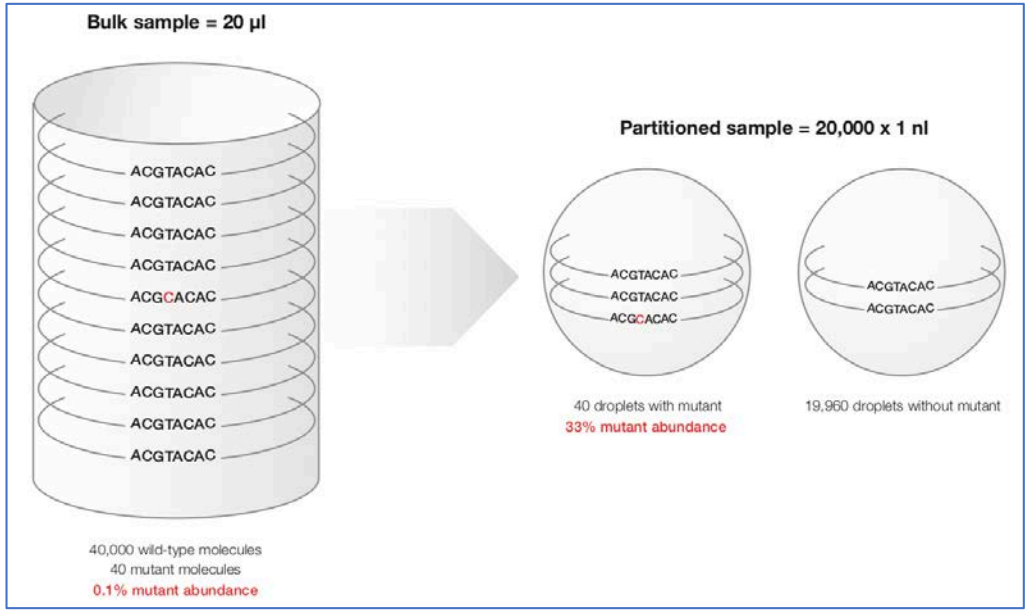
eg: CPD = .25

Targets	Droplets, %
0	78
1	19.5
2	2.4
3	0.2
4	0.01

Statistical Considerations for Different Applications

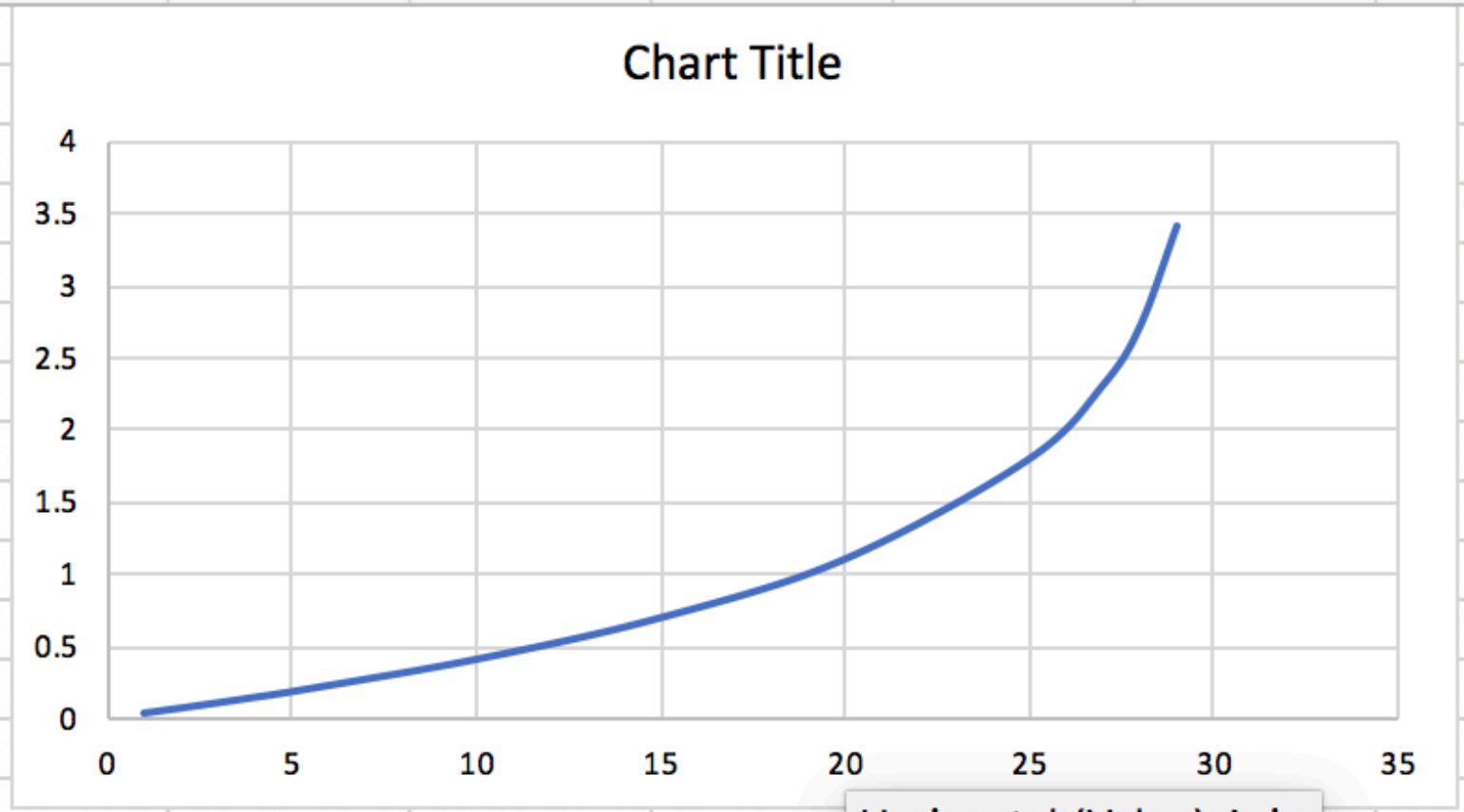
- **Rare mutation detection**:

  ddPCR partitions the sample into droplets which dilutes the wt alleles in droplets containing both wt & mutant alleles. Therefore, the mutant allele is present in droplets at a greater *relative* abundance than in bulk solution. Since the LoD is determined primarily by the number of wt molecules screened, the LoD can be increased by increasing the number of wells (molecules) screened. LoD may be up to 1 target molecule among 100k molecules.



=-LN(1-(A2/30))

	B
	total count is 30
1	0.033901552
2	0.068992871
3	0.105360516
4	0.143100844
5	0.182321557
10	0.405465108
15	0.693147181
20	1.098612289
25	1.791759469
27	2.302585093
28	2.708050201
29	3.401197382
30	



The impact of rare and low-frequency genetic variants in common disease

Lorenzo Bomba,

Klaudia Walter and

Nicole SoranzoEmail author

*Genome Biology*2017**18**:77

https://doi.org/10.1186/s13059-017-1212-4

Genotype Phasing

https://youtu.be/15NPZCGP_e4

Abstract

Despite thousands of genetic loci identified to date, a large proportion of genetic variation predisposing to complex disease and traits remains unaccounted for. Advances in sequencing technology enable focused explorations on the contribution of low-frequency and rare variants to human traits. Here we review experimental approaches and current knowledge on the contribution of these genetic variants in complex disease and discuss challenges and opportunities for personalised medicine.

Genomic tools for assessing low-frequency and rare variants

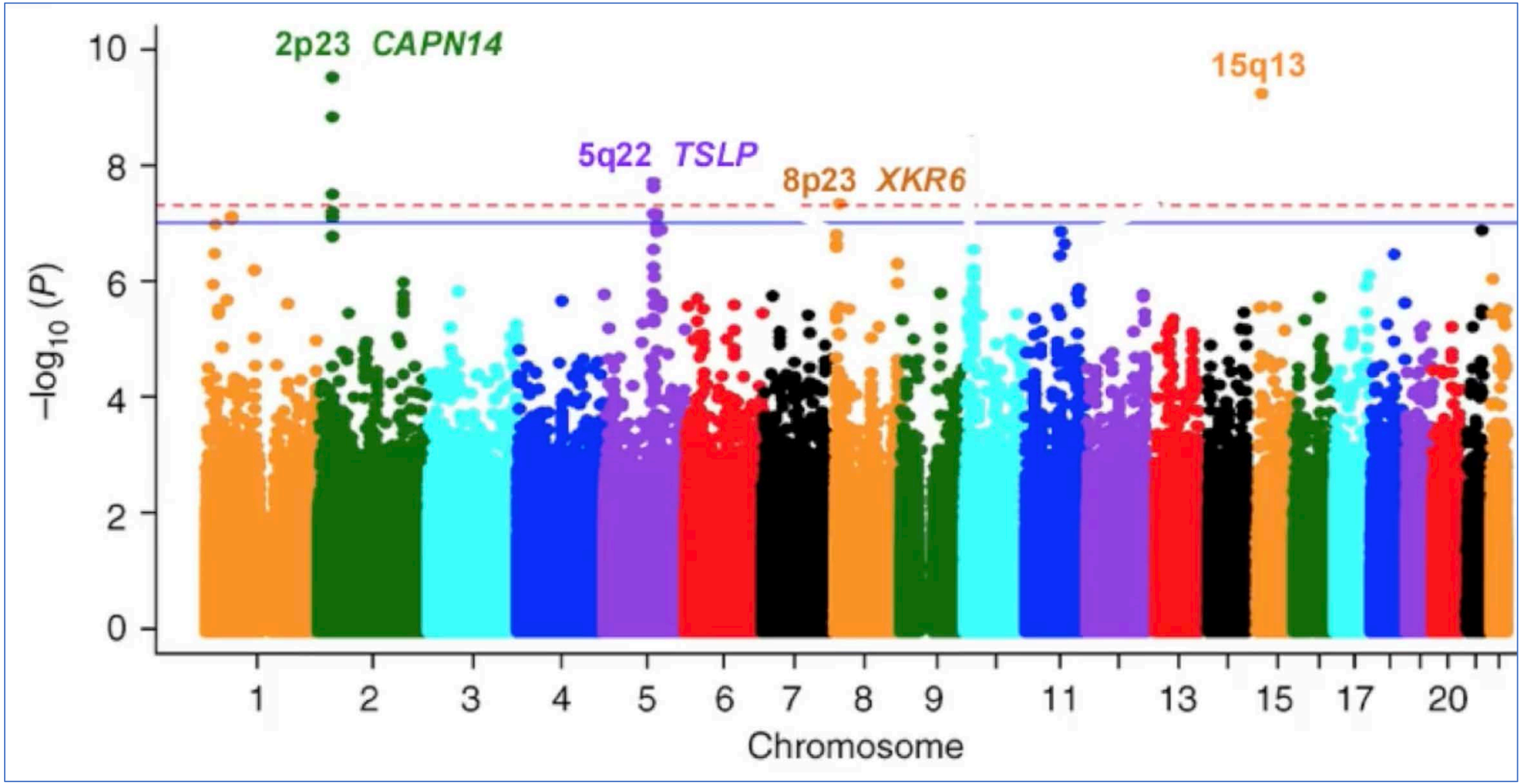
Three broad strategies are available to access low-frequency and rare variants:

- genotype ***imputation***
- custom genotyping arrays
- whole-exome or whole-genome sequencing.

Imputation

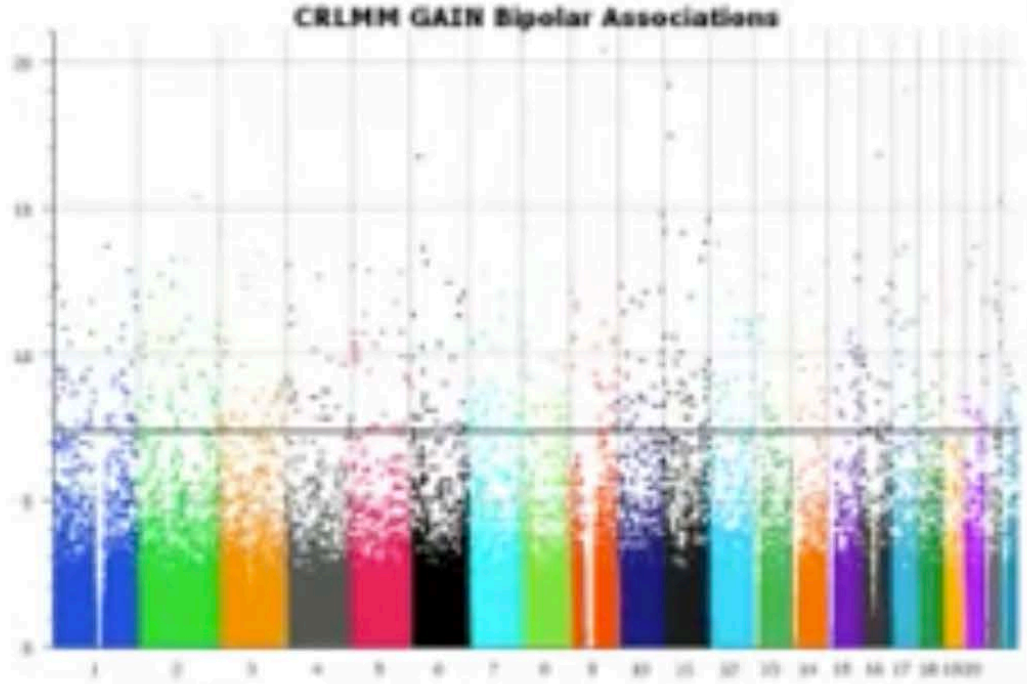
- Genotype imputation provides a cost-effective strategy for expanding the SNP content of genome-wide genotyping arrays.
- It relies on the availability of **reference panels** of phased haplotypes that can be used to impute genotypes into sparse datasets generated by commercial genotyping arrays.
- Multiple **different reference panels** have been generated since 2005, enabled by expanding collections of polymorphisms in human populations.

Manhattan Plots

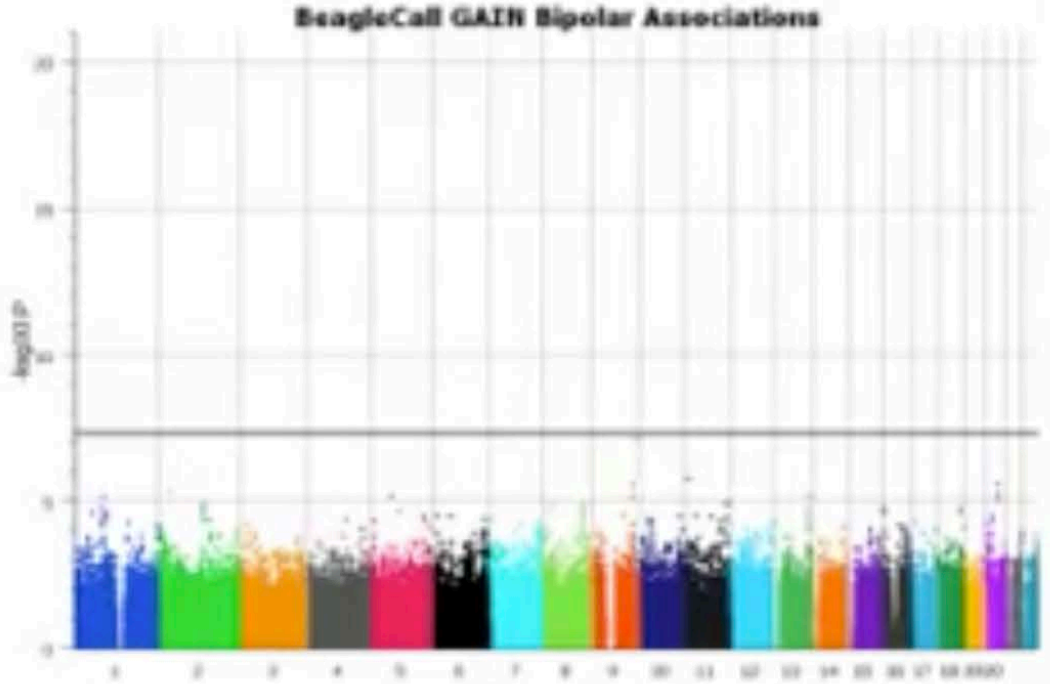


pvalue	"-LOG10(pvalue)"
0.01	2
0.005	2.301029996
0.001	3

Genotype Calling and Imputation with BEAGLE and BEAGLECALL Genetic Analysis Tools



- Many spurious associations without additional filtration



- All false positives removed
- Made mega-analysis with WTCCC 500k BD study artifact-free after imputation

Custom genotyping arrays

- An alternative strategy to imputation to survey low-frequency and rare variants in association studies takes advantage of **custom genotyping arrays**.
- These arrays are often **disease focused** and aim to enrich standard haplotype tagging SNP panels with variants of interest identified through sequencing and fine-mapping efforts.

Exome or whole-genome sequencing

- Historically, candidate gene sequencing studies have been used to explore sequence variation through relatively small-scale sequencing efforts.

your first writing assignment

Optimal methods for association analysis with low-frequency and rare variants

…. several statistical methods have been proposed **to increase statistical power in association studies**, typically by seeking to **combine information across multiple rare variants** within a specific genomic functional unit (e.g. gene, exon).

Burden tests, Variance-component tests, Combined tests (SKAT), Other tests

Other tests have been developed to account for signal sparsity across the tested region and include least absolute shrinkage and selection operator (LASSO) and the exponential combination (EC) test

Power, replication and confounding affecting rare variant association tests

RVAS's : rare variant association studies (RVASs)

An ongoing challenge is to systematically **evaluate the** relative merit, assumptions, implementation and **statistical power of different analyses**.

Study designs for enriching or prioritising rare variants

loss-of-function (LoF) variants : LoF variants provide powerful tools to understand the impact of “knocking out” human genes, akin to gene knockout experiments commonly conducted in model organisms. Understanding the phenotypic and clinical consequences of carrying LoF alleles, particularly when they are carried in the homozygous (i.e. complete knockout) state, has been shown to provide crucial insights into the identification of new disease genes and druggable pathways

Initial results from associations from large-scale sequencing projects

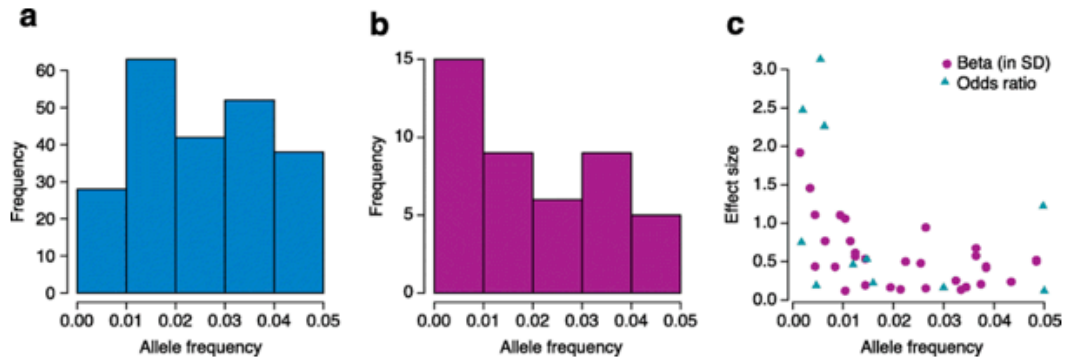
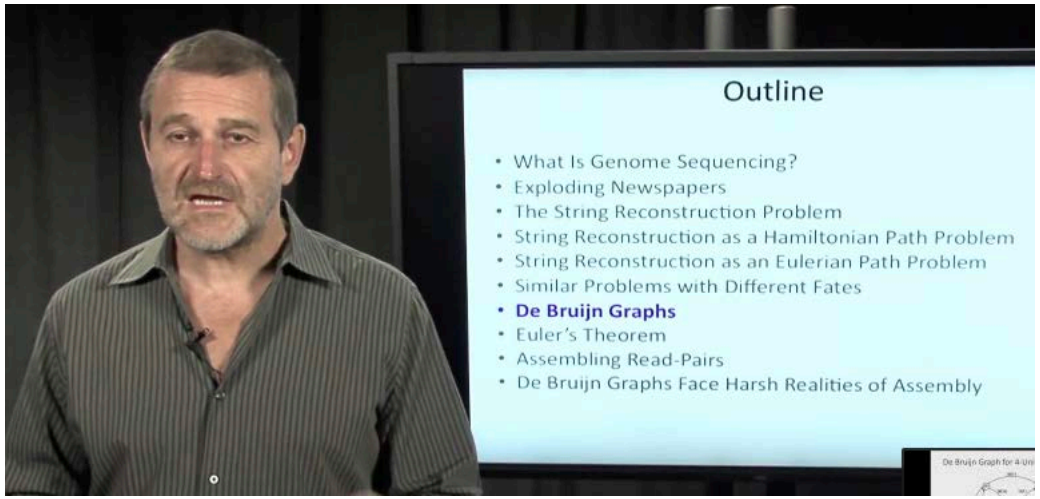


Fig. 1
The allele frequency spectrum for **a** genome-wide association study variants (Additional file 1) and **b** sequenced variants that were associated with a variety of traits (Table 3 and Additional file 1). There is a clear shift to lower allele frequencies for variants discovered in sequencing studies. **c** The effect size versus allele frequency for sequenced variants; i.e. to detect associations that involve variants with lower allele frequencies, higher effect sizes are needed or large sample sizes. Effect size is usually measured as "beta" for quantitative traits and as "odds ratio" for dichotomous traits

De Bruijn Graphs

https://youtu.be/f-ecmECK7lw



BROAD INSTITUTE

Cambridge, Massachusetts, USA.

Introduction to *De novo* RNA-Seq Assembly using Trinity

Brian Haas
Broad Institute

https://youtu.be/D3PSaxhOVlU

The impact of rare and low-frequency genetic variants in common disease

Lorenzo Bomba,

Klaudia Walter and

Nicole SoranzoEmail author

*Genome Biology*2017**18**:77

https://doi.org/10.1186/s13059-017-1212-4

Log Transformation of Genomic Data

Probability Density Plots

Cumulative Density Plots

Quantile Normalization

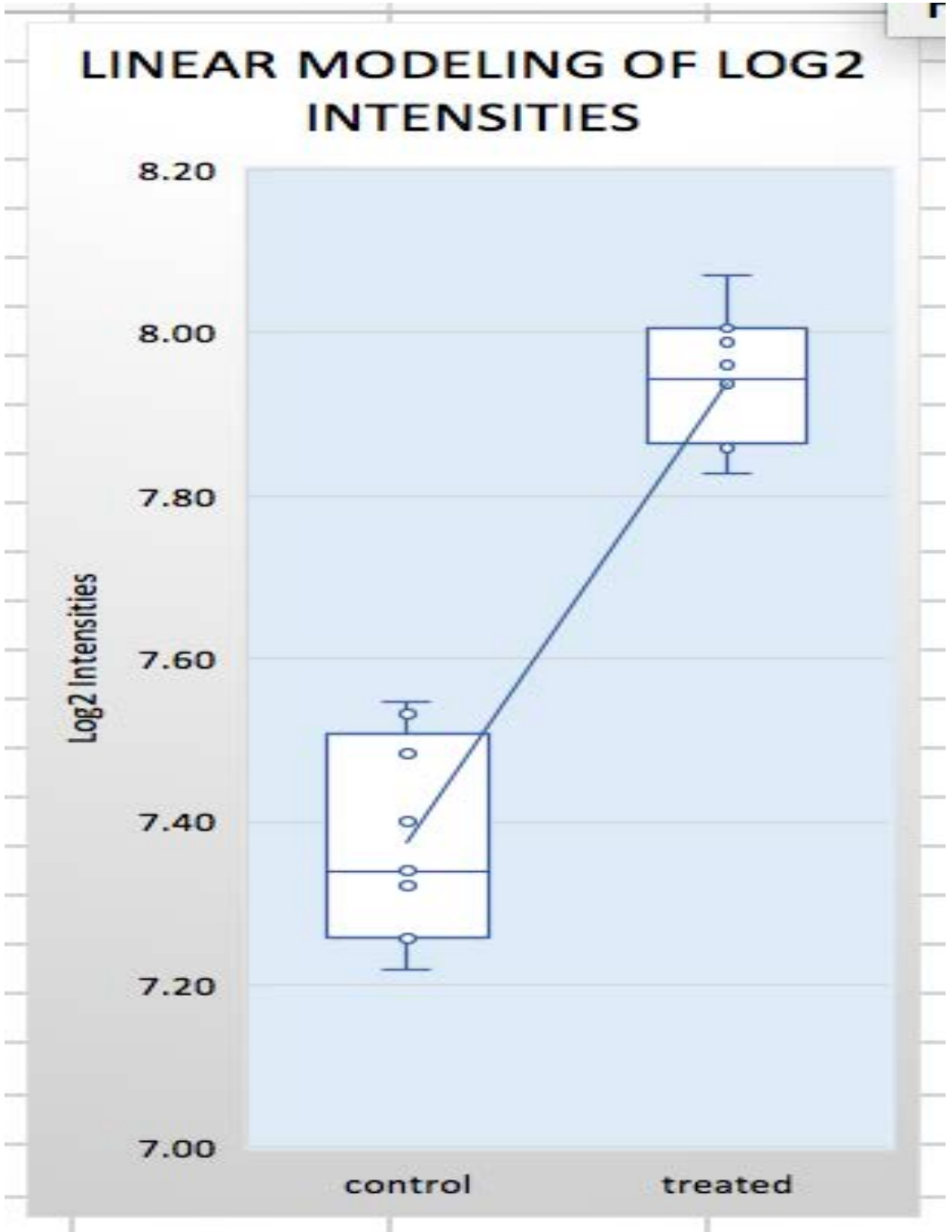
Quantile Quantile Plots

actual intensity	Log2 intensity	Log10 intensity
1	0.00	0.00
2	1.00	0.30
3	1.58	0.48
4	2.00	0.60
5	2.32	0.70
6	2.58	0.78
7	2.81	0.85
8	3.00	0.90
9	3.17	0.95
10	3.32	1.00
20	4.32	1.30
30	4.91	1.48
40	5.32	1.60
50	5.64	1.70
60	5.91	1.78
70	6.13	1.85
80	6.32	1.90
90	6.49	1.95
100	6.64	2.00
200	7.64	2.30
300	8.23	2.48
400	8.64	2.60
500	8.97	2.70
600	9.23	2.78
700	9.45	2.85
800	9.64	2.90
900	9.81	2.95
1000	9.97	3.00

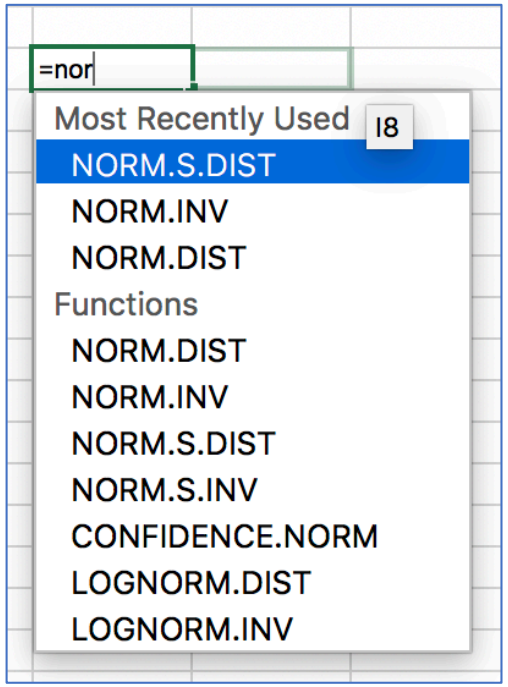
CLASS	INTENSITY	log2 intensity
control	187	7.55
control	179	7.48
control	153	7.26
control	169	7.40
control	162	7.34
control	160	7.32
control	153	7.26
control	185	7.53
control	149	7.22
treated	233	7.86
treated	245	7.94
treated	249	7.96
treated	233	7.86
treated	246	7.94
treated	254	7.99
treated	269	8.07
treated	257	8.01
treated	258	8.01
treated	232	7.86
treated	227	7.83

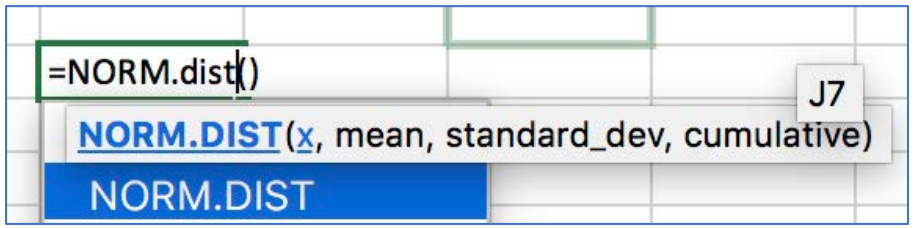
Log Transformation of Genomic Data

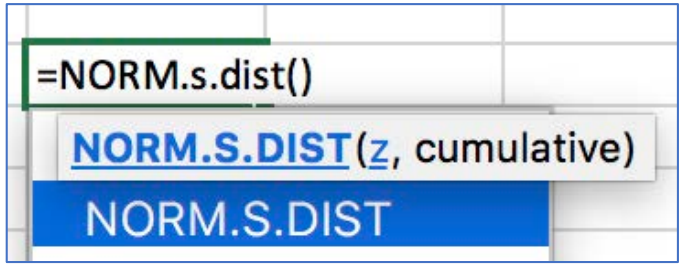
CLASS	INTENSITY	log2 intensity
control	187	7.55
control	179	7.48
control	153	7.26
control	169	7.40
control	162	7.34
control	160	7.32
control	153	7.26
control	185	7.53
control	149	7.22
treated	233	7.86
treated	245	7.94
treated	249	7.96
treated	233	7.86
treated	246	7.94
treated	254	7.99
treated	269	8.07
treated	257	8.01
treated	258	8.01
treated	232	7.86
treated	227	7.83

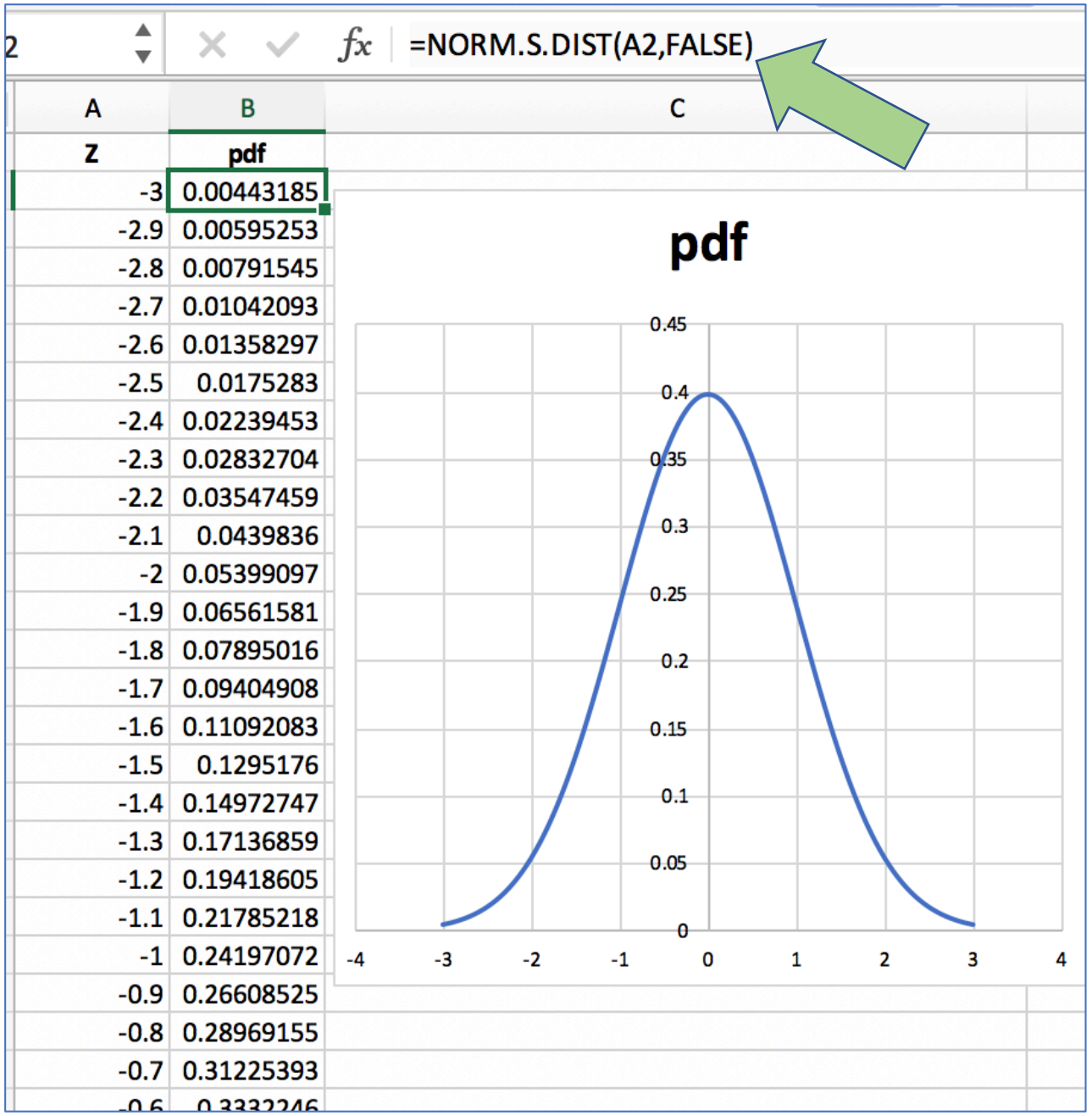


Probability Density Plots



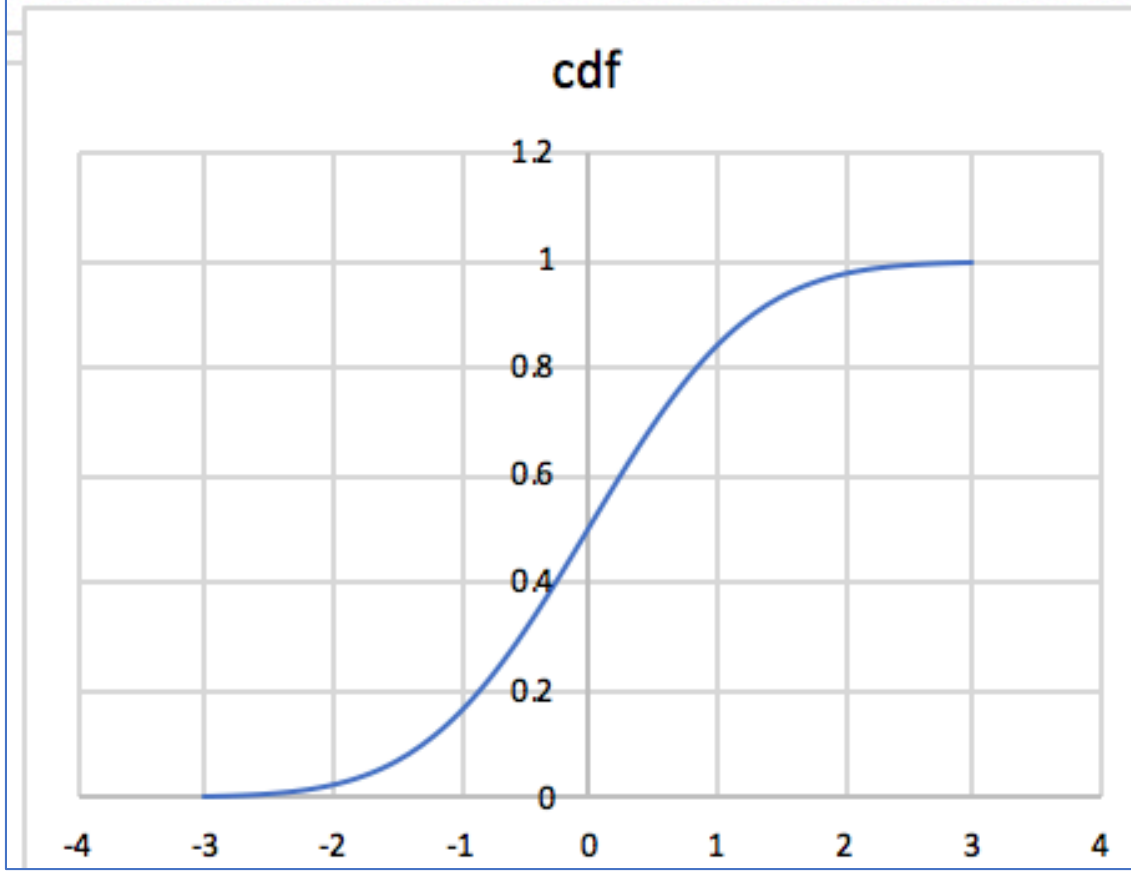




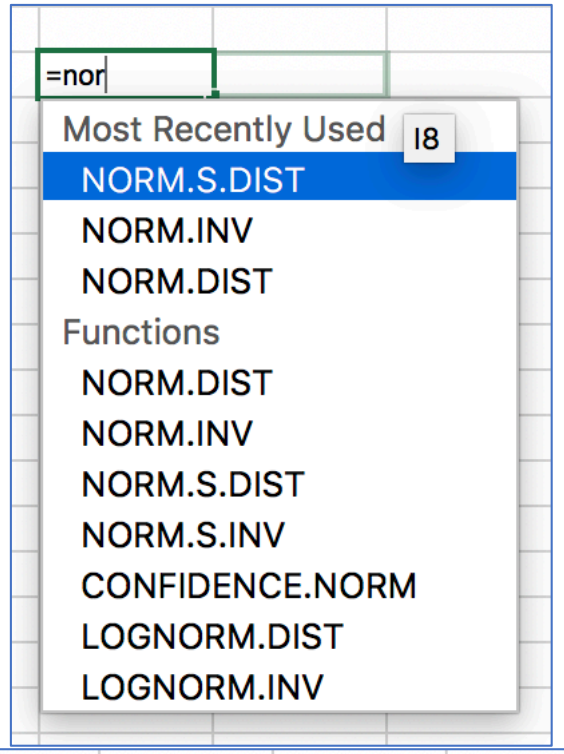


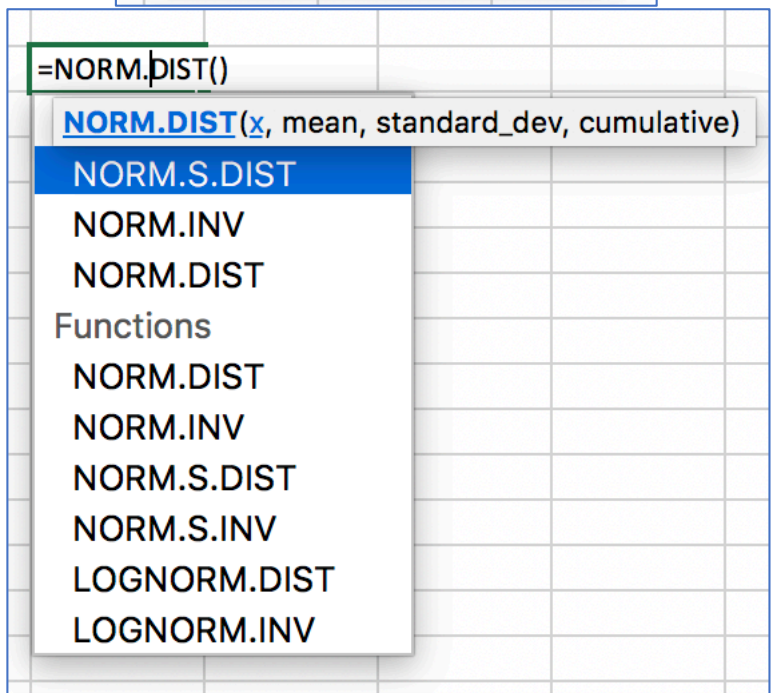
Cumulative Distribution

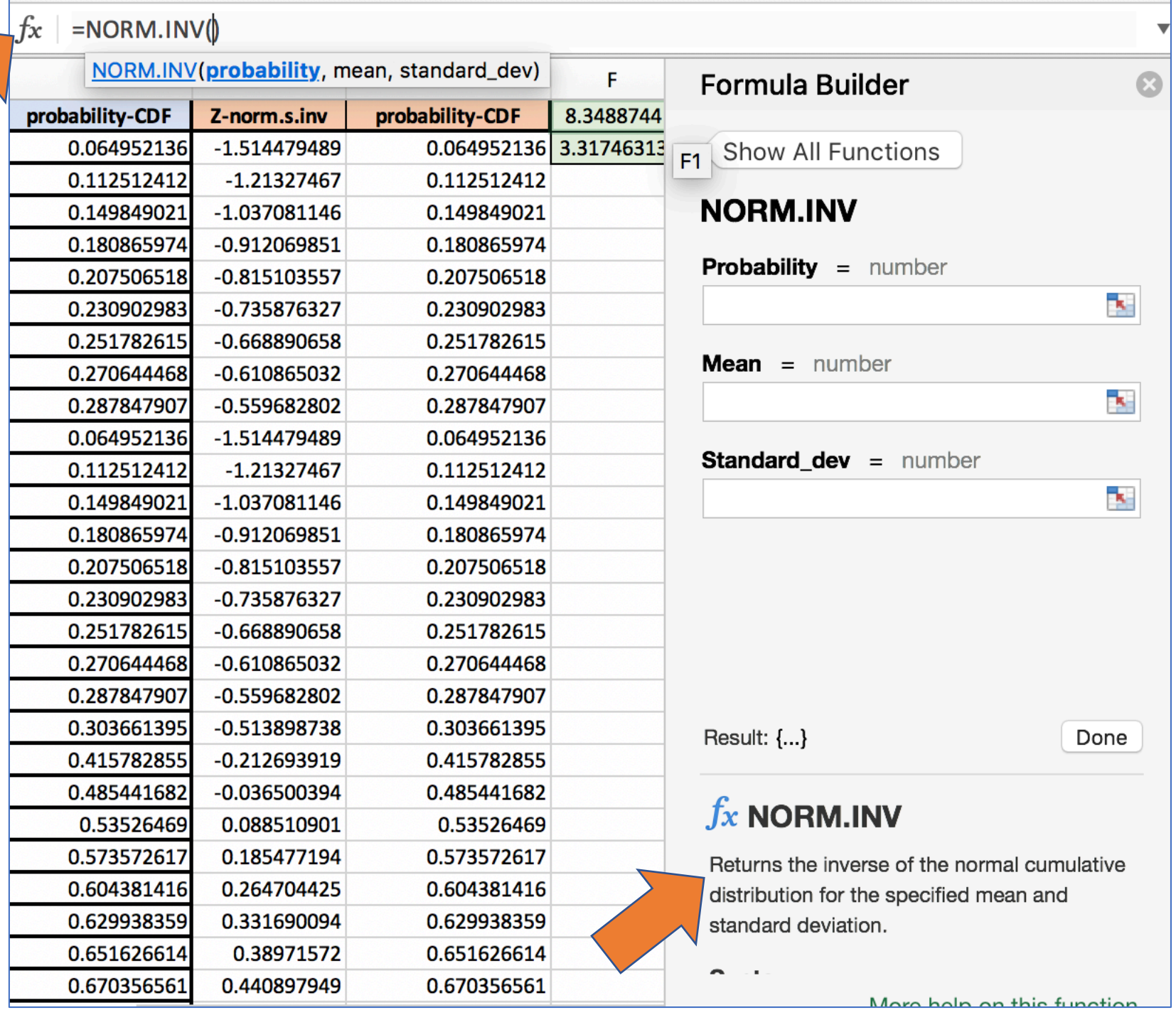
=NORM.S.DIST(D2,TRUE)



Z	cdf
-3	0.0013499
-2.9	0.0018658
-2.8	0.0025551
-2.7	0.003467
-2.6	0.0046612
-2.5	0.0062097
-2.4	0.0081975
-2.3	0.0107241
-2.2	0.0139034
-2.1	0.0178644
-2	0.0227501
-1.9	0.0287166
-1.8	0.0359303


=nor
Most Recently Used
NORM.S.DIST
NORM.INV
NORM.DIST
Functions
NORM.DIST
NORM.INV
NORM.S.DIST
NORM.S.INV
CONFIDENCE.NORM
LOGNORM.DIST
LOGNORM.INV


=NORM.DIST()
NORM.DIST(x, mean, standard_dev, cumulative)
NORM.S.DIST
NORM.INV
NORM.DIST
Functions
NORM.DIST
NORM.INV
NORM.S.DIST
NORM.S.INV
LOGNORM.DIST
LOGNORM.INV


=NORM.INV()
NORM.INV(probability, mean, standard_dev)
F
8.3488744
3.31746313
probability-CDF
Z-norm.s.inv
probability-CDF
0.064952136 -1.514479489 0.064952136
0.112512412 -1.21327467 0.112512412
0.149849021 -1.037081146 0.149849021
0.180865974 -0.912069851 0.180865974
0.207506518 -0.815103557 0.207506518
0.230902983 -0.735876327 0.230902983
0.251782615 -0.668890658 0.251782615
0.270644468 -0.610865032 0.270644468
0.287847907 -0.559682802 0.287847907
0.064952136 -1.514479489 0.064952136
0.112512412 -1.21327467 0.112512412
0.149849021 -1.037081146 0.149849021
0.180865974 -0.912069851 0.180865974
0.207506518 -0.815103557 0.207506518
0.230902983 -0.735876327 0.230902983
0.251782615 -0.668890658 0.251782615
0.270644468 -0.610865032 0.270644468
0.287847907 -0.559682802 0.287847907
0.303661395 -0.513898738 0.303661395
0.415782855 -0.212693919 0.415782855
0.485441682 -0.036500394 0.485441682
0.53526469 0.088510901 0.53526469
0.573572617 0.185477194 0.573572617
0.604381416 0.264704425 0.604381416
0.629938359 0.331690094 0.629938359
0.651626614 0.38971572 0.651626614
0.670356561 0.440897949 0.670356561
Formula Builder
F1
Show All Functions
NORM.INV
Probability = number
Mean = number
Standard_dev = number
Result: {...}
Done
NORM.INV
Returns the inverse of the normal cumulative distribution for the specified mean and standard deviation.

A **Z-Score** is :

(Observed – Expected)/ Standard Deviation

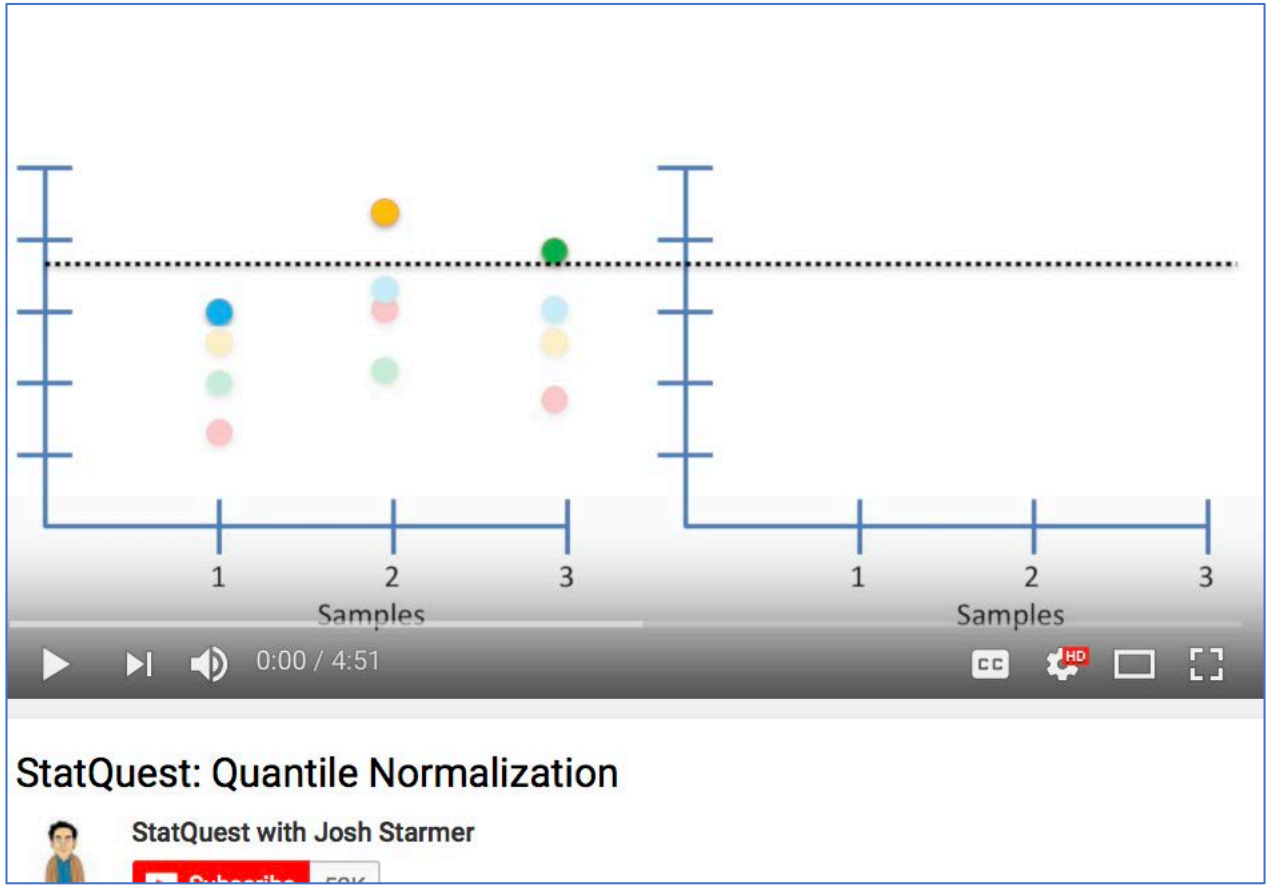
(Data point – Mean)/ Standard Deviation

I2 =(B2-F1)/F2

	A	B	C	D	E	F	G	H	I
	I counts	log2-Icounts	probability-CDF	Z-norm.s.inv		8.3488744	avg-log2-I		"Z=(data-mean)/STDEV"
2	10	3.321928095	0.064952136	-1.514479489		3.31746313	stdev		-1.515298319
3	20	4.321928095	0.112512412	-1.21327467					-1.213863168
4	30	4.906890596	0.149849021	-1.037081146					-1.037534908
5	40	5.321928095	0.180865974	-0.912069851					-0.912428017
6	50	5.64385619	0.207506518	-0.815103557					-0.815387573
7	60	5.906890596	0.230902983	-0.735876327					-0.736099758
8	70	6.129283017	0.251782615	-0.668890658					-0.669062864
9	80	6.321928095	0.270644468	-0.610865032					-0.610992866
10	90	6.491853096	0.287847907	-0.559682802					-0.559771498
11	100	3.321928095	0.064952136	-1.514479489					-1.515298319
12	200	4.321928095	0.112512412	-1.21327467					-1.213863168
13	300	4.906890596	0.149849021	-1.037081146					-1.037534908
14	400	5.321928095	0.180865974	-0.912069851					-0.912428017
15	500	5.64385619	0.207506518	-0.815103557					-0.815387573
16	600	5.906890596	0.230902983	-0.735876327					-0.736099758
17	700	6.129283017	0.251782615	-0.668890658					-0.669062864
18	800	6.321928095	0.270644468	-0.610865032					-0.610992866
19	900	6.491853096	0.287847907	-0.559682802					-0.559771498
20	1000	6.64385619	0.303661395	-0.513898738					-0.513952422
21	2000	7.64385619	0.415782855	-0.212693919					-0.212517271
22	3000	8.22881869	0.485441682	-0.036500394					-0.036189012
23	4000	8.64385619	0.53526469	0.088510901					0.08891788
24	5000	8.965784285	0.573572617	0.185477194					0.185958323
25	6000	9.22881869	0.604381416	0.264704425					0.265246139
26	7000	9.451211112	0.629938359	0.331690094					0.332283032

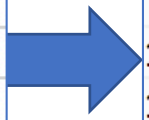
Quantile Normalization of Data from Different Experiments *(e.g. different microarrays)*

https://youtu.be/ecjN6Xpv6SE



Copy and paste the GENE_IDs column in front of each sample column:

	GENE_IDs	SAMPLE_A	SAMPLE_B	SAMPLE_C
2	Gene01	121.7	160.0	142.2
3	Gene02	141.5	169.0	150.2
4	Gene03	117.3	121.0	155.0
5	Gene04	143.7	127.0	124.6
6	Gene05	154.7	142.0	147.0
7	Gene06	161.3	124.0	137.4
8	Gene07	119.5	154.0	143.8
9	Gene08	130.5	118.0	131.0
10	Gene09	156.9	145.0	148.6
11	Gene10	132.7	112.0	153.4
12	Gene11	134.9	148.0	127.8
13	Gene12	150.3	163.0	139.0
14	Gene13	137.1	130.0	129.4
15	Gene14	145.9	166.0	145.4
16	Gene15	128.3	133.0	132.6
17	Gene16	139.3	136.0	151.8
18	Gene17	126.1	157.0	134.2
19	Gene18	148.1	139.0	135.8
20	Gene19	152.5	151.0	126.2
21	Gene20	159.1	172.0	140.6
22	Gene21	123.9	115.0	123.0



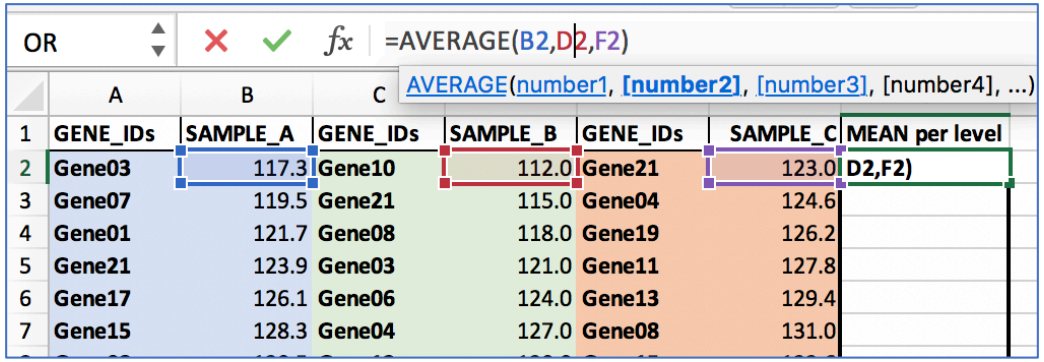
	GENE_IDs	SAMPLE_A	GENE_IDs	SAMPLE_B	GENE_IDs	SAMPLE_C
2	Gene01	121.7	Gene01	160.0	Gene01	142.2
3	Gene02	141.5	Gene02	169.0	Gene02	150.2
4	Gene03	117.3	Gene03	121.0	Gene03	155.0
5	Gene04	143.7	Gene04	127.0	Gene04	124.6
6	Gene05	154.7	Gene05	142.0	Gene05	147.0
7	Gene06	161.3	Gene06	124.0	Gene06	137.4
8	Gene07	119.5	Gene07	154.0	Gene07	143.8
9	Gene08	130.5	Gene08	118.0	Gene08	131.0
10	Gene09	156.9	Gene09	145.0	Gene09	148.6
11	Gene10	132.7	Gene10	112.0	Gene10	153.4
12	Gene11	134.9	Gene11	148.0	Gene11	127.8
13	Gene12	150.3	Gene12	163.0	Gene12	139.0
14	Gene13	137.1	Gene13	130.0	Gene13	129.4
15	Gene14	145.9	Gene14	166.0	Gene14	145.4
16	Gene15	128.3	Gene15	133.0	Gene15	132.6
17	Gene16	139.3	Gene16	136.0	Gene16	151.8
18	Gene17	126.1	Gene17	157.0	Gene17	134.2
19	Gene18	148.1	Gene18	139.0	Gene18	135.8
20	Gene19	152.5	Gene19	151.0	Gene19	126.2
21	Gene20	159.1	Gene20	172.0	Gene20	140.6
22	Gene21	123.9	Gene21	115.0	Gene21	123.0

Sort each “SAMPLE SET” by signal values, smallest to largest

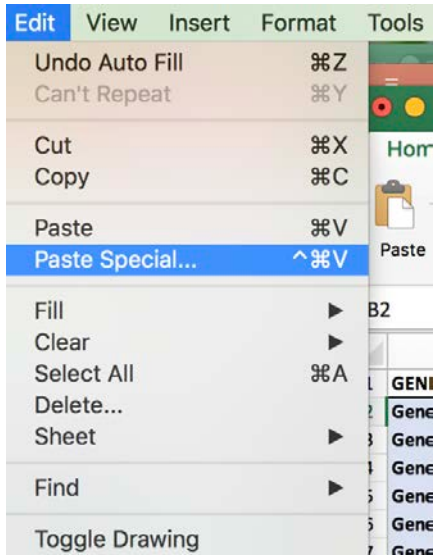
	GENE_IDs	SAMPLE_A	GENE_IDs	SAMPLE_B	GENE_IDs	SAMPLE_C
2	Gene01	121.7	Gene01	160.0	Gene01	142.2
3	Gene02	141.5	Gene02	169.0	Gene02	150.2
4	Gene03	117.3	Gene03	121.0	Gene03	155.0
5	Gene04	143.7	Gene04	127.0	Gene04	124.6
6	Gene05	154.7	Gene05	142.0	Gene05	147.0
7	Gene06	161.3	Gene06	124.0	Gene06	137.4
8	Gene07	119.5	Gene07	154.0	Gene07	143.8
9	Gene08	130.5	Gene08	118.0	Gene08	131.0
10	Gene09	156.9	Gene09	145.0	Gene09	148.6
11	Gene10	132.7	Gene10	112.0	Gene10	153.4
12	Gene11	134.9	Gene11	148.0	Gene11	127.8
13	Gene12	150.3	Gene12	163.0	Gene12	139.0
14	Gene13	137.1	Gene13	130.0	Gene13	129.4
15	Gene14	145.9	Gene14	166.0	Gene14	145.4
16	Gene15	128.3	Gene15	133.0	Gene15	132.6
17	Gene16	139.3	Gene16	136.0	Gene16	151.8
18	Gene17	126.1	Gene17	157.0	Gene17	134.2
19	Gene18	148.1	Gene18	139.0	Gene18	135.8
20	Gene19	152.5	Gene19	151.0	Gene19	126.2
21	Gene20	159.1	Gene20	172.0	Gene20	140.6
22	Gene21	123.9	Gene21	115.0	Gene21	123.0

	GENE_IDs	SAMPLE_A	GENE_IDs	SAMPLE_B	GENE_IDs	SAMPLE_C
2	Gene03	117.3	Gene10	112.0	Gene21	123.0
3	Gene07	119.5	Gene21	115.0	Gene04	124.6
4	Gene01	121.7	Gene08	118.0	Gene19	126.2
5	Gene21	123.9	Gene03	121.0	Gene11	127.8
6	Gene17	126.1	Gene06	124.0	Gene13	129.4
7	Gene15	128.3	Gene04	127.0	Gene08	131.0
8	Gene08	130.5	Gene13	130.0	Gene15	132.6
9	Gene10	132.7	Gene15	133.0	Gene17	134.2
10	Gene11	134.9	Gene16	136.0	Gene18	135.8
11	Gene13	137.1	Gene18	139.0	Gene06	137.4
12	Gene16	139.3	Gene05	142.0	Gene12	139.0
13	Gene02	141.5	Gene09	145.0	Gene20	140.6
14	Gene04	143.7	Gene11	148.0	Gene01	142.2
15	Gene14	145.9	Gene19	151.0	Gene07	143.8
16	Gene18	148.1	Gene07	154.0	Gene14	145.4
17	Gene12	150.3	Gene17	157.0	Gene05	147.0
18	Gene19	152.5	Gene01	160.0	Gene09	148.6
19	Gene05	154.7	Gene12	163.0	Gene02	150.2
20	Gene09	156.9	Gene14	166.0	Gene16	151.8
21	Gene20	159.1	Gene02	169.0	Gene10	153.4
22	Gene06	161.3	Gene20	172.0	Gene03	155.0

Now calculate the mean for each level. *Copy* & *Paste Special* this column of formulas into values:



	A	B	C	D	E	F	G
1	GENE_IDs	SAMPLE_A	GENE_IDs	SAMPLE_B	GENE_IDs	SAMPLE_C	MEAN per level
2	Gene03	117.3	Gene10	112.0	Gene21	123.0	
3	Gene07	119.5	Gene21	115.0	Gene04	124.6	
4	Gene01	121.7	Gene08	118.0	Gene19	126.2	
5	Gene21	123.9	Gene03	121.0	Gene11	127.8	
6	Gene17	126.1	Gene06	124.0	Gene13	129.4	
7	Gene15	128.3	Gene04	127.0	Gene08	131.0	



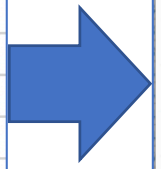
	A	B	C	D	E	F	G
1	GENE_IDs	SAMPLE_A	GENE_IDs	SAMPLE_B	GENE_IDs	SAMPLE_C	MEAN per level
2	Gene03	117.3	Gene10	112.0	Gene21	123.0	117.4
3	Gene07	119.5	Gene21	115.0	Gene04	124.6	119.7
4	Gene01	121.7	Gene08	118.0	Gene19	126.2	122.0
5	Gene21	123.9	Gene03	121.0	Gene11	127.8	124.2
6	Gene17	126.1	Gene06	124.0	Gene13	129.4	126.5
7	Gene15	128.3	Gene04	127.0	Gene08	131.0	128.8
8	Gene08	130.5	Gene13	130.0	Gene15	132.6	131.0
9	Gene10	132.7	Gene15	133.0	Gene17	134.2	133.3
10	Gene11	134.9	Gene16	136.0	Gene18	135.8	135.6
11	Gene13	137.1	Gene18	139.0	Gene06	137.4	137.8
12	Gene16	139.3	Gene05	142.0	Gene12	139.0	140.1
13	Gene02	141.5	Gene09	145.0	Gene20	140.6	142.4
14	Gene04	143.7	Gene11	148.0	Gene01	142.2	144.6
15	Gene14	145.9	Gene19	151.0	Gene07	143.8	146.9
16	Gene18	148.1	Gene07	154.0	Gene14	145.4	149.2
17	Gene12	150.3	Gene17	157.0	Gene05	147.0	151.4
18	Gene19	152.5	Gene01	160.0	Gene09	148.6	153.7
19	Gene05	154.7	Gene12	163.0	Gene02	150.2	156.0
20	Gene09	156.9	Gene14	166.0	Gene16	151.8	158.2
21	Gene20	159.1	Gene02	169.0	Gene10	153.4	160.5
22	Gene06	161.3	Gene20	172.0	Gene03	155.0	162.8
23							

Copy the *Mean per level* data column to each replace each sample data column:

	GENE_IDs	SAMPLE_A	GENE_IDs	SAMPLE_B	GENE_IDs	SAMPLE_C	MEAN per level
2	Gene03	117.4	Gene10	117.4	Gene21	117.4	117.4
3	Gene07	119.7	Gene21	119.7	Gene04	119.7	119.7
4	Gene01	122.0	Gene08	122.0	Gene19	122.0	122.0
5	Gene21	124.2	Gene03	124.2	Gene11	124.2	124.2
6	Gene17	126.5	Gene06	126.5	Gene13	126.5	126.5
7	Gene15	128.8	Gene04	128.8	Gene08	128.8	128.8
8	Gene08	131.0	Gene13	131.0	Gene15	131.0	131.0
9	Gene10	133.3	Gene15	133.3	Gene17	133.3	133.3
10	Gene11	135.6	Gene16	135.6	Gene18	135.6	135.6
11	Gene13	137.8	Gene18	137.8	Gene06	137.8	137.8
12	Gene16	140.1	Gene05	140.1	Gene12	140.1	140.1
13	Gene02	142.4	Gene09	142.4	Gene20	142.4	142.4
14	Gene04	144.6	Gene11	144.6	Gene01	144.6	144.6
15	Gene14	146.9	Gene19	146.9	Gene07	146.9	146.9
16	Gene18	149.2	Gene07	149.2	Gene14	149.2	149.2
17	Gene12	151.4	Gene17	151.4	Gene05	151.4	151.4
18	Gene19	153.7	Gene01	153.7	Gene09	153.7	153.7
19	Gene05	156.0	Gene12	156.0	Gene02	156.0	156.0
20	Gene09	158.2	Gene14	158.2	Gene16	158.2	158.2
21	Gene20	160.5	Gene02	160.5	Gene10	160.5	160.5
22	Gene06	162.8	Gene20	162.8	Gene03	162.8	162.8
23							

Now sort each sample set by the GENE_IDs:

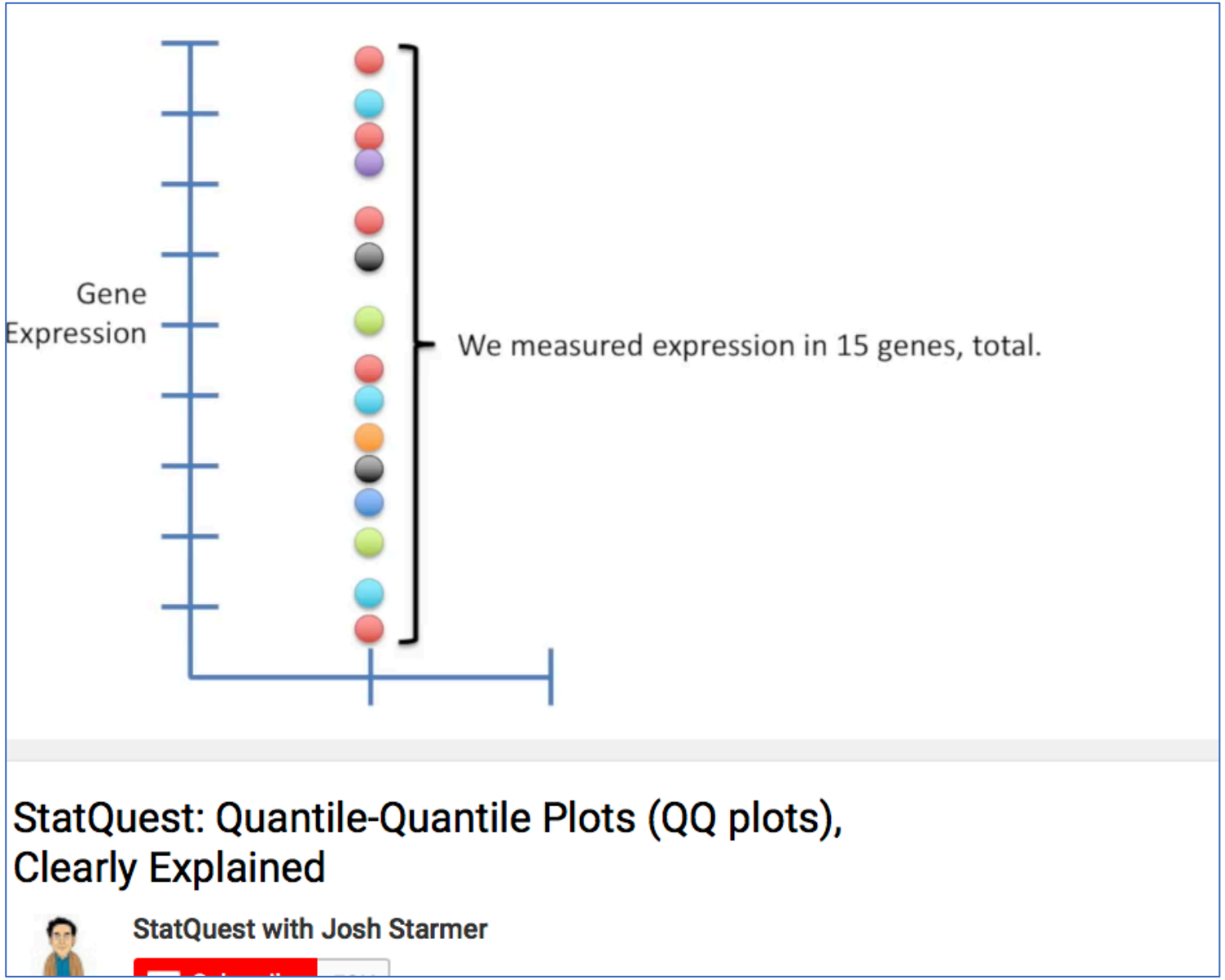
	GENE_IDs	SAMPLE_A	GENE_IDs	SAMPLE_B	GENE_IDs	SAMPLE_C	MEAN per level
2	Gene03	117.4	Gene10	117.4	Gene21	117.4	117.4
3	Gene07	119.7	Gene21	119.7	Gene04	119.7	119.7
4	Gene01	122.0	Gene08	122.0	Gene19	122.0	122.0
5	Gene21	124.2	Gene03	124.2	Gene11	124.2	124.2
6	Gene17	126.5	Gene06	126.5	Gene13	126.5	126.5
7	Gene15	128.8	Gene04	128.8	Gene08	128.8	128.8
8	Gene08	131.0	Gene13	131.0	Gene15	131.0	131.0
9	Gene10	133.3	Gene15	133.3	Gene17	133.3	133.3
10	Gene11	135.6	Gene16	135.6	Gene18	135.6	135.6
11	Gene13	137.8	Gene18	137.8	Gene06	137.8	137.8
12	Gene16	140.1	Gene05	140.1	Gene12	140.1	140.1
13	Gene02	142.4	Gene09	142.4	Gene20	142.4	142.4
14	Gene04	144.6	Gene11	144.6	Gene01	144.6	144.6
15	Gene14	146.9	Gene19	146.9	Gene07	146.9	146.9
16	Gene18	149.2	Gene07	149.2	Gene14	149.2	149.2
17	Gene12	151.4	Gene17	151.4	Gene05	151.4	151.4
18	Gene19	153.7	Gene01	153.7	Gene09	153.7	153.7
19	Gene05	156.0	Gene12	156.0	Gene02	156.0	156.0
20	Gene09	158.2	Gene14	158.2	Gene16	158.2	158.2
21	Gene20	160.5	Gene02	160.5	Gene10	160.5	160.5
22	Gene06	162.8	Gene20	162.8	Gene03	162.8	162.8
23							



	GENE_IDs	SAMPLE_A	GENE_IDs	SAMPLE_B	GENE_IDs	SAMPLE_C
2	Gene01	122.0	Gene01	153.7	Gene01	144.6
3	Gene02	142.4	Gene02	160.5	Gene02	156.0
4	Gene03	117.4	Gene03	124.2	Gene03	162.8
5	Gene04	144.6	Gene04	128.8	Gene04	119.7
6	Gene05	156.0	Gene05	140.1	Gene05	151.4
7	Gene06	162.8	Gene06	126.5	Gene06	137.8
8	Gene07	119.7	Gene07	149.2	Gene07	146.9
9	Gene08	131.0	Gene08	122.0	Gene08	128.8
10	Gene09	158.2	Gene09	142.4	Gene09	153.7
11	Gene10	133.3	Gene10	117.4	Gene10	160.5
12	Gene11	135.6	Gene11	144.6	Gene11	124.2
13	Gene12	151.4	Gene12	156.0	Gene12	140.1
14	Gene13	137.8	Gene13	131.0	Gene13	126.5
15	Gene14	146.9	Gene14	158.2	Gene14	149.2
16	Gene15	128.8	Gene15	133.3	Gene15	131.0
17	Gene16	140.1	Gene16	135.6	Gene16	158.2
18	Gene17	126.5	Gene17	151.4	Gene17	133.3
19	Gene18	149.2	Gene18	137.8	Gene18	135.6
20	Gene19	153.7	Gene19	146.9	Gene19	122.0
21	Gene20	160.5	Gene20	162.8	Gene20	142.4
22	Gene21	124.2	Gene21	119.7	Gene21	117.4

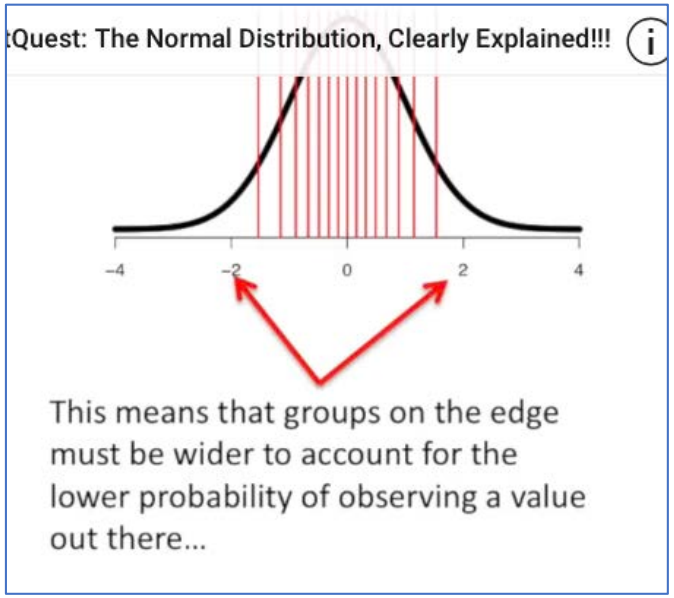
Quantile Quantile PLOTS

https://youtu.be/okjYjClSjOg



D2 | =1/(B2+1)

A	B	C	D	E	F	G
INPUT your sorted data list in this column (auto-copied to col F)	INPUT in cell A2 below the desired number of quantiles (1000max):	Select B3 (says "2") & drag down to fill desired # of QUANTILES (*matching column "A" rows*)	uniformly spaced QUANTILES (use these as "standardized" probabilities)	QUANTILES in Z-scores (# STDEVs from the mean)	copy of col A	Now select the values in columns E and F and insert a scatter plot
7.64385619	10	1	0.090909091	-1.335177736	7.64385619	
8.22881869		2	0.181818182	-0.908457869	8.22881869	
8.64385619		3	0.272727273	-0.604585347	8.64385619	
8.965784285		4	0.363636364	-0.348755696	8.96578428	
9.22881869		5	0.454545455	-0.114185294	9.22881869	
9.451211112		6	0.545454545	0.114185294	9.45121111	
9.64385619		7	0.636363636	0.348755696	9.64385619	
9.813781191		8	0.727272727	0.604585347	9.81378119	
9.965784285		9	0.818181818	0.908457869	9.96578428	
10.96578428		10	0.909090909	1.335177736	10.9657843	

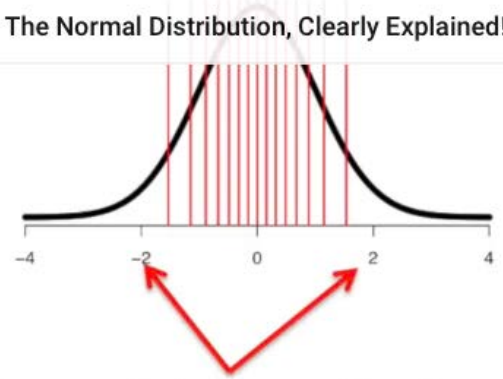


The formula for the rest of the column is a little different. Drag it fill down :

=D2+1/(B2+1)

	D	
ays "2") & fill desired NTILES *olumn "A"* s)	uniformly spaced QUANTILES (use these as "standardized" probabilities)	QUA scor fro
	0.090909091	
	0.181818182	
	0.272727273	

Quest: The Normal Distribution, Clearly Explained!!!



This means that groups on the edge must be wider to account for the lower probability of observing a value out there...

E2 =NORM.S.INV(D2)

	A	B	C	D	E	F
1	INPUT your sorted data list in this column (auto-copied to col F)	INPUT in cell A2 below the desired number of quantiles (1000max):	Select B3 (says "2") & drag down to fill desired # of QUANTILES (*matching column "A" rows*)	uniformly spaced QUANTILES (use these as "standardized" probabilities)	QUANTILES in Z-scores (# STDEVs from the mean)	copy of col A
2	7.64385619	10	1	0.090909091	-1.335177736	7.64385619
3	8.22881869		2	0.181818182	-0.908457869	8.22881869
4	8.64385619		3	0.272727273	-0.604585347	8.64385619
5	8.965784285		4	0.363636364	-0.348755696	8.96578428
6	9.22881869		5	0.454545455	-0.114185294	9.22881869
7	9.451211112		6	0.545454545	0.114185294	9.45121111
8	9.64385619		7	0.636363636	0.348755696	9.64385619
9	9.813781191		8	0.727272727	0.604585347	9.81378119
10	9.965784285		9	0.818181818	0.908457869	9.96578428
11	10.96578428		10	0.909090909	1.335177736	10.9657843

Select B3 (says "2") & drag down to fill desired # of QUANTILES (*matching column "A" rows*)	uniformly spaced QUANTILES (use these as "standardized" probabilities)	QUANTILES in Z-scores (# STDEVs from the mean)	copy of col A	Now select the values in columns E and F and insert a scatter plot
1	0.090909091	-1.335177736	7.64385619	
2	0.181818182	-0.908457869	8.22881869	
3	0.272727273	-0.604585347	8.64385619	
4	0.363636364	-0.348755696	8.96578428	
5	0.454545455	-0.114185294	9.22881869	
6	0.545454545	0.114185294	9.45121111	
7	0.636363636	0.348755696	9.64385619	
8	0.727272727	0.604585347	9.81378119	
9	0.818181818	0.908457869	9.96578428	
10	0.909090909	1.335177736	10.9657843