Topological methods for gene correlation analysis of RNA-seq data sets

Derek Covert, Kyle Houfek

University of South Florida

kdhoufek@mail.usf.edu

December 1, 2017

Overview

RNA Seq

RNA Seq is a data analysis method used in biology laboratories to analyze genetic data sets. It measures the expression of each gene at some given time points.

	0 hr	**12 hr**	**24 hr**	**36 hr**
Gene A	a_1	a_2	a_3	a_4
Gene B	b_1	b_2	b_3	b_4
Gene C	c_1	c_2	c_3	c_4

We're interested in genetic data sets from the ciliate known as *Oxytricha Trifallax*. Our data was obtained and normalized by the Landweber Lab at Columbia University.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1		t12hr_A	t12hr_B	t12hr_C	t24hr_A	t24hr_B	t24hr_C	t36hr_A	t36hr_B	t36hr_C	t48hr_A	t48hr_B	t48hr_C	t60hr_A	t60hr_B	t60hr_C	t72hr_A	t72hr_B	t72hr_C
2	g1[OXTMA	16.19855	14.05026	7.041078	7.039844	2.087251	7.284308	11.53404	15.09253	15.8564	9.664042	24.54635	8.710736	14.49826	13.63514	17.44929	11.23406	14.74323	8.862117
3	g12909[O	161.9855	183.434	177.2005	201.1384	227.5104	186.2702	205.4501	173.5641	210.5074	173.9528	116.5952	176.3924	293.0719	286.338	273.3722	244.9024	409.4081	370.7319
4	g19660[O	0	0.78057	0	2.011384	2.087251	3.121846	33.16036	37.73133	36.08699	32.97144	27.17632	42.82779	1.03559	5.244286	3.87762	0	2.268189	2.954039
5	g20729[O	23.3979	35.12565	39.89944	38.21629	39.65777	18.73108	18.02194	23.58208	9.841906	13.64335	13.14983	15.96968	24.85416	27.27028	21.32691	35.94898	20.4137	22.15529
6	g21768[O	3.599677	3.12228	11.73513	5.02846	5.218127	10.40615	7.208774	5.6597	0.546773	5.116257	12.27318	7.258947	14.49826	15.73286	14.54107	51.67666	24.95008	26.58635
7	g22771[O	111.59	241.9767	213.5794	228.2921	202.4633	180.0265	116.7821	100.9313	141.6141	155.1931	118.3485	206.88	107.7013	95.446	86.27704	71.89796	71.44795	50.21866
8	g22772[O	6851.985	1602.51	2193.296	2247.721	1947.405	2227.958	1041.668	1152.692	833.2813	1032.916	1470.151	865.2665	2615.9	3106.715	3330.875	9041.169	3826.435	3918.533
9	g23846[O	28.79741	7.025131	7.041078	26.14799	7.305378	26.01539	9.371407	12.26268	6.56127	8.527096	8.766555	7.984842	4.14236	4.195428	3.87762	0	4.536378	1.47702
10	g23847[O	1.799838	2.34171	2.347026	0	1.043625	0	2.162632	0.943283	0.546773	0	0.876655	1.451789	5.177949	0	2.908215	0	2.268189	1.47702
11	g23848[O	26.99758	31.2228	37.55242	7.039844	12.52351	9.365539	27.39334	24.52537	25.69831	19.89656	14.02649	26.8581	24.85416	19.92829	23.26572	11.23406	18.14551	17.72423
12	g24886[O	100.791	103.8158	145.5156	186.053	195.158	129.0363	156.4304	174.5074	185.3559	198.3971	219.1639	222.8497	109.7725	114.3254	126.0226	150.5364	130.4209	175.7653
13	g25913[O	26.99758	75.7153	70.41078	14.07969	19.82888	38.50277	30.99773	16.9791	16.94995	32.40296	28.92963	26.13221	41.4236	25.17257	22.29631	6.740434	21.54779	22.15529
14	g4[OXTMA	12.59887	29.66166	23.47026	9.051227	4.174502	8.324923	16.58018	6.602983	8.74836	16.48572	18.40977	19.59916	54.88626	34.61228	31.99036	6.740434	10.20685	28.06337
15	g1350[OX	2014.019	15868.99	972.8423	16222.82	22912.8	16323.09	17475.51	19663.68	22436.26	19332.06	17781.2	20369.33	23207.57	27928.97	28487.9	34135.81	35773.87	32837.1
16	g2704[OX	3860.653	1639.978	2145.182	376.1288	512.4201	629.5723	565.1679	640.4894	414.4536	928.885	1539.407	811.5503	3871.035	4837.329	4418.548	13359.54	9322.256	7902.055
17	g2705[OX	269.9758	165.4809	140.8216	285.6165	353.789	248.7071	233.5643	300.9074	320.4087	212.0404	203.3841	211.9613	70.42011	75.51771	68.82775	94.36608	122.4822	97.48329
18	g2706[OX	1031.307	691.5851	708.8018	309.7531	260.9064	284.088	163.6392	149.0388	124.1174	654.3125	1136.146	715.0063	463.9443	538.0637	570.9795	1828.904	990.0644	1038.345
19	g5362[OX	44.99596	63.22617	53.9816	61.34721	92.88267	45.78708	68.48336	58.48357	88.57715	79.58623	56.10595	76.21894	82.84719	73.42	74.64418	58.4171	55.57063	38.40251
20	g6697[OX	64.79418	33.56451	71.58429	68.38705	79.31554	61.39631	59.11195	81.12237	72.72075	58.55272	37.69619	56.61979	43.49477	38.80771	37.80679	20.2213	43.09559	22.15529
21	g7887[OX	32.39709	58.54275	66.89024	38.21629	34.43964	45.78708	48.29879	54.71043	50.30307	125.6325	149.0314	182.1996	173.9791	212.918	200.6668	148.2896	153.1027	193.4896
22	g9062[OX	30.59725	102.2547	36.3789	105.5977	89.75179	96.77723	140.5711	129.2298	111.5416	90.38721	96.4321	66.05642	177.0859	159.4263	146.3801	92.11927	179.1869	227.461
23	g10346[O	77.39305	227.1459	109.1367	51.29029	36.52689	82.20862	102.3646	83.00893	74.90784	92.09263	62.24254	83.47789	165.6944	77.61543	94.03228	67.40434	90.72755	110.7765
24	g10347[O	669.5399	309.1057	526.9073	253.4344	210.8123	229.976	286.1883	337.6954	258.0766	563.9253	769.7035	533.5326	542.6491	726.858	755.1664	1213.278	773.4524	760.6651
25	g11603[O	255.5771	34.34508	85.66645	72.40982	58.44303	61.39631	27.39334	34.90148	24.05799	54.00494	78.02234	38.47242	155.3385	212.918	220.0549	862.7756	310.7419	336.7605
26	g12911[O	75.59321	58.54275	89.18699	102.5806	57.3994	43.70585	77.85476	69.80297	101.6997	39.22464	31.5596	47.18315	23.81857	16.78171	22.29631	42.68942	30.62055	28.06337
27	g14155[O	10.79903	32.00337	19.94972	87.4952	72.01016	72.84308	67.0416	83.95222	60.14498	42.63548	64.87251	50.81263	181.2282	94.39714	133.7779	29.20855	62.37519	62.03482
28	g15457[O	142.1872	192.0202	146.6891	423.3963	445.6281	495.3329	218.4259	244.3104	250.4218	200.671	209.5207	211.9613	417.3427	535.966	496.3353	1246.98	1101.206	939.3844
29	g16656[O	59.39467	104.5964	70.41078	22.12522	28.17789	30.17785	70.64599	54.71043	49.20953	101.1882	54.35264	63.15284	108.7369	47.19857	58.16429	22.46811	30.62055	45.78761
30	g18549[O	48.59564	106.9381	38.72593	173.9847	169.0673	164.4172	178.0567	170.7343	199.572	140.9813	130.6217	137.1941	293.0719	304.1686	322.8118	80.88521	201.8688	208.2598

RNA Seq

We view each gene as a vector, and its expression at each time point as the components.

$$\textbf{Gene A} = <a_1, a_2, a_3, a_4>$$

$$\textbf{Gene B} = <b_1, b_2, b_3, b_4>$$

$$\textbf{Gene C} = <c_1, c_2, c_3, c_4>$$

We can now take the distance between any two vectors using a defined distance metric.

Distances

For the following definitions[1], assume we have two points $p = (p_1, p_2, \ldots, p_n)$ and $q = (q_1, q_2, \ldots, q_n)$ in $\mathbb{R}^n$.

Definition

For any two points p and q, we say the *Euclidean distance* $d(p, q)$ between them is defined by:

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

[1] Adams, C. and Franzosa, R. *Introduction to Topology: Pure and Applied* (144), Pearson (2008)

Distances

Definition

For any two points p and q, we say the *Taxicab distance* $d_T(p, q)$ between them is defined by:

$$d_T(p, q) = |p_1 - q_1| + |p_2 - q_2| + \cdots + |p_n - q_n|$$

Definition

For any two points p and q, we say the *Max distance* $d_M(p, q)$ between them is defined by:

$$d_M(p, q) = max\{|p_1 - q_1|, |p_2 - q_2|, \ldots, |p_n - q_n|\}$$

Normalization

Furthermore, we will also want to group together vectors which are co-linear in $\mathbb{R}^n$. To do this, we perform a linear normalization on each of the vectors.

Definition

Let $\vec{v} = <v_1, v_2, \ldots, v_n>$ be an n-dimensional vector. We say the *magnitude* of $\vec{v}$ is $M = \sum_{i=1}^{n} v_i^2$. Then the *normalized vector* of $\vec{v}$ is $\vec{u} = \frac{\vec{v}}{M}$

Distance Matrix

To obtain the pairwise distances of all possible pairs of vectors, we create a distance matrix. The entry a_{ij} of this matrix will correspond to the distance between the i^{th} and j^{th} vectors.

These distances can now be analyzed using a persistent homology.

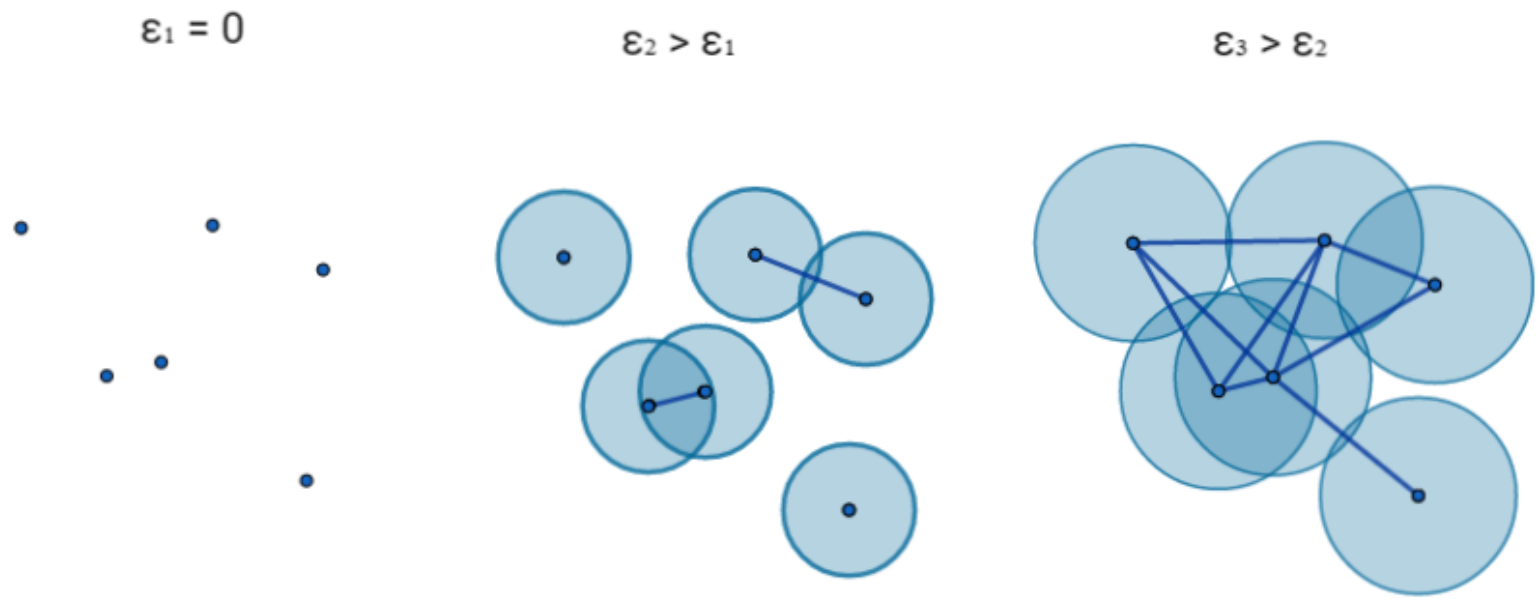
Persistent Homology

Persistent Homology is a topological method used in data analysis to find correlations within large sets of data.

To investigate a set of points in $\mathbb{R}^n$, we define a distance function on $\mathbb{R}^n$ and find the distance between any two points.

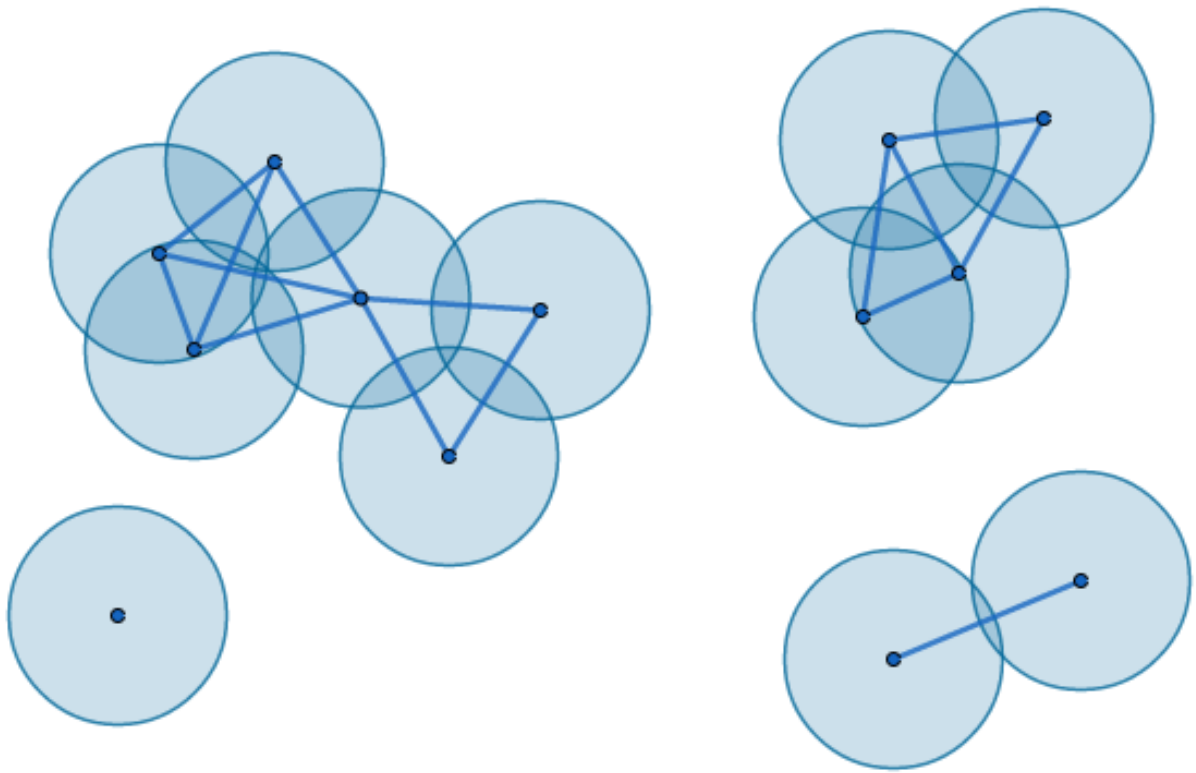
Persistent Homology

For $\varepsilon \geq 0$ we consider the ε neighborhood around each point in $\mathbb{R}^n$. Points are connected if their ε neighborhoods overlap. As ε grows larger, neighborhoods overlap and components are connected.



Clusters

As more components are connected, we begin to generate clusters.



Birth and Death

We are interested in those clusters which persist with large ε values, but we need some way of identifying them. To do this, we introduce the concept of birth and death.

In an n-dimensional persistent homology, we have a *birth* when an n-dimensional simplex is formed, and we have a *death* when an $(n+1)$-dimensional simplex is formed.[2]

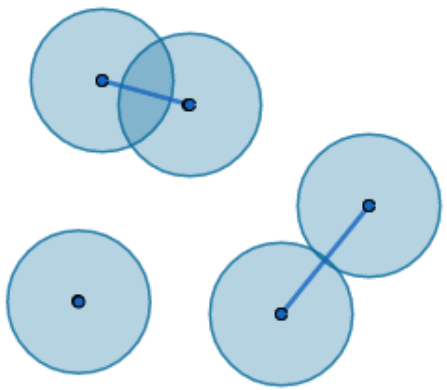
[2] Nanda, V. and Sazdanovic, R. Simplicial Models and Topological Inference in Biological Systems. *Discrete and Topological Models in Molecular Biology* (109-141), Springer Berlin Heidelberg (2014)

Dimension 0

In the 0-dimensional persistent homology h_0, a birth occurs with every point in h_0, while the death of a component occurs when two components are connected.

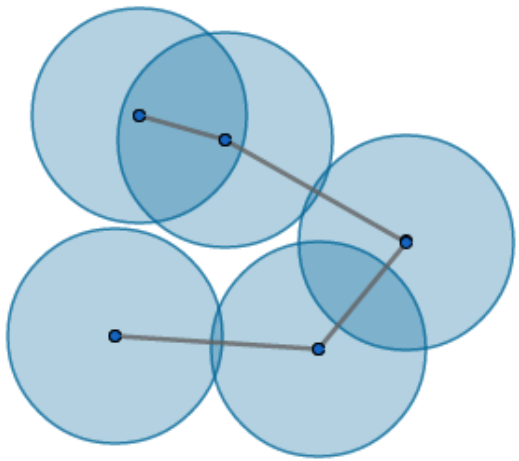
n = 5

Hence, 5 “births”



n = 3

Hence, 2 “deaths”



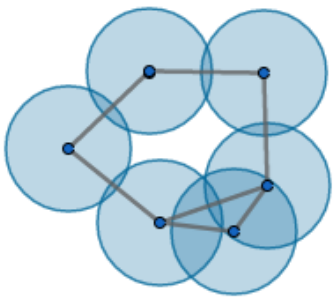
n = 1

Hence, 2 more “deaths”

Dimension 1

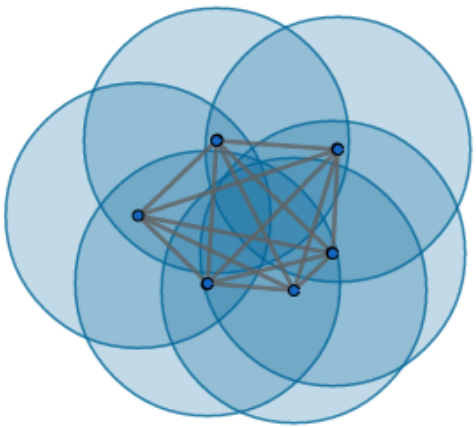
In the 1-dimensional persistent homology h_1, a birth occurs when a cycle with a "hole" appears, while a death occurs when this "hole" is covered by the ε neighborhoods.

n = 0



n = 1

One cycle with a "hole"



n = 0

Cycle no longer has a hole, hence a death

Lifespan

Definition[3]

Let ε_1 be the ε value at which point a persistent homology was birthed, and let ε_2 be the ε value at which point that persistent homology died. Then the *lifespan* of the persistent homology is defined as $L = \varepsilon_2 - \varepsilon_1$.

Those persistent homologies with large lifespans will help us identify characteristics within our data set.

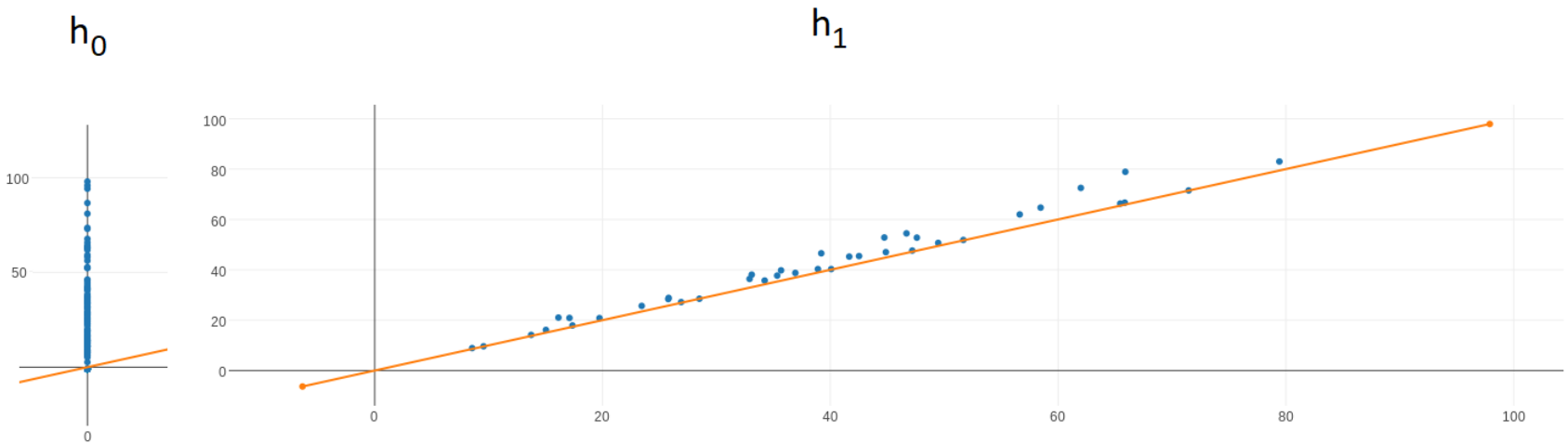
[3] Nanda, V. and Sazdanovic, R. Simplicial Models and Topological Inference in Biological Systems. *Discrete and Topological Models in Molecular Biology* (109-141), Springer Berlin Heidelberg (2014)

Back to RNA Seq

Now, we can use persistent homologies to study correlations in genes.

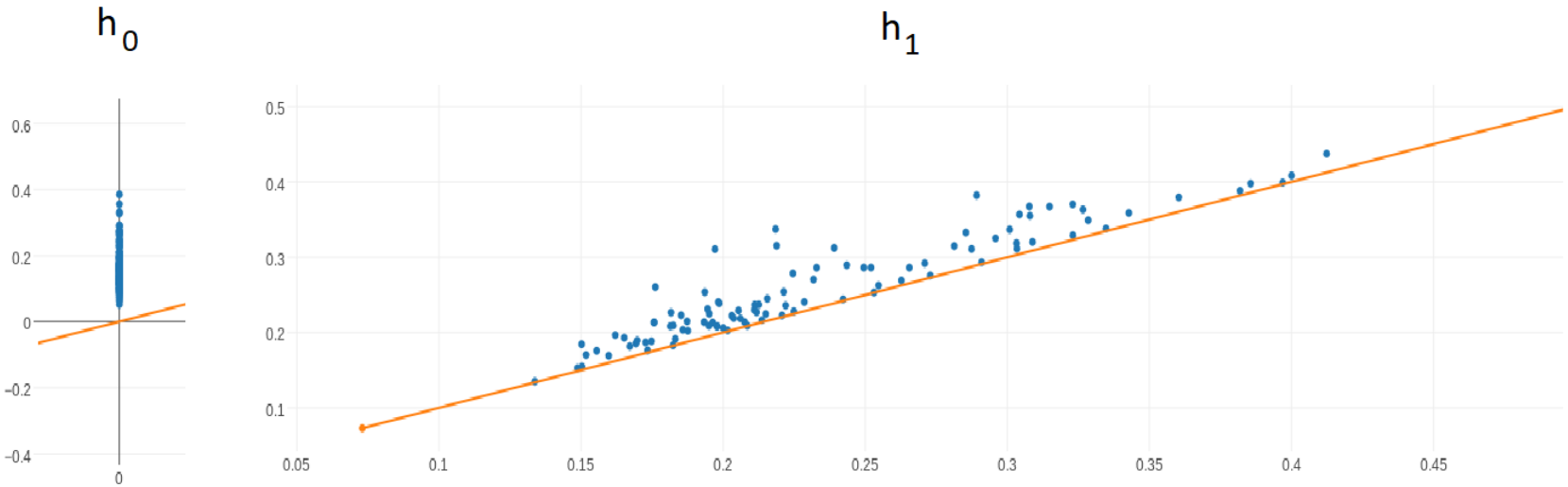
Results

Thus far, we have only analyzed small samples of a data set from the Landweber Lab for *O. Trifallax*, using about 200 points.



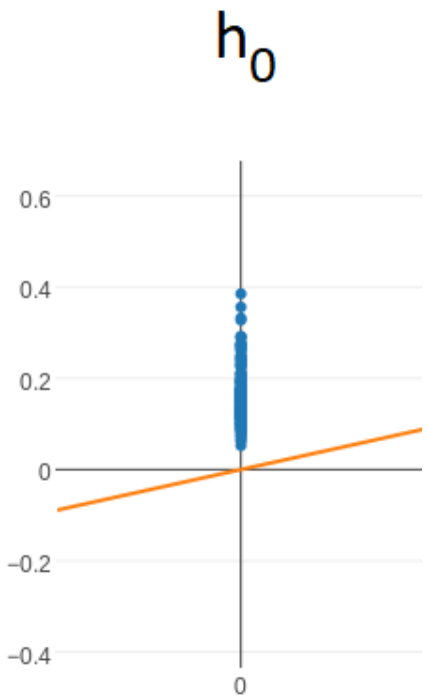
Results

We also have calculated the normalized results of these 200 points.



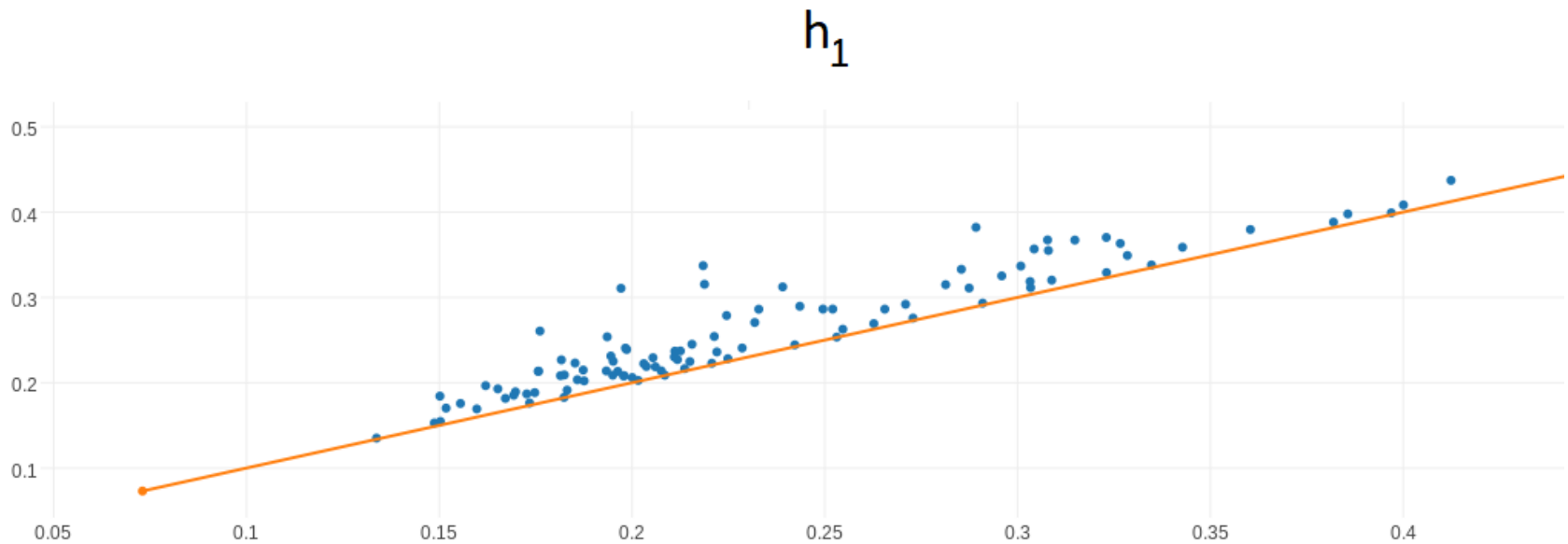
Interpreting Results

The data from h_0 will help us identify genes which are correlated.



Interpreting Results

We are also interested in studying rearrangement patterns in *O. Trifallax*. The data from h_1 can help us identify rearrangement pathways.



Next Steps

As we move forward, the first step is to analyze our full data set, which consists of approximately 20,000 points.

Additionally, we want to develop our own graphing program, which will allow us to quickly identify the exact lifespan of a given cluster, as well as tell us which genes are within that cluster.

Conclusion

Thank you for joining!