Efficient Statistical Significance Approximation for Local Association Analysis of High-Throughput Time Series Data

Li C. Xia *, Dongmei Ai *, Jacob Cram , Jed A. Fuhrman , Fengzhu Sun †

September 7, 2012

1 Supplementary Results

		The number of time points n											
x	Theory	10	20	30	40	60	80	100	120	140	160	180	200
2	0.1815	0.0848	0.0987	0.1062	0.1122	0.1201	0.1235	0.1290	0.1294	0.1355	0.1375	0.1430	0.1379
2.2	0.1111	0.0541	0.0621	0.0645	0.0665	0.0699	0.0771	0.0767	0.0783	0.0798	0.0798	0.0861	0.0831
2.4	0.0656	0.0341	0.0367	0.0392	0.0416	0.0411	0.0435	0.0457	0.0451	0.0477	0.0483	0.0526	0.0498
2.6	0.0373	0.0223	0.0221	0.0252	0.0235	0.0232	0.0249	0.0261	0.0253	0.0261	0.0276	0.0301	0.0275
2.8	0.0204	0.0147	0.0128	0.0154	0.0131	0.0129	0.0138	0.0163	0.0129	0.0141	0.0159	0.0159	0.0152
3.0	0.0108	0.0093	0.0082	0.0088	0.0074	0.0069	0.0071	0.0090	0.0072	0.0066	0.0087	0.0083	0.0074
3.2	0.0055	0.0056	0.0051	0.0038	0.0036	0.0030	0.0035	0.0054	0.0040	0.0042	0.0043	0.0043	0.0038
3.4	0.0027	0.0033	0.0031	0.0017	0.0022	0.0009	0.0016	0.0027	0.0019	0.0018	0.0027	0.0028	0.0017
3.6	0.0013	0.0019	0.0020	0.0011	0.0014	0.0004	0.0006	0.0012	0.0011	0.0007	0.0012	0.0015	0.0008
3.8	0.0006	0.0007	0.0008	0.0006	0.0010	0.0002	0.0004	0.0009	0.0007	0.0002	0.0008	0.0008	0.0004
4.0	0.0003	0.0004	0.0005	0.0003	0.0005	0.0000	0.0003	0.0004	0.0004	0.0001	0.0002	0.0002	0.0001
4.2	0.0001	0.0002	0.0004	0.0002	0.0005	0.0000	0.0001	0.0002	0.0003	0.0000	0.0001	0.0001	0.0001
4.4	0.0000	0.0001	0.0003	0.0001	0.0002	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
4.6	0.0000	0.0000	0.0003	0.0001	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
4.8	0.0000	0.0000	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5.0	0.0000	0.0000	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table S1: Theoretical approximation for local similarity analysis p-values versus the simulated probability $P(LS(D)/\sqrt{n} \geq x)$. The theoretical approximate probability based on equation (10) with $\sigma = 1$ is given in the 2nd column and the simulated probability that $LS(D)/\sqrt{n} \geq x$ is given in the 3rd to the 14th columns. $\mathbf{D = 0.}$

*The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

†to whom correspondence should be addressed

		The number of time points n											
x	Theory	10	20	30	40	60	80	100	120	140	160	180	200
2	0.4516	0.1692	0.2289	0.2464	0.2698	0.2972	0.3084	0.3220	0.3354	0.3393	0.3410	0.3531	0.3491
2.2	0.2977	0.1139	0.1485	0.1636	0.1708	0.1871	0.1978	0.2043	0.2169	0.2150	0.2119	0.2300	0.2254
2.4	0.1841	0.0757	0.0976	0.0988	0.1050	0.1123	0.1176	0.1261	0.1352	0.1313	0.1292	0.1401	0.1393
2.6	0.1077	0.0513	0.0606	0.0608	0.0626	0.0650	0.0692	0.0748	0.0813	0.0776	0.0745	0.0768	0.0832
2.8	0.0601	0.0318	0.0385	0.0357	0.0379	0.0374	0.0398	0.0464	0.0460	0.0426	0.0447	0.0434	0.0470
3.0	0.0320	0.0186	0.0224	0.0187	0.0233	0.0208	0.0222	0.0276	0.0245	0.0220	0.0246	0.0249	0.0248
3.2	0.0164	0.0127	0.0124	0.0113	0.0133	0.0117	0.0115	0.0139	0.0122	0.0129	0.0128	0.0127	0.0133
3.4	0.0081	0.0078	0.0074	0.0062	0.0073	0.0058	0.0063	0.0078	0.0065	0.0066	0.0066	0.0060	0.0066
3.6	0.0038	0.0044	0.0039	0.0030	0.0045	0.0025	0.0032	0.0043	0.0039	0.0035	0.0035	0.0034	0.0034
3.8	0.0017	0.0032	0.0023	0.0015	0.0023	0.0012	0.0019	0.0023	0.0017	0.0015	0.0018	0.0015	0.0014
4.0	0.0008	0.0013	0.0013	0.0005	0.0011	0.0007	0.0010	0.0009	0.0004	0.0005	0.0007	0.0006	0.0005
4.2	0.0003	0.0010	0.0009	0.0003	0.0003	0.0002	0.0004	0.0004	0.0003	0.0005	0.0003	0.0003	0.0003
4.4	0.0001	0.0007	0.0003	0.0003	0.0001	0.0001	0.0004	0.0000	0.0002	0.0004	0.0002	0.0002	0.0002
4.6	0.0001	0.0004	0.0002	0.0003	0.0000	0.0001	0.0002	0.0000	0.0001	0.0002	0.0002	0.0001	0.0000
4.8	0.0000	0.0002	0.0002	0.0001	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000
5.0	0.0000	0.0002	0.0001	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table S2: Theoretical approximation for local similarity analysis p-values versus Simulated probability $P(LS(D)/\sqrt{n} \geq x)$. The theoretical approximate probability based on equation (10) with $\sigma = 1$ is given in the 2nd column and the simulated probability that $LS(D)/\sqrt{n} \geq x$ is given in the 3rd to the 14th columns. **D = 1.**

		The number of time points n											
x	Theory	10	20	30	40	60	80	100	120	140	160	180	200
2	0.6326	0.2199	0.2914	0.3429	0.3704	0.4167	0.4447	0.4709	0.4792	0.4855	0.4935	0.5072	0.5178
2.2	0.4452	0.1514	0.1947	0.2316	0.2437	0.2759	0.3004	0.3132	0.3245	0.3264	0.3305	0.3429	0.3456
2.4	0.2876	0.1053	0.1228	0.1478	0.1524	0.1738	0.1884	0.1964	0.2032	0.2020	0.2043	0.2145	0.2141
2.6	0.1730	0.0713	0.0765	0.0887	0.0949	0.1036	0.1119	0.1146	0.1194	0.1203	0.1211	0.1226	0.1301
2.8	0.0981	0.0447	0.0453	0.0534	0.0558	0.0602	0.0660	0.0681	0.0657	0.0704	0.0690	0.0721	0.0732
3.0	0.0528	0.0288	0.0256	0.0303	0.0316	0.0347	0.0356	0.0385	0.0352	0.0370	0.0360	0.0421	0.0412
3.2	0.0272	0.0185	0.0139	0.0172	0.0170	0.0188	0.0199	0.0187	0.0179	0.0213	0.0185	0.0227	0.0192
3.4	0.0134	0.0121	0.0084	0.0100	0.0099	0.0097	0.0099	0.0089	0.0089	0.0092	0.0103	0.0120	0.0091
3.6	0.0063	0.0077	0.0046	0.0060	0.0051	0.0048	0.0059	0.0053	0.0044	0.0048	0.0045	0.0064	0.0053
3.8	0.0029	0.0053	0.0021	0.0037	0.0026	0.0021	0.0021	0.0031	0.0030	0.0023	0.0024	0.0035	0.0024
4.0	0.0013	0.0031	0.0011	0.0018	0.0019	0.0013	0.0007	0.0017	0.0017	0.0011	0.0014	0.0015	0.0010
4.2	0.0005	0.0020	0.0003	0.0009	0.0006	0.0006	0.0003	0.0011	0.0012	0.0003	0.0004	0.0006	0.0005
4.4	0.0002	0.0008	0.0003	0.0005	0.0005	0.0001	0.0001	0.0006	0.0006	0.0002	0.0003	0.0003	0.0001
4.6	0.0001	0.0006	0.0001	0.0004	0.0004	0.0000	0.0001	0.0003	0.0002	0.0000	0.0001	0.0001	0.0000
4.8	0.0000	0.0002	0.0000	0.0002	0.0001	0.0000	0.0001	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000
5.0	0.0000	0.0001	0.0000	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table S3: Theoretical approximation for local similarity analysis p-values versus the simulated probability $P(LS(D)/\sqrt{n} \geq x)$. The theoretical approximate probability based on equation (10) with $\sigma = 1$ is given in the 2nd column and the simulated probability that $LS(D)/\sqrt{n} \geq x$ is given in the 3rd to the 14th columns. **D = 2.**

		The number of time points n										
x	Theory	10	20	30	40	60	80	100	120	140	160	180	200
2	0.7539	0.2509	0.3331	0.4103	0.4544	0.5120	0.5402	0.5571	0.5804	0.5972	0.6128	0.6059	0.6259
2.2	0.5616	0.1731	0.2301	0.2772	0.3124	0.3533	0.3707	0.3917	0.4077	0.4109	0.4368	0.4256	0.4406
2.4	0.3779	0.1177	0.1486	0.1772	0.2000	0.2293	0.2344	0.2539	0.2613	0.2679	0.2819	0.2722	0.2805
2.6	0.2336	0.0785	0.0952	0.1122	0.1236	0.1366	0.1401	0.1578	0.1562	0.1573	0.1756	0.1678	0.1669
2.8	0.1346	0.0513	0.0583	0.0679	0.0736	0.0767	0.0813	0.0918	0.0875	0.0894	0.0977	0.0998	0.0947
3.0	0.0732	0.0320	0.0343	0.0379	0.0416	0.0443	0.0453	0.0534	0.0469	0.0481	0.0523	0.0517	0.0509
3.2	0.0379	0.0199	0.0205	0.0207	0.0210	0.0237	0.0264	0.0278	0.0246	0.0261	0.0259	0.0275	0.0271
3.4	0.0187	0.0123	0.0129	0.0116	0.0110	0.0124	0.0122	0.0142	0.0107	0.0133	0.0145	0.0145	0.0128
3.6	0.0089	0.0083	0.0073	0.0055	0.0059	0.0058	0.0061	0.0076	0.0052	0.0067	0.0068	0.0069	0.0058
3.8	0.0040	0.0047	0.0046	0.0030	0.0029	0.0029	0.0036	0.0048	0.0023	0.0031	0.0028	0.0034	0.0027
4.0	0.0018	0.0028	0.0032	0.0013	0.0015	0.0015	0.0013	0.0024	0.0005	0.0014	0.0011	0.0010	0.0011
4.2	0.0007	0.0019	0.0021	0.0004	0.0007	0.0010	0.0005	0.0008	0.0001	0.0005	0.0008	0.0003	0.0002
4.4	0.0003	0.0009	0.0015	0.0002	0.0004	0.0007	0.0002	0.0002	0.0000	0.0000	0.0003	0.0003	0.0001
4.6	0.0001	0.0006	0.0009	0.0002	0.0003	0.0003	0.0000	0.0001	0.0000	0.0000	0.0000	0.0001	0.0001
4.8	0.0000	0.0003	0.0002	0.0000	0.0002	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
5.0	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000

Table S4: Theoretical approximation for local similarity analysis p-values versus the simulated probability $P(LS(D)/\sqrt{n} \geq x)$. The theoretical approximate probability based on equation (10) with $\sigma = 1$ is given in the 2nd column and the simulated probability that $LS(D)/\sqrt{n} \geq x$ is given in the 3rd to the 14th columns. **D = 3.**

(a) D=0	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	233	0
P_{perm} ≤0.05	19	48
(b) D=1	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	225	0
P_{perm} ≤0.05	37	38
(c) D=2	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	228	0
P_{perm} ≤0.05	36	36
(d) D=3	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	228	0
P_{perm} ≤0.05	36	36

Table S5: The comparison of significant gene pairs using P_{theo} and P_{perm} given type-I error 0.05 for local similarity analysis of 25 randomly selected factors from the CDC dataset: (a) D=0, (b) D=1, (c) D=2, (d) D=3. The total number of comparisons is 300.

(a) D=0	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	488	0
P_{perm} ≤0.05	31	261
(b) D=1	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	492	0
P_{perm} ≤0.05	47	241
(c) D=2	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	505	0
P_{perm} ≤0.05	57	218
(d) D=3	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	516	0
P_{perm} ≤0.05	53	211

Table S6: The comparison of significant OTU pairs using P_{theo} and P_{perm} given type-I error 0.05 for local similarity analysis of 40 selected OTUs from SPOT dataset: (a) D=0, (b) D=1, (c) D=2, (d) D=3. The total number of OTU pairs is 780.

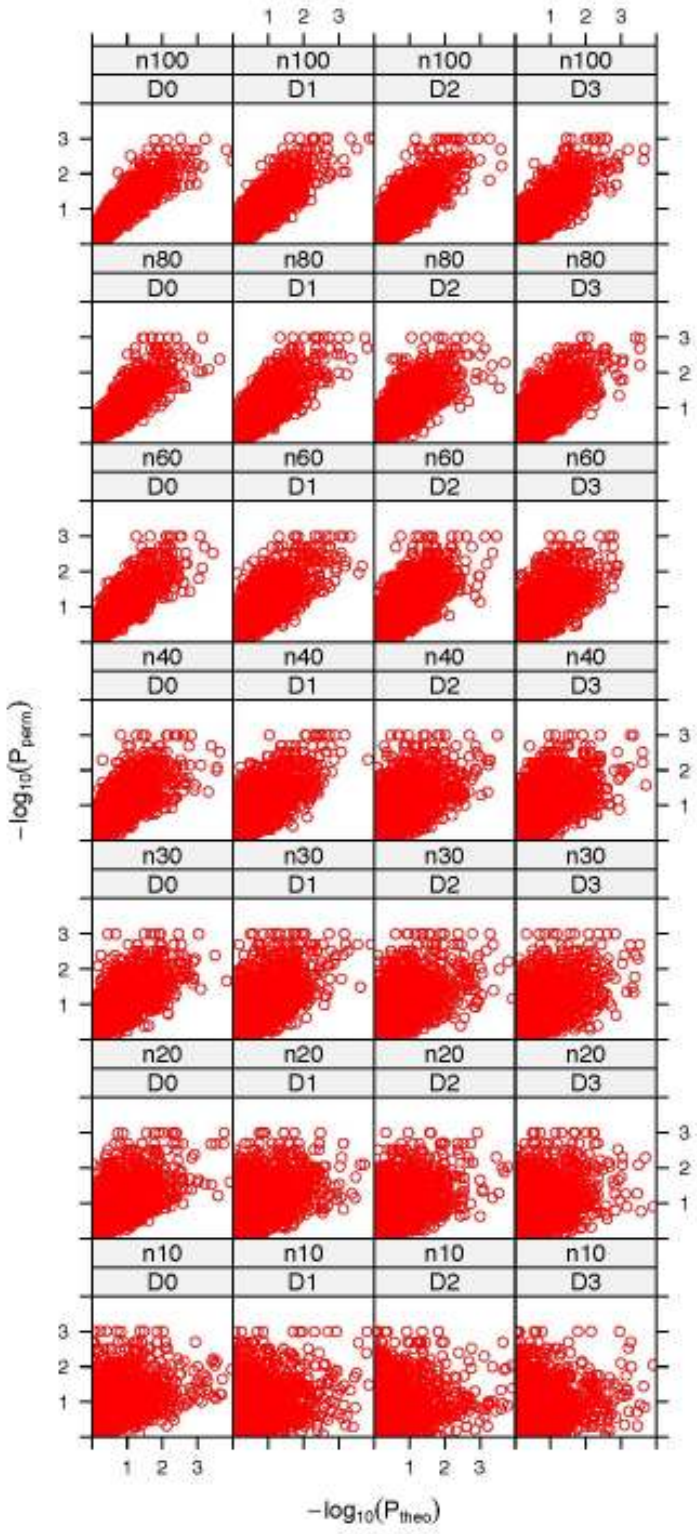

Figure S1: Local similarity analysis P_{theo} vs P_{perm} for 10,000 pairs of simulated data. Columns D0 to D3 are for $D = 0, 1, 2, 3$. Rows n10 to n100 are for $n = 10, 20, 30, 40, 60, 80, 100$.

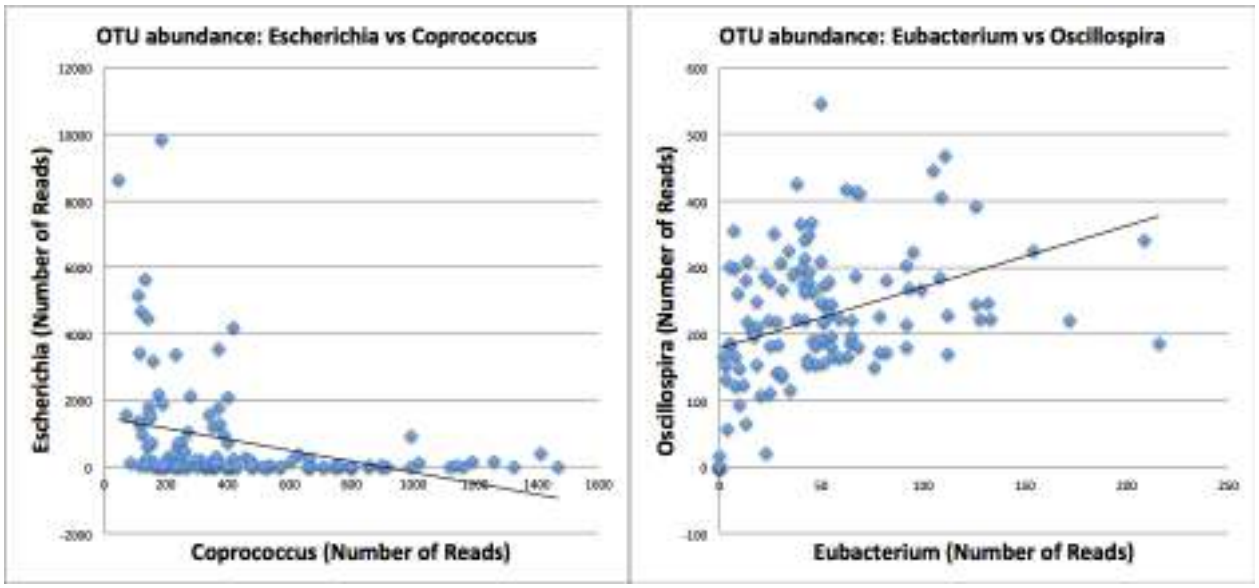

Figure S2: Examples of significant OTU local associations from the 'F4' feces sample of the MPH dataset. OTU abundance profiles (measured in number of read counts) are shifted to synchronize the co-occurrence according to local similarity analysis. (left) *Coprococcus* and *Escherichia* (LS=-0.3179, P=0.0002; r=-0.3314, P=0.0001); (right) *Eubacterium* and *Oscillospira* (LS=0.3862, P=0.0001; r=0.3525, P=0.0001)

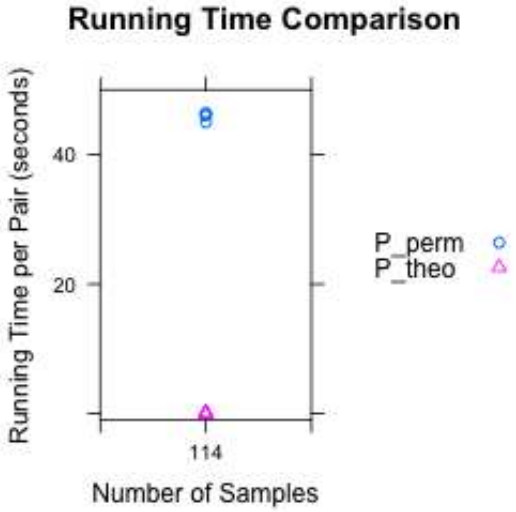

Figure S3: Running time comparison between theoretical approach and 1000 times permutation for four sets of 40 OTUs, 114 time points data. Note that the constant computation time using the theoretical approach that is independent of sample size as compared to sample-size dependent computation time of permutation approaches.