Web-based Supplementary Materials for "Bayesian Semiparametric Regression Models for Evaluating Pathway Effects on Continuous and Binary Clinical Outcomes"

Inyoung Kim[1,*], Herbert Pang[2], and Hongyu Zhao[3,4]

1 Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA.

2 Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27705, USA.

3 Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, USA.

4 Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA.

*To whom correspondence should be addressed:

Inyoung Kim, Ph.D

Department of Statistics, Virginia Tech., Blacksburg, VA 24061, USA.

Tel: (540) 231-5366

Fax: (540) 231-3863

E-Mail: inyoungk@vt.edu

A Non-identifiability between τ and ρ when $\rho \to 0$

If $\tau \sim O(\frac{1}{\rho^m})$ for any positive value m and $\rho \sim O\{E(||z - z'||^2)\}$, Liu's estimator $\hat{\Theta}$ is asymptotically normally distributed with mean Θ and covariance matrix $I(\Theta)$, where

$$I_{\theta_l,\theta_{l'}} = \frac{1}{2} tr\{P\Sigma_{\theta_l} P\Sigma_{\theta_{l'}}\}.$$

Define some notations

$$\begin{aligned} \Sigma &= \sigma^2 I + \tau \mathbf{K} \\ P &= \Sigma^{-1} - \Sigma^{-1} X (X'\Sigma^{-1} X)^{-1} X'\Sigma^{-1} \end{aligned}$$

and Σ_{θ_l} are the first derivative of Σ with respect to θ_l. Further define P_τ and $P_{\tau,\tau}$ are the first and second derivatives of P with respect to τ. Other notations are similarly defined.

Based on this setting, if $\rho \to 0$, $P_\tau = P_{\tau\tau} = 0$ and $P_\rho = P_{\rho\rho} = 0$, then the asymptotic distributions of $\hat{\tau}$ and $\hat{\rho}$ are the same and degenerated.

B The proof of equivalence of the two testings

The testing of H_0: $\{\mathbf{r}(\mathbf{z})$ is a point mass at zero$\}\cup\{\mathbf{r}(\mathbf{z})$ has a constant covariance matrix as a function of $\mathbf{z}\}$ is equivalent to the testing of $\frac{\partial \mathbf{K}(\mathbf{Z})}{\partial \mathbf{Z}} = \mathbf{0}$.

$$\begin{aligned} K_{ij}(\mathbf{z}_i, \mathbf{z}_j) &= \tau \exp(-\frac{||\mathbf{z}_i - \mathbf{z}_j||^2}{\rho}) \\ \frac{\partial K_{ij}(\mathbf{z}_i, \mathbf{z}_j)}{\partial \mathbf{z}_i} &= -2(||\mathbf{z}_i - \mathbf{z}_j||)(\frac{\tau}{\rho}) \exp(-\frac{||\mathbf{z}_i - \mathbf{z}_j||^2}{\rho}) \end{aligned}$$

If $\rho \to 0$ and $\tau \to 0$ with faster rate $O(\rho^m)$, then $\frac{\partial \mathbf{K}(\mathbf{Z})}{\partial \mathbf{Z}} = \mathbf{0}$. That is, if $\frac{\tau}{\rho} \to 0$, then $\frac{\partial \mathbf{K}(\mathbf{Z})}{\partial \mathbf{Z}} = \mathbf{0}$.

If $\rho \to \infty$, then $0 \leq \exp(-\frac{||\mathbf{z}_i - \mathbf{z}_j||^2}{\rho}) \leq 1$. Therefore, $0 \leq \frac{\tau}{\rho}\exp(-\frac{||\mathbf{z} - \mathbf{z}_j||^2}{\rho}) \leq \frac{\tau}{\rho}$.

Hence, if $\frac{\tau}{\rho} \to 0$, $\frac{\partial \mathbf{K}(\mathbf{Z})}{\partial \mathbf{Z}} = \mathbf{0}$.

If $\rho \to 0$ and $\tau \sim O(\frac{1}{\rho^m})$, then $\exp(-\frac{||\mathbf{z}_i - \mathbf{z}_j||^2}{\rho}) \to 0$. Therefore, $\frac{\partial \mathbf{K}(\mathbf{Z})}{\partial \mathbf{Z}} = \mathbf{0}$.

Therefore, if $\frac{\tau}{\rho} \to 0$ or $\rho \to 0$, then $\frac{\partial \mathbf{K}(\mathbf{Z})}{\partial \mathbf{Z}} = \mathbf{0}$.

		Kernel(K)	GK LIKE	BHM	BHM (BF)
		Case			
$n = 60$	Type I	0	0.04 (0.00003)	0.04(0.00003)	0.05(0.00003)
$p = 5$	Power	1	0.99 (0.00001)	0.99(0.00001)	1(0.00001)
		2	0.90 (0.00003)	0.94(0.00003)	0.97(0.00002)
		3	0.86 (0.00004)	0.89(0.00004)	0.91(0.00003)
		4	0.84 (0.00004)	0.86(0.00004)	0.89(0.00002)
		5	0.77 (0.00004)	0.78(0.00004)	0.81(0.00004)
$n = 60$	Type I	0	0.03 (0.00003)	0.03(0.00003)	0.04(0.00003)
$p = 200$	Power	1	0.84 (0.00003)	0.85(0.00003)	0.89(0.00003)
		2	0.81 (0.00004)	0.82(0.00004)	0.87(0.00003)
		3	0.77 (0.00004)	0.79(0.00004)	0.81(0.00004)
		4	0.73 (0.00005)	0.73(0.00005)	0.77(0.00005)
		5	0.72 (0.00005)	0.73(0.00005)	0.76(0.00005)

Table 1: Estimated type I error rate and power on a continuous clinical outcome. The number within parenthesis is standard deviation. LIKE=LLG's likelihood-based approach with kernel K using a resampling based inference; BHM=Bayesian approach on hierarchical model using a resampling based inference; BHM(BF)=Bayesian approach on hierarchical model using a inference based on Bayes factor; GK=Gaussian kernel.

		Kernel	Method	Simulation	model			
	Parameters	K		Case 0	Case 1	Case 2	Case 3	Case 4
n=60	β	GK	LIKE	0.99	0.99	1.01	1.01	0.99
p=5	σ			0.93	1.07	1.26	0.94	0.92
	τ			0.24	227.65	1.34	5.67	1.12
	ρ			0.79	4.10	1.22	2.33	1.53
	R^2			0.02	0.99	0.95	0.87	0.91
	β	GK	BHM	1.01	0.99	0.99	0.99	0.99
	σ			0.85	1.05	0.99	0.96	0.94
	τ			0.37	240.29	1.22	4.67	1.76
	ρ			0.99	4.30	1.31	2.86	0.95
	R^2			0.02	0.98	0.98	0.89	0.93
n=60	β	GK	LIKE	1.00	2.98	1.00	1.54	1.17
p=200	σ			0.31	2.23	1.00	2.04	4.13
	τ			0.92	5.81	1.020	4.18	2.72
	ρ			1.79	0.35	1.01	0.06	0.12
	R^2			0.03	0.85	0.82	0.77	0.74
	β	GK	BHM	1.01	2.79	1.00	1.87	1.01
	σ			0.48	2.37	0.99	2.40	3.98
	τ			0.94	6.07	1.02	4.38	2.60
	ρ			1.29	0.35	1.02	0.06	0.12
	R^2			0.02	0.87	0.83	0.80	0.74

Table 2: Average estimated parameter values on a continuous clinical outcome. LIKE=LLG's likelihood-based approach with kernel K; BHM=Bayesian approach on hierarchical model with kernel K; GK=Gaussian kernel.

		Kernel	Method	Simulation	model			
	Parameters	K		case 0	Case 1	Case 2	Case 3	Case 4
n=60	β	P2K	LIKE	1.98	1.01	1.00	0.99	1.01
p=5	σ			1.07	1.06	1.03	0.99	0.99
	τ			0.13	30.66	0.52	1.04	0.41
	ρ			N/A	N/A	N/A	N/A	N/A
	R^2			0.016	0.90	0.80	0.91	0.94
	β	P2K	BHM	1.25	0.99	1.02	1.01	0.97
	σ			1.07	1.04	1.05	0.99	1.01
	τ			0.14	29.71	0.45	0.93	1.08
	ρ			N/A	N/A	N/A	N/A	N/A
	R^2			0.015	0.91	0.81	0.92	0.95
n=60	β	P2K	LIKE	1.09	1.02	0.94	1.03	1.01
n=200	σ			1.19	1.20	1.24	2.48	1.21
	τ			0.01	1.01	0.49	0.96	0.50
	ρ			N/A	N/A	N/A	N/A	N/A
	R^2			0.023	0.88	0.77	0.81	0.66
	β	P2K	BHM	0.97	1.05	1.00	0.97	0.99
	σ			1.19	1.20	1.06	2.17	1.22
	τ			0.01	1.01	0.86	1.21	0.53
	ρ			N/A	N/A	N/A	N/A	N/A
	R^2			0.02	0.89	0.77	0.81	0.66

Table 3: Average estimated parameter values on a continuous clinical outcome. LIKE=LLG's likelihood-based approach with kernel K; BHM=Bayesian approach on hierarchical model with kernel K; P2K=Quadratic kernel.

		Kernel	Method	Simulation	model			
	Parameters	K		Case 0	Case 1	Case 2	Case3	Case 4
n=60	β	P1K	LIKE	0.95	1.09	1.01	0.99	1.01
p=5	σ			0.98	2.32	1.11	1.07	1.00
	τ			0.17	65.19	0.94	4.28	1.18
	ρ			N/A	N/A	N/A	N/A	N/A
	R^2			0.01	0.87	0.72	0.91	0.92
	β	P1K	BHM	1.02	0.9908	0.99	0.99	1.01
	σ			1.00	2.3109	1.07	1.08	0.99
	τ			0.21	64.07	0.95	4.84	1.01
	ρ			N/A	N/A	N/A	N/A	N/A
	R^2			0.01	0.87	0.74	0.92	0.93
n=60	β	P1K	LIKE	1.25	0.92	1.07	0.94	1.04
p=200	σ			1.22	1.69	1.35	2.03	1.55
	τ			0.28	1.59	0.36	0.76	0.91
	ρ			N/A	N/A	N/A	N/A	N/A
	R^2			0.02	0.84	0.60	0.78	0.82
	β	P1K	BHM	0.91	0.92	1.01	1.34	0.98
	σ			1.18	1.68	1.32	2.78	1.39
	τ			0.25	1.47	0.35	0.89	0.89
	ρ			N/A	N/A	N/A	N/A	N/A
	R^2			0.02	0.84	0.61	0.79	0.82

Table 4: Average estimated parameter values on a continuous clinical outcome. LIKE=LLG's likelihood-based approach with kernel K; BHM=Bayesian approach on hierarchical model with kernel K; P1K=Linear kernel

		Kernel(K)	GK LIKE	BHM	BHM (BF)
		Case			
$n = 60$	Type I	6	0.04 (0.00002)	0.044(0.00002)	0.053(0.00002)
$p = 5$	Power	7	1 (0.00001)	1(0.00001)	1(0.00001)
		8	0.92 (0.00003)	0.95(0.00003)	1(0.00002)
		9	0.78 (0.00003)	0.78(0.00003)	0.82(0.00002)
$n = 60$	Type I	6	0.037 (0.00003)	0.037(0.00003)	0.042(0.00003)
$p = 200$	Power	7	0.85 (0.00003)	0.86(0.00003)	0.89(0.00003)
		8	0.82 (0.00003)	0.82(0.00003)	0.87(0.00003)
		9	0.73 (0.00005)	0.73(0.00005)	0.77(0.00005)

Table 5: Estimated type I error rate and power on a binary outcome. The number within parenthesis are standard error. LIKE=LLG's likelihood-based approach with kernel K using a resampling based inference; BHLM=Bayesian approach on hierarchical latent model with kernel K using a resampling based inference; BHLM(BF)=Bayesian approach on hierarchical latent model using Bayes Factor; GK=Gaussian kernel.

		BHM(BF)			
		GK	P1K	P2K	NNK
LLG	GK	0.39			
	P1K		0.46		
	P2K			0.46	
	NNK				0.34

Table 6: The proportions of overlap between the Bayesian inference based Bayes factor and re-sampling based inference: we selected the top 50 pathways from the Bayesian approach and the likelihood based approach for each kernel. LLG=Liu et al's likelihood-based approach with kernel K using a resampling based inference; BHM(BF)=Bayesian approach on hierarchical model using Bayes Factor; GK=Gaussian kernel; P1K=Linear kernel; P2K=Quadratic kernel; NNK=Neural Network kernel.

	GSEA	BHM(BF)			
		P1K	P2K	GK	NNK
Global	0.36	0.42	0.42	0.12	0.4
GSEA		0.43	0.41	0.22	0.28

Table 7: The proportions of overlap among Global, GSEA, and the Bayesian inference based Bayes factor: we selected the top 50 pathways from each method. BHM(BF)=Bayesian approach on hierarchical model using Bayes Factor; GK=Gaussian kernel; P1K=Linear kernel; P2K=Quadratic kernel; NNK=Neural Network kernel. Global=global test by Goeman *et al.* (2004); GSEA=Gene Set Enrichment Analysis by Subramanian *et al.* (2005).

Gene Id	Gene Name	Gene Id	Gene Name	$\frac{1}{S}\sum_{s=1}^{S}\lvert\hat{\tau}^{b}_{Bay,-(g,g')}-\hat{\tau}\rvert$	$\frac{1}{B}\sum_{s=1}^{S}\lvert\hat{\rho}^{s}_{Bay,-(g,g')}-\hat{\rho}\rvert$	$\frac{1}{S}\sum_{s=1}^{S}\lvert(\frac{\hat{\tau}}{\hat{\rho}})^{s}_{Bay,-(g,g')}-(\frac{\hat{\tau}}{\hat{\rho}})\rvert$
205963_s_at	DNAJA3	215000_s_at	FEZ2	42.55	1.13	32.11
208903_at	RPS28	217852_s_at	SBNO1	39.57	2.53	53.99
203312_x_at	ARF6	218188_s_at	RPS5P1	32.69	1.13	40.76
211999_at	H3F3B	217150_s_at	NF2	31.38	0.26	51.29
205679_x_at	AGC1	220966_x_at	ARPC5L	31.35	3.28	60.07
201739_at	SGK	204746_s_at	PRKCABP	30.08	0.99	39.49
207152_at	NTRK2	211998_at	H3F3B	29.69	3.86	61.95
210835_s_at	CTBP2	218188_s_at	RPS5P1	28.14	3.33	61.18
203138_at	HAT1	210835_s_at	CTBP2	26.62	0.99	42.96
212397_at	RDX	215000_s_at	FEZ2	26.29	1.13	46.43
203312_x_at	ARF6	211998_at	H3F3B	24.13	3.03	61.68
215148_s_at	RAB3D	218188_s_at	RPS5P1	23.56	1.13	48.82
201475_x_at	MARS	212397_at	RDX	22.41	1.24	51.54
201739_at	SGK	32699_s_at	PVR	22.09	0.83	43.03
210330_at	SGCD	218869_at	MLYCD	17.61	1.29	55.99
201218_at	CTBP2	214443_at	PVR	17.31	1.48	57.95
202569_s_at	MARK3	205679_x_at	AGC1	16.20	1.69	60.00
201170_s_at	BHLHB2	218869_at	MLYCD	15.89	0.89	51.93
203311_s_at	ARF6	215489_x_at	HOMER3	15.29	0.70	47.69
201170_s_at	BHLHB2	31807_at	WDR5B	15.02	1.13	56.34
209407_s_at	DEAF1	209736_at	SOX13	14.69	0.16	22.49
202569_s_at	MARK3	217852_s_at	SBNO1	14.35	1.13	56.94
203149_at	PVRL2	209199_s_at	MEF2C	13.82	1.13	57.39
209341_s_at	IKBKB	218458_at	GCL	13.58	0.08	90.16
214182_at	ARF6	214245_at	RPS14	13.46	1.14	57.75
201218_at	CTBP2	201908_at	DVL3	12.86	1.13	58.26
207968_s_at	MEF2C	217780_at	PTD008	12.48	1.13	58.57
207968_s_at	MEF2C	221600_s_at	PTD012	12.46	1.13	58.60
204746_s_at	PRKCABP	217269_s_at	RAB7	12.43	0.65	50.46
202305_s_at	FEZ2	205698_s_at	MAP2K6	11.82	1.13	59.16
207968_s_at	MEF2C	208645_s_at	RPS14	11.77	1.13	59.20
203149_at	PVRL2	207152_at	NTRK2	11.69	1.13	59.27
207152_at	NTRK2	217150_s_at	NF2	11.55	1.13	59.38

Table 8: Top gene pairs, within pathway 36 *c17_U133_probes*, ranked by the mean of the absolute values of the difference $\frac{1}{S}\sum_{s=1}^{S}|\hat{\tau}-\hat{\tau}^{s}_{Bay,-(g,g')}|$, where $\hat{\tau}$ and $\hat{\rho}$ are the mean of posterior samples obtained by fitting Bayesian hierarchical model (3) using the observed data and $\hat{\tau}^{s}_{Bay,-(g,g')}$ and $\hat{\tau}^{s}_{Bay,-(g,g')}$ are the sth posterior sample obtained by fitting Bayesian hierarchical model (3) after removing a gene pair g and g'.

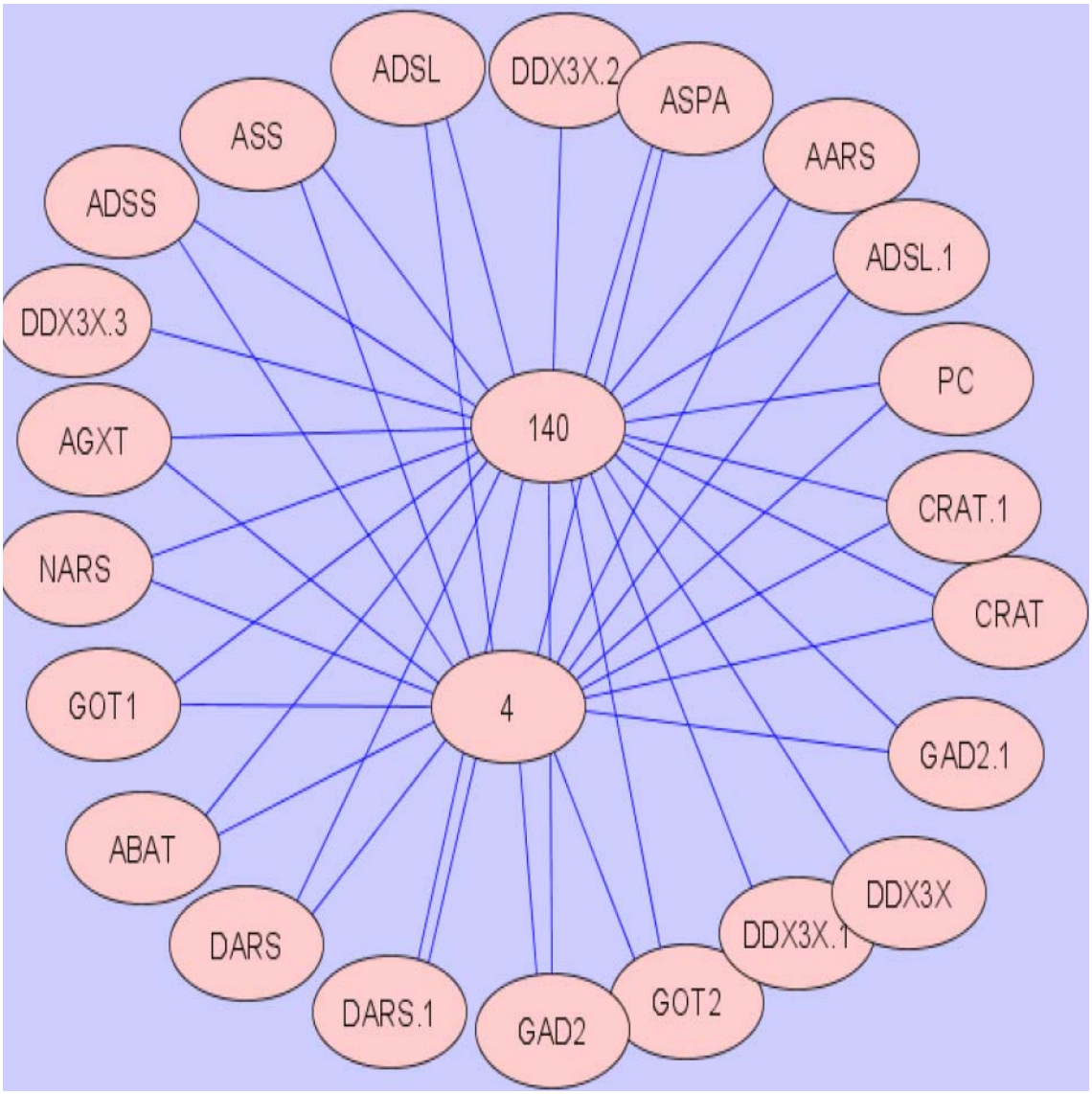

Figure 1: Common genes between pathway 4, *Alanine and aspartate metabolism*, and pathway 140, *MAP00252_Alanine_and_aspartate_metabolism*.

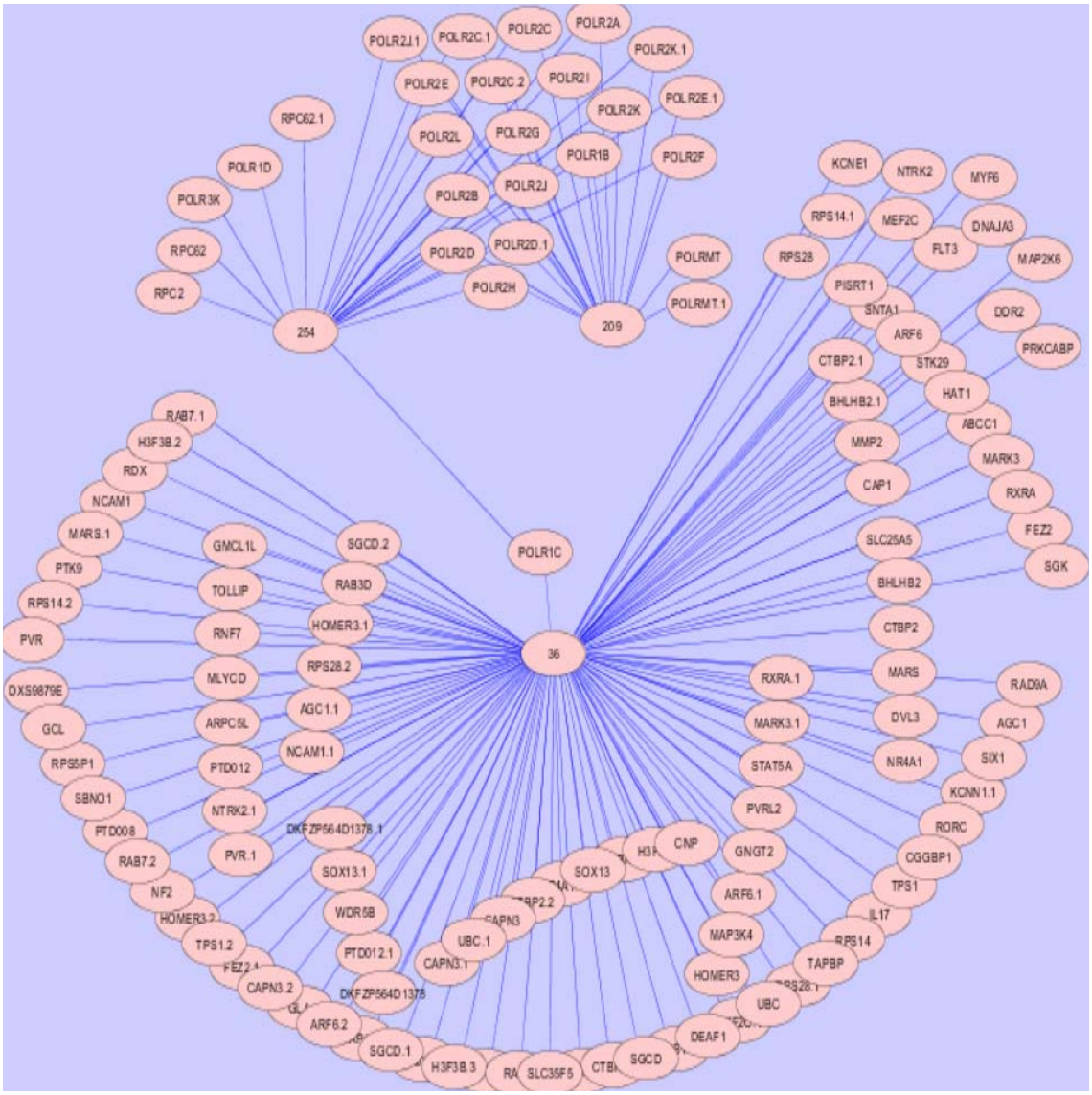

Figure 2: Common genes among pathway 36, *c17_U133_probes*, pathway 209, *MAP03020_RNA_polymerase*, pathway 254 *RNA polymerase*.

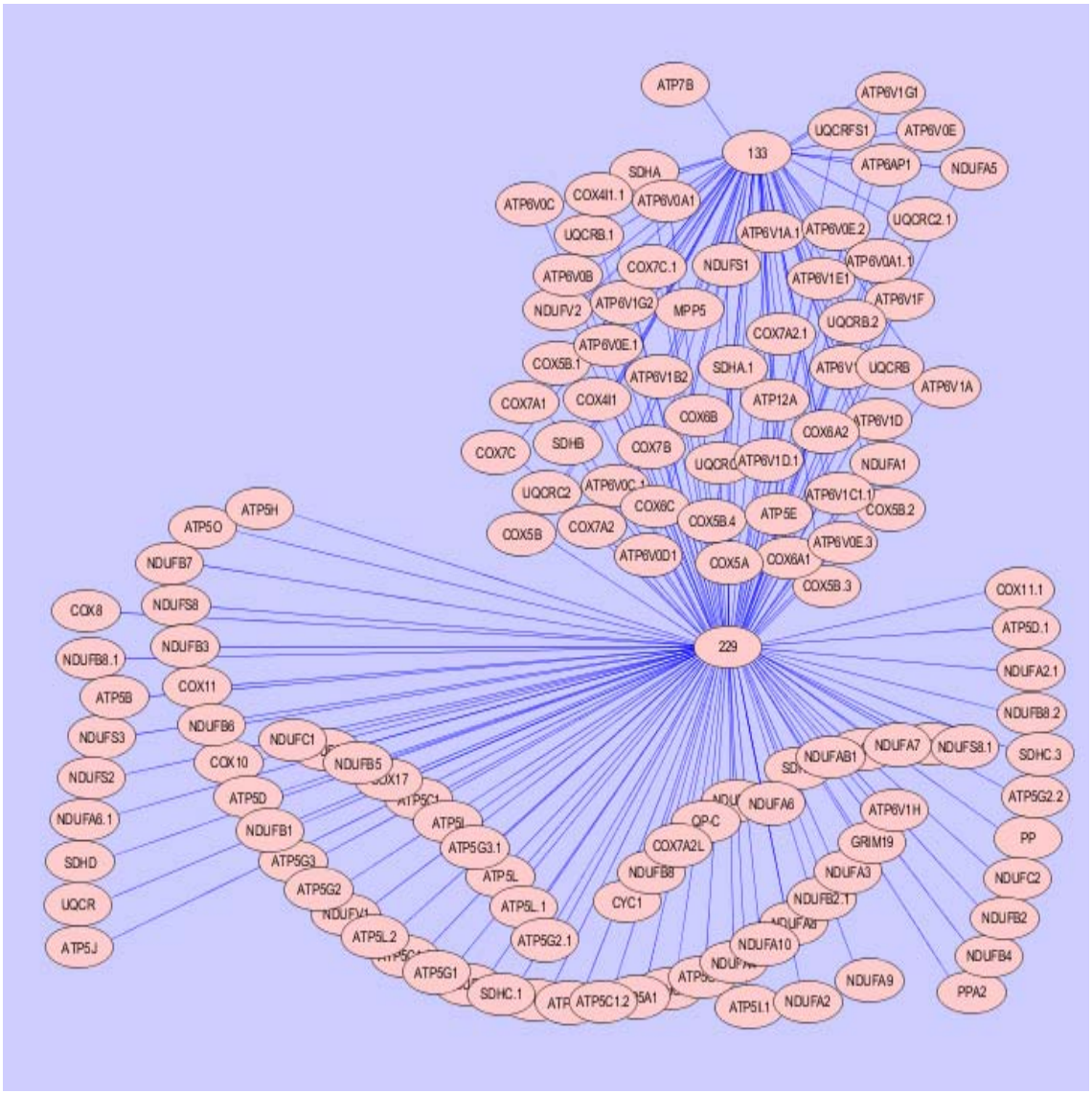

Figure 3: Common genes between pathway 133, *MAP00190_Oxidative_phosphorylation*, and pathway 229, *Oxidative phosphorylation*.