On Error Models for Misclassification Events on 16S and WGS sequences

Anthony Fodor
UNC Charlotte
Bioinformatics and Genomics

Copy of the talk: https://afodor.github.io/ (top link)

How do we measure and think about richness?

How do we distinguish rare mis-classifications from low abundance taxa?

Is “everything everywhere”?

A common problem in bioinformatics

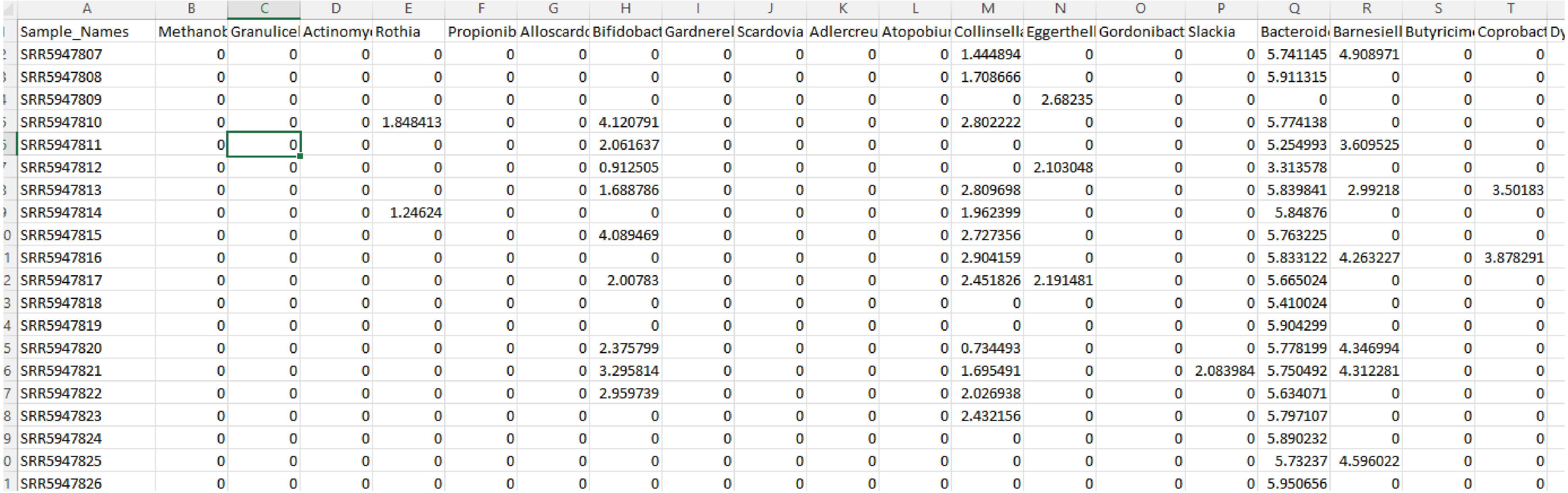
You detect a ASV (sequence variant) in a 16S sequence dataset.

You detect a taxa as being present in a metagenomic WGS sequence dataset

What is the probability that that sequence variant is really there “biologically” and does not reflect sequencing (or some other kind of) error

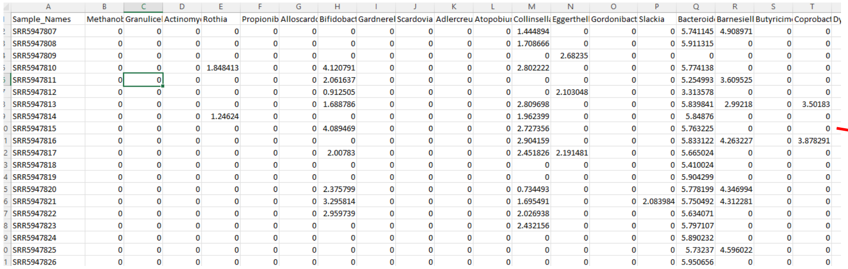
How we view prevalence and richness is very algorithm and method dependent

Metaphlan tends to produce very sparse spreadsheets with a few dominant taxa and lots of zeros…

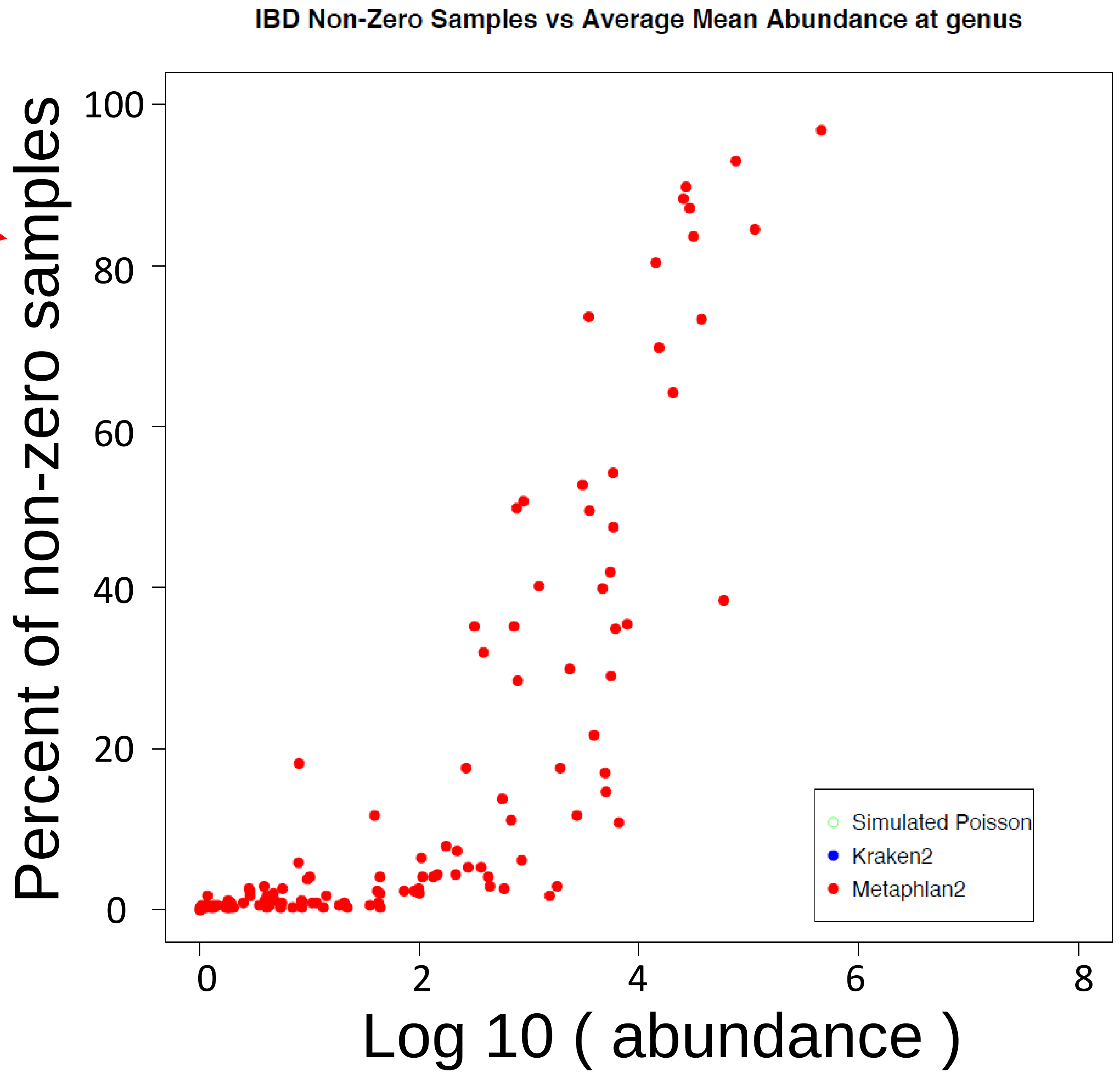


A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	
Sample_Names	Methanot	Granulicel	Actinomy	Rothia	Propionib	Alloscardc	Bifidobact	Gardnerel	Scardovia	Adlercreu	Atopobiu	Collinsella	Eggerthell	Gordonibact	Slackia	Bacteroid	Barnesiell	Butyricim	Coprobact	Dy
SRR5947807	0	0	0	0	0	0	0	0	0	0	0	1.444894	0	0	0	5.741145	4.908971	0	0	
SRR5947808	0	0	0	0	0	0	0	0	0	0	0	1.708666	0	0	0	5.911315	0	0	0	
SRR5947809	0	0	0	0	0	0	0	0	0	0	0	0	2.68235	0	0	0	0	0	0	
SRR5947810	0	0	0	1.848413	0	0	4.120791	0	0	0	0	2.802222	0	0	0	5.774138	0	0	0	
SRR5947811	0	0	0	0	0	0	2.061637	0	0	0	0	0	0	0	0	5.254993	3.609525	0	0	
SRR5947812	0	0	0	0	0	0	0.912505	0	0	0	0	0	2.103048	0	0	3.313578	0	0	0	
SRR5947813	0	0	0	0	0	0	1.688786	0	0	0	0	2.809698	0	0	0	5.839841	2.99218	0	3.50183	
SRR5947814	0	0	0	1.24624	0	0	0	0	0	0	0	1.962399	0	0	0	5.84876	0	0	0	
SRR5947815	0	0	0	0	0	0	4.089469	0	0	0	0	2.727356	0	0	0	5.763225	0	0	0	
SRR5947816	0	0	0	0	0	0	0	0	0	0	0	2.904159	0	0	0	5.833122	4.263227	0	3.878291	
SRR5947817	0	0	0	0	0	0	2.00783	0	0	0	0	2.451826	2.191481	0	0	5.665024	0	0	0	
SRR5947818	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.410024	0	0	0	
SRR5947819	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.904299	0	0	0	
SRR5947820	0	0	0	0	0	0	2.375799	0	0	0	0	0.734493	0	0	0	5.778199	4.346994	0	0	
SRR5947821	0	0	0	0	0	0	3.295814	0	0	0	0	1.695491	0	0	2.083984	5.750492	4.312281	0	0	
SRR5947822	0	0	0	0	0	0	2.959739	0	0	0	0	2.026938	0	0	0	5.634071	0	0	0	
SRR5947823	0	0	0	0	0	0	0	0	0	0	0	2.432156	0	0	0	5.797107	0	0	0	
SRR5947824	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.890232	0	0	0	
SRR5947825	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.73237	4.596022	0	0	
SRR5947826	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.950656	0	0	0	

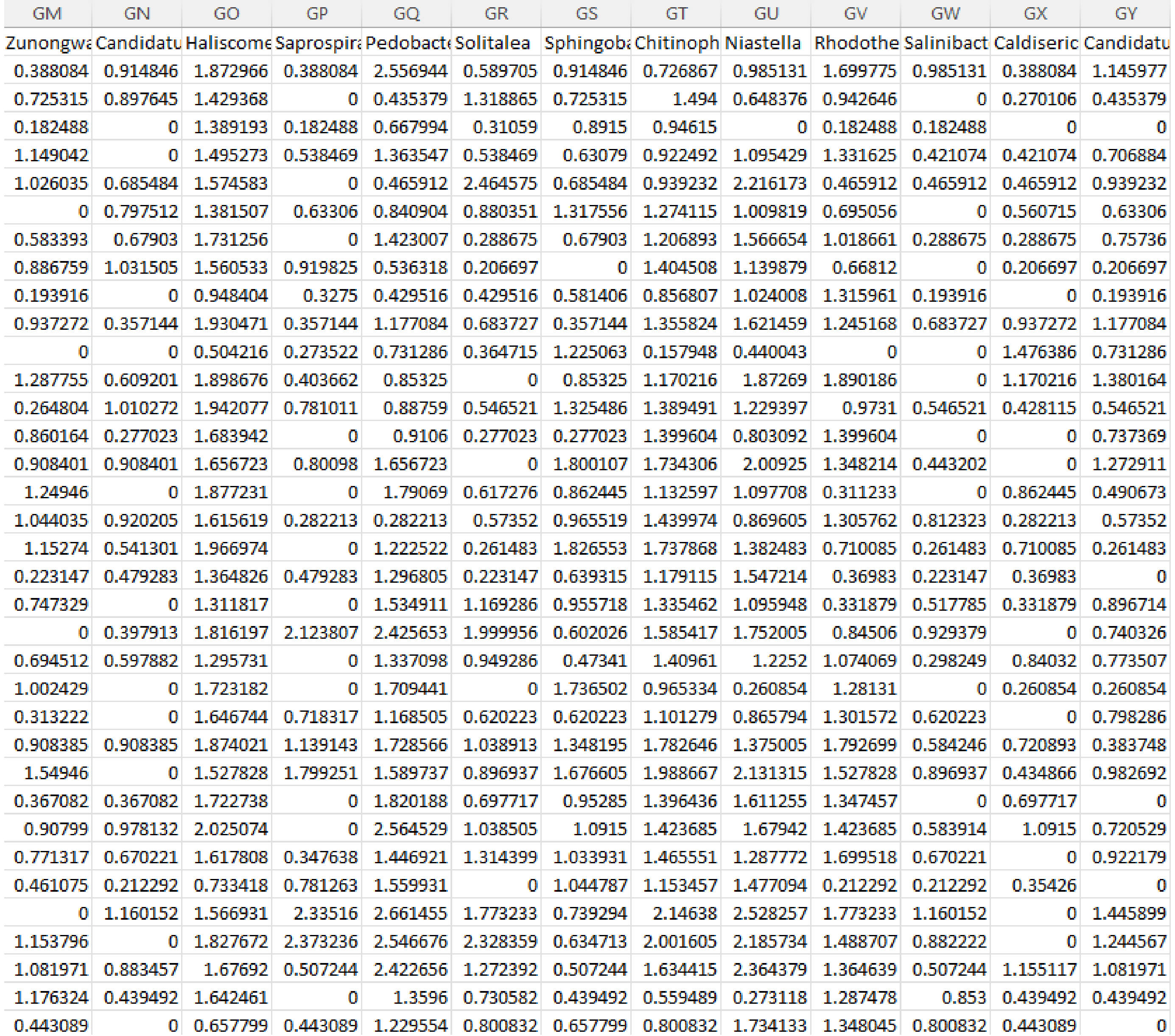
How we view prevalence and richness is very algorithm and method dependent



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	
Sample_Names	Methanob	Granulicel	Actinomy	Rothia	Propionib	Alloscardo	Bifidobact	Gardnerel	Scardovia	Adlercreu	Atopobiu	Collinsella	Eggerthell	Gordonibact	Slackia	Bacteroid	Barnesiell	Butyricim	Coprobact	Dy
SRR5947807	0	0	0	0	0	0	0	0	0	0	0	1.444894	0	0	0	5.741145	4.908971	0	0	
SRR5947808	0	0	0	0	0	0	0	0	0	0	0	1.708666	0	0	0	5.911315	0	0	0	
SRR5947809	0	0	0	0	0	0	0	0	0	0	0	0	2.68235	0	0	0	0	0	0	
SRR5947810	0	0	0	1.848413	0	0	4.120791	0	0	0	0	2.802222	0	0	0	5.774138	0	0	0	
SRR5947811	0	0	0	0	0	0	2.061637	0	0	0	0	0	0	0	0	5.254993	3.609525	0	0	
SRR5947812	0	0	0	0	0	0	0.912505	0	0	0	0	0	2.103048	0	0	3.313578	0	0	0	
SRR5947813	0	0	0	0	0	0	1.688786	0	0	0	0	2.809698	0	0	0	5.839841	2.99218	0	3.50183	
SRR5947814	0	0	0	1.24624	0	0	0	0	0	0	0	1.962399	0	0	0	5.84876	0	0	0	
SRR5947815	0	0	0	0	0	0	4.089469	0	0	0	0	2.727356	0	0	0	5.763225	0	0	0	
SRR5947816	0	0	0	0	0	0	0	0	0	0	0	2.904159	0	0	0	5.833122	4.263227	0	3.878291	
SRR5947817	0	0	0	0	0	0	2.00783	0	0	0	0	2.451826	2.191481	0	0	5.665024	0	0	0	
SRR5947818	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.410024	0	0	0	
SRR5947819	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.904299	0	0	0	
SRR5947820	0	0	0	0	0	0	2.375799	0	0	0	0	0.734493	0	0	0	5.778199	4.346994	0	0	
SRR5947821	0	0	0	0	0	0	3.295814	0	0	0	0	1.695491	0	0	2.083984	5.750492	4.312281	0	0	
SRR5947822	0	0	0	0	0	0	2.959739	0	0	0	0	2.026938	0	0	0	5.634071	0	0	0	
SRR5947823	0	0	0	0	0	0	0	0	0	0	0	2.432156	0	0	0	5.797107	0	0	0	
SRR5947824	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.890232	0	0	0	
SRR5947825	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.73237	4.596022	0	0	
SRR5947826	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.950656	0	0	0	

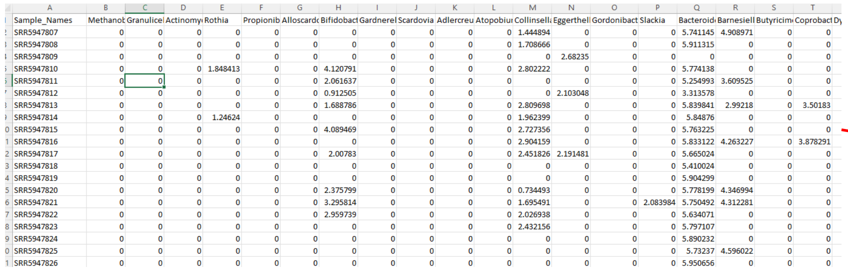


Kraken gives us a much less sparse view of the world…

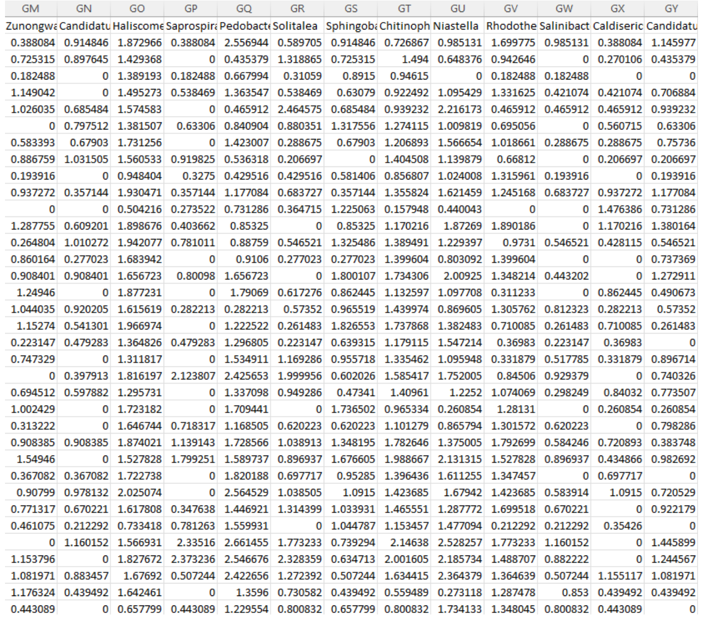


GM	GN	GO	GP	GQ	GR	GS	GT	GU	GV	GW	GX	GY
Zunongwa	Candidatu	Haliscome	Saprospira	Pedobacte	Solitalea	Sphingoba	Chitinoph	Niastella	Rhodothe	Salinibact	Caldiseric	Candidatu
0.388084	0.914846	1.872966	0.388084	2.556944	0.589705	0.914846	0.726867	0.985131	1.699775	0.985131	0.388084	1.145977
0.725315	0.897645	1.429368	0	0.435379	1.318865	0.725315	1.494	0.648376	0.942646	0	0.270106	0.435379
0.182488	0	1.389193	0.182488	0.667994	0.31059	0.8915	0.94615	0	0.182488	0.182488	0	0
1.149042	0	1.495273	0.538469	1.363547	0.538469	0.63079	0.922492	1.095429	1.331625	0.421074	0.421074	0.706884
1.026035	0.685484	1.574583	0	0.465912	2.464575	0.685484	0.939232	2.216173	0.465912	0.465912	0.465912	0.939232
0	0.797512	1.381507	0.63306	0.840904	0.880351	1.317556	1.274115	1.009819	0.695056	0	0.560715	0.63306
0.583393	0.67903	1.731256	0	1.423007	0.288675	0.67903	1.206893	1.566654	1.018661	0.288675	0.288675	0.75736
0.886759	1.031505	1.560533	0.919825	0.536318	0.206697	0	1.404508	1.139879	0.66812	0	0.206697	0.206697
0.193916	0	0.948404	0.3275	0.429516	0.429516	0.581406	0.856807	1.024008	1.315961	0.193916	0	0.193916
0.937272	0.357144	1.930471	0.357144	1.177084	0.683727	0.357144	1.355824	1.621459	1.245168	0.683727	0.937272	1.177084
0	0	0.504216	0.273522	0.731286	0.364715	1.225063	0.157948	0.440043	0	0	1.476386	0.731286
1.287755	0.609201	1.898676	0.403662	0.85325	0	0.85325	1.170216	1.87269	1.890186	0	1.170216	1.380164
0.264804	1.010272	1.942077	0.781011	0.88759	0.546521	1.325486	1.389491	1.229397	0.9731	0.546521	0.428115	0.546521
0.860164	0.277023	1.683942	0	0.9106	0.277023	0.277023	1.399604	0.803092	1.399604	0	0	0.737369
0.908401	0.908401	1.656723	0.80098	1.656723	0	1.800107	1.734306	2.00925	1.348214	0.443202	0	1.272911
1.24946	0	1.877231	0	1.79069	0.617276	0.862445	1.132597	1.097708	0.311233	0	0.862445	0.490673
1.044035	0.920205	1.615619	0.282213	0.282213	0.57352	0.965519	1.439974	0.869605	1.305762	0.812323	0.282213	0.57352
1.15274	0.541301	1.966974	0	1.222522	0.261483	1.826553	1.737868	1.382483	0.710085	0.261483	0.710085	0.261483
0.223147	0.479283	1.364826	0.479283	1.296805	0.223147	0.639315	1.179115	1.547214	0.36983	0.223147	0.36983	0
0.747329	0	1.311817	0	1.534911	1.169286	0.955718	1.335462	1.095948	0.331879	0.517785	0.331879	0.896714
0	0.397913	1.816197	2.123807	2.425653	1.999956	0.602026	1.585417	1.752005	0.84506	0.929379	0	0.740326
0.694512	0.597882	1.295731	0	1.337098	0.949286	0.47341	1.40961	1.2252	1.074069	0.298249	0.84032	0.773507
1.002429	0	1.723182	0	1.709441	0	1.736502	0.965334	0.260854	1.28131	0	0.260854	0.260854
0.313222	0	1.646744	0.718317	1.168505	0.620223	0.620223	1.101279	0.865794	1.301572	0.620223	0	0.798286
0.908385	0.908385	1.874021	1.139143	1.728566	1.038913	1.348195	1.782646	1.375005	1.792699	0.584246	0.720893	0.383748
1.54946	0	1.527828	1.799251	1.589737	0.896937	1.676605	1.988667	2.131315	1.527828	0.896937	0.434866	0.982692
0.367082	0.367082	1.722738	0	1.820188	0.697717	0.95285	1.396436	1.611255	1.347457	0	0.697717	0
0.90799	0.978132	2.025074	0	2.564529	1.038505	1.0915	1.423685	1.67942	1.423685	0.583914	1.0915	0.720529
0.771317	0.670221	1.617808	0.347638	1.446921	1.314399	1.033931	1.465551	1.287772	1.699518	0.670221	0	0.922179
0.461075	0.212292	0.733418	0.781263	1.559931	0	1.044787	1.153457	1.477094	0.212292	0.212292	0.35426	0
0	1.160152	1.566931	2.33516	2.661455	1.773233	0.739294	2.14638	2.528257	1.773233	1.160152	0	1.445899
1.153796	0	1.827672	2.373236	2.546676	2.328359	0.634713	2.001605	2.185734	1.488707	0.882222	0	1.244567
1.081971	0.883457	1.67692	0.507244	2.422656	1.272392	0.507244	1.634415	2.364379	1.364639	0.507244	1.155117	1.081971
1.176324	0.439492	1.642461	0	1.3596	0.730582	0.439492	0.559489	0.273118	1.287478	0.853	0.439492	0.439492
0.443089	0	0.657799	0.443089	1.229554	0.800832	0.657799	0.800832	1.734133	1.348045	0.800832	0.443089	0

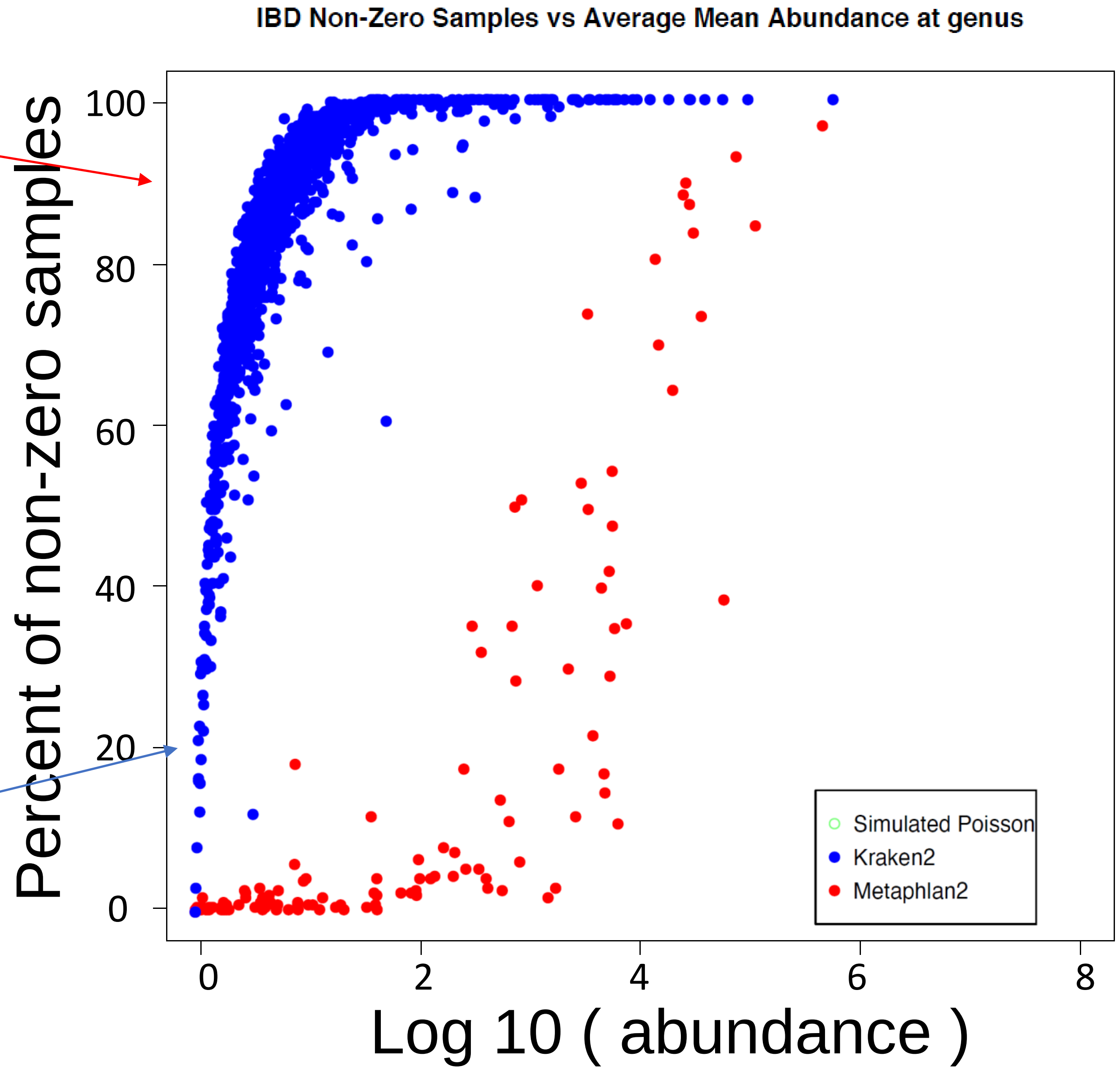
How we view prevalence and richness is very algorithm and method dependent



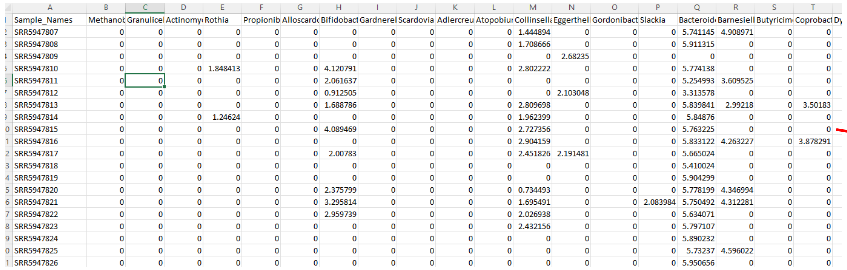
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	
Sample_Names	Methanob	Granulicel	Actinomy	Rothia	Propionib	Alloscardo	Bifidobact	Gardnerel	Scardovia	Adlercreu	Atopobiu	Collinsella	Eggerthel	Gordonibact	Slackia	Bacteroid	Barnesiell	Butyricim	Coprobact	Dy
SRR5947807	0	0	0	0	0	0	0	0	0	0	0	1.444894	0	0	0	5.741145	4.908971	0	0	
SRR5947808	0	0	0	0	0	0	0	0	0	0	0	1.708666	0	0	0	5.911315	0	0	0	
SRR5947809	0	0	0	0	0	0	0	0	0	0	0	0	2.68235	0	0	0	0	0	0	
SRR5947810	0	0	0	1.848413	0	0	4.120791	0	0	0	0	2.802222	0	0	0	5.774138	0	0	0	
SRR5947811	0	0	0	0	0	0	2.061637	0	0	0	0	0	0	0	0	5.254993	3.609525	0	0	
SRR5947812	0	0	0	0	0	0	0.912505	0	0	0	0	0	2.103048	0	0	3.313578	0	0	0	
SRR5947813	0	0	0	0	0	0	1.688786	0	0	0	0	2.809698	0	0	0	5.839841	2.99218	0	3.50183	
SRR5947814	0	0	0	1.24624	0	0	0	0	0	0	0	1.962399	0	0	0	5.84876	0	0	0	
SRR5947815	0	0	0	0	0	0	4.089469	0	0	0	0	2.727356	0	0	0	5.763225	0	0	0	
SRR5947816	0	0	0	0	0	0	0	0	0	0	0	2.904159	0	0	0	5.833122	4.263227	0	3.878291	
SRR5947817	0	0	0	0	0	0	2.00783	0	0	0	0	2.451826	2.191481	0	0	5.665024	0	0	0	
SRR5947818	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.410024	0	0	0	
SRR5947819	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.904299	0	0	0	
SRR5947820	0	0	0	0	0	0	2.375799	0	0	0	0	0.734493	0	0	0	5.778199	4.346994	0	0	
SRR5947821	0	0	0	0	0	0	3.295814	0	0	0	0	1.695491	0	0	2.083984	5.750492	4.312281	0	0	
SRR5947822	0	0	0	0	0	0	2.959739	0	0	0	0	2.026938	0	0	0	5.634071	0	0	0	
SRR5947823	0	0	0	0	0	0	0	0	0	0	0	2.432156	0	0	0	5.797107	0	0	0	
SRR5947824	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.890232	0	0	0	
SRR5947825	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.73237	4.596022	0	0	
SRR5947826	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.950656	0	0	0	



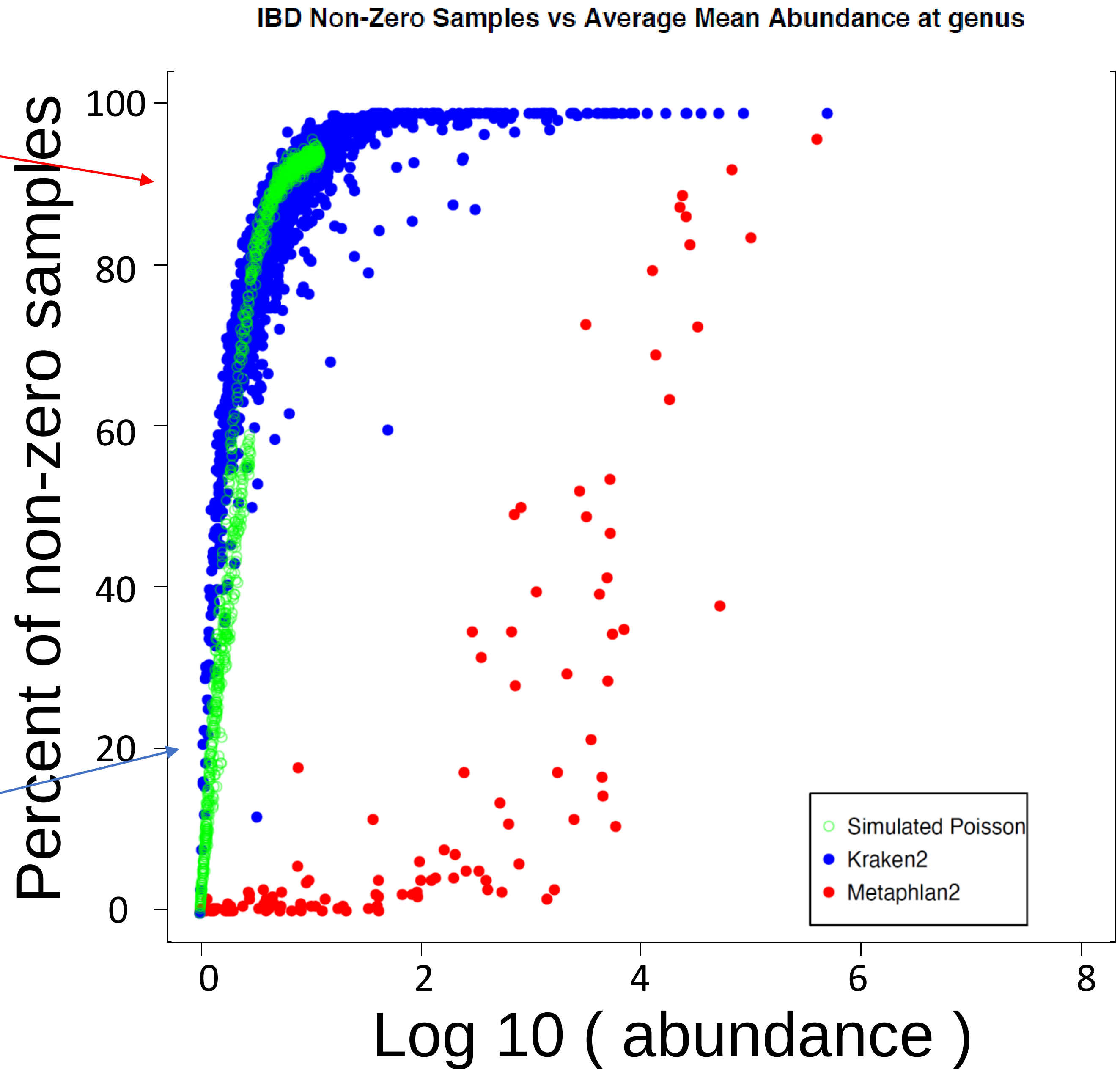
GM	GN	GO	GP	GQ	GR	GS	GT	GU	GV	GW	GX	GY
Zunongwa	Candidatu	Haliscome	Saprospira	Pedobacte	Solitalea	Sphingoba	Chitinoph	Niastella	Rhodothe	Salinibact	Caldiseric	Candidatu
0.388084	0.914846	1.872966	0.388084	2.556944	0.589705	0.914846	0.726867	0.985131	1.699775	0.985131	0.388084	1.145977
0.725315	0.897645	1.429368	0	0.435379	1.318865	0.725315	1.494	0.648376	0.942646	0	0.270106	0.435379
0.182488	0	1.389193	0.182488	0.667994	0.31059	0.8915	0.94615	0	0.182488	0.182488	0	0
1.149042	0	1.495273	0.538469	1.363547	0.538469	0.63079	0.922492	1.095429	1.331625	0.421074	0.421074	0.706884
1.026035	0.685484	1.574583	0	0.465912	2.464575	0.685484	0.939232	2.216173	0.465912	0.465912	0.465912	0.939232
0	0.797512	1.381507	0.63306	0.840904	0.880351	1.317556	1.274115	1.009819	0.695056	0	0.560715	0.63306
0.583393	0.67903	1.731256	0	1.423007	0.288675	0.67903	1.206893	1.566654	1.018661	0.288675	0.288675	0.75736
0.886759	1.031505	1.560533	0.919825	0.536318	0.206697	0	1.404508	1.139879	0.66812	0	0.206697	0.206697
0.193916	0	0.948404	0.3275	0.429516	0.429516	0.581406	0.856807	1.024008	1.315961	0.193916	0	0.193916
0.937272	0.357144	1.930471	0.357144	1.177084	0.683727	0.357144	1.355824	1.621459	1.245168	0.683727	0.937272	1.177084
0	0	0.504216	0.273522	0.731286	0.364715	1.225063	0.157948	0.440043	0	0	1.476386	0.731286
1.287755	0.609201	1.898676	0.403662	0.85325	0	0.85325	1.170216	1.87269	1.890186	0	1.170216	1.380164
0.264804	1.010272	1.942077	0.781011	0.88759	0.546521	1.325486	1.389491	1.229397	0.9731	0.546521	0.428115	0.546521
0.860164	0.277023	1.683942	0	0.9106	0.277023	0.277023	1.399604	0.803092	1.399604	0	0	0.737369
0.908401	0.908401	1.656723	0.80098	1.656723	0	1.800107	1.734306	2.00925	1.348214	0.443202	0	1.272911
1.24946	0	1.877231	0	1.79069	0.617276	0.862445	1.132597	1.097708	0.311233	0	0.862445	0.490673
1.044035	0.920205	1.615619	0.282213	0.282213	0.57352	0.965519	1.439974	0.869605	1.305762	0.812323	0.282213	0.57352
1.15274	0.541301	1.966974	0	1.222522	0.261483	1.826553	1.737868	1.382483	0.710085	0.261483	0.710085	0.261483
0.223147	0.479283	1.364826	0.479283	1.296805	0.223147	0.639315	1.179115	1.547214	0.36983	0.223147	0.36983	0
0.747329	0	1.311817	0	1.534911	1.169286	0.955718	1.335462	1.095948	0.331879	0.517785	0.331879	0.896714
0	0.397913	1.816197	2.123807	2.425653	1.999956	0.602026	1.585417	1.752005	0.84506	0.929379	0	0.740326
0.694512	0.597882	1.295731	0	1.337098	0.949286	0.47341	1.40961	1.2252	1.074069	0.298249	0.84032	0.773507
1.002429	0	1.723182	0	1.709441	0	1.736502	0.965334	0.260854	1.28131	0	0.260854	0.260854
0.313222	0	1.646744	0.718317	1.168505	0.620223	0.620223	1.101279	0.865794	1.301572	0.620223	0	0.798286
0.908385	0.908385	1.874021	1.139143	1.728566	1.038913	1.348195	1.782646	1.375005	1.792699	0.584246	0.720893	0.383748
1.54946	0	1.527828	1.799251	1.589737	0.896937	1.676605	1.988667	2.131315	1.527828	0.896937	0.434866	0.982692
0.367082	0.367082	1.722738	0	1.820188	0.697717	0.95285	1.396436	1.611255	1.347457	0	0.697717	0
0.90799	0.978132	2.025074	0	2.564529	1.038505	1.0915	1.423685	1.67942	1.423685	0.583914	1.0915	0.720529
0.771317	0.670221	1.617808	0.347638	1.446921	1.314399	1.033931	1.465551	1.287772	1.699518	0.670221	0	0.922179
0.461075	0.212292	0.733418	0.781263	1.559931	0	1.044787	1.153457	1.477094	0.212292	0.212292	0.35426	0
0	1.160152	1.566931	2.33516	2.661455	1.773233	0.739294	2.14638	2.528257	1.773233	1.160152	0	1.445899
1.153796	0	1.827672	2.373236	2.546676	2.328359	0.634713	2.001605	2.185734	1.488707	0.882222	0	1.244567
1.081971	0.883457	1.67692	0.507244	2.422656	1.272392	0.507244	1.634415	2.364379	1.364639	0.507244	1.155117	1.081971
1.176324	0.439492	1.642461	0	1.3596	0.730582	0.439492	0.559489	0.273118	1.287478	0.853	0.439492	0.439492
0.443089	0	0.657799	0.443089	1.229554	0.800832	0.657799	0.800832	1.734133	1.348045	0.800832	0.443089	0

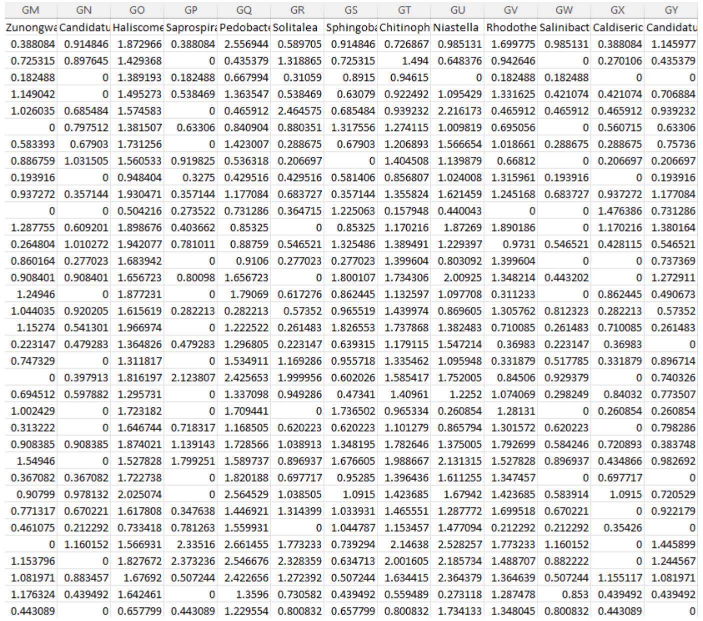


We seek a null model that is unlikely to be explained by biology



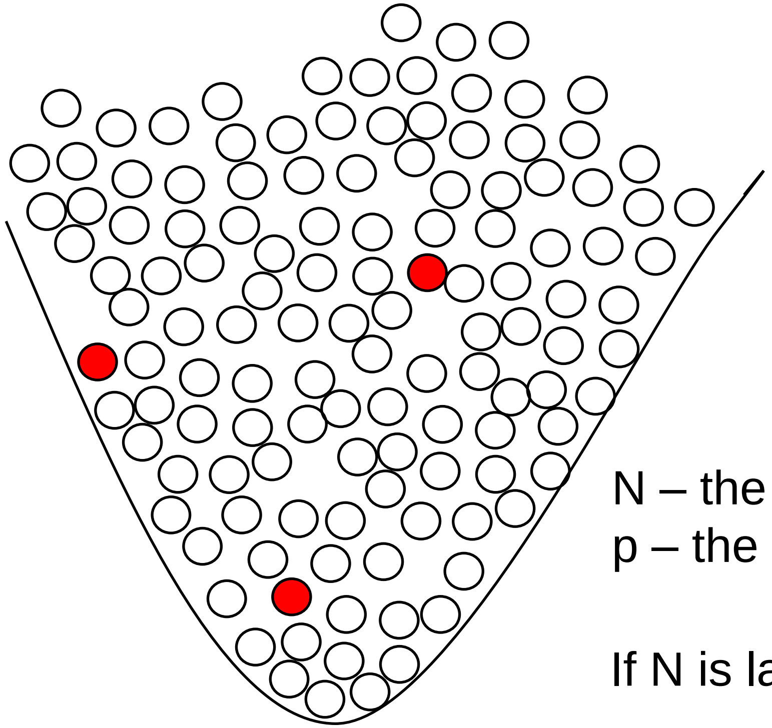
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	
Sample_Names	Methanot	Granulicel	Actinomy	Rothia	Propionib	Alloscardo	Bifidobact	Gardnerel	Scardovia	Adlercreu	Atopobiur	Collinsella	Eggerthel	Gordonibact	Slackia	Bacteroid	Barnesiell	Butyricim	Coprobact	Dy
SRR5947807	0	0	0	0	0	0	0	0	0	0	0	1.444894	0	0	0	5.741145	4.908971	0	0	
SRR5947808	0	0	0	0	0	0	0	0	0	0	0	1.708666	0	0	0	5.911315	0	0	0	
SRR5947809	0	0	0	0	0	0	0	0	0	0	0	0	2.68235	0	0	0	0	0	0	
SRR5947810	0	0	0	1.848413	0	0	4.120791	0	0	0	0	2.802222	0	0	0	5.774138	0	0	0	
SRR5947811	0	0	0	0	0	0	2.061637	0	0	0	0	0	0	0	0	5.254993	3.609525	0	0	
SRR5947812	0	0	0	0	0	0	0.912505	0	0	0	0	0	2.103048	0	0	3.313578	0	0	0	
SRR5947813	0	0	0	0	0	0	1.688786	0	0	0	0	2.809698	0	0	0	5.839841	2.99218	0	3.50183	
SRR5947814	0	0	0	1.24624	0	0	0	0	0	0	0	1.962399	0	0	0	5.84876	0	0	0	
SRR5947815	0	0	0	0	0	0	4.089469	0	0	0	0	2.727356	0	0	0	5.763225	0	0	0	
SRR5947816	0	0	0	0	0	0	0	0	0	0	0	2.904159	0	0	0	5.833122	4.263227	0	3.878291	
SRR5947817	0	0	0	0	0	0	2.00783	0	0	0	0	2.451826	2.191481	0	0	5.665024	0	0	0	
SRR5947818	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.410024	0	0	0	
SRR5947819	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.904299	0	0	0	
SRR5947820	0	0	0	0	0	0	2.375799	0	0	0	0	0.734493	0	0	0	5.778199	4.346994	0	0	
SRR5947821	0	0	0	0	0	0	3.295814	0	0	0	0	1.695491	0	0	2.083984	5.750492	4.312281	0	0	
SRR5947822	0	0	0	0	0	0	2.959739	0	0	0	0	2.026938	0	0	0	5.634071	0	0	0	
SRR5947823	0	0	0	0	0	0	0	0	0	0	0	2.432156	0	0	0	5.797107	0	0	0	
SRR5947824	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.890232	0	0	0	
SRR5947825	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.73237	4.596022	0	0	
SRR5947826	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.950656	0	0	0	





GM	GN	GO	GP	GQ	GR	GS	GT	GU	GV	GW	GX	GY
Zunongwa	Candidatu	Haliscome	Saprospira	Pedobacte	Solitalea	Sphingoba	Chitinoph	Niastella	Rhodothe	Salinibact	Caldiseric	Candidatu
0.388084	0.914846	1.872966	0.388084	2.556944	0.589705	0.914846	0.726867	0.985131	1.699775	0.985131	0.388084	1.145977
0.725315	0.897645	1.429368	0	0.435379	1.318865	0.725315	1.494	0.648376	0.942646	0	0.270106	0.435379
0.182488	0	1.389193	0.182488	0.667994	0.31059	0.8915	0.94615	0	0.182488	0.182488	0	0
1.149042	0	1.495273	0.538469	1.363547	0.538469	0.63079	0.922492	1.095429	1.331625	0.421074	0.421074	0.706884
1.026035	0.685484	1.574583	0	0.465912	2.464575	0.685484	0.939232	2.216173	0.465912	0.465912	0.465912	0.939232
0	0.797512	1.381507	0.63306	0.840904	0.880351	1.317556	1.274115	1.009819	0.695056	0	0.560715	0.63306
0.583393	0.67903	1.731256	0	1.423007	0.288675	0.67903	1.206893	1.566654	1.018661	0.288675	0.288675	0.75736
0.886759	1.031505	1.560533	0.919825	0.536318	0.206697	0	1.404508	1.139879	0.66812	0	0.206697	0.206697
0.193916	0	0.948404	0.3275	0.429516	0.429516	0.581406	0.856807	1.024008	1.315961	0.193916	0	0.193916
0.937272	0.357144	1.930471	0.357144	1.177084	0.683727	0.357144	1.355824	1.621459	1.245168	0.683727	0.937272	1.177084
0	0	0.504216	0.273522	0.731286	0.364715	1.225063	0.157948	0.440043	0	0	1.476386	0.731286
1.287755	0.609201	1.898676	0.403662	0.85325	0	0.85325	1.170216	1.87269	1.890186	0	1.170216	1.380164
0.264804	1.010272	1.942077	0.781011	0.88759	0.546521	1.325486	1.389491	1.229397	0.9731	0.546521	0.428115	0.546521
0.860164	0.277023	1.683942	0	0.9106	0.277023	0.277023	1.399604	0.803092	1.399604	0	0	0.737369
0.908401	0.908401	1.656723	0.80098	1.656723	0	1.800107	1.734306	2.00925	1.348214	0.443202	0	1.272911
1.24946	0	1.877231	0	1.79069	0.617276	0.862445	1.132597	1.097708	0.311233	0	0.862445	0.490673
1.044035	0.920205	1.615619	0.282213	0.282213	0.57352	0.965519	1.439974	0.869605	1.305762	0.812323	0.282213	0.57352
1.15274	0.541301	1.966974	0	1.222522	0.261483	1.826553	1.737868	1.382483	0.710085	0.261483	0.710085	0.261483
0.223147	0.479283	1.364826	0.479283	1.296805	0.223147	0.639315	1.179115	1.547214	0.36983	0.223147	0.36983	0
0.747329	0	1.311817	0	1.534911	1.169286	0.955718	1.335462	1.095948	0.331879	0.517785	0.331879	0.896714
0	0.397913	1.816197	2.123807	2.425653	1.999956	0.602026	1.585417	1.752005	0.84506	0.929379	0	0.740326
0.694512	0.597882	1.295731	0	1.337098	0.949286	0.47341	1.40961	1.2252	1.074069	0.298249	0.84032	0.773507
1.002429	0	1.723182	0	1.709441	0	1.736502	0.965334	0.260854	1.28131	0	0.260854	0.260854
0.313222	0	1.646744	0.718317	1.168505	0.620223	0.620223	1.101279	0.865794	1.301572	0.620223	0	0.798286
0.908385	0.908385	1.874021	1.139143	1.728566	1.038913	1.348195	1.782646	1.375005	1.792699	0.584246	0.720893	0.383748
1.54946	0	1.527828	1.799251	1.589737	0.896937	1.676605	1.988667	2.131315	1.527828	0.896937	0.434866	0.982692
0.367082	0.367082	1.722738	0	1.820188	0.697717	0.95285	1.396436	1.611255	1.347457	0	0.697717	0
0.90799	0.978132	2.025074	0	2.564529	1.038505	1.0915	1.423685	1.67942	1.423685	0.583914	1.0915	0.720529
0.771317	0.670221	1.617808	0.347638	1.446921	1.314399	1.033931	1.465551	1.287772	1.699518	0.670221	0	0.922179
0.461075	0.212292	0.733418	0.781263	1.559931	0	1.044787	1.153457	1.477094	0.212292	0.212292	0.35426	0
0	1.160152	1.566931	2.33516	2.661455	1.773233	0.739294	2.14638	2.528257	1.773233	1.160152	0	1.445899
1.153796	0	1.827672	2.373236	2.546676	2.328359	0.634713	2.001605	2.185734	1.488707	0.882222	0	1.244567
1.081971	0.883457	1.67692	0.507244	2.422656	1.272392	0.507244	1.634415	2.364379	1.364639	0.507244	1.155117	1.081971
1.176324	0.439492	1.642461	0	1.3596	0.730582	0.439492	0.559489	0.273118	1.287478	0.853	0.439492	0.439492
0.443089	0	0.657799	0.443089	1.229554	0.800832	0.657799	0.800832	1.734133	1.348045	0.800832	0.443089	0

Our simplest possible model for classification: “Binomial” or “Poisson” models



You have a very (infinitely) large vat of perfectly mixed ping-pong balls

We model a sequence classification event
(for example, labeling a read represents a sequence variant)
as the probability of randomly drawing red balls from a set of sequences

N – the total number of sequences in a sample (the red balls + the white balls observed)
p – the fraction of all balls in the vat that belong to a variant

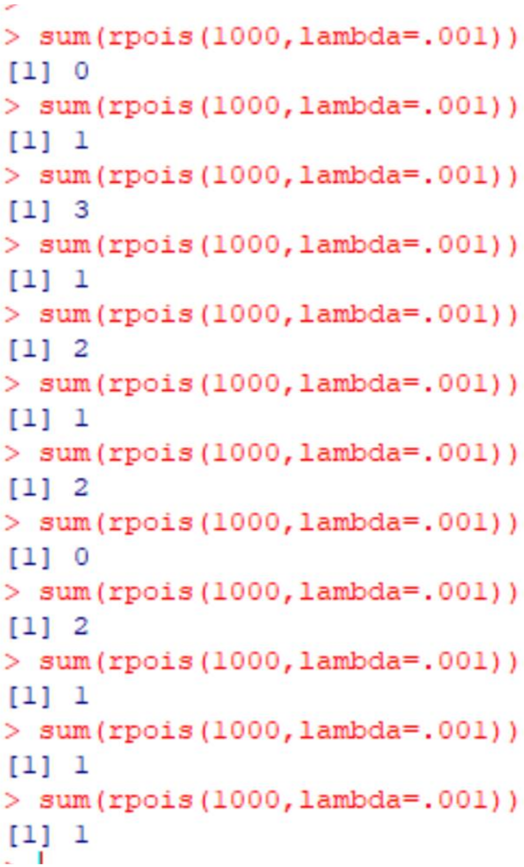
If N is large and p is small we say this is a Poisson process.

We draw a 1,000 ping pong balls (sequences) with a “true”
relative abundance of 0.1%, we would expect 1 red ping pong balls.

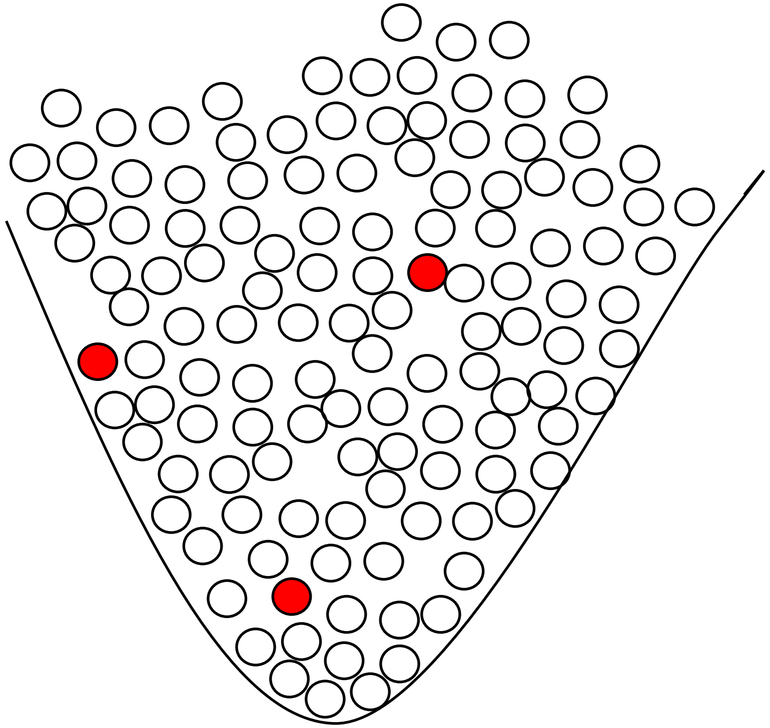
If N is large and p is small we say this is a Poisson process.

We draw a 1,000 ping pong balls (sequences) with a “true”
relative abundance of 0.1%, we would expect 1 red ping pong balls.

We can produce our expectation under such a null model very easily in a language such as R



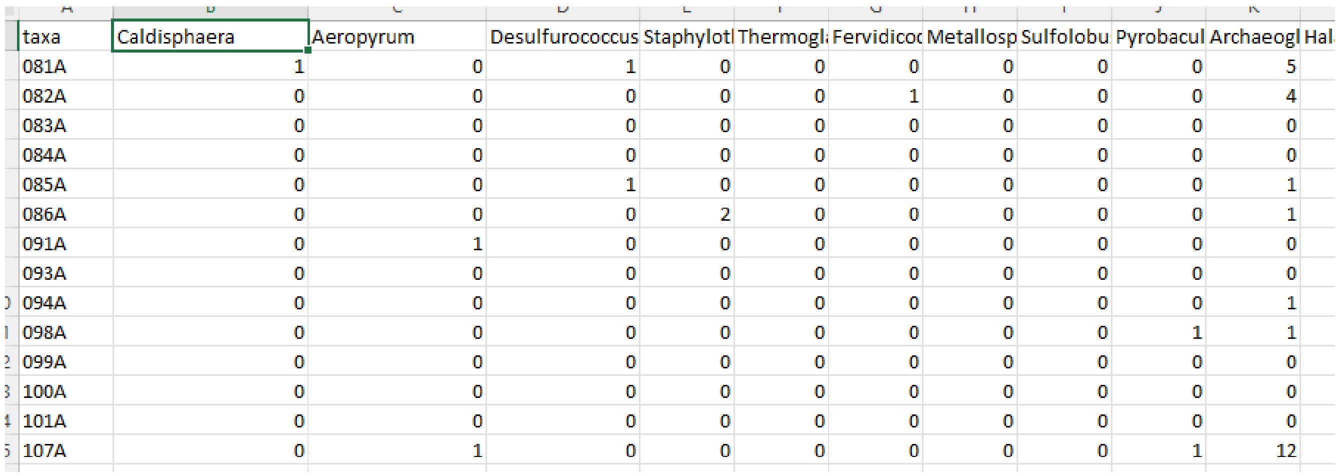
```
> sum(rpois(1000,lambda=.001))
[1] 0
> sum(rpois(1000,lambda=.001))
[1] 1
> sum(rpois(1000,lambda=.001))
[1] 3
> sum(rpois(1000,lambda=.001))
[1] 1
> sum(rpois(1000,lambda=.001))
[1] 2
> sum(rpois(1000,lambda=.001))
[1] 1
> sum(rpois(1000,lambda=.001))
[1] 2
> sum(rpois(1000,lambda=.001))
[1] 0
> sum(rpois(1000,lambda=.001))
[1] 2
> sum(rpois(1000,lambda=.001))
[1] 1
> sum(rpois(1000,lambda=.001))
[1] 1
> sum(rpois(1000,lambda=.001))
[1] 1
```



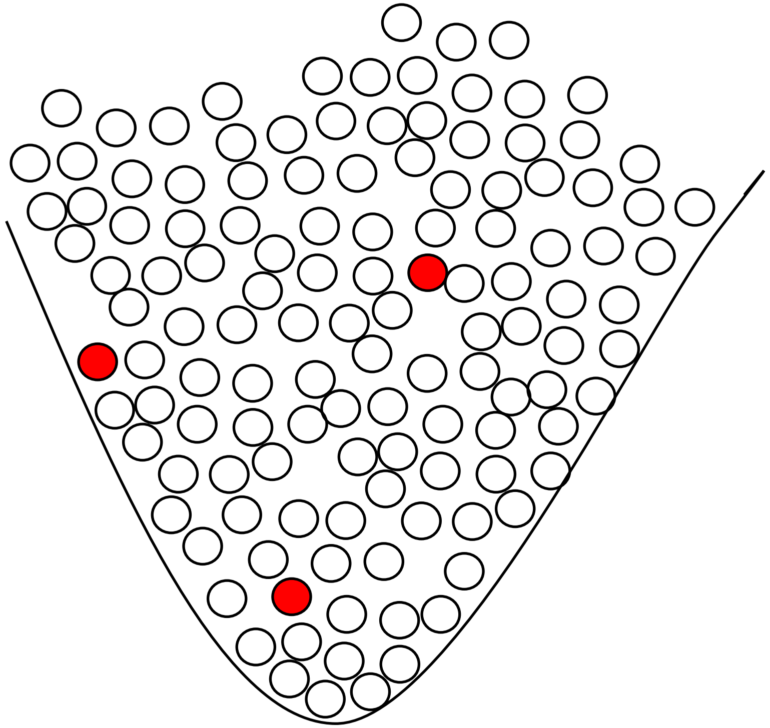
We can generate an entire “simulated” dataset under the Poisson null

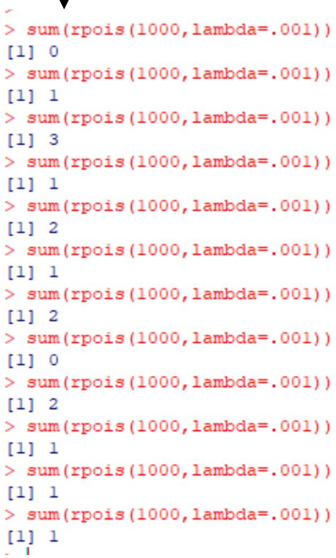
Taxa

Samples



taxa	Caldisphaera	Aeropyrum	Desulfurococcus	Staphylotl	Thermogl:	Fervidicoc	Metallosp	Sulfolobu	Pyrobacul	Archaeogl	Hal
081A	1	0	1	0	0	0	0	0	0	5	
082A	0	0	0	0	0	1	0	0	0	4	
083A	0	0	0	0	0	0	0	0	0	0	
084A	0	0	0	0	0	0	0	0	0	0	
085A	0	0	1	0	0	0	0	0	0	1	
086A	0	0	0	2	0	0	0	0	0	1	
091A	0	1	0	0	0	0	0	0	0	0	
093A	0	0	0	0	0	0	0	0	0	0	
094A	0	0	0	0	0	0	0	0	0	1	
098A	0	0	0	0	0	0	0	0	1	1	
099A	0	0	0	0	0	0	0	0	0	0	
100A	0	0	0	0	0	0	0	0	0	0	
101A	0	0	0	0	0	0	0	0	0	0	
107A	0	1	0	0	0	0	0	0	1	12	





```
> sum(rpois(1000,lambda=.001))
[1] 0
> sum(rpois(1000,lambda=.001))
[1] 1
> sum(rpois(1000,lambda=.001))
[1] 3
> sum(rpois(1000,lambda=.001))
[1] 1
> sum(rpois(1000,lambda=.001))
[1] 2
> sum(rpois(1000,lambda=.001))
[1] 1
> sum(rpois(1000,lambda=.001))
[1] 2
> sum(rpois(1000,lambda=.001))
[1] 0
> sum(rpois(1000,lambda=.001))
[1] 2
> sum(rpois(1000,lambda=.001))
[1] 1
> sum(rpois(1000,lambda=.001))
[1] 1
> sum(rpois(1000,lambda=.001))
[1] 1
```

Each cell counts how many “red balls” (that taxa) from all the balls in the sample.

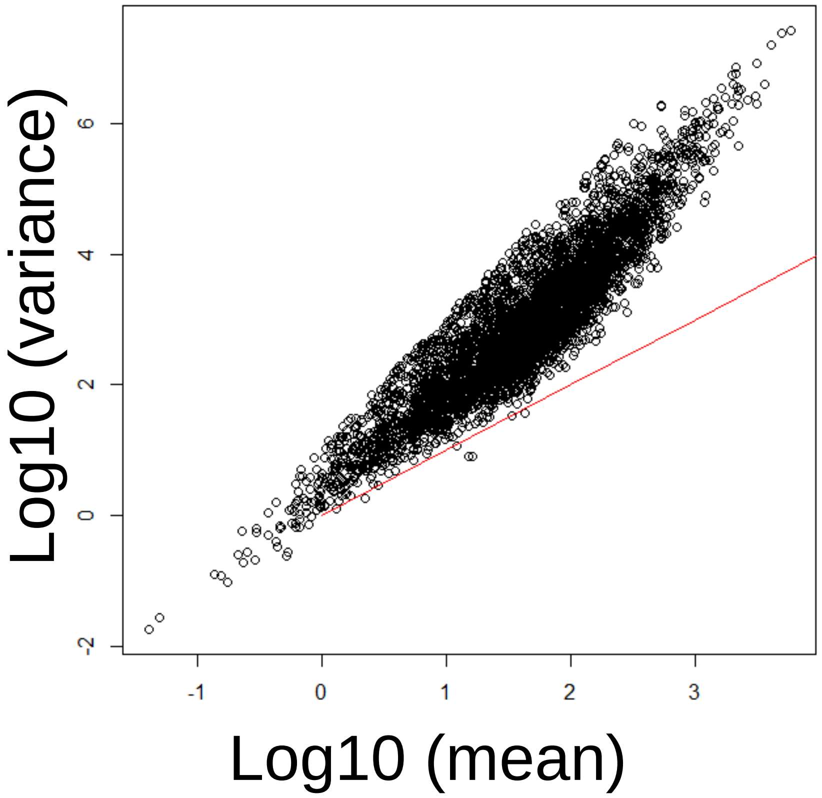
Features of a Poisson null model:

No biology (constant background error rate irrespective of sample type of phenotype)

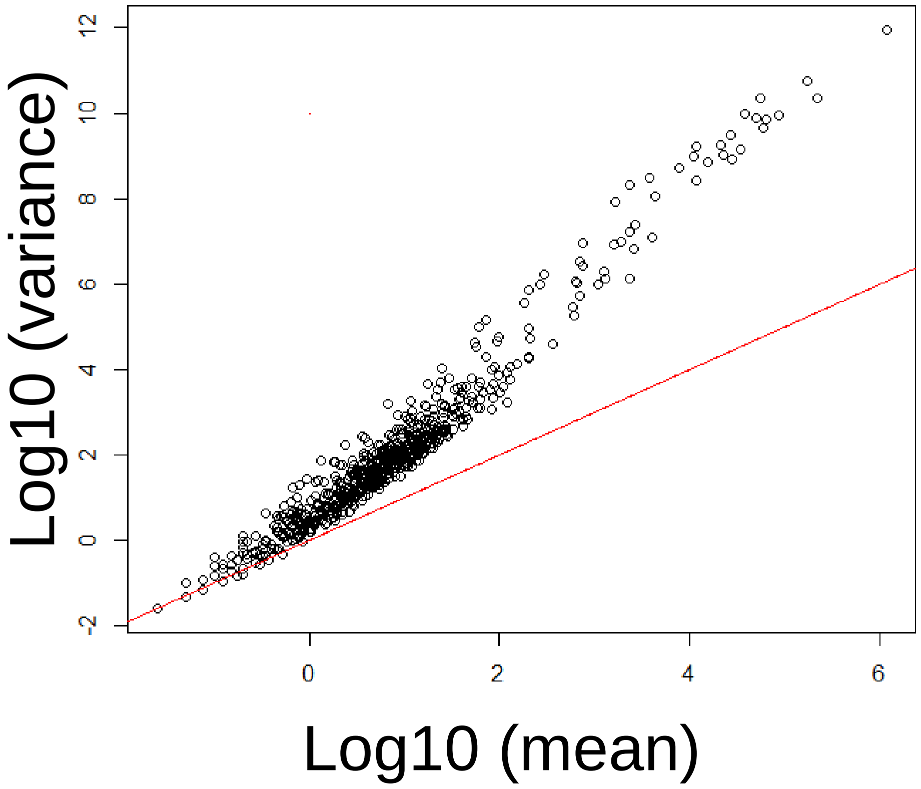
Mean = variance

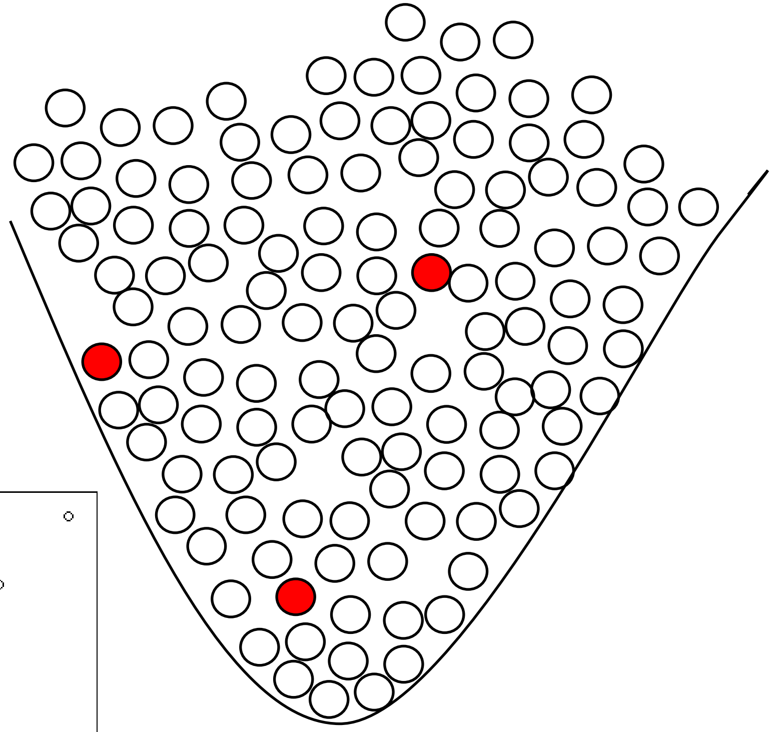
Poisson processes are reliably inadequate for modeling counts tables in genomics experiments

RNA- seq dataset



WGS dataset (through Kraken)

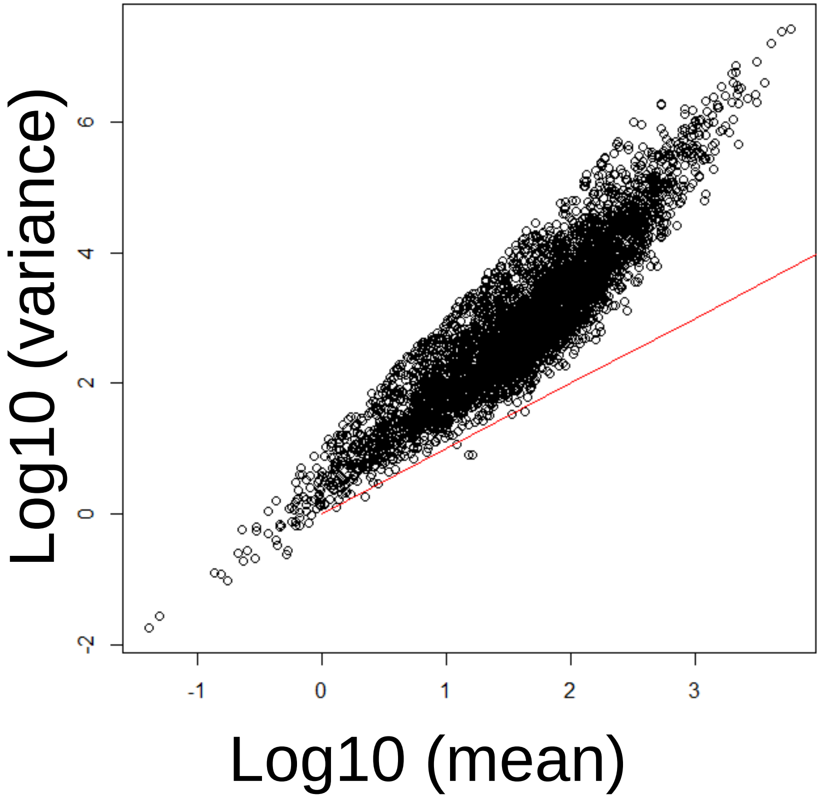




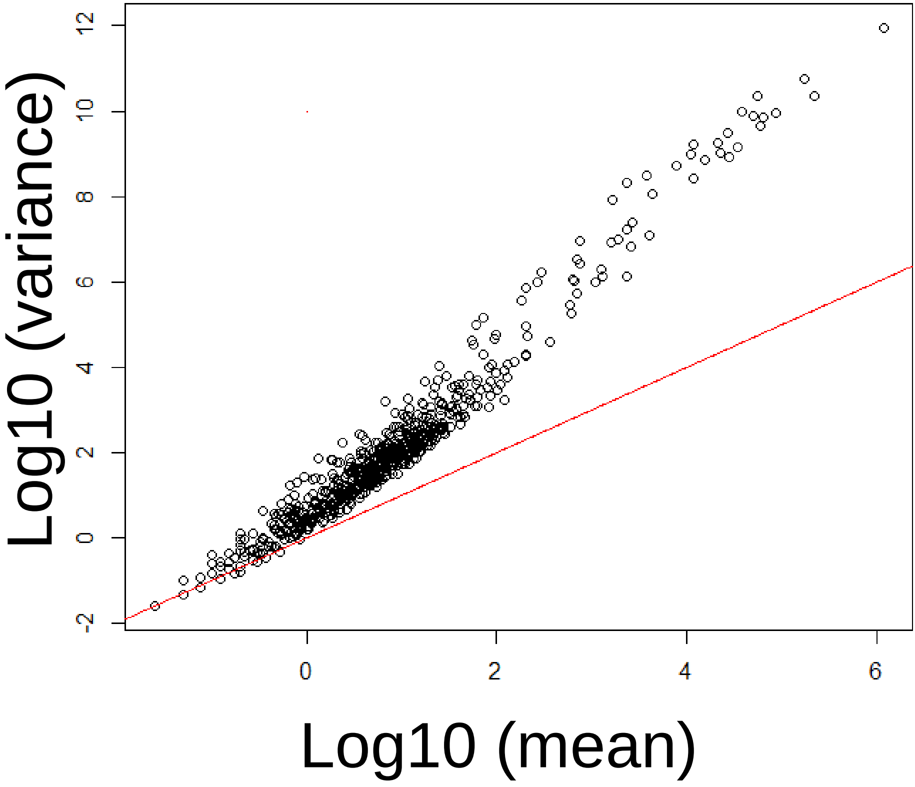
Independent draws across samples with a constant relative abundance

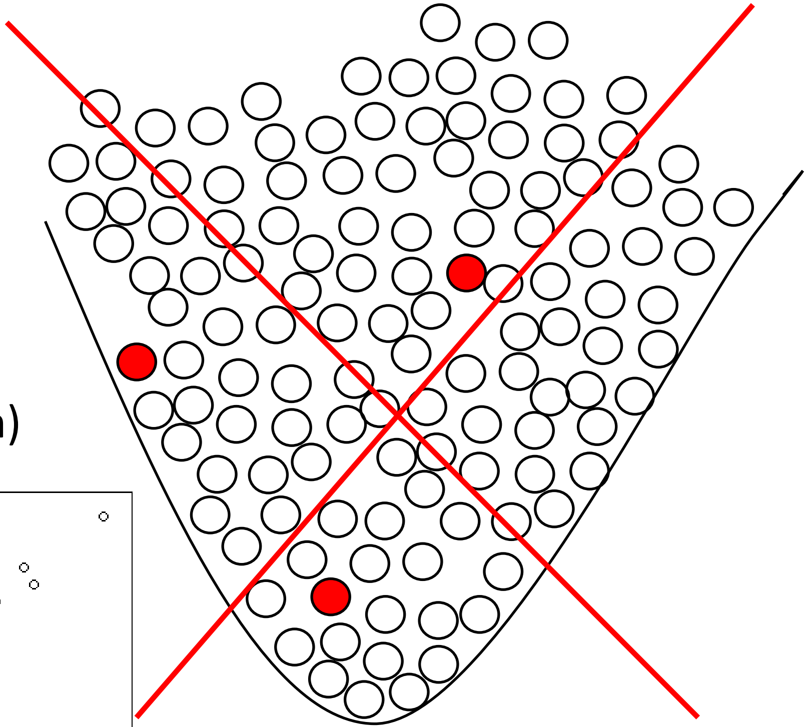
Poisson processes are reliably inadequate for modeling counts tables in genomics experiments

RNA- seq dataset



WGS dataset (through Kraken)

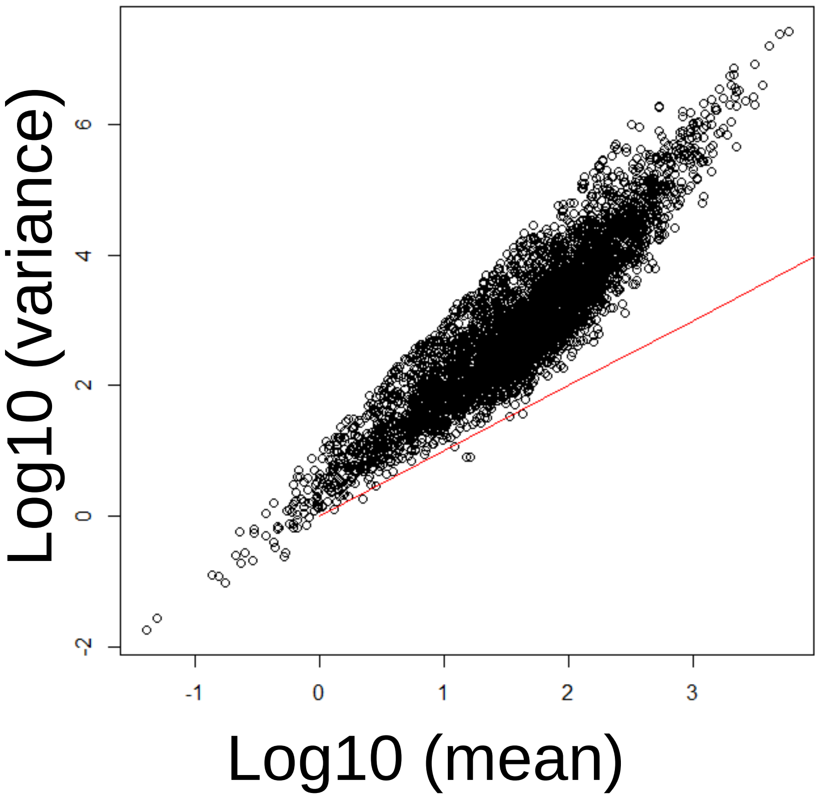




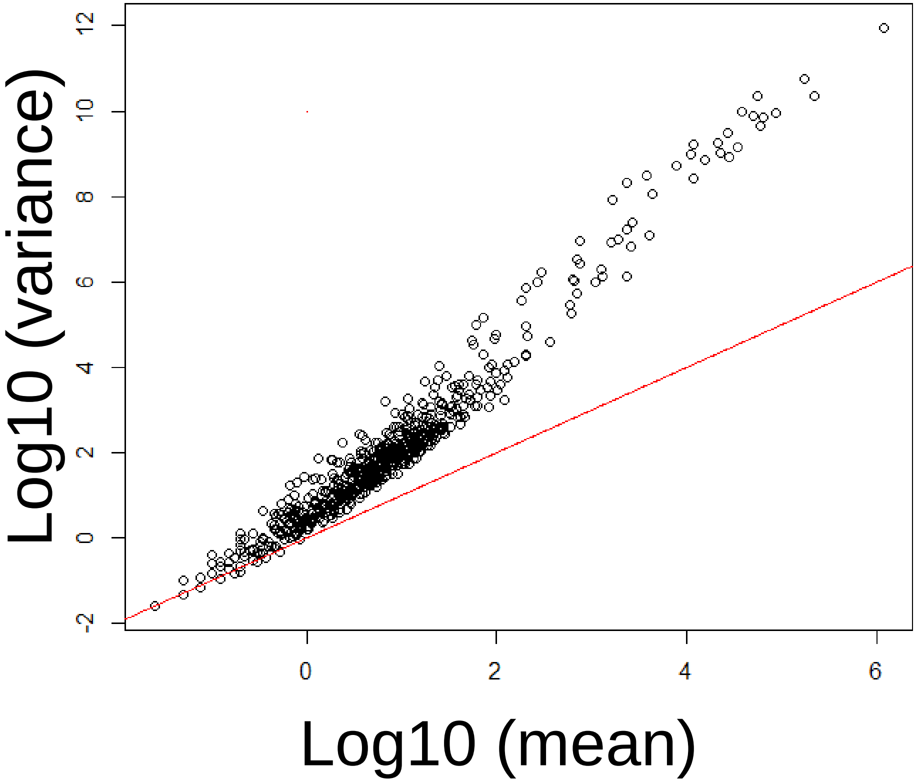
Independent draws across samples with a constant relative abundance

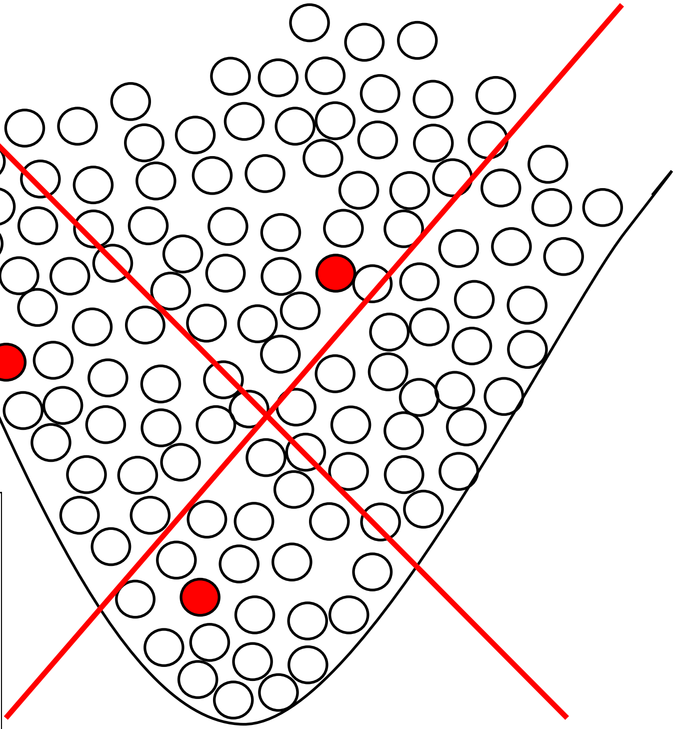
Poisson processes are reliably inadequate for modeling counts tables in genomics experiments

RNA- seq dataset



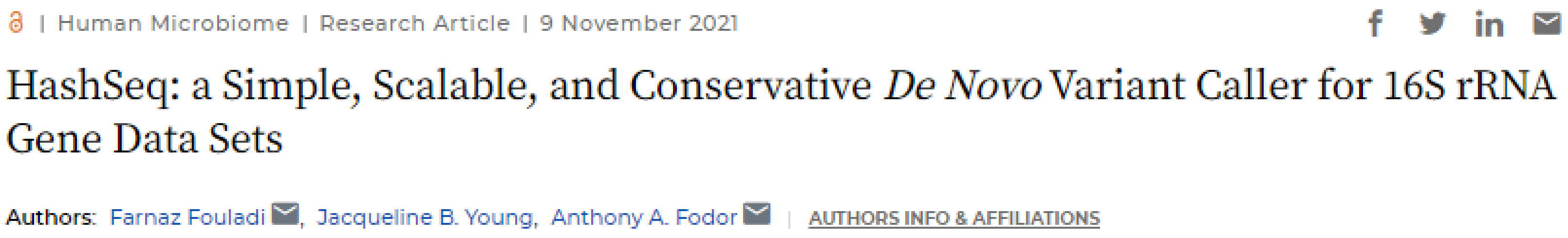
WGS dataset (through Kraken)





~~Independent draws across samples with a constant relative abundance~~

Poisson processes can be surprisingly useful in describing accumulation (richness) of sequence variants



| Human Microbiome | Research Article | 9 November 2021

HashSeq: a Simple, Scalable, and Conservative *De Novo* Variant Caller for 16S rRNA Gene Data Sets

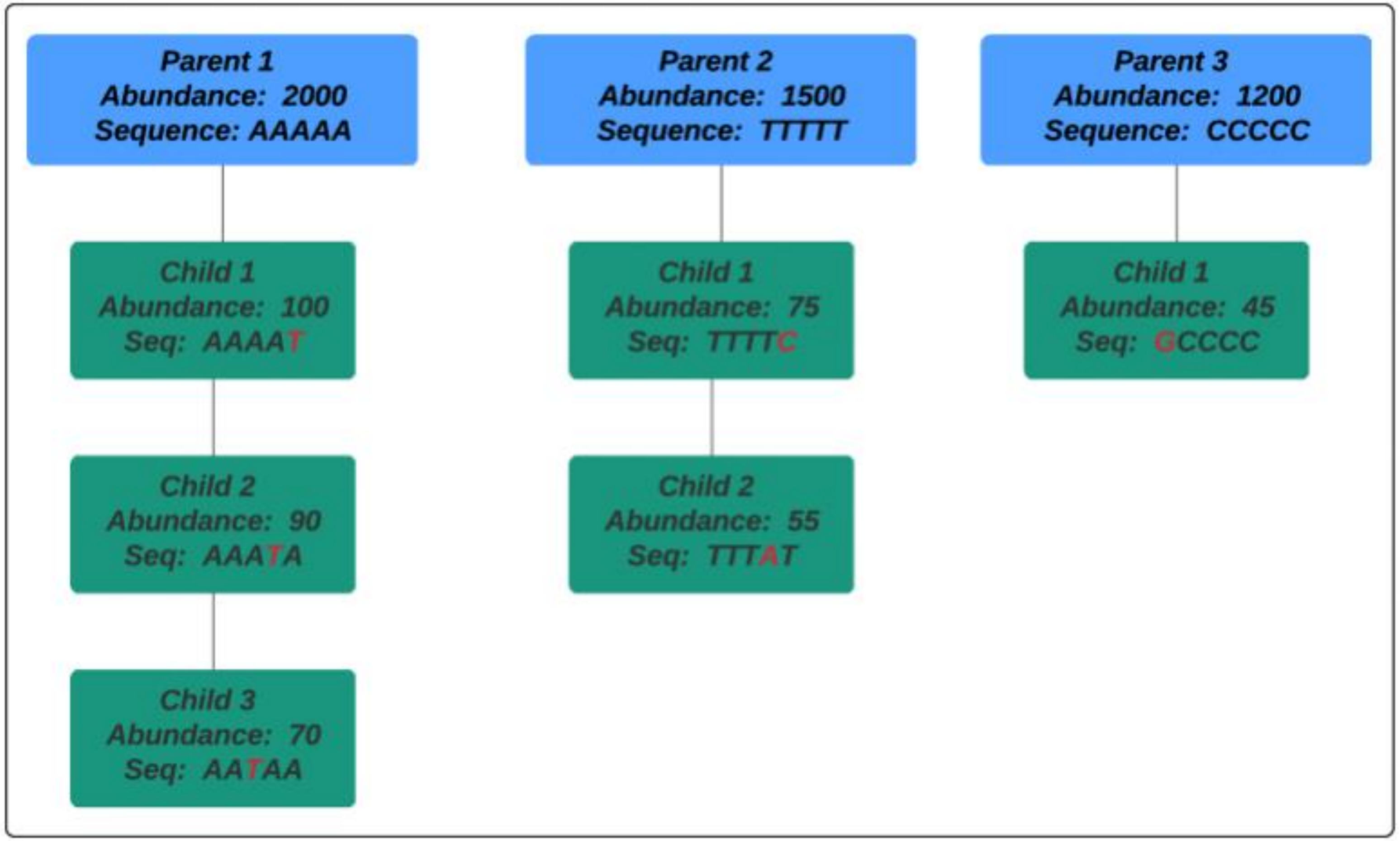
Authors: Farnaz Fouladi, Jacqueline B. Young, Anthony A. Fodor | AUTHORS INFO & AFFILIATIONS

mSystems

Farnaz Fouladi

Jack Young

Identify all unique 16S sequences in a dataset.
Sort by abundance – find children (“single mismatch”) variants



Parent 1
Abundance: 2000
Sequence: AAAAA

Child 1
Abundance: 100
Seq: AAAAT

Child 2
Abundance: 90
Seq: AAATA

Child 3
Abundance: 70
Seq: AATAA

Parent 2
Abundance: 1500
Sequence: TTTTT

Child 1
Abundance: 75
Seq: TTTTC

Child 2
Abundance: 55
Seq: TTTAT

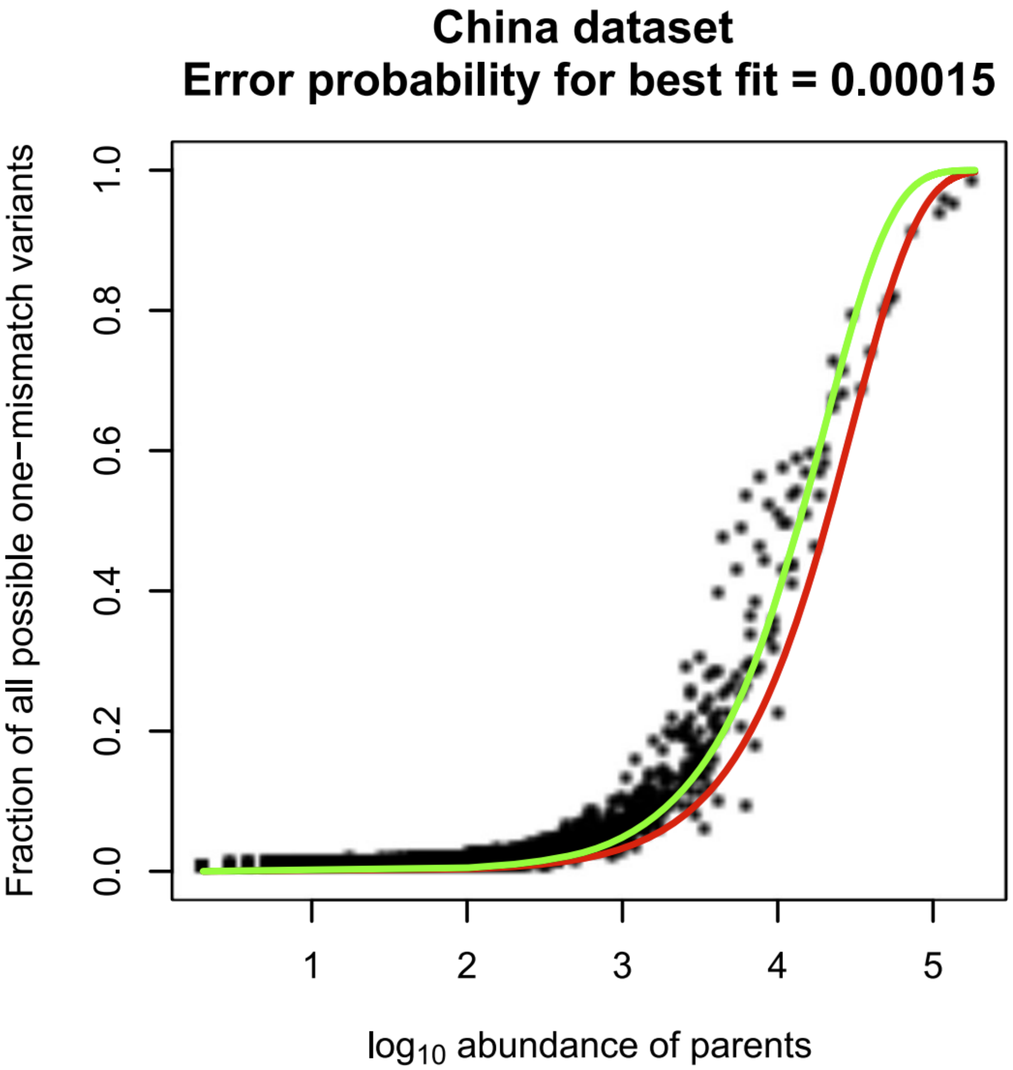
Parent 3
Abundance: 1200
Sequence: CCCCC

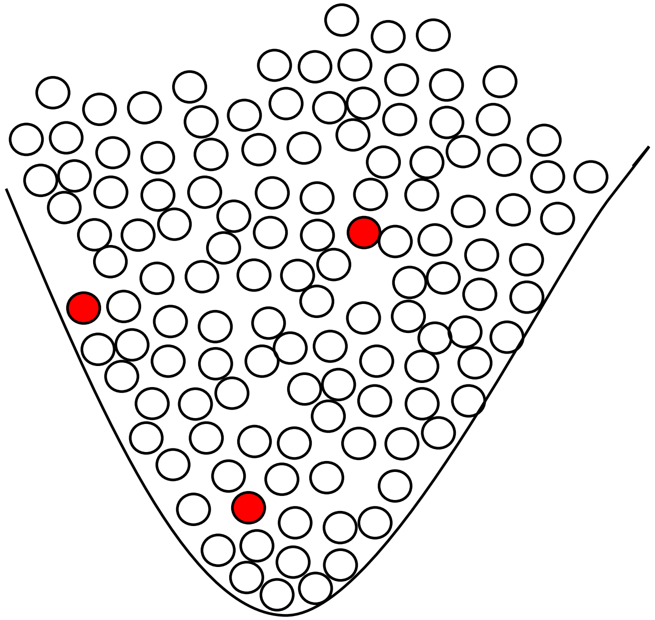
Child 1
Abundance: 45
Seq: GCCCC

FIG 1 Cluster formation of parents and their one-mismatch children in the HashSeq algorithm. In this clustering strategy, sequence variants are sorted according to their abundances. Starting with the most abundant sequence variant, considered the first parent sequence, clusters are formed by adding all the one-mismatch variants (one-mismatch children) to each cluster.

https://journals.asm.org/doi/full/10.1128/mSystems.00697-21

Richness is well described across datasets by a simple Poisson process with a constant error rate





Richness is well described across datasets by a simple Poisson process with a constant error rate

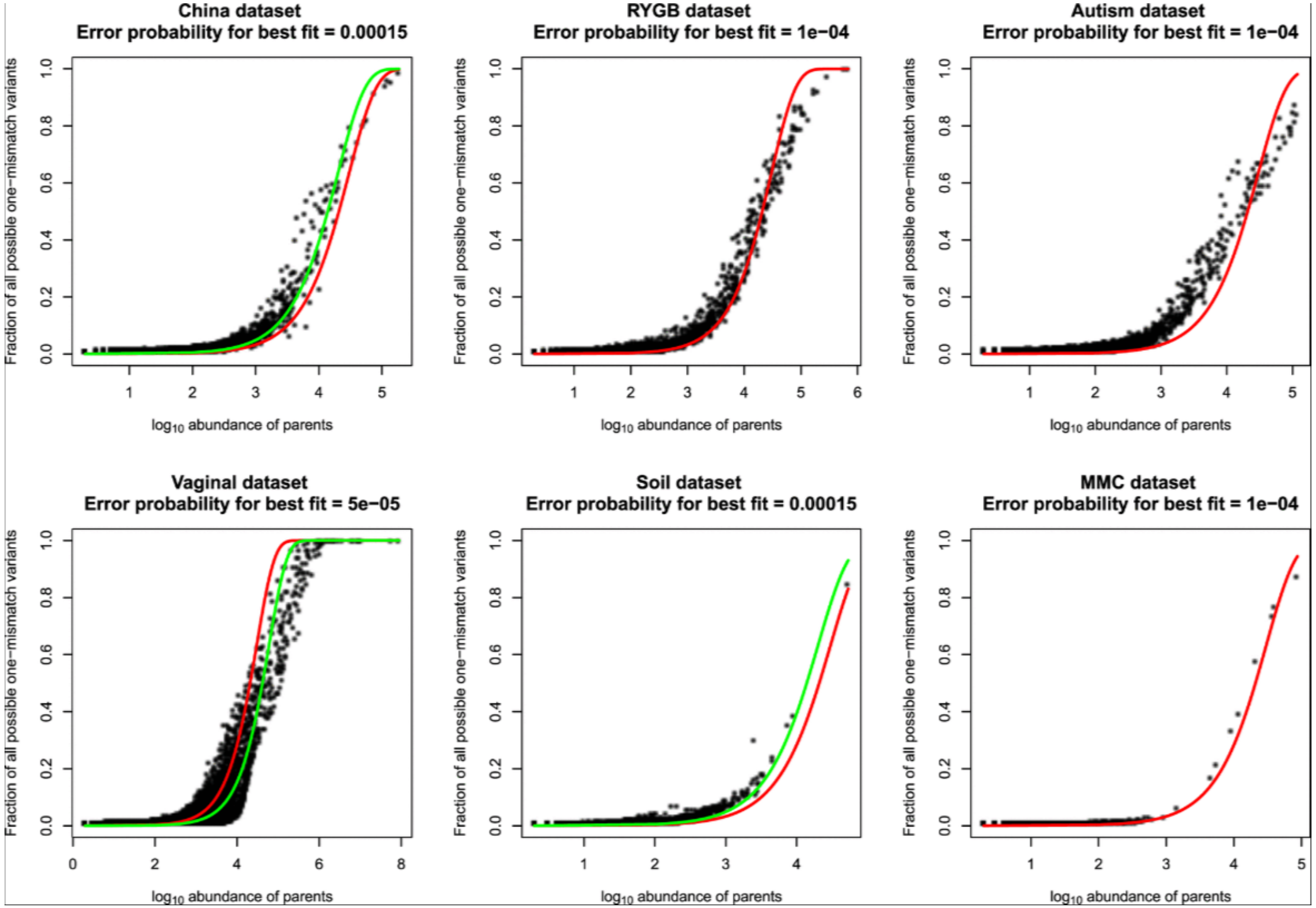
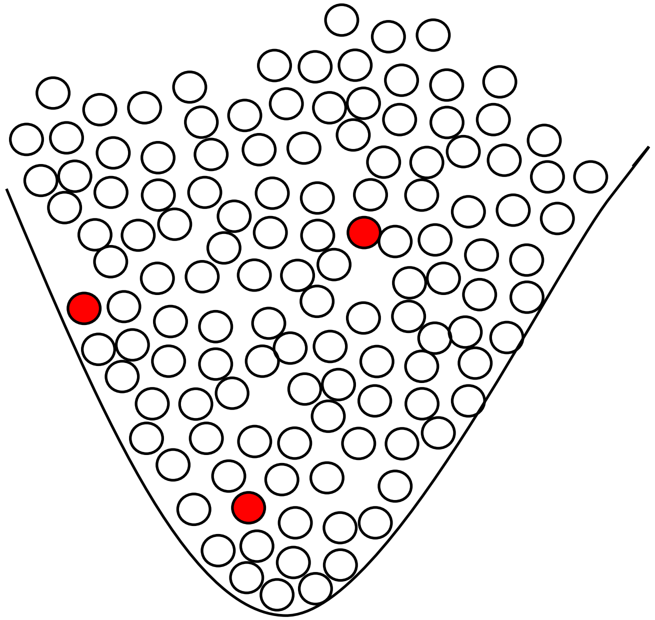


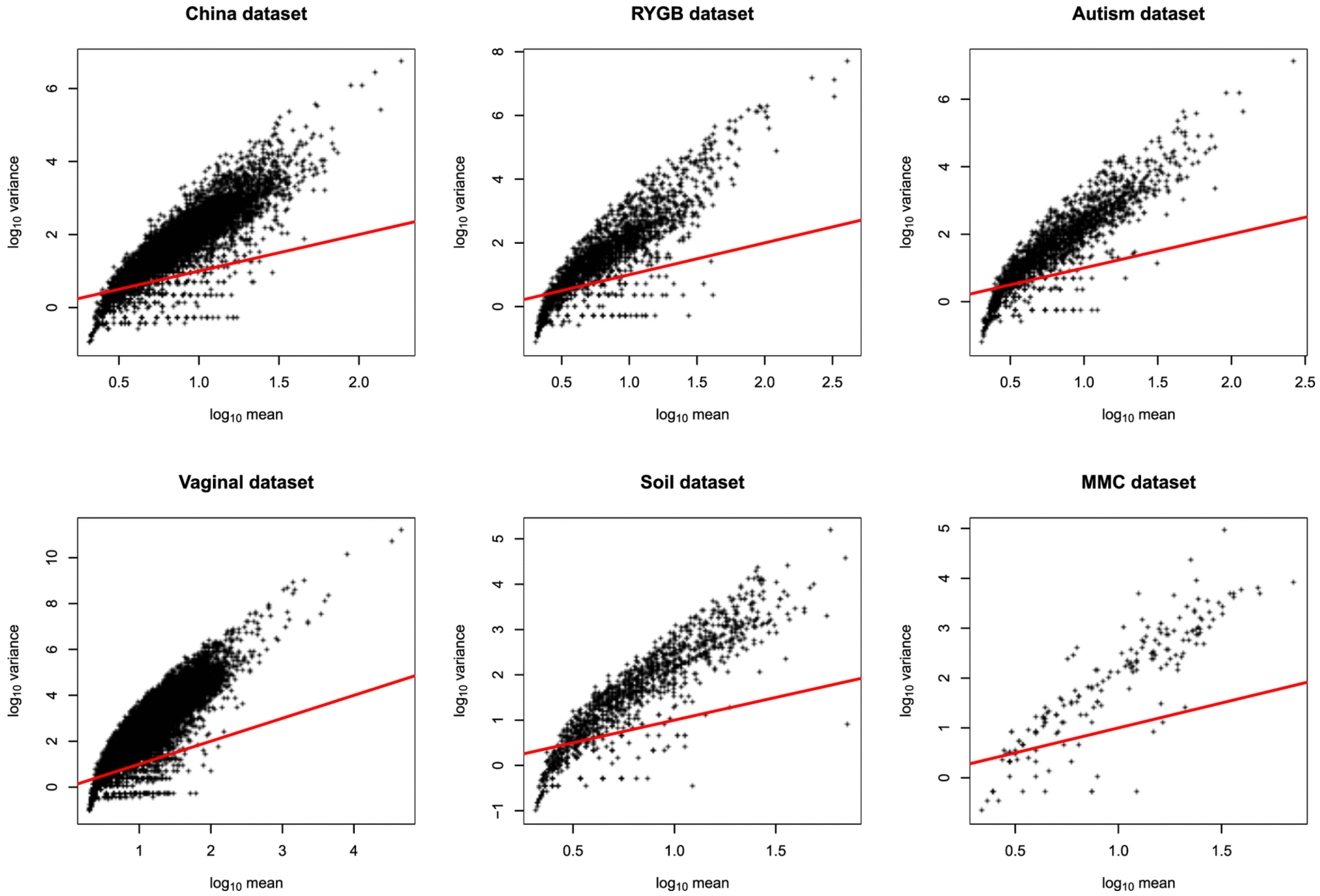
FIG 2 The presence or absence of unique one-mismatch variants can be well modeled with a simple one-parameter Poisson distribution with an almost constant error rate across six independent 16S rRNA gene Illumina data sets. Plots show the relationship between the abundance of parent sequences on the $\log_{10}$ scale and the fraction of all possible unique one-mismatch variants for the parent sequences. These data are we modeled by a simple one-parameter Poisson distribution. The red line corresponds to an error rate P of 10^{-4}. The China, vaginal, and soil data sets were best modeled using slightly different error rates for each data set (green lines, China and soil $P = 1.5 \times 10^{-4}$ and soil $P = 5 \times 10^{-5}$, respectively).

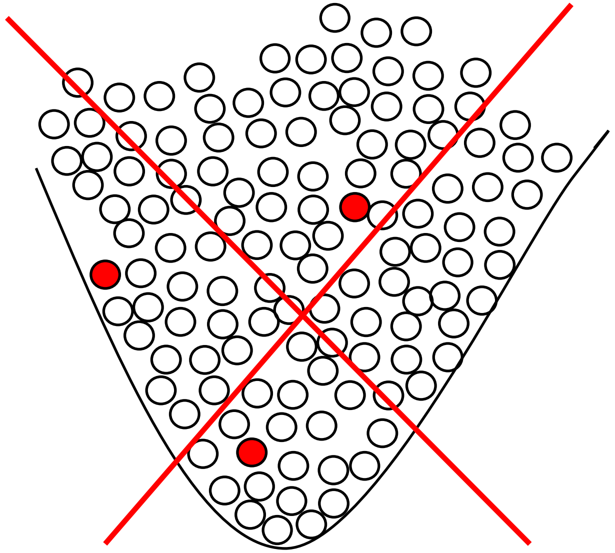


🟥 Constant error rate

🟩 Best error rate for that data set

Abundance (as usual) is not well fit by Poisson assumptions





Richness is well described across datasets by a simple Poisson process with a constant error rate

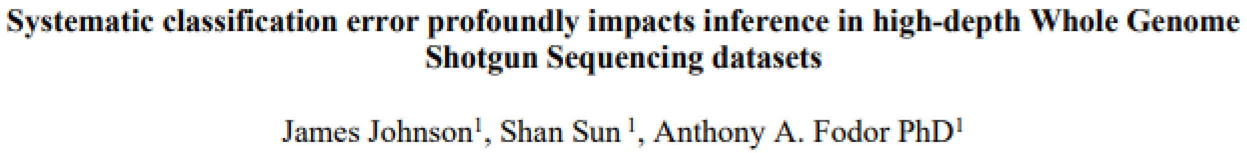
Abundance (as usual) is not well fit by Poisson assumptions

One hypothesis:

In 16S experiments, initial errors accumulate by taq sequencing error (a Poisson process)

The final abundance is dependent on PCR amplification (not a Poisson process)

Surprisingly, Poisson algorithms can also be of utility in shotgun sequencing datasets



Systematic classification error profoundly impacts inference in high-depth Whole Genome Shotgun Sequencing datasets

James Johnson[1], Shan Sun [1], Anthony A. Fodor PhD[1]

James Johnson

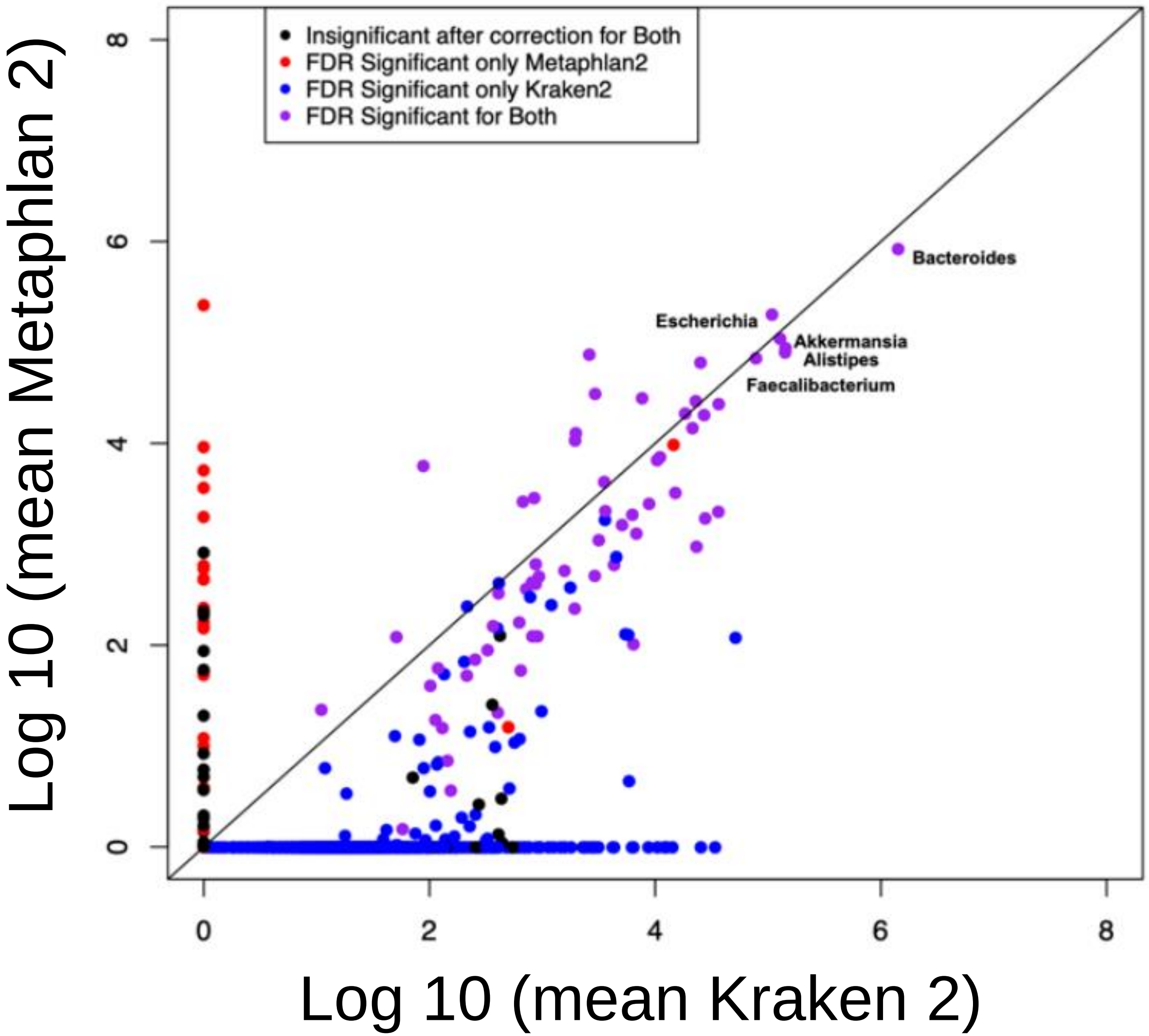
Shan Sun

Bioarchive (and unpublished!)

https://www.biorxiv.org/content/10.1101/2022.04.04.487034v2.abstract

Kraken and Metaphlan agree on high-abundance taxa but not on low-abundance taxa

Kraken finds not only more taxa but more taxa significantly associated with metadata



Inference is case/control
For IBD at a 5% FDR threshold

Can we evaluate the algorithms even though we don’t know the “correct” answer…

Examine the correlation structure of predictions

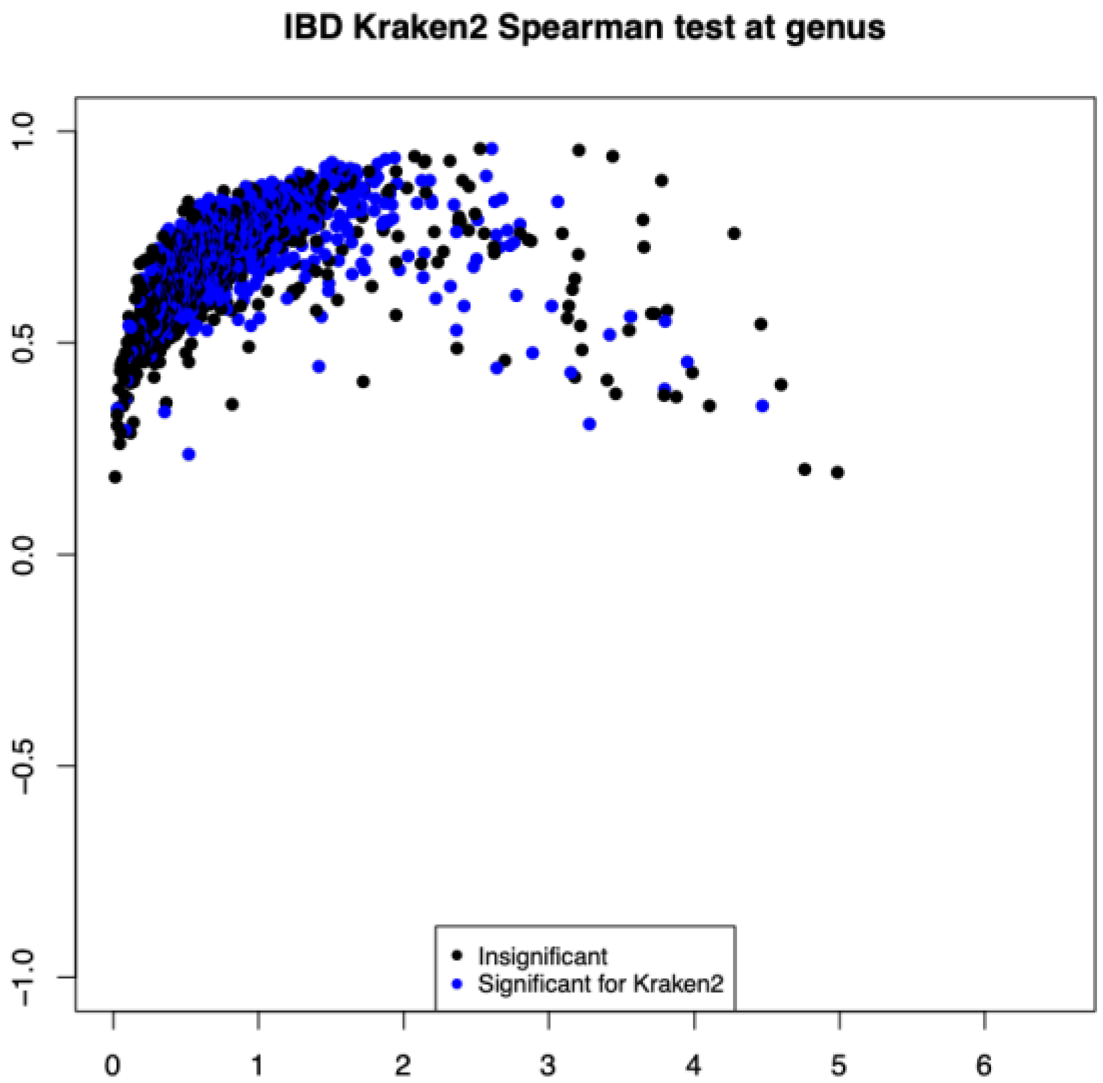
Sort all taxa by abundance

→ 1st most abundance – Bacteroides
→ 2nd most abundance – Escherichia
3rd most abundant – Akkermansia
4th most abundant – Alistipes
…
…

For each taxa, report the highest correlation coefficient among all more abundant taxa

So for the 2nd most abundant, this will be the correlation with the 1st most abundance
For the 3rd most abundant, this will be the max(cor(3rd, 2nd), cor(3rd , 1st))
For the 4th most abundance, this be the max(cor(4th, 1st), cor(4th, 2nd), cor(4th, 3rd))
And so forth…

Many of taxa for Kraken are highly correlated with a more abundant “parent” taxa



Rho

Log 10 (mean Kraken 2)

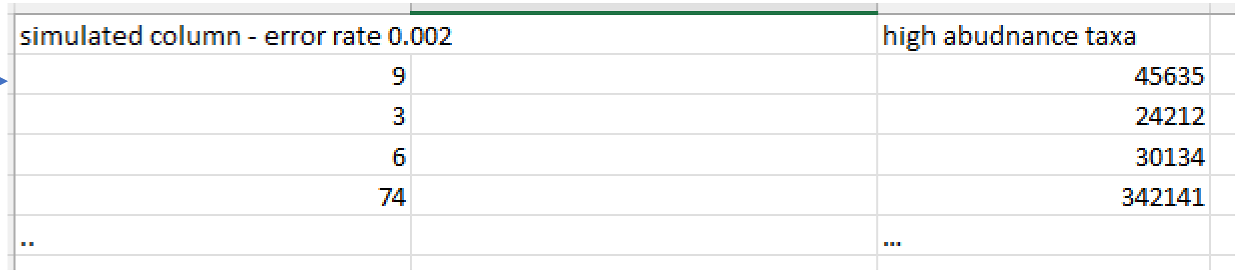
We can model this behavior with a simple Poisson-based procedure with a small # of free parameters

Assume the top 10 taxa are “real”.

Simulate the rest of the dataset as Poisson based sampling error:

for each “simulated” taxa
randomly choose one of
randomly choose an error rate over some range (e.g. 0 < error <= 0.002)

sum(rpois(45635, lambda = .0002))



simulated column - error rate 0.002		high abudnance taxa
9		45635
3		24212
6		30134
74		342141
..		...

We can model this behavior with a simple Poisson-based procedure with a small # of free parameters

Assume the top 10 taxa are “real”.

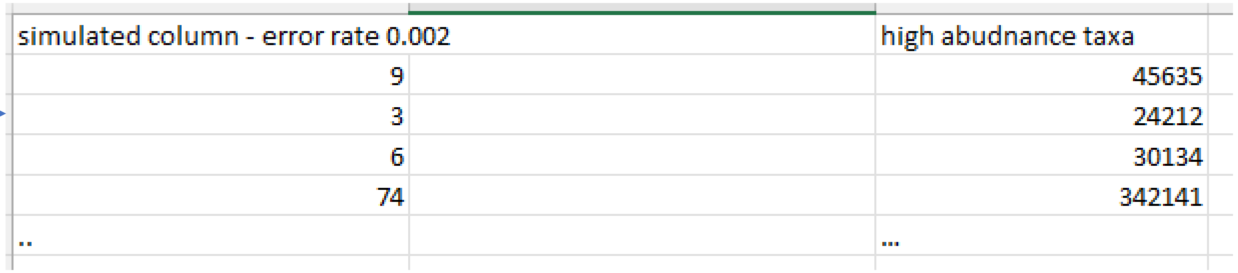
Simulate the rest of the dataset as Poisson based sampling error:

for each “simulated” taxa

randomly choose one of

randomly choose an error rate over some range (e.g. 0 < error <= 0.002)

sum(rpois(24212, lambda = .0002))



simulated column - error rate 0.002		high abudnance taxa
9		45635
3		24212
6		30134
74		342141
..		...

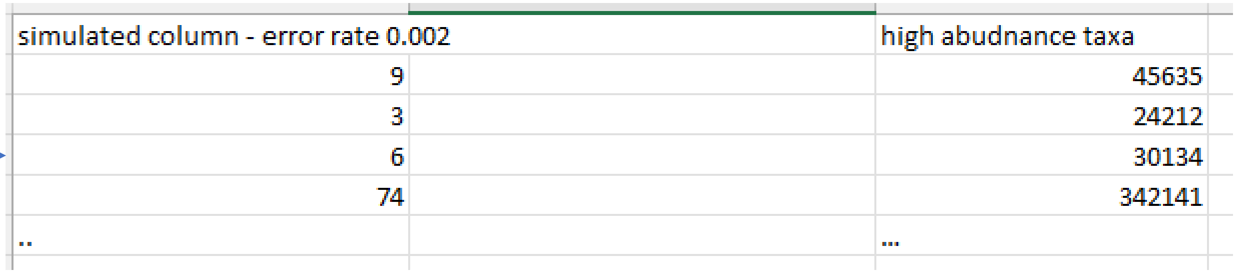
We can model this behavior with a simple Poisson-based procedure with a small # of free parameters

Assume the top 10 taxa are “real”.

Simulate the rest of the dataset as Poisson based sampling error:

for each “simulated” taxa
randomly choose one of
randomly choose an error rate over some range (e.g. 0 < error <= 0.002)

sum(rpois(30134, lambda = .0002))



simulated column - error rate 0.002		high abudnance taxa
9		45635
3		24212
6		30134
74		342141
..		...

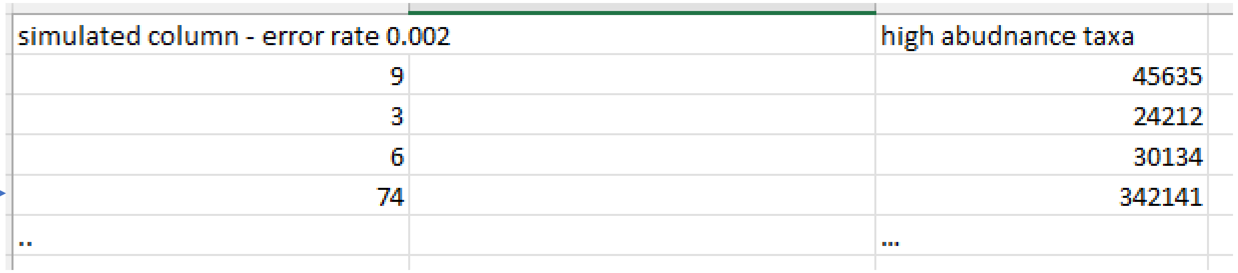
We can model this behavior with a simple Poisson-based procedure with a small # of free parameters

Assume the top 10 taxa are “real”.

Simulate the rest of the dataset as Poisson based sampling error:

for each “simulated” taxa
randomly choose one of
randomly choose an error rate over some range (e.g. 0 < error <= 0.002)

sum(rpois(342141, lambda = .0002))



simulated column - error rate 0.002		high abudnance taxa
9		45635
3		24212
6		30134
74		342141
..		...

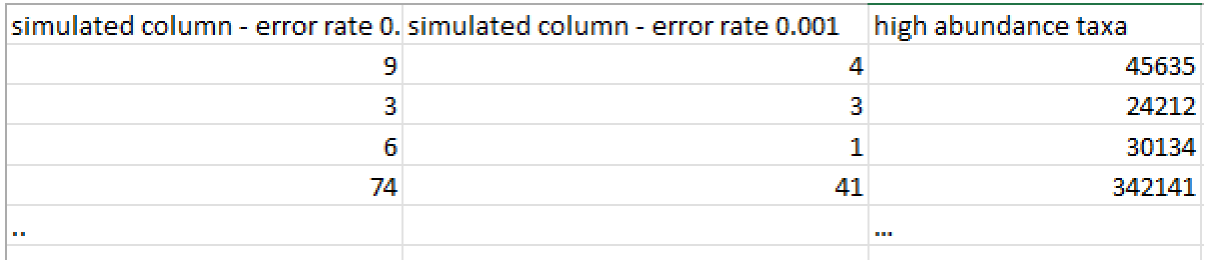
We can model this behavior with a simple Poisson-based procedure with a small # of free parameters

Assume the top 10 taxa are “real”.

Simulate the rest of the dataset as Poisson based sampling error:

for each “simulated” taxa
randomly choose one of
randomly choose an error rate over some range (e.g. 0 < error <= 0.002)

sum(rpois(45635, lambda = .0001))



simulated column - error rate 0.	simulated column - error rate 0.001	high abundance taxa
9	4	45635
3	3	24212
6	1	30134
74	41	342141
..		...

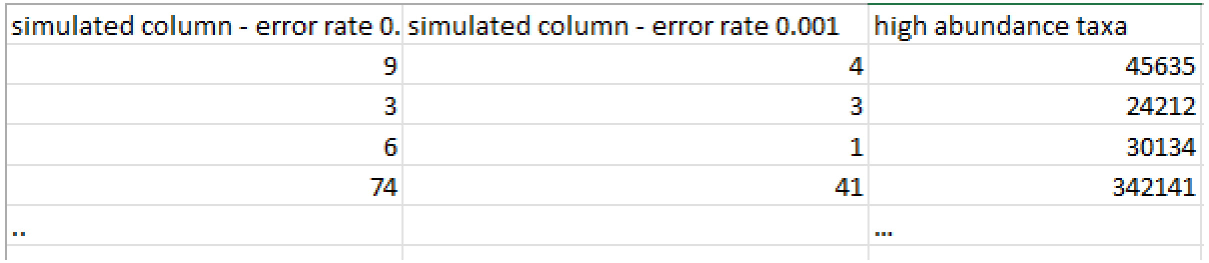
We can model this behavior with a simple Poisson-based procedure with a small # of free parameters

Assume the top 10 taxa are “real”.

Simulate the rest of the dataset as Poisson based sampling error:

    for each “simulated” taxa
    randomly choose one of
    randomly choose an error rate over some range (e.g. 0 < error <= 0.002)

sum(rpois(24212, lambda = .0001))



simulated column - error rate 0.	simulated column - error rate 0.001	high abundance taxa
9	4	45635
3	3	24212
6	1	30134
74	41	342141
..		...

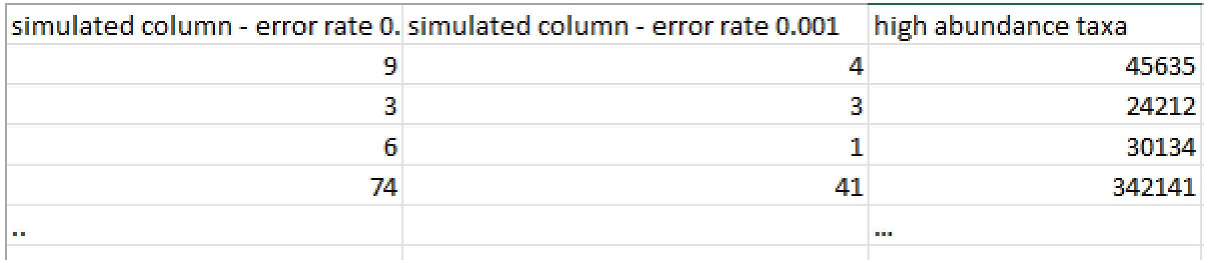
We can model this behavior with a simple Poisson-based procedure with a small # of free parameters

Assume the top 10 taxa are “real”.

Simulate the rest of the dataset as Poisson based sampling error:

```
    for each “simulated” taxa
    randomly choose one of
    randomly choose an error rate over some range (e.g. 0 < error <= 0.002 )
```

sum(rpois(30134, lambda = .0001))



simulated column - error rate 0.	simulated column - error rate 0.001	high abundance taxa
9	4	45635
3	3	24212
6	1	30134
74	41	342141
..		...

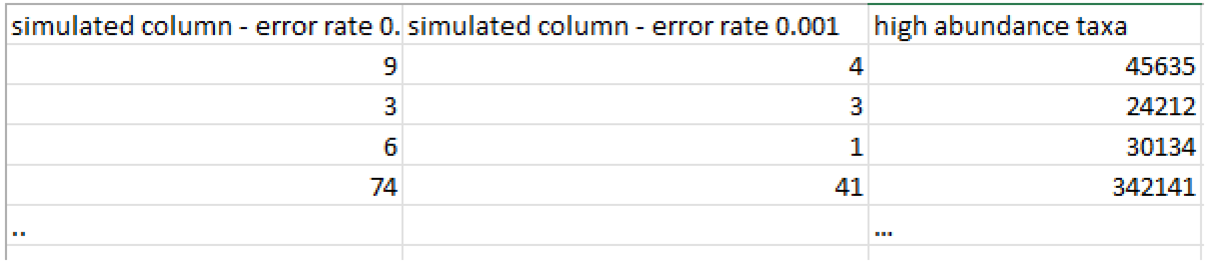
We can model this behavior with a simple Poisson-based procedure with a small # of free parameters

Assume the top 10 taxa are “real”.

Simulate the rest of the dataset as Poisson based sampling error:

for each “simulated” taxa
randomly choose one of
randomly choose an error rate over some range (e.g. 0 < error <= 0.002)

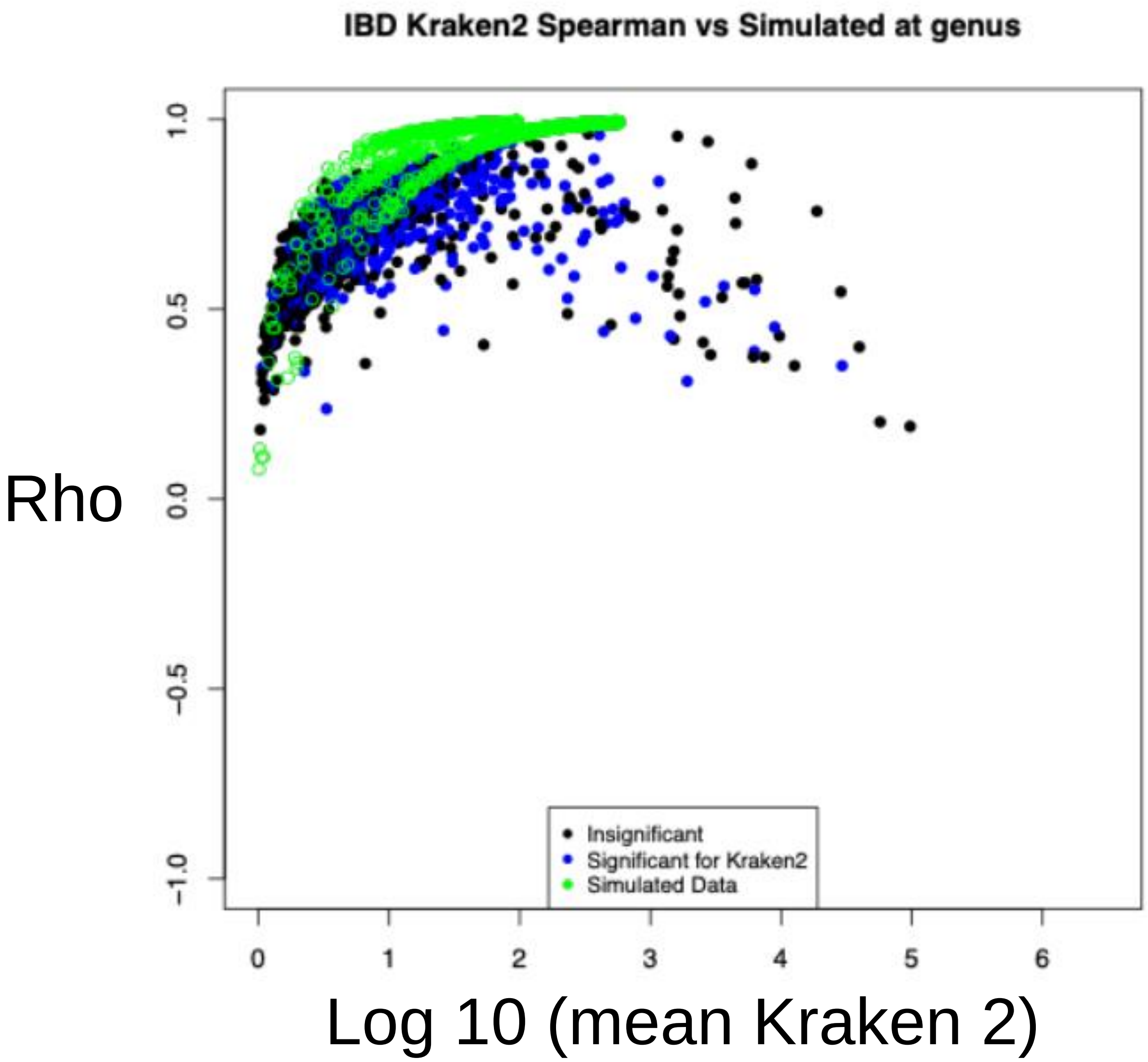
sum(rpois(342141, lambda = .0001))



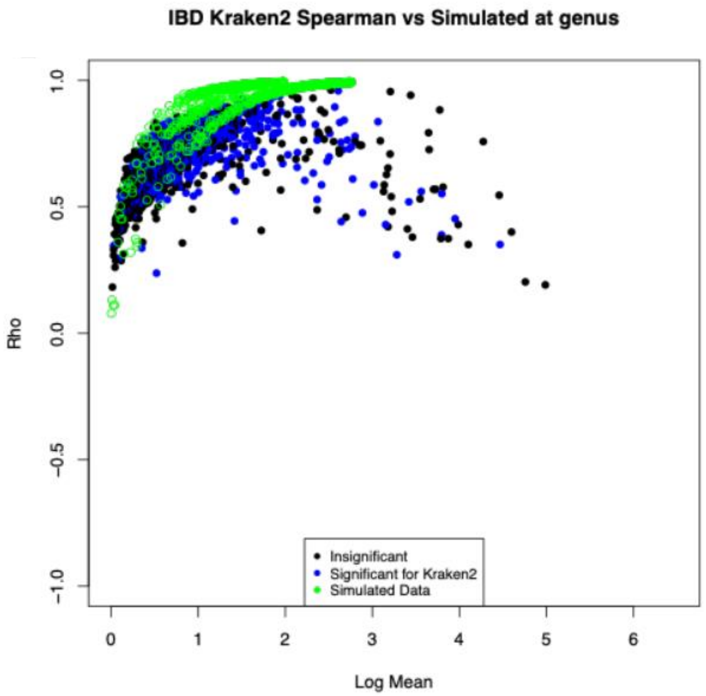
simulated column - error rate 0.	simulated column - error rate 0.001	high abundance taxa
9	4	45635
3	3	24212
6	1	30134
74	41	342141
..		...

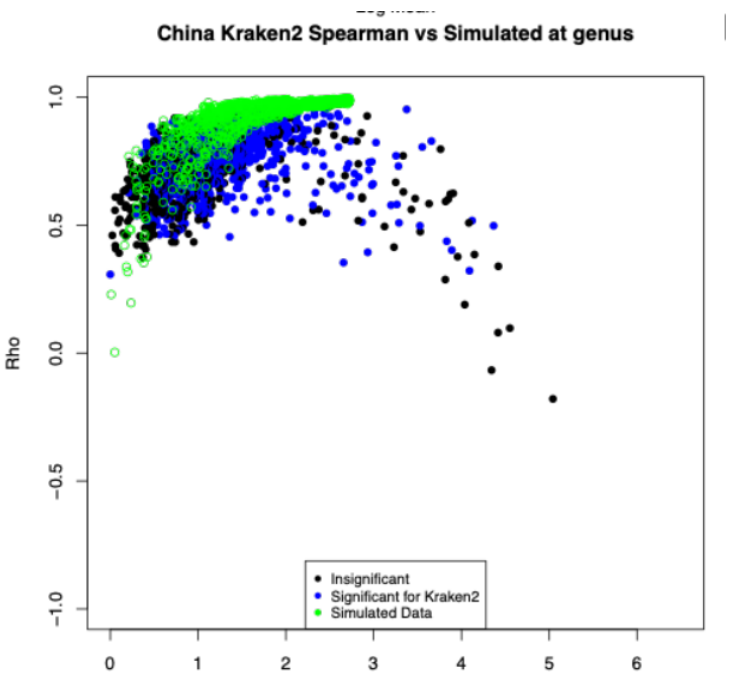
In this way we simulate the entire dataset assuming that everything except the most abundant taxa is Poisson-based classification error of the most abundant taxa

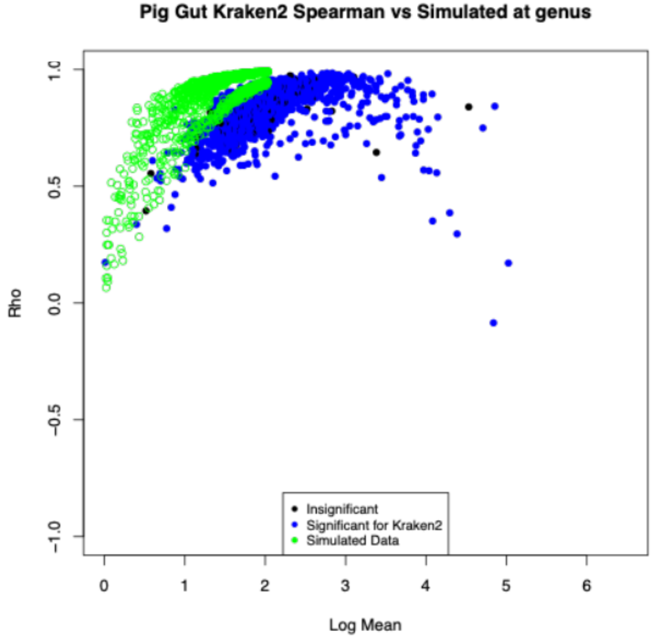
Somewhat remarkably, this simple model captures much of the behavior of low-abundance Kraken taxa

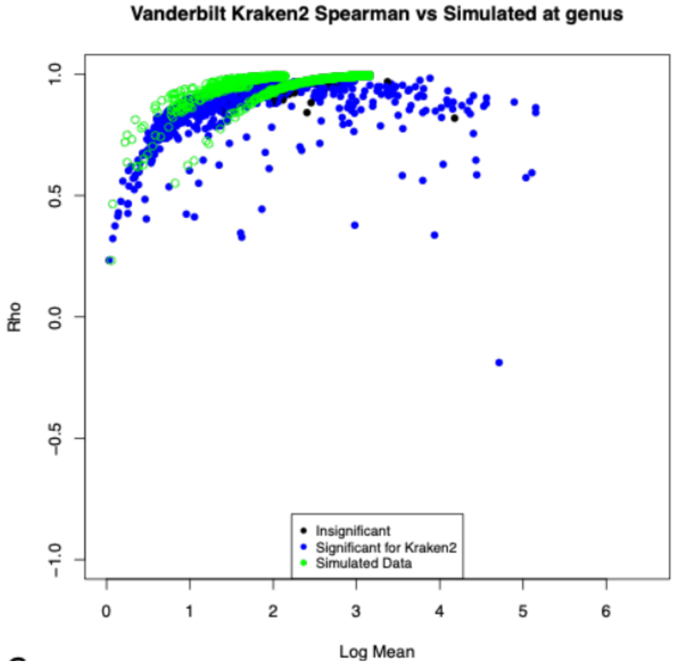


A constant error rate fits three of four datasets very well

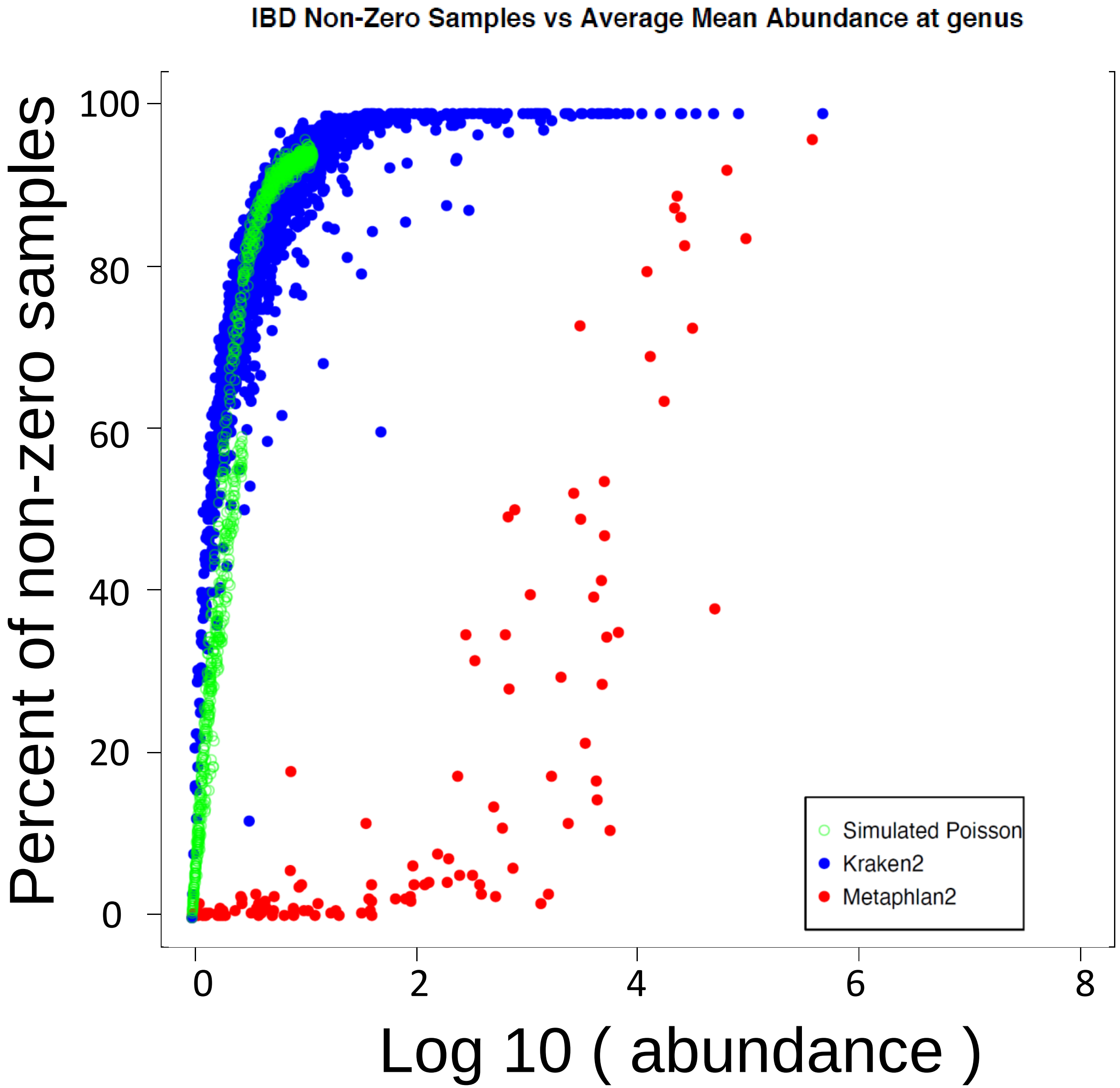

IBD Kraken2 Spearman vs Simulated at genus


China Kraken2 Spearman vs Simulated at genus


Pig Gut Kraken2 Spearman vs Simulated at genus


Vanderbilt Kraken2 Spearman vs Simulated at genus

We can explain much of the prevalence relationship from Kraken with our Poisson model (alas, with a different background error rate….)



Mis-classification events from k-mer classifiers of WGS can be well modeled
with a Poisson distribution with no biology in the null model

Spurious correlations can be problematic for network analysis

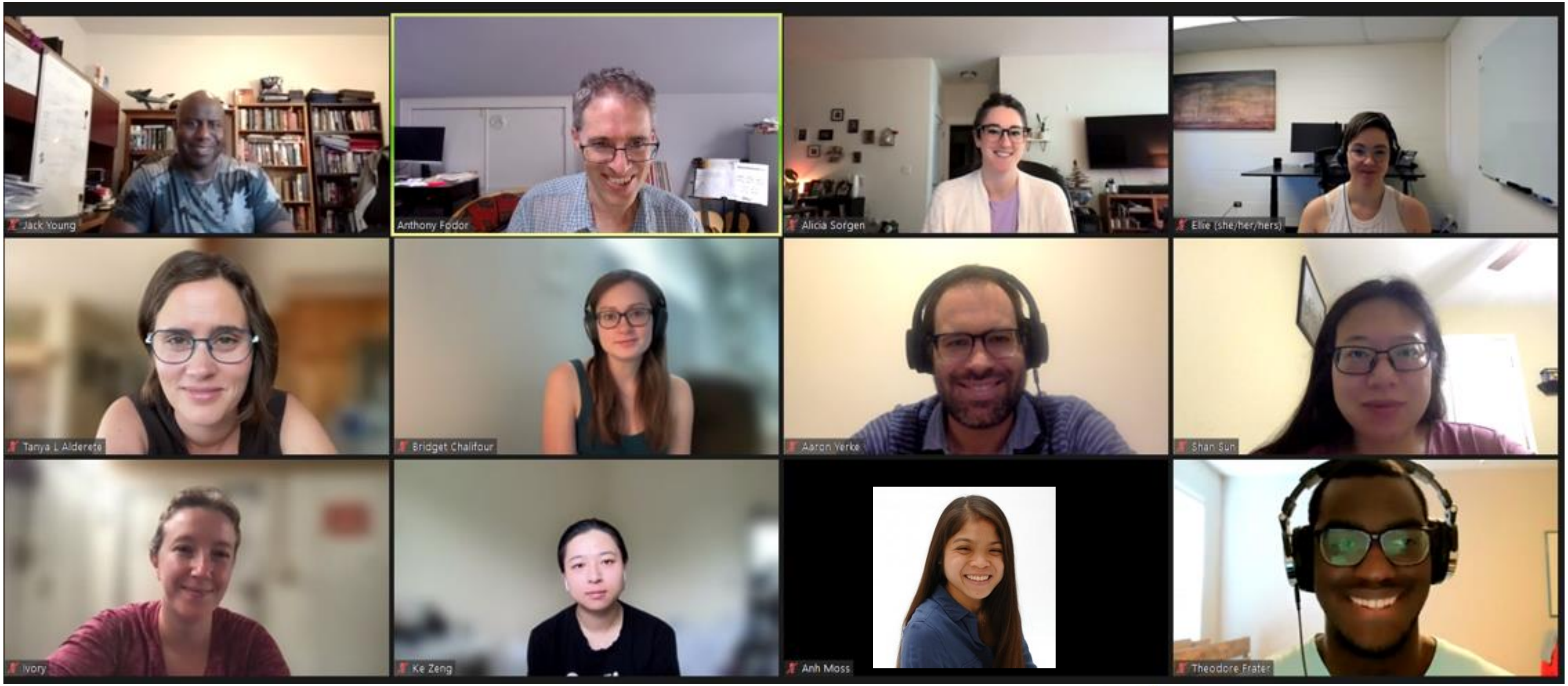
Low abundance taxa with high correlations to high abundance taxa should be independently confirmed
as being actually present and not “phantom”

Filtering thresholds for WGS datasets should be set from abundance (not prevalence!)

Mis-classification events become more likely as sequencing depth and database density increase!

Error models may allow us to capture background expectations and evaluate null hypotheses
that a given observation of a taxa can be explained by background error rate calculations…

We have such a model for 16S ASVs and are working towards that in WGS



UNC CHARLOTTE

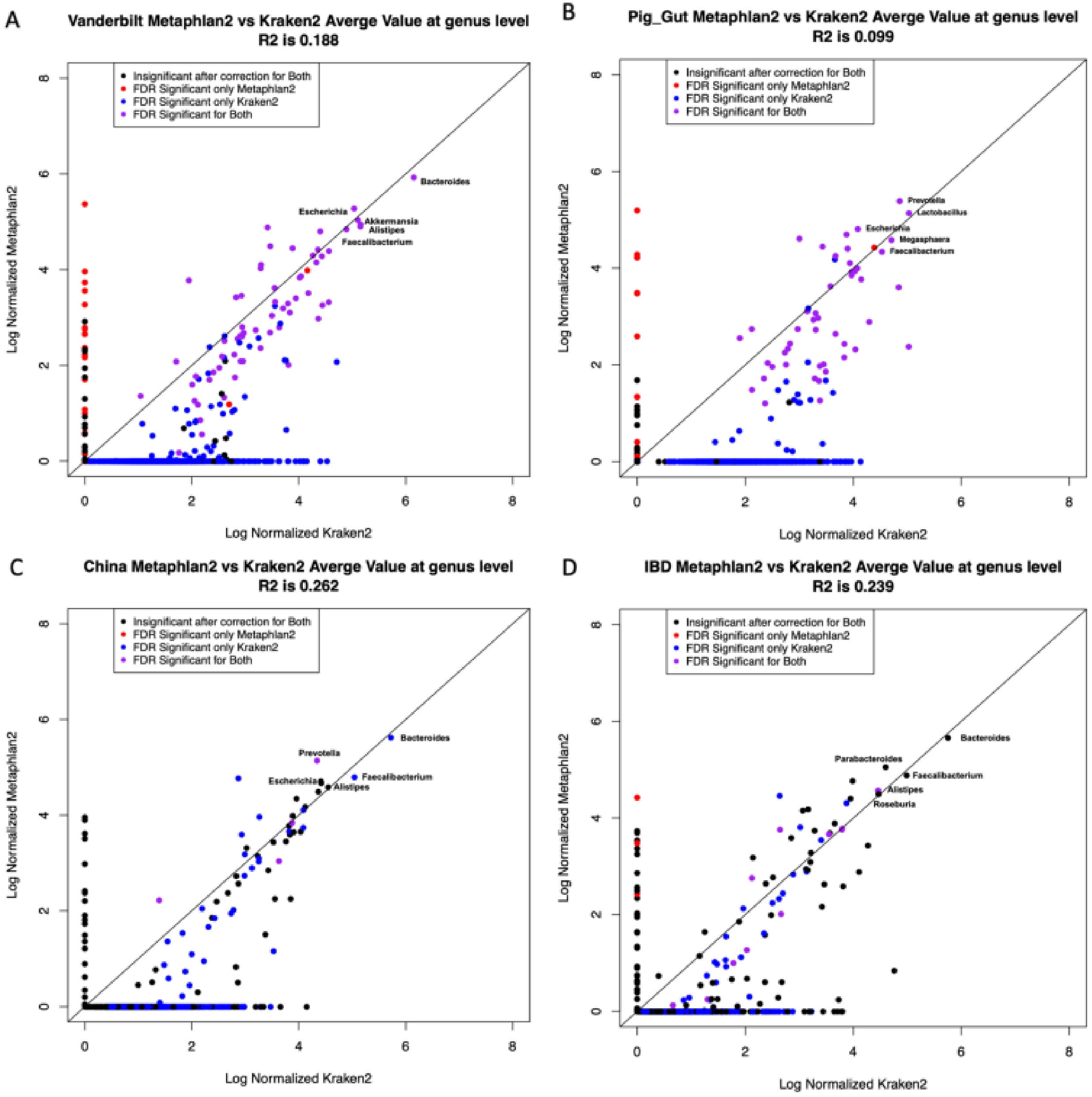
NATIONAL INSTITUTES OF HEALTH

Arthritis Foundation®

NSF

Kraken and Metaphlan agree on high-abundance taxa but not on low-abundance taxa

Kraken finds not only more taxa but more taxa significantly associated with metadata


A
Vanderbilt Metaphlan2 vs Kraken2 Averge Value at genus level
R2 is 0.188
● Insignificant after correction for Both
● FDR Significant only Metaphlan2
● FDR Significant only Kraken2
● FDR Significant for Both
Bacteroides
Escherichia
Akkermansia
Alistipes
Faecalibacterium
Log Normalized Metaphlan2
Log Normalized Kraken2
B
Pig_Gut Metaphlan2 vs Kraken2 Averge Value at genus level
R2 is 0.099
● Insignificant after correction for Both
● FDR Significant only Metaphlan2
● FDR Significant only Kraken2
● FDR Significant for Both
Prevotella
Lactobacillus
Escherichia
Megasphaera
Faecalibacterium
Log Normalized Metaphlan2
Log Normalized Kraken2
C
China Metaphlan2 vs Kraken2 Averge Value at genus level
R2 is 0.262
● Insignificant after correction for Both
● FDR Significant only Metaphlan2
● FDR Significant only Kraken2
● FDR Significant for Both
Bacteroides
Prevotella
Faecalibacterium
Escherichia
Alistipes
Log Normalized Metaphlan2
Log Normalized Kraken2
D
IBD Metaphlan2 vs Kraken2 Averge Value at genus level
R2 is 0.239
● Insignificant after correction for Both
● FDR Significant only Metaphlan2
● FDR Significant only Kraken2
● FDR Significant for Both
Bacteroides
Parabacteroides
Faecalibacterium
Alistipes
Roseburia
Log Normalized Metaphlan2
Log Normalized Kraken2