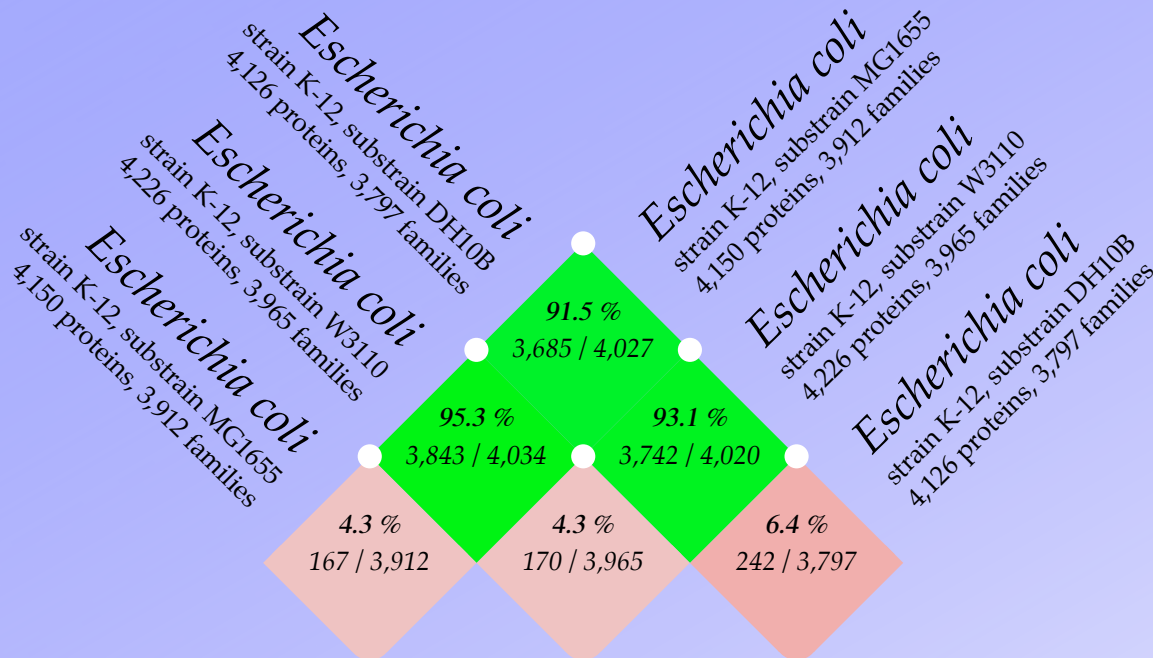


# Introduction to BLAST Matrices and BLAST Atlases

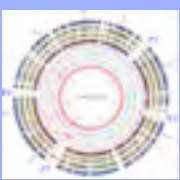


Dave Ussery

Workshop on Comparative Genomics  
King Mongkut's University of Technology Thonburi  
Bangkok, Thailand

Computer Exercises for Wednesday  
10 March, 2010





# When are two proteins the same??

50% length of query

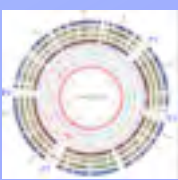


Query sequence (protein)

Subject sequence (protein)

50% identity of match

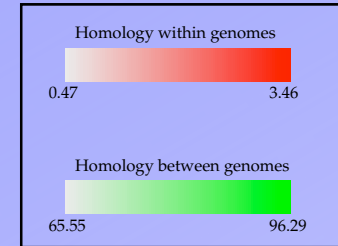
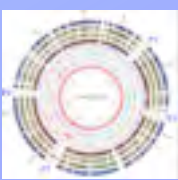
Disclaimer - the "50-50 rule"



## Proteome comparison of *Escherichia coli*

ALR: 0.75, e-value 1e-10

<i>E. coli</i> CFT073 5379 genes	<i>E. coli</i> K-12 DH108 4126 genes	<i>E. coli</i> K-12 W3110 4226 genes	<i>E. coli</i> K-12 MG1655 4232 genes	
3997 / 5379 66.9%	3979 / 4126 96.4%	4059 / 4226 96.0%	310 / 4232 7.5%	<i>E. coli</i> K-12 MG1655 4232 genes
3616 / 5379 67.2%	4036 / 4126 97.8%	356 / 4226 8.4%	4016 / 4232 97.2%	<i>E. coli</i> K-12 W3110 4226 genes
3486 / 5379 64.8%	534 / 4126 12.9%	3738 / 4226 88.5%	3633 / 4232 87.9%	<i>E. coli</i> K-12 DH108 4126 genes
642 / 5379 11.9%	3513 / 4126 85.1%	3378 / 4226 79.9%	3291 / 4232 79.6%	<i>E. coli</i> CFT073 5379 genes



<i>Escherichia coli</i> 042 4898 genes, bp	<i>Escherichia coli</i> E2348 3849 genes, bp	<i>Escherichia coli</i> O157.RIMD0509952 5361 genes, bp	<i>Escherichia coli</i> O157.EDL93 5349 genes, bp	<i>Escherichia coli</i> K-12.W3110 4337 genes, bp	<i>Escherichia coli</i> K-12.MG1655 4254 genes, bp	<i>Escherichia coli</i> CFT073 5379 genes, bp
4039 / 5379 75.1%	3526 / 5379 65.6%	3894 / 5379 72.4%	3890 / 5379 72.3%	3792 / 5379 70.5%	3799 / 5379 70.6%	50 / 5379 0.9%
3749 / 4254 88.1%	3185 / 4254 74.9%	3775 / 4254 88.7%	3786 / 4254 89.0%	4096 / 4254 96.3%	36 / 4254 0.8%	3675 / 4254 86.4%
3763 / 4337 86.8%	3208 / 4337 74.0%	3745 / 4337 86.4%	3755 / 4337 86.6%	50 / 4337 1.2%	4120 / 4337 95.0%	3686 / 4337 85.0%
4097 / 5349 76.6%	3583 / 5349 67.0%	4920 / 5349 92.0%	185 / 5349 3.5%	3857 / 5349 72.1%	3888 / 5349 72.7%	3962 / 5349 74.1%
4126 / 5361 77.0%	3636 / 5361 67.8%	134 / 5361 2.5%	4920 / 5361 91.8%	3860 / 5361 72.0%	3884 / 5361 72.4%	3970 / 5361 74.1%
3133 / 3849 81.4%	46 / 3849 1.2%	3189 / 3849 82.9%	3185 / 3849 82.7%	3041 / 3849 79.0%	3029 / 3849 78.7%	3217 / 3849 83.6%
23 / 4898 0.5%	3364 / 4898 68.7%	4001 / 4898 81.7%	3979 / 4898 81.2%	3824 / 4898 78.1%	3824 / 4898 78.1%	3987 / 4898 81.4%

*Escherichia coli*  
CFT073  
5379 genes, 5,231,428 bp

*Escherichia coli*  
K-12.MG1655  
4254 genes, 4,639,675 bp

*Escherichia coli*  
K-12.W3110  
4337 genes, 4,641,433 bp

*Escherichia coli*  
O157.EDL93  
5349 genes, 5,528,445 bp

*Escherichia coli*  
O157.RIMD0509952  
5361 genes, 5,498,450 bp

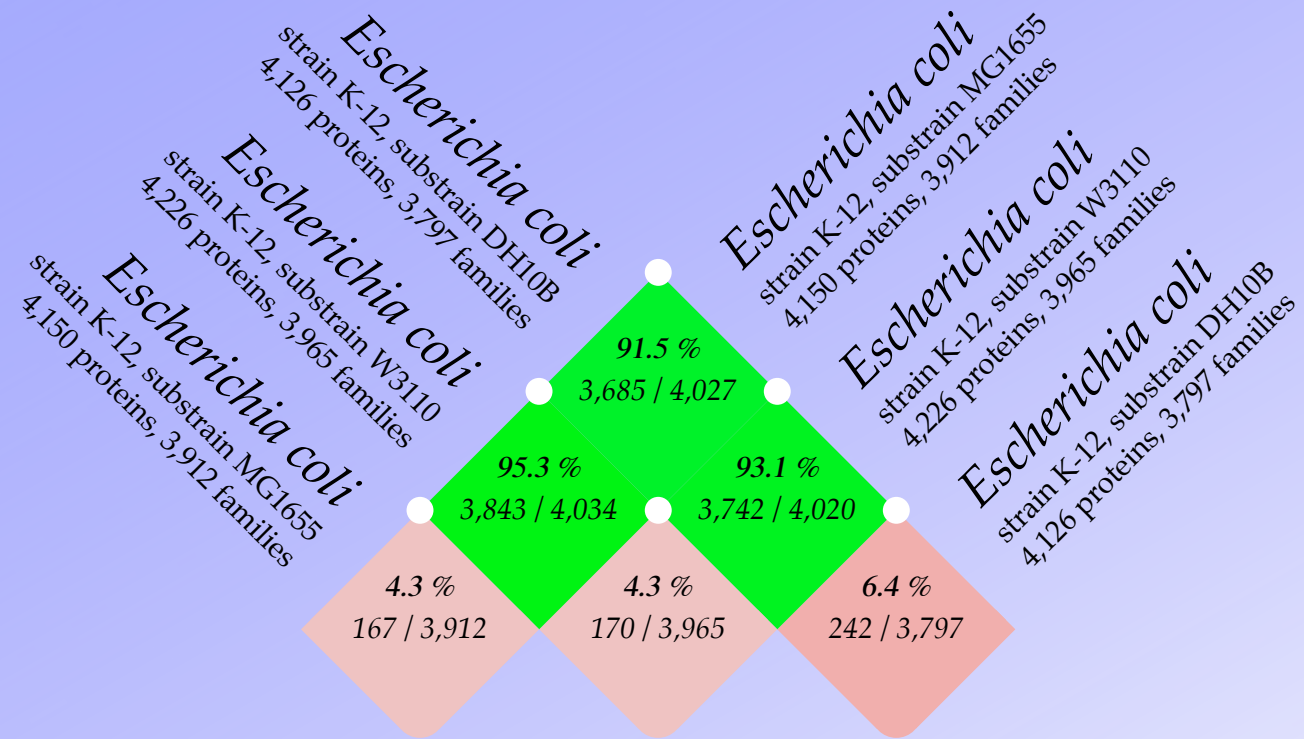
*Escherichia coli*  
E2348  
3849 genes, 4,227,846 bp

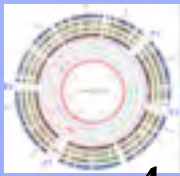
*Escherichia coli*  
042  
4898 genes, 5,241,977 bp



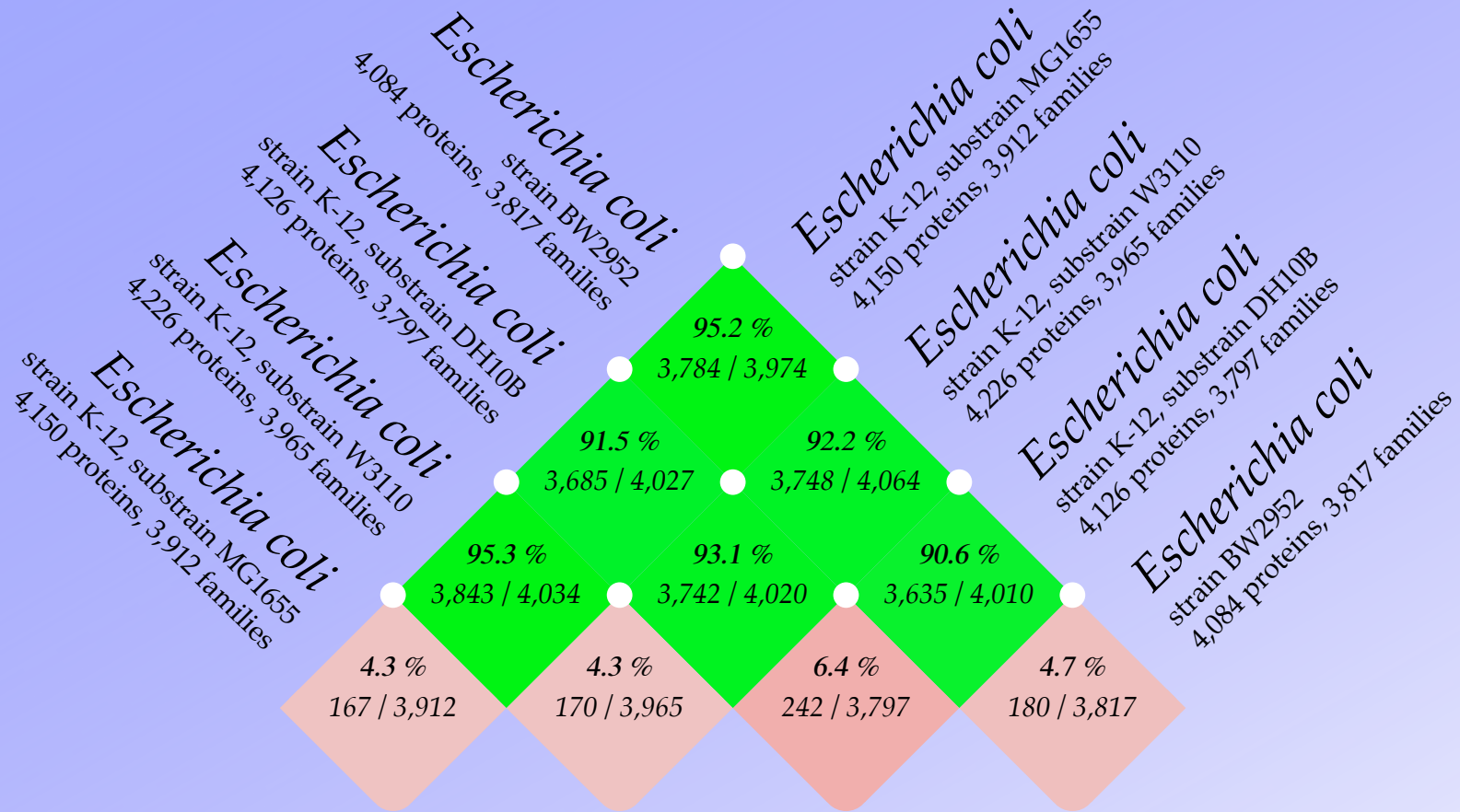


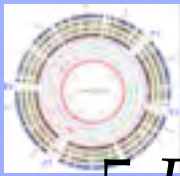
# 3 *E. coli* K-12 genomes



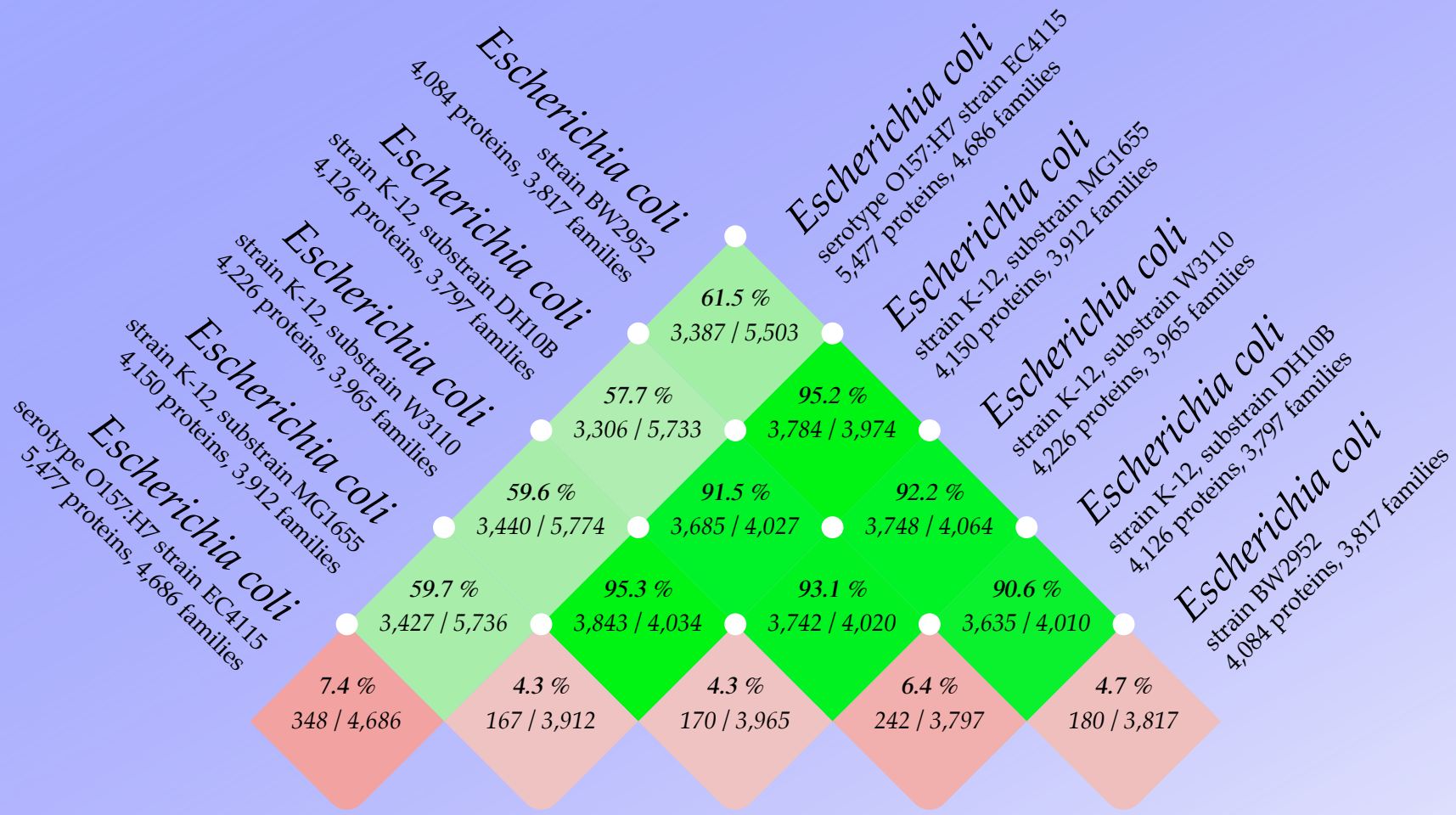


# 4 *E. coli* genomes





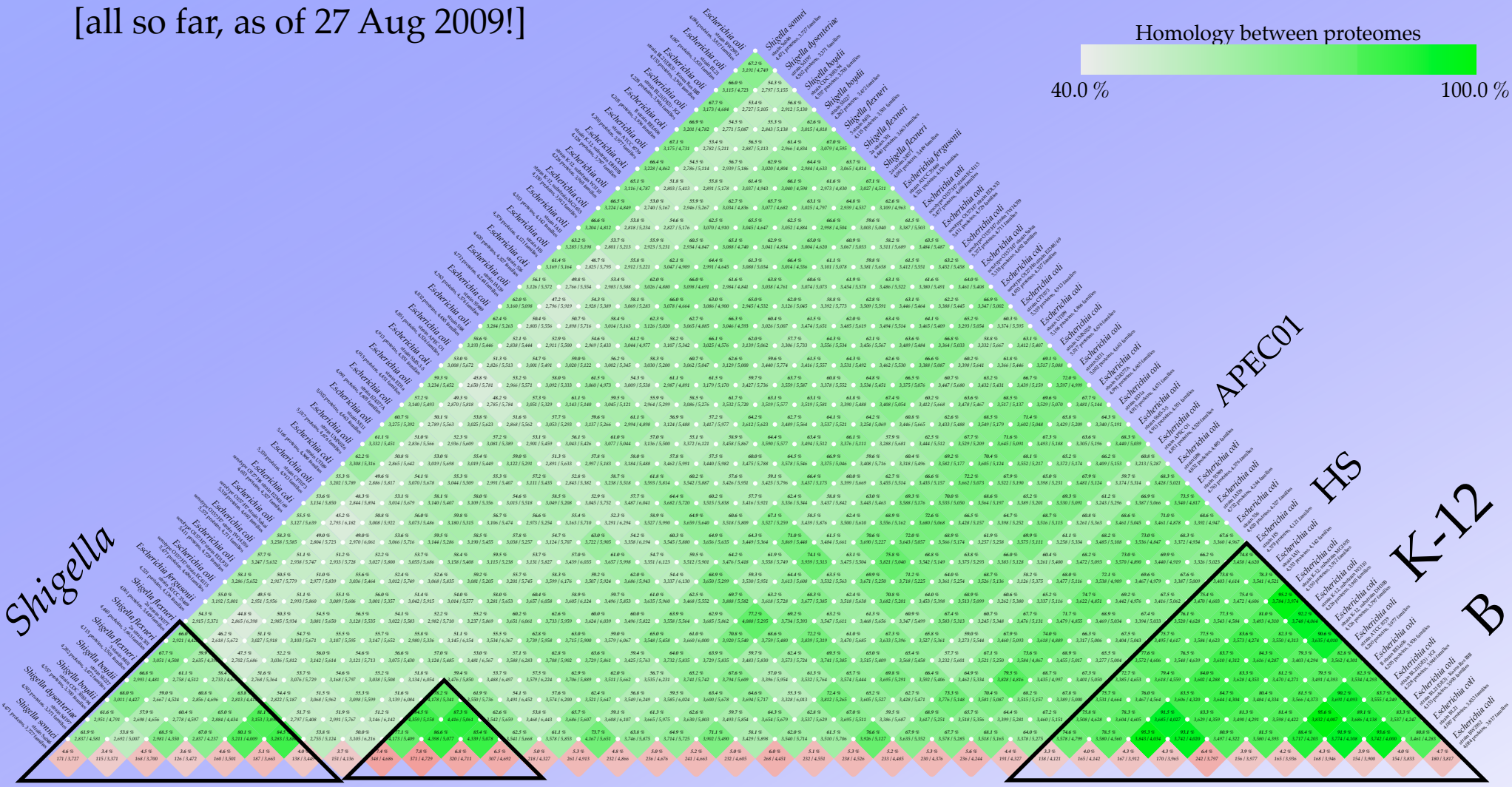
# 5 *E. coli* genomes





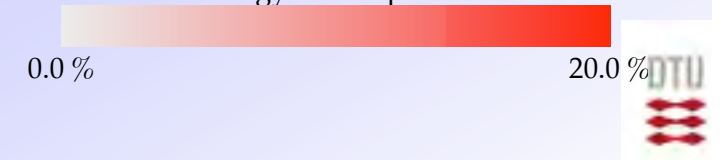
# 28 E. coli genomes

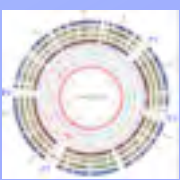
[all so far, as of 27 Aug 2009!]



## O157:H7

## Homology within proteomes



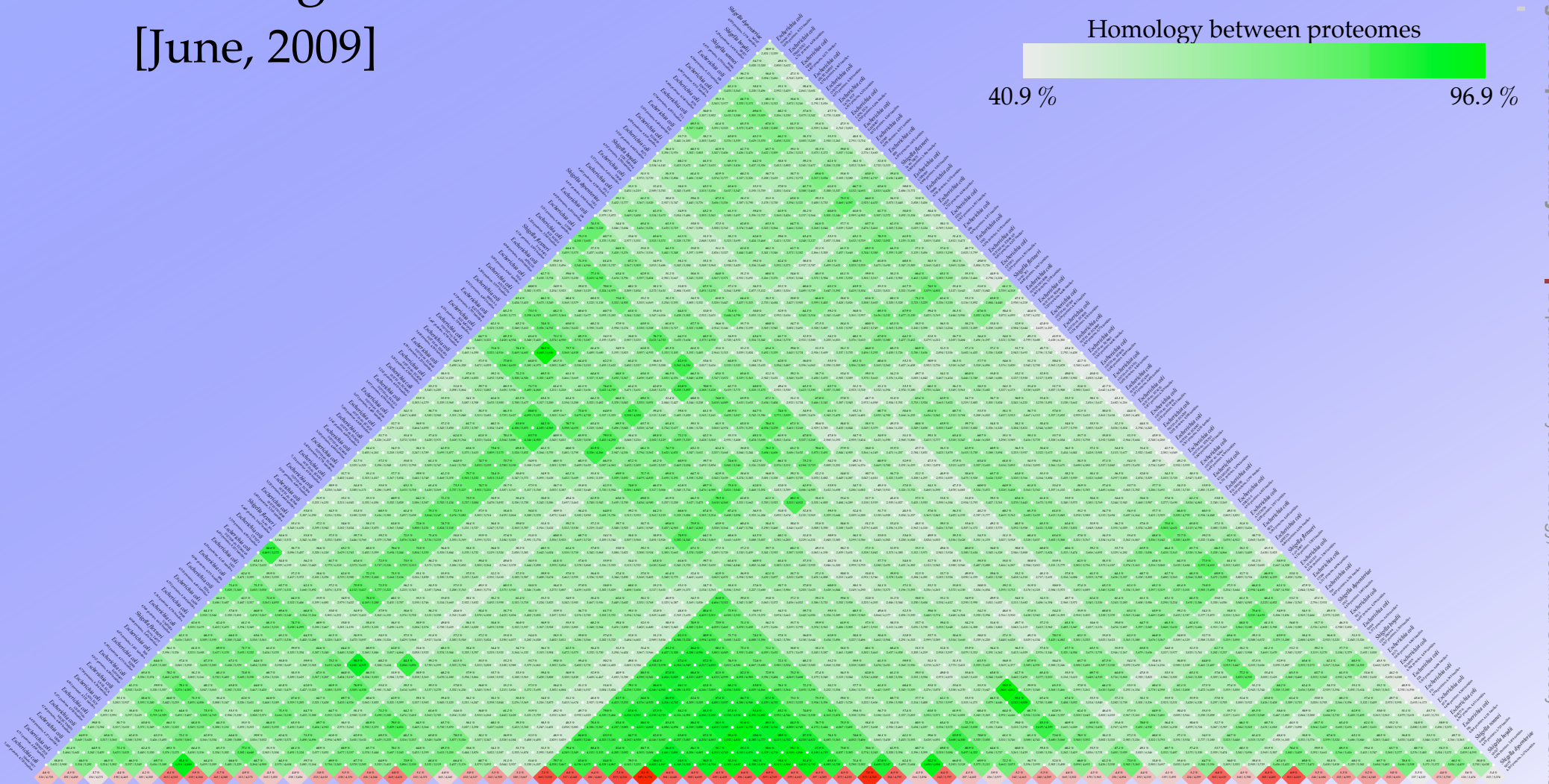


# 60 *E. coli* genomes [June, 2009]

Homology between proteomes

40.9%

96.9%



Homology within proteomes

3.3%

8.1%

