Estimate Gene Family Novelty

Why?

1. identify novel families
2. identify organisms, even clades with more share of novel genes

How to quantify novelty

Strategy:

1. Gene Novelty
2. Family Novelty: Percentage of Novel genes in a family

(gene family novelty calculation is gene based so it is easy for different family classification)

Gene that is not covered by **Known** Pfam and TIGRfam is a novel gene

1. GO based
   Some Pfam, TIGRfam and INTEPROT have Gene Ontology assignments (GO)
2. The rest need manual decision

Manual Decision Keywords

Function

utilization factor coat assembly kinase esterase antigen receptor isomerase dehydrogenase Fibronectin regulators permease hemolysin activity reductase dehydratase enzyme binding nodulation phosphatase ATPase rubredoxin hydrolase polymerase synthesis regulate transport repair hexon lyase transfer induced ipoprotein export portal terminase proteolysis transposon transposition sporulation GTPase myosin Translocator transposase phatase processing bind ThiF toxin repressor mobilisation protease tol-pal secretion immunoglobulin glcoprotein sorting ligase Recombinase hydratase exosortase sortase exonuclease CRISPR nuclease xylanase helicase peptidase integrase mutase carboxypetidase aminase dextrinase cyclase carboxylase proteinase synthetase synthase primase ribosome lipase invertase oxygenase nitrogenase translocase amidase oxidase methylases excisionase hydantoinase elongation deacetylase disaggregatase anhydrase urease sialidase transacetylase interaction DegV stabilisation regulatory Antiporter exchanger adhesin secretory catalyse translation homeobox signal expressed tail phosphoprotein shock death FlhB acceptor fixation core metabolism YoaH essential glycoprotein resistance replication catalytic cytochrome associated SlyX SlyD division initiation finger photosynthe packag spherulation apopto redox structural production adhesive stress pathway P51 matrix movement iron-sulphur immediate-early calponin Autophagy RPE-encoded adhesion activator endospore cleavage modification methylation symporter flagellar regulator migration DNA-K methanogenesis PsbQ interact biogenesis integrin arabinosidase development ribosomal photosystem cytoskele collagen golgi nexin ubiquitination pheromone metal-dependent mucin complex ferritin HBS1-like effector syndrome differentiation SepQ Retinin radical suppressor silk KaiC IPR001233 IPR006487 IPR003366 IPR005616 IPR005187 IPR004145 IPR015400 prophage lysis morphogenesis spliceosome bZIP cataboli conjugative SAFF antimicrobial IPR006791 IPR006798 IPR006803 IPR006853 IPR006872 IPR017508 protect formation propeller Gp37 nodulin capsid Flo11 acclimation AaeX tegument Pkr1 YecM accessory virulence channel Dnaj ubiquitin Facilitator prion IPR012983 IPR010931 IPR011101 IPR010292 IPR014907

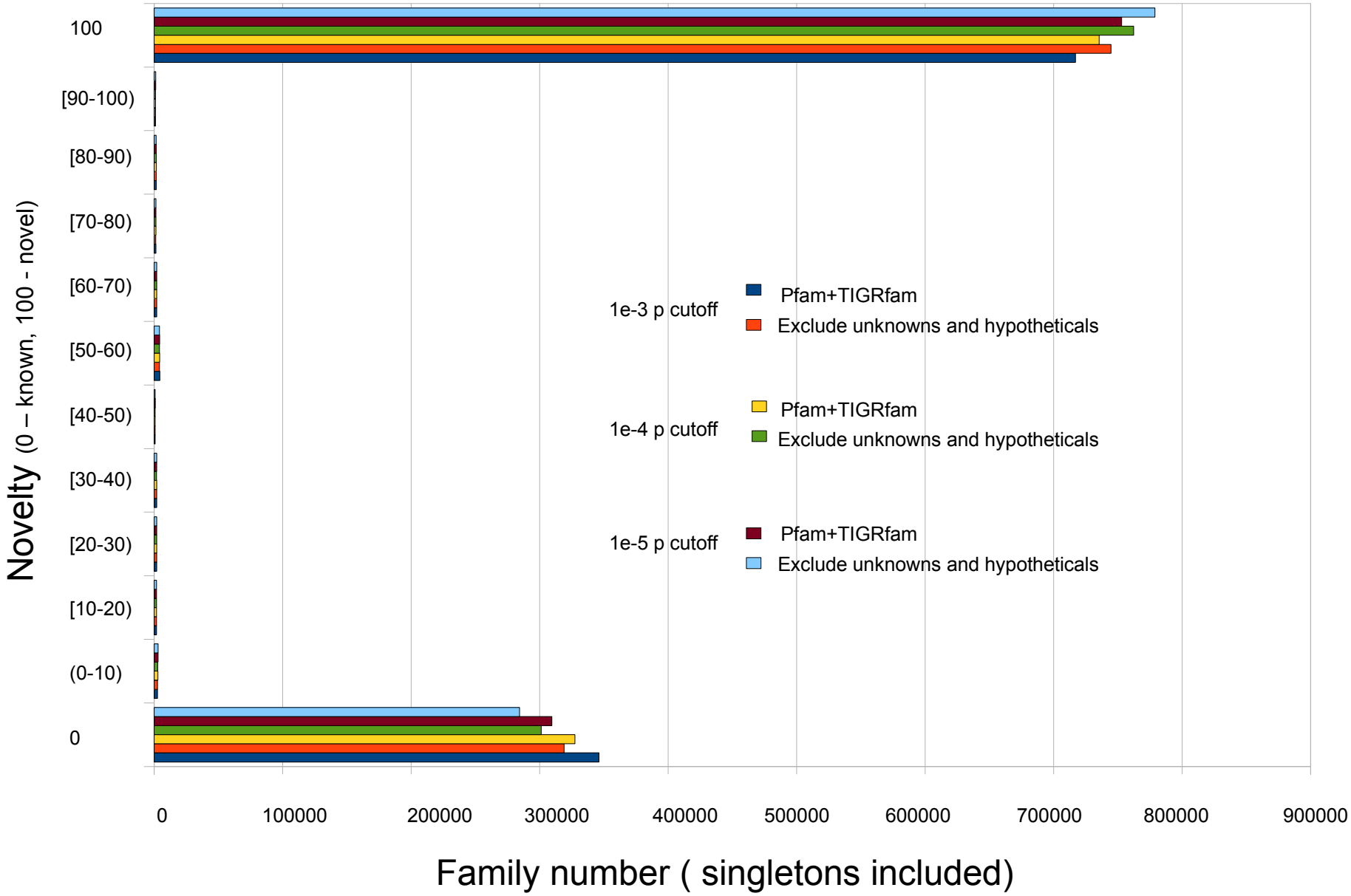
Structure

Zinc-ribbon GrpB barrel structure bridge knottin signature feature zipper unique hairpin 4FE-4S helix loop helices strand coil helical plastic IPR002791 IPR005303 IPR006491 IPR006970

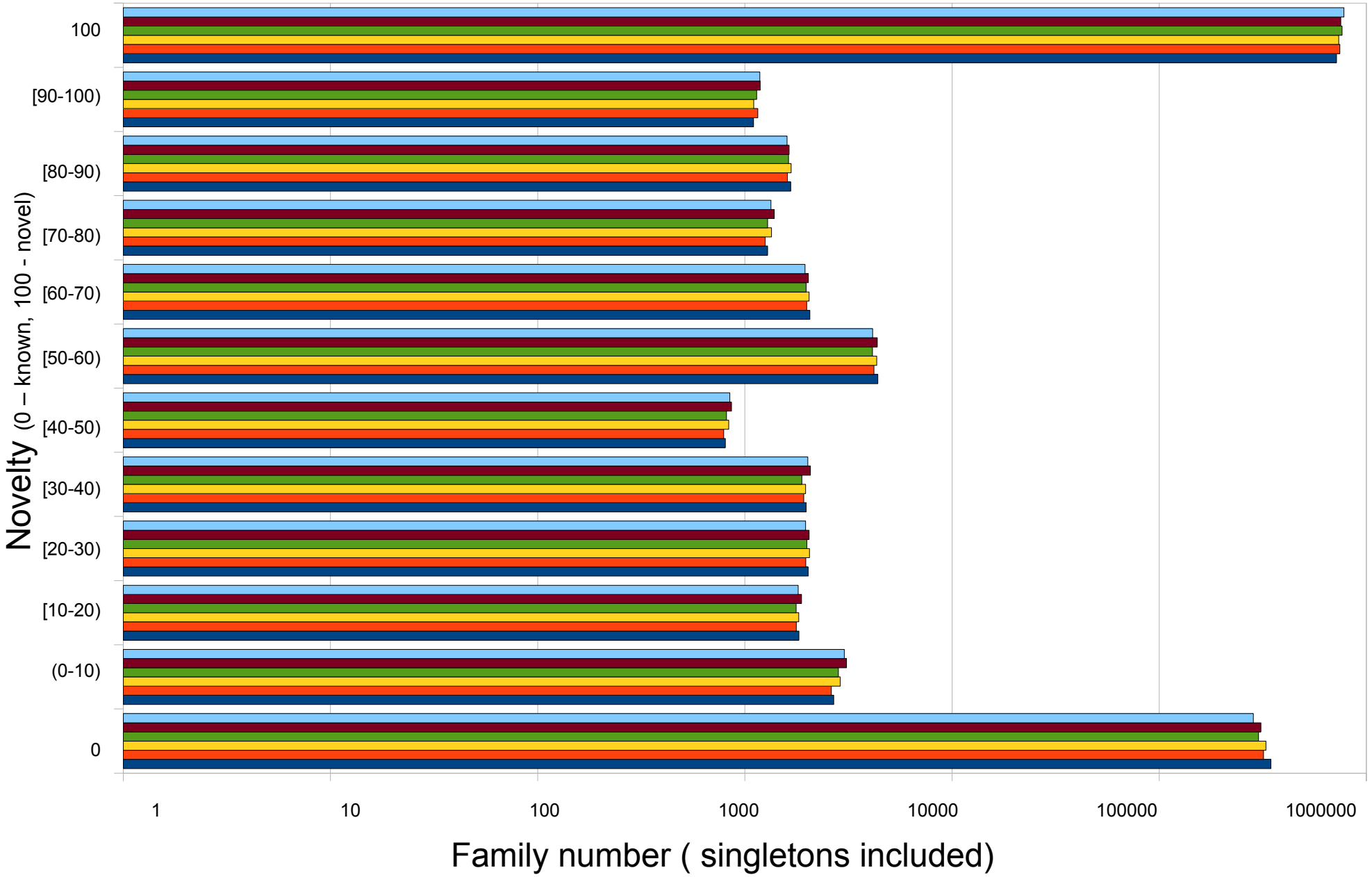
Target

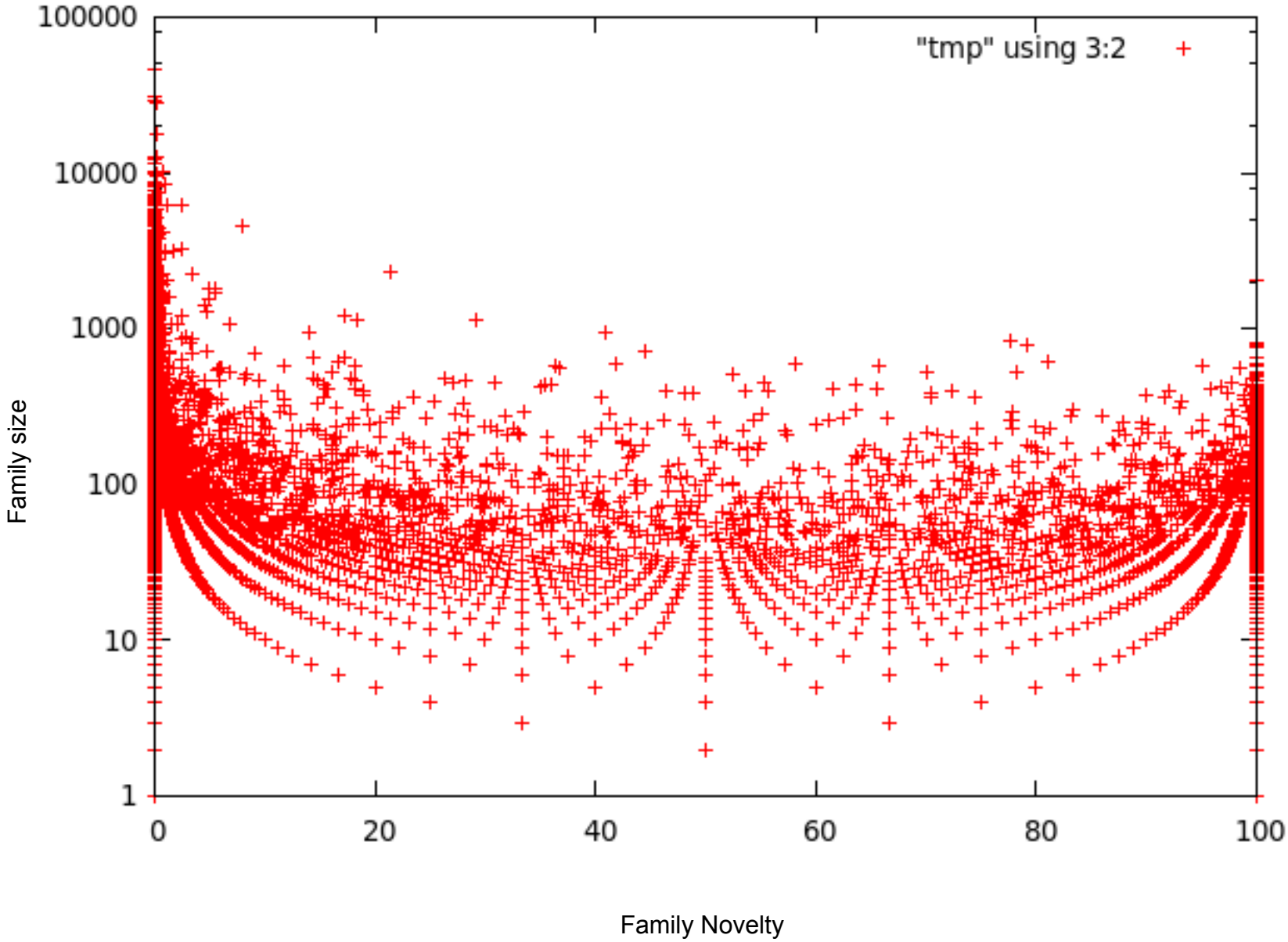
nucleocapsid envelope membrane surface extracellular layer secreted phycobilisome

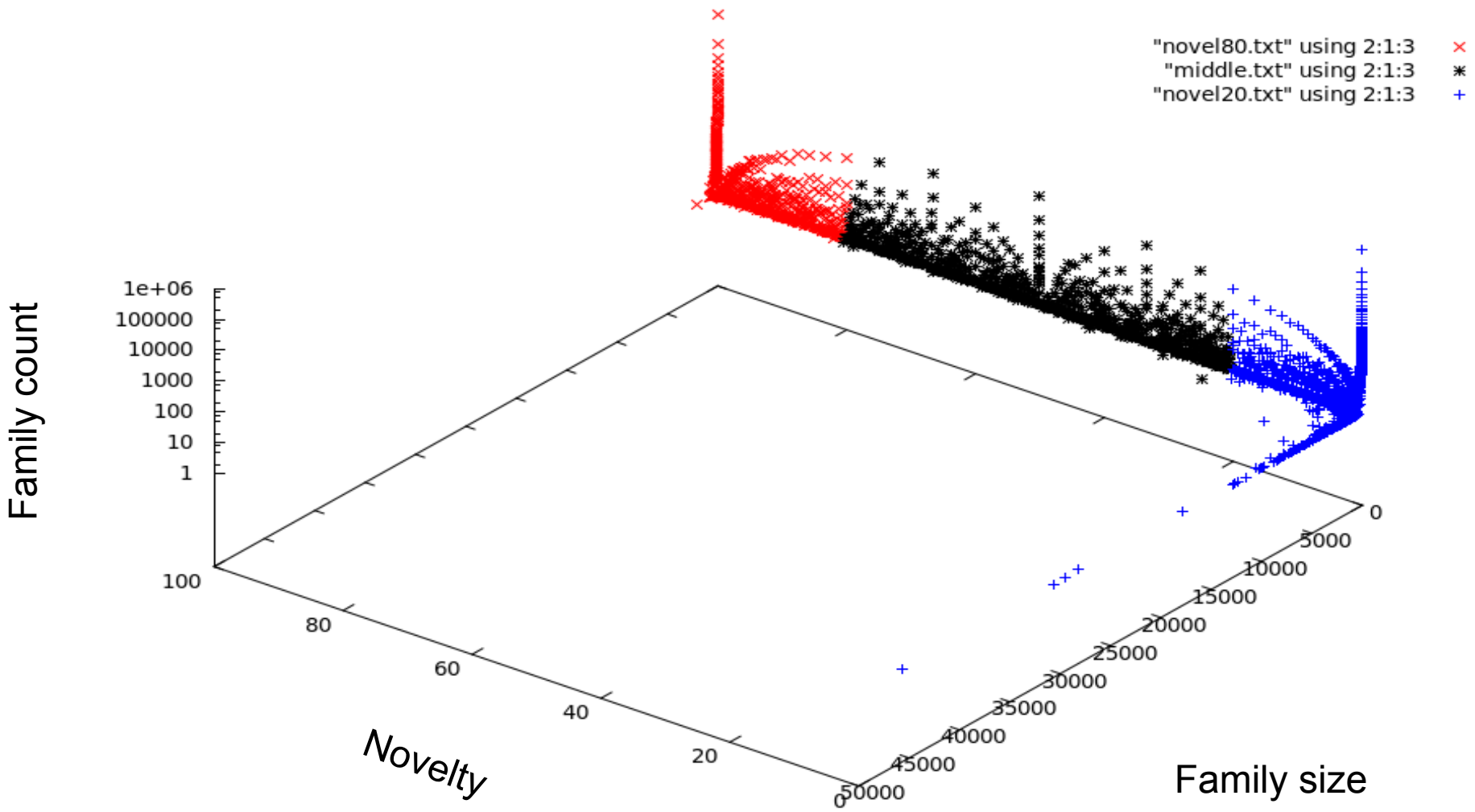
Pfam 11912 (9785 are regarded as being studied)
TIGRfam 3808 (3749 are regared as being stidied)

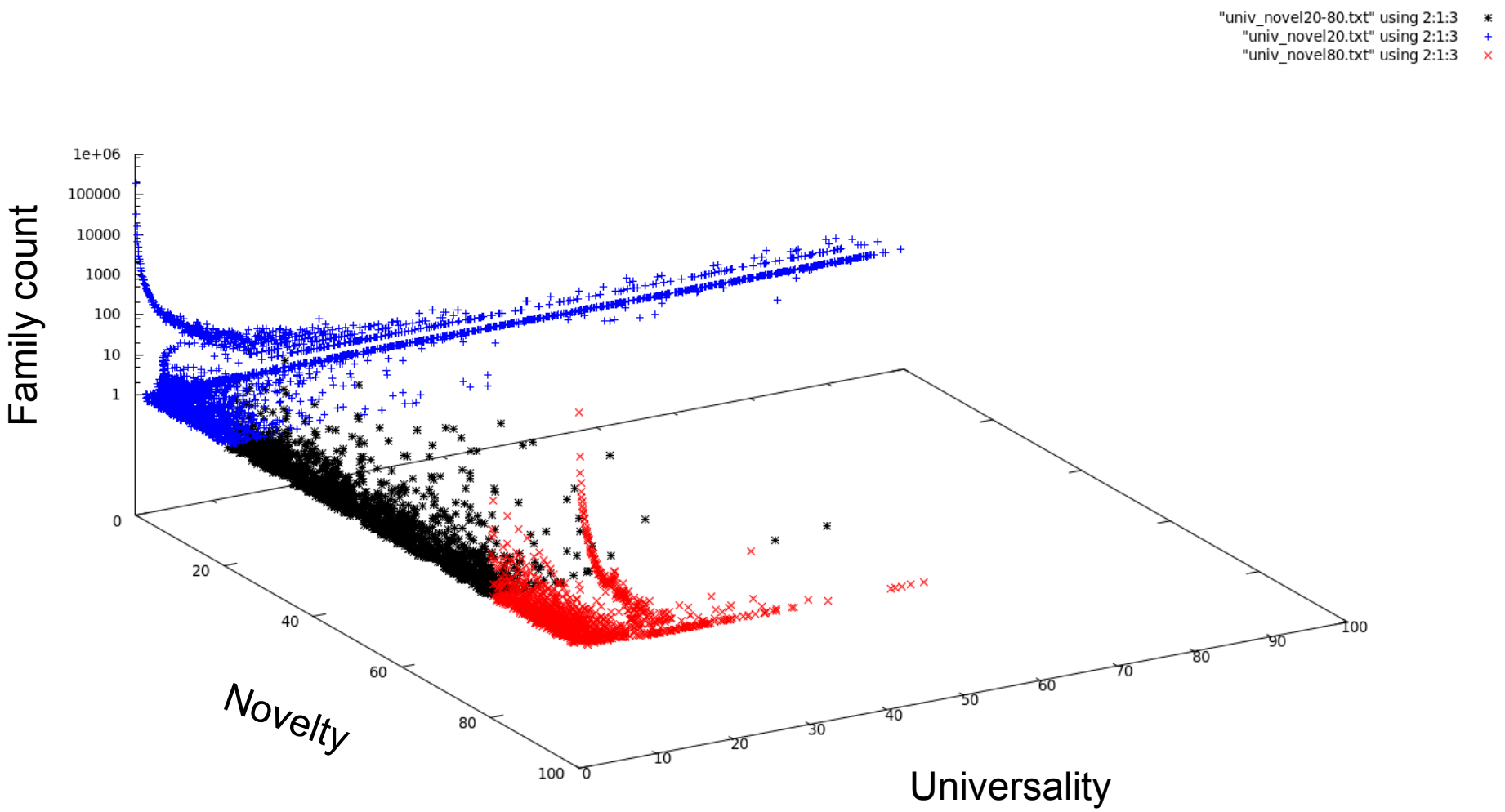


Guillaume built and hmmscan the families


Novelty (0 – known, 100 - novel)
100
[90-100)
[80-90)
[70-80)
[60-70)
[50-60)
[40-50)
[30-40)
[20-30)
[10-20)
(0-10)
0
1
10
100
1000
10000
100000
1000000
Family number (singletons included)

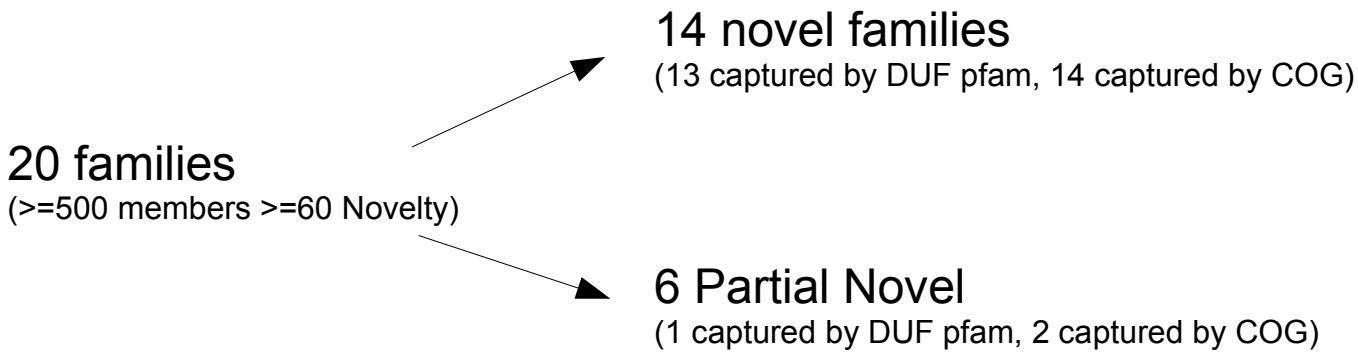

"tmp" using 3:2
100000
10000
1000
100
10
1
Family size
0
20
40
60
80
100
Family Novelty


"novel80.txt" using 2:1:3
"middle.txt" using 2:1:3
"novel20.txt" using 2:1:3
Family count
1e+06
100000
10000
1000
100
10
1
100
80
60
40
20
0
Novelty
0
5000
10000
15000
20000
25000
30000
35000
40000
45000
50000
Family size


"univ_novel20-80.txt" using 2:1:3
"univ_novel20.txt" using 2:1:3
"univ_novel80.txt" using 2:1:3
Family count
1e+06
100000
10000
1000
100
10
1
0
Novelty
0
20
40
60
80
100
Universality
0
10
20
30
40
50
60
70
80
90
100

Novel Large Families

FamilyID	Size	Novelty	Universality	Pfam DUF	COG	Verdict
2047	652	100	17.11	DUF421	COG2323	Novel
2331	522	78.16	11.94	-	COG1396	Novel, weak PF01381, transcriptional regulator only comes from the COG
2297	533	100	25.63	PF02583	COG1937	Novel
1543	2040	100	32.45	DUF2179 PF02588	COG1284	Novel
1908	755	100	41.09	PF02639	COG1671	Novel
2359	513	100	27.61	DUF488	COG3189	Novel
2124	601	100	28	DUF159	COG2135	Novel
1886	770	100	41.97	DUF328	COG3022	Novel
2145	560	98.57	30.69	PF02636	COG1565	Novel
2386	503	100	15.29	DUF849	COG3246	Novel
1851	755	100	40.54	PF02696	COG0397	Novel
2192	570	100	24.09	DUF74	COG0393	Novel
1838	815	100	44.83	DUF520	COG1666	Novel, nucleotide-binding proteins came from one paper about YajQ
1867	791	100	43.18	DUF179	COG1678	Novel, transcriptional regulator only comes from the COG
2299	532	70.11	1.93	-	-	Novel, week PF01381(helix-turn-helix domain)
346762	784	79.21	37.51	DUF45	COG1451(25.9%)	25.9% metal dependent hydrolase (COG) 17.6% metallopeptidase (PF08325)
2025	586	65.7	28.33	-	COG0435(28.7%)	34.8% PF00043 Glutathione S-transferase, 28.7% COG0435
1787	827	77.63	45.1	-	COG1956(36.3%)	22.4% weak PF01590 GAF domain, 36.3% COG1956
2182	571	95.1	25.47	-	COG3655	Novel, transcriptional regulator only comes from the COG
2121	607	81.05	33.33	-	COG2996	COG2996 Uncharacterized protein, 115 weak PF00575 hits (RNA binding)



Why 191205 singletons are not novel?

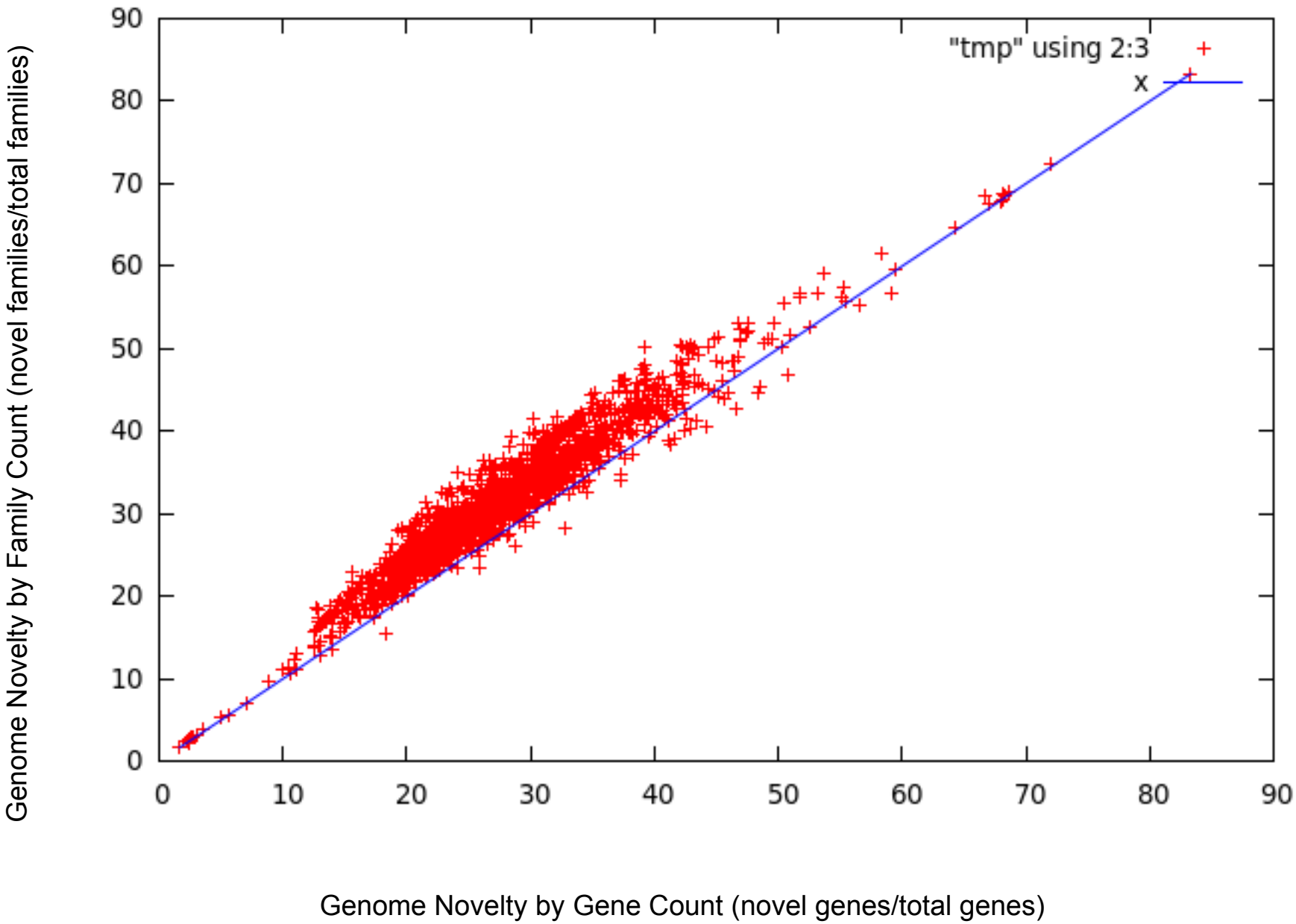
I looked at 20 random such singletons:

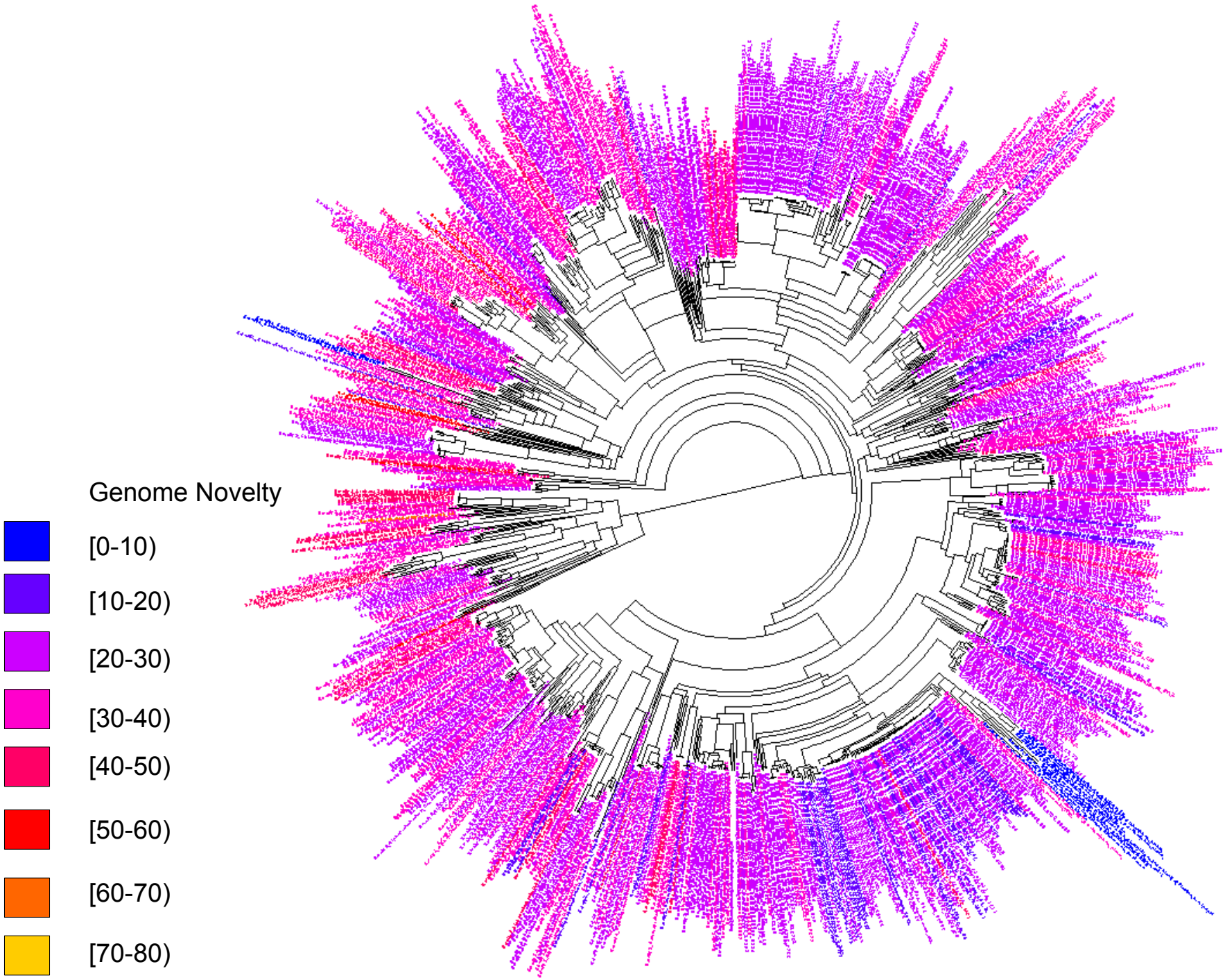
16 partial Pfam hits
2 multi-domains
2 complete domain hits

(the 80% span cutoff in family building is the main reason)

Genome Novelty?

The average of gene family novelties in a genome


"tmp" using 2:3
x
Genome Novelty by Family Count (novel families/total families)
Genome Novelty by Gene Count (novel genes/total genes)
0
10
20
30
40
50
60
70
80
90

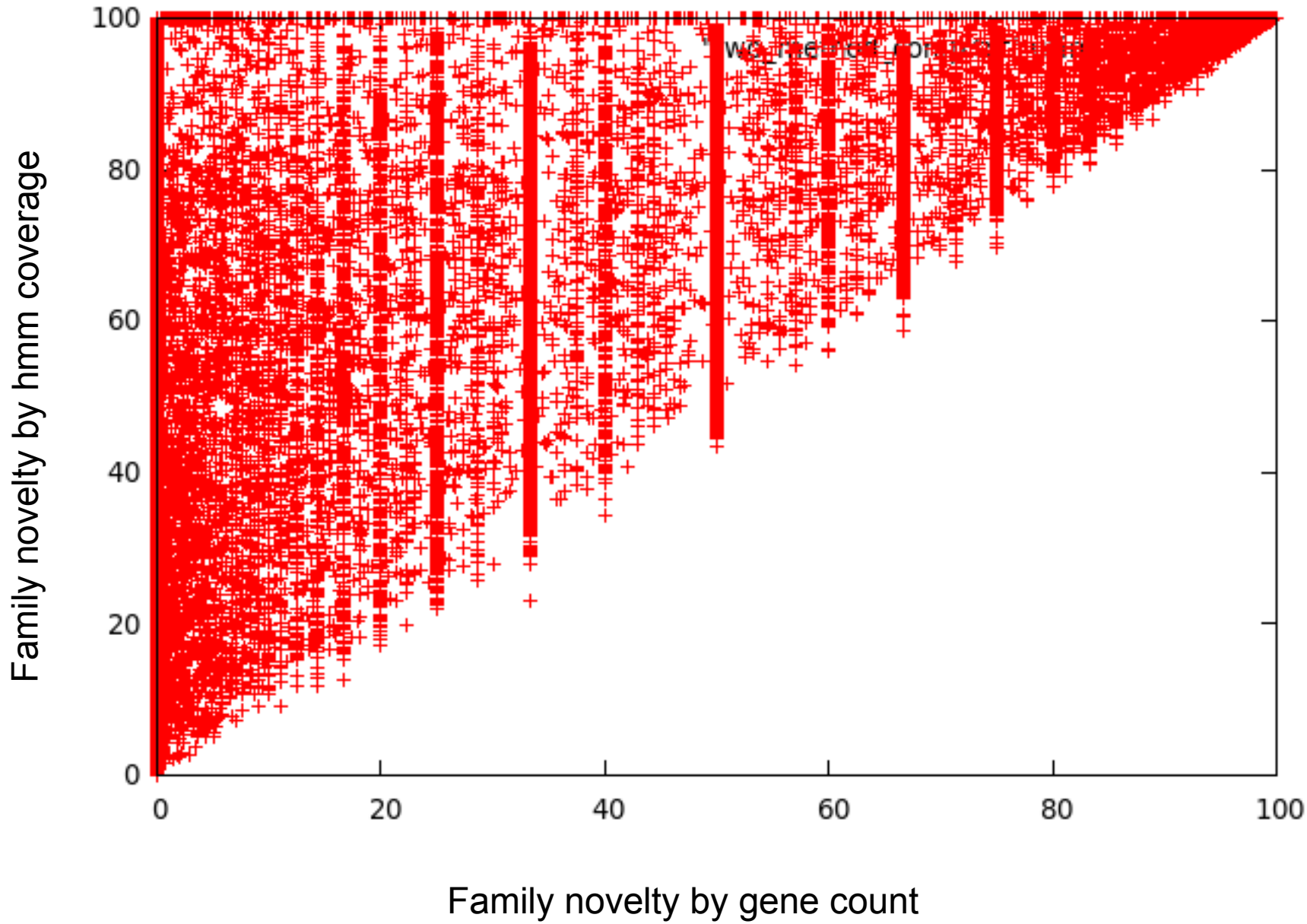

Genome Novelty
[0-10)
[10-20)
[20-30)
[30-40)
[40-50)
[50-60)
[60-70)
[70-80)

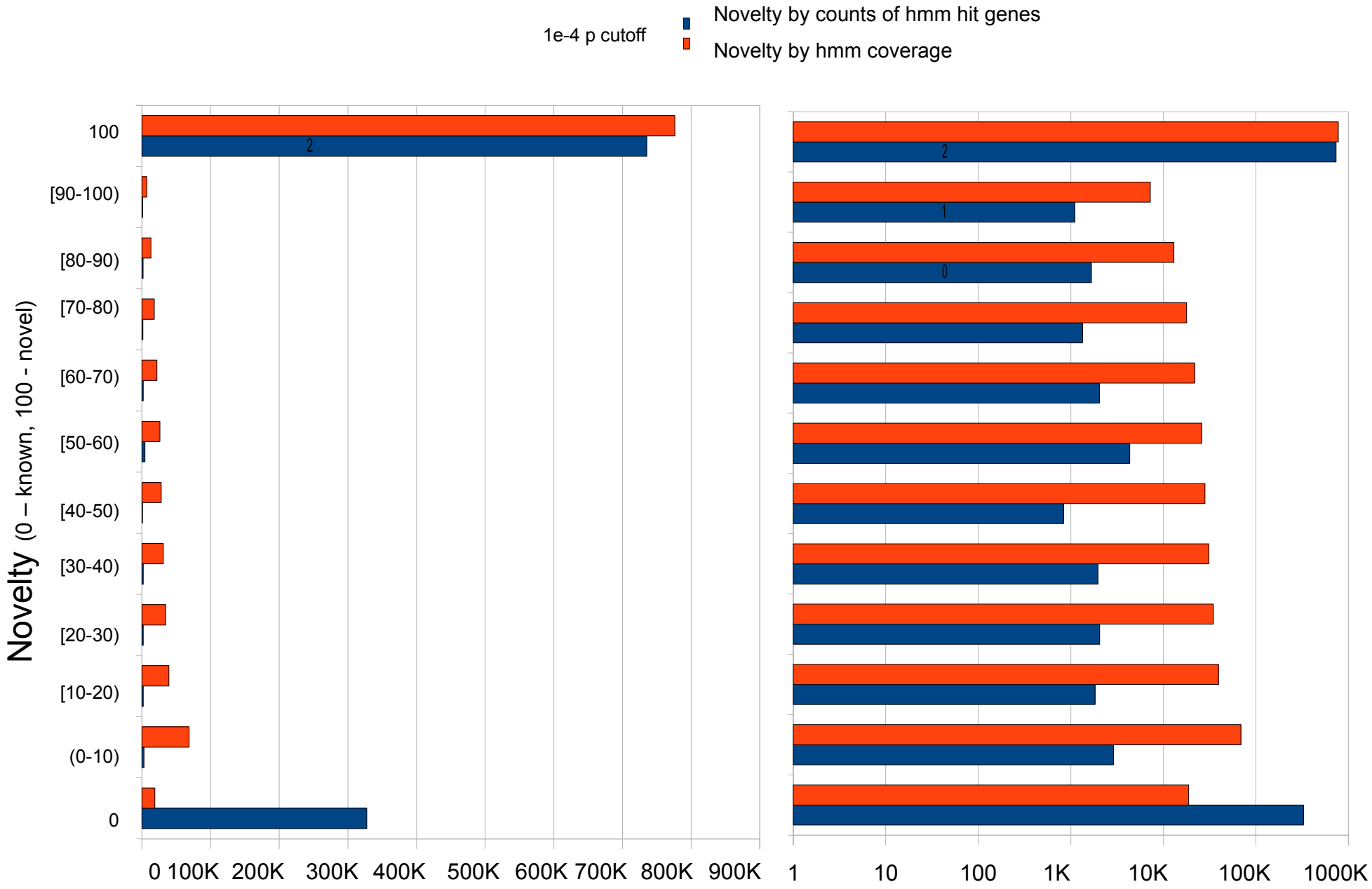
Estimate Family Novelties by HMM coverage

Previous Equation regard the gene as non-novel, with genome novelty set as 0

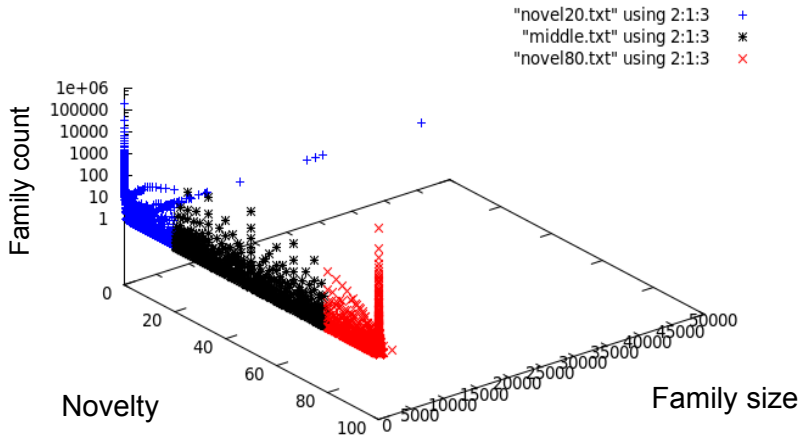
If we only consider the region covered by HMM, this gene novelty is 0.90

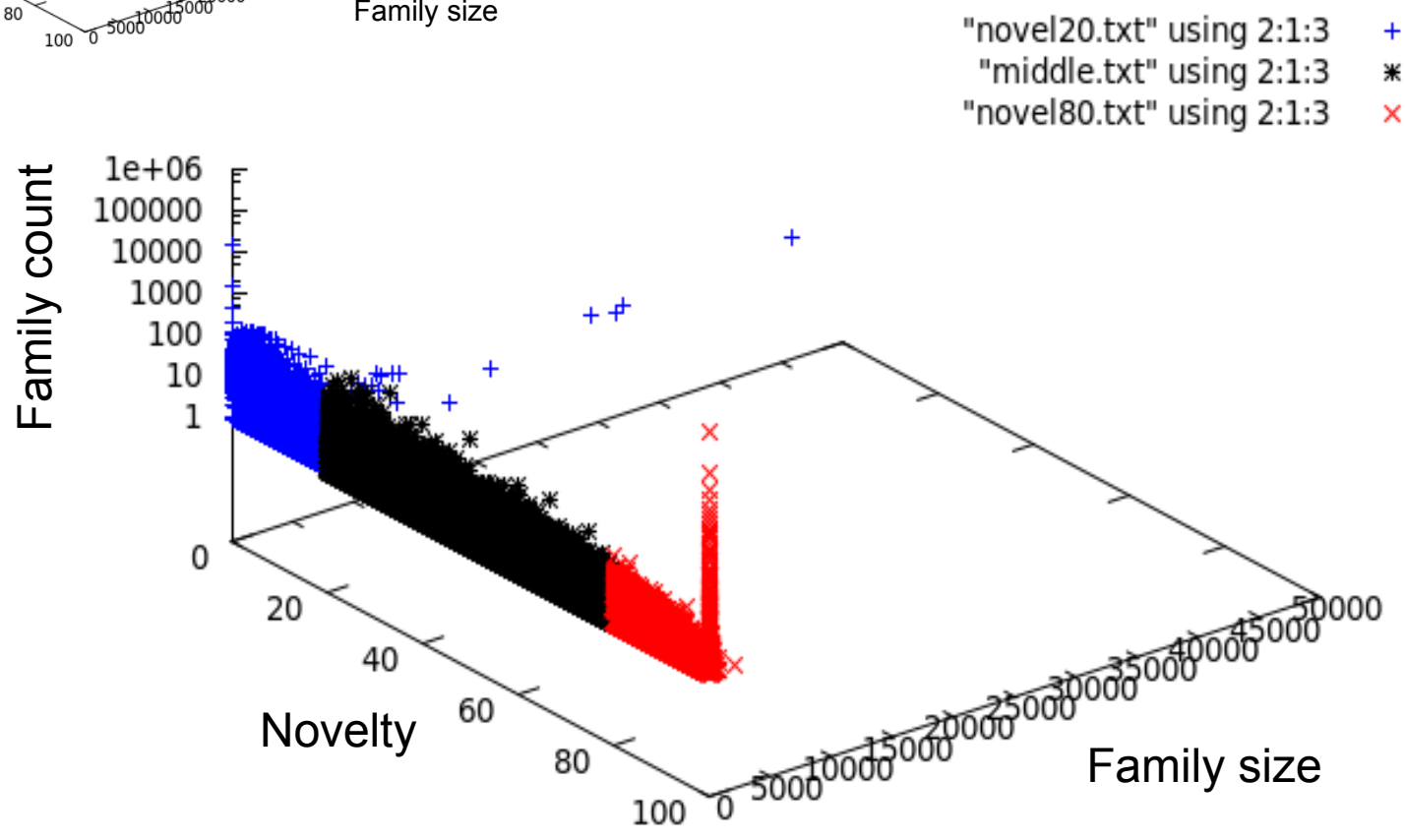
Gene Family Novelty = HMM covered sequence lengths/total sequence lengths


Family novelty by hmm coverage
Family novelty by gene count
0
20
40
60
80
100


1e-4 p cutoff
Novelty by counts of hmm hit genes
Novelty by hmm coverage
Novelty (0 – known, 100 - novel)
100
[90-100)
[80-90)
[70-80)
[60-70)
[50-60)
[40-50)
[30-40)
[20-30)
[10-20)
(0-10)
0
2
0 100K 200K 300K 400K 500K 600K 700K 800K 900K
1
0
1 10 100 1K 10K 100K 1000K

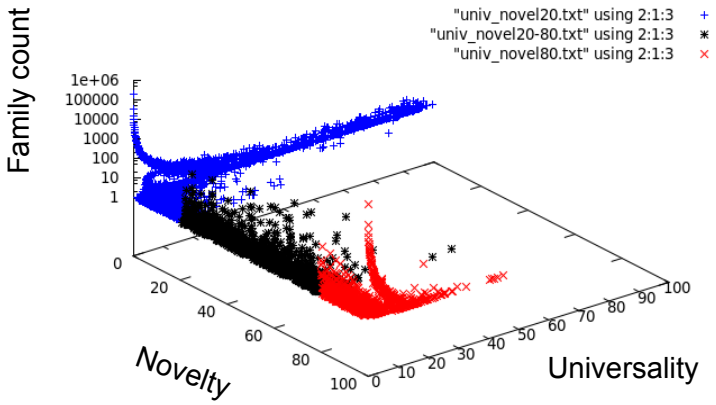
Novelty by counts of hmm hit genes

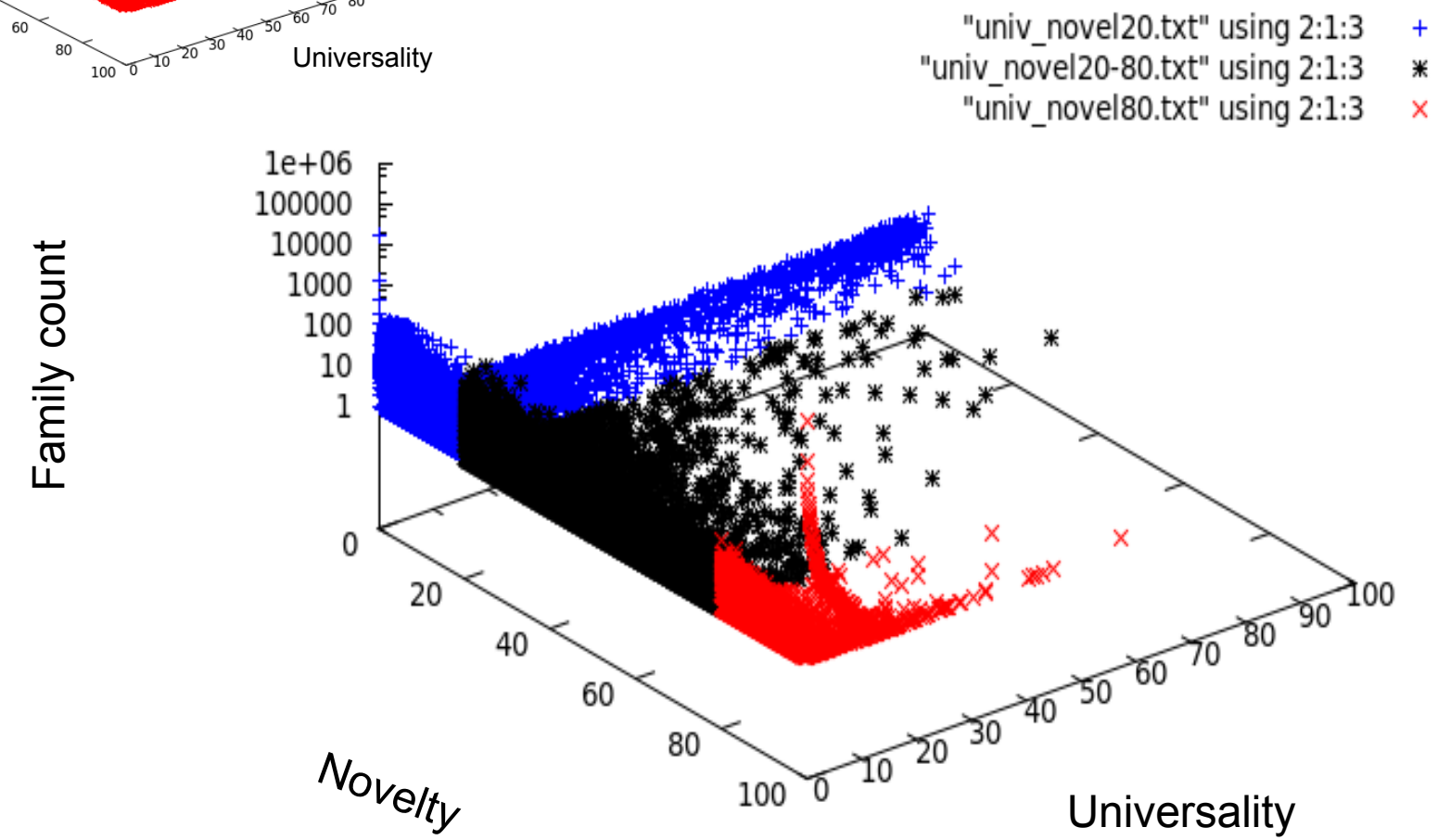




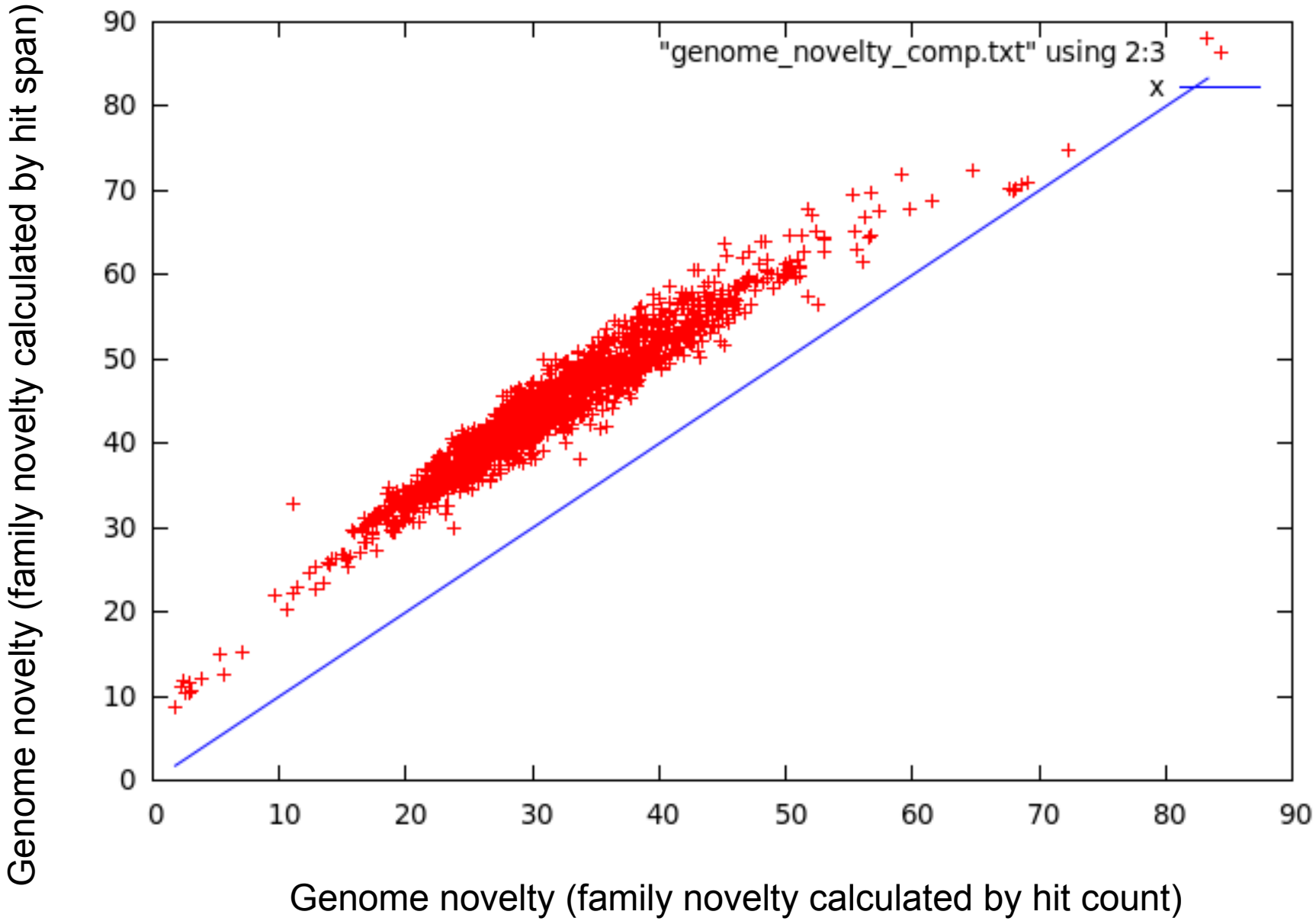
Novelty by hmm coverage

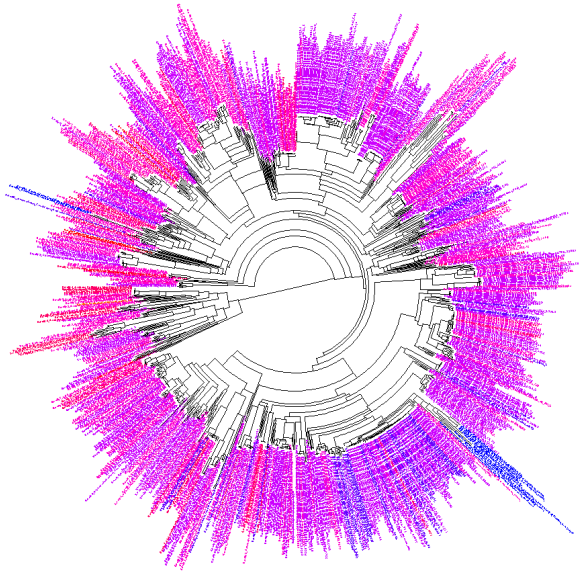
Novelty by counts of hmm hit genes





Novelty by hmm coverage

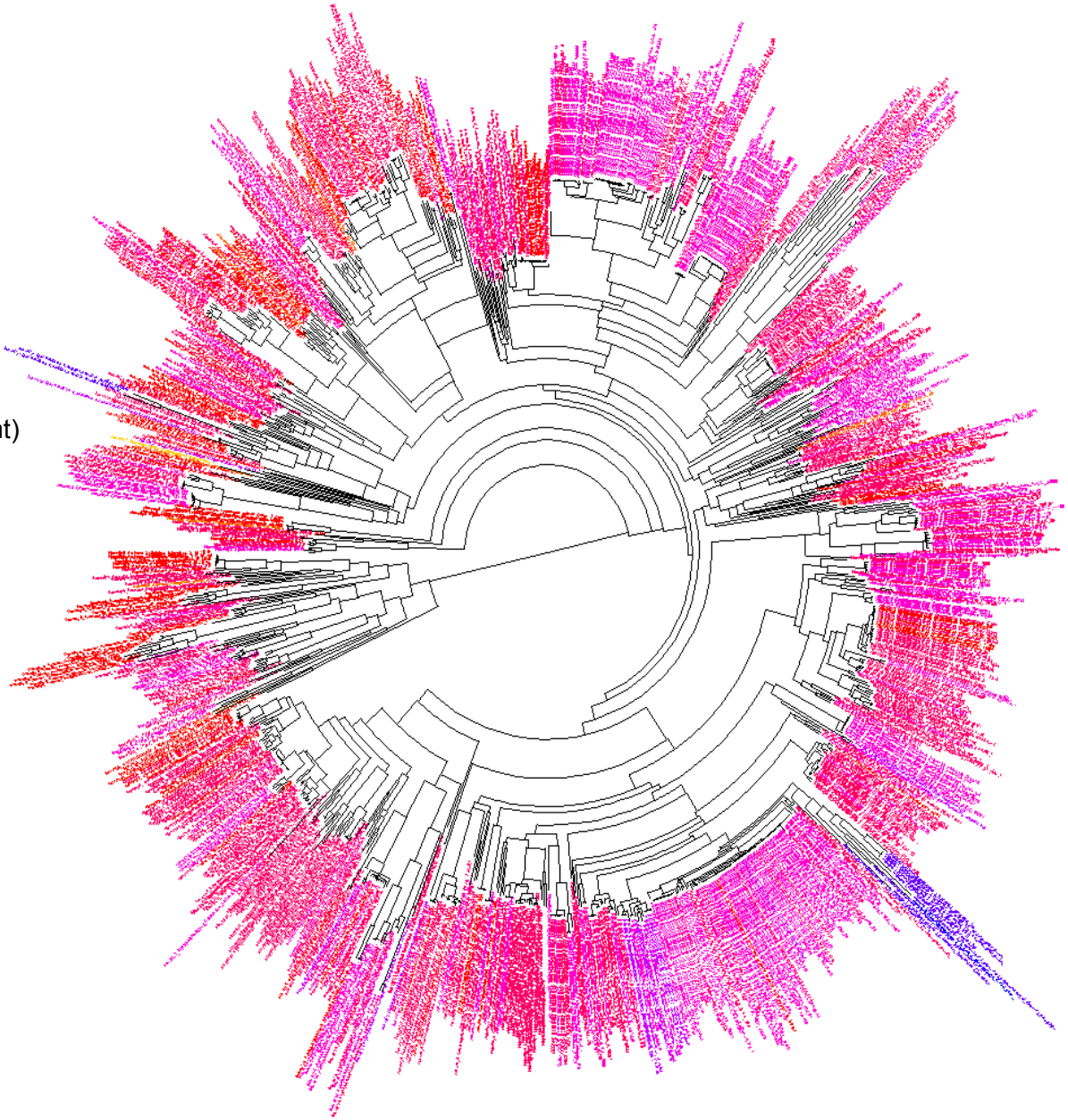

"genome_novelty_comp.txt" using 2:3
x
Genome novelty (family novelty calculated by hit span)
Genome novelty (family novelty calculated by hit count)
0
10
20
30
40
50
60
70
80
90



Genome novelty (family novelty calculated by hit count)

Genome Novelty

- [0-10)
- [10-20)
- [20-30)
- [30-40)
- [40-50)
- [50-60)
- [60-70)
- [70-80)



Genome novelty (family novelty calculated by hit span)

Gene Family Novelty and Phylogenetic Profiling

965 complete genomes

60763 families that span at least 5 genomes

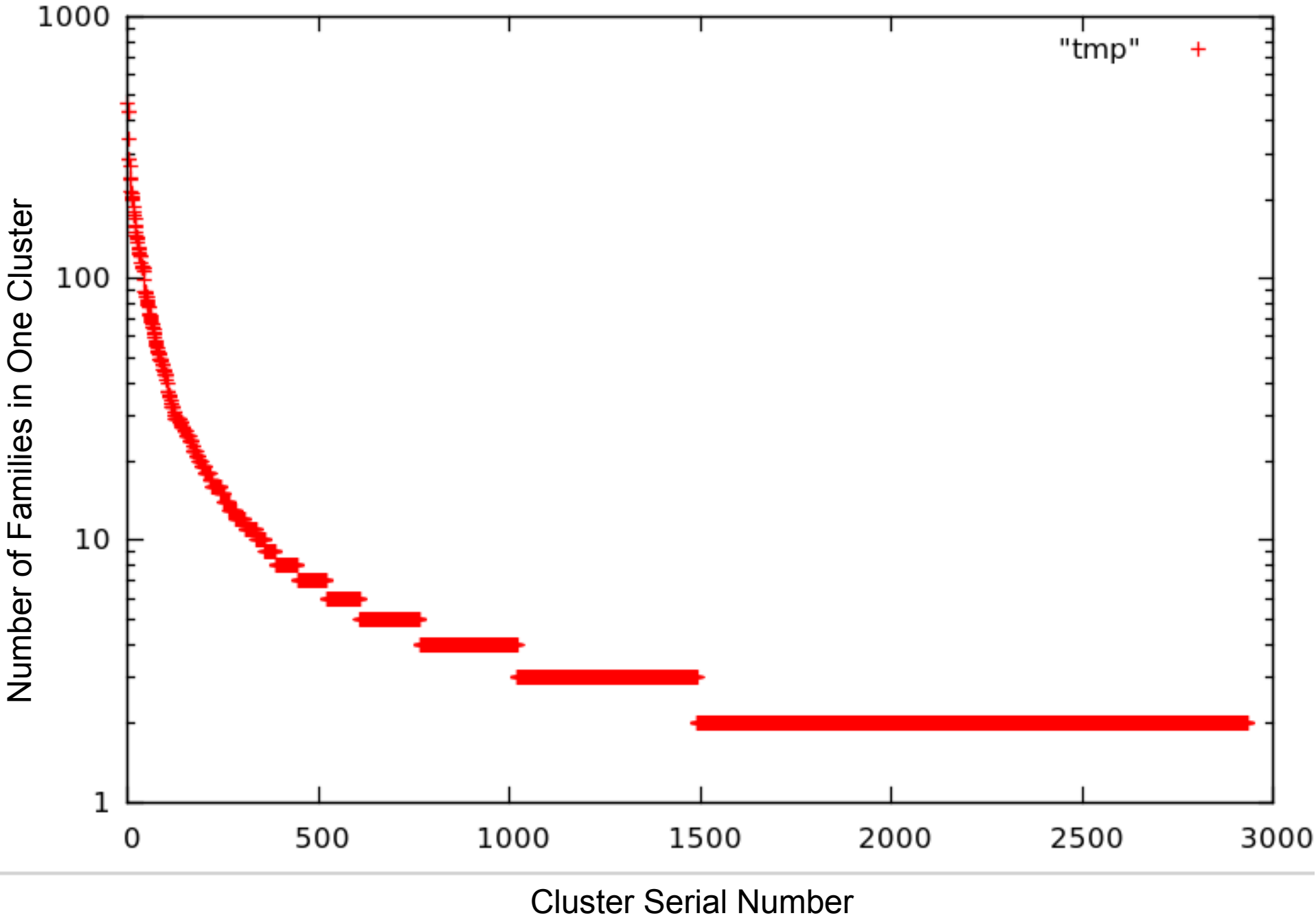
$$\text{Family genome sharing value} = \frac{\text{genome number shared by two families}}{\text{genome number covered by two families}}$$

Family pair (genome sharing value >=90%)

↓ MCL (I=2)

36380 cannot be clustered

The rest are clustered into 2834 clusters


"tmp"
Number of Families in One Cluster
Cluster Serial Number
1000
100
10
1
0
500
1000
1500
2000
2500
3000

Understand Novel families based upon Known Families linked by Phylogenetic profile clusters

Example PG Cluster:

Family ID	Novelty_by_hmm_span	
FAM8287ID	100.00	Protein of unknown function (DUF327)
FAM8129ID	100.00	hypothetical
FAM8711ID	100.00	hypothetical
FAM8688ID	52.95	PF12181, MogR_DNAbind, DNA binding domain of the motility gene repressor (MogR)
FAM11996ID	37.51	flagellar assembly protein H
FAM8324ID	36.10	flagellar motor switch protein
FAM8309ID	29.42	Flagellar hook-associated protein
FAM7910ID	6.89	flagellar basal body rod protein FlgB
FAM7856ID	3.35	hypothetical DUF1798 tructure of one of the proteins in this family has been solved
FAM7949ID	3.06	flagellar biosynthesis protein FliQ
FAM7983ID	2.92	flagellar hook-basal body protein FliE
FAM7695ID	2.39	flagellar biosynthesis protein FliR
FAM7613ID	2.14	ferrous iron transport protein A
FAM8456ID	1.74	flagellar protein FliS
FAM7752ID	1.73	phospholipase C
FAM7996ID	1.13	flagellar basal body rod protein FlgG