### PLIER:Pathway-Level Information Extractor

Maria Chikina Computational and Systems Biology University of Pittsburgh

 Interpretation 1: A plot of the first 2-3 principal components (PCs) should separate samples according to the phenotype of interest.



- Interpretation 1: A plot of the first 2-3 principal components (PCs) should separate samples according to the phenotype of interest.
  - Controlled model system experiments.
  - Phenotypes with large effects, such as tissues or cancer subtypes.



- Interpretation 1: A plot of the first 2-3 principal components (PCs) should separate samples according to the phenotype of interest.
  - Controlled model system experiments.
  - Phenotypes with large effects, such as tissues or cancer subtypes.
- Interpretation 2: The first few PCs are nuisance variables that should be removed from the data.



- Interpretation 1: A plot of the first 2-3 principal components (PCs) should separate samples according to the phenotype of interest.
  - Controlled model system experiments.
  - Phenotypes with large effects, such as tissues or cancer subtypes.
- Interpretation 2: The first few PCs are nuisance variables that should be removed from the data.
  - eQTL discovery: removing PCs improves results.



- Interpretation 1: A plot of the first 2-3 principal components (PCs) should separate samples according to the phenotype of interest.
  - Controlled model system experiments.
  - Phenotypes with large effects, such as tissues or cancer subtypes.
- Interpretation 2: The first few PCs are nuisance variables that should be removed from the data.
  - eQTL discovery: removing PCs improves results.



# Generative model for gene expression data

 Gene expression is driven by upstream factors that give rise to the observed data structure.

 PCA gives us a representation of these upstream factors but not a one-to-one correspondence.



Low rank matrix approximations (such as ones given by PCA) are effective because a limited number of upstream factors explain a large fraction of measurement variance.

 Low rank matrix approximations (such as ones given by PCA) are effective because a limited number of upstream factors explain a large fraction of measurement variance.

Given gene expression matrix  $Y_{g \times s}$ :

**MINIMIZE**  $||Y_{g \times s} - Z_{g \times k}B_{k \times s}||_F^2$ 

 Low rank matrix approximations (such as ones given by PCA) are effective because a limited number of upstream factors explain a large fraction of measurement variance.

Given gene expression matrix  $Y_{g \times s}$ :

**MINIMIZE** 
$$||Y_{g \times s} - Z_{g \times k}B_{k \times s}||_F^2$$

 B contains the principal components (PCs) or more generally latent variables (LVs). Z contains the "loadings" (effect of each LV on the genes).

Low rank matrix approximations (such as ones given by PCA) are effective because a limited number of upstream factors explain a large fraction of measurement variance.

Given gene expression matrix  $Y_{g \times s}$ :

**MINIMIZE** 
$$||Y_{g \times s} - Z_{g \times k}B_{k \times s}||_F^2$$

- B contains the principal components (PCs) or more generally latent variables (LVs). Z contains the "loadings" (effect of each LV on the genes).
- We hope that the individual vectors B<sub>i</sub> (latent variables) are meaningful. SVD (PCA) only guarantees minimum error—it doesn't guarantee anything about the interpretability of B.

 Low rank matrix approximations (such as ones given by PCA) are effective because a limited number of upstream factors explain a large fraction of measurement variance.

Given gene expression matrix  $Y_{g \times s}$ :

**MINIMIZE** 
$$||Y_{g \times s} - Z_{g \times k}B_{k \times s}||_F^2$$

- B contains the principal components (PCs) or more generally latent variables (LVs). Z contains the "loadings" (effect of each LV on the genes).
- We hope that the individual vectors B<sub>i</sub> (latent variables) are meaningful. SVD (PCA) only guarantees minimum error—it doesn't guarantee anything about the interpretability of B.
- Other methods that constrain Z to be sparse or positive may recover more meaningful structure.

 Low rank matrix approximations (such as ones given by PCA) are effective because a limited number of upstream factors explain a large fraction of measurement variance.

Given gene expression matrix  $Y_{g \times s}$ :

**MINIMIZE** 
$$||Y_{g \times s} - Z_{g \times k}B_{k \times s}||_F^2$$

- B contains the principal components (PCs) or more generally latent variables (LVs). Z contains the "loadings" (effect of each LV on the genes).
- We hope that the individual vectors B<sub>i</sub> (latent variables) are meaningful. SVD (PCA) only guarantees minimum error—it doesn't guarantee anything about the interpretability of B.
- Other methods that constrain Z to be sparse or positive may recover more meaningful structure.

$$\begin{split} \text{MINIMIZE} \quad ||Y - ZB||_F^2 + \lambda ||Z||_{L^1} \\ \text{SUBJECT TO} \quad Z > 0. \end{split}$$

We construct an example with 7 upstream factors; can we recover them?

- We construct an example with 7 upstream factors; can we recover them?
- We can make this problem quite hard by making some upstream factors have low variance

- We construct an example with 7 upstream factors; can we recover them?
- We can make this problem quite hard by making some upstream factors have low variance (very realistic: e.g., some cell-types have low abundance).

- We construct an example with 7 upstream factors; can we recover them?
- We can make this problem quite hard by making some upstream factors have low variance (very realistic: e.g., some cell-types have low abundance).



- We construct an example with 7 upstream factors; can we recover them?
- We can make this problem quite hard by making some upstream factors have low variance (very realistic: e.g., some cell-types have low abundance).
- PCA is very restrictive: each component is orthogonal.



- We construct an example with 7 upstream factors; can we recover them?
- We can make this problem quite hard by making some upstream factors have low variance (very realistic: e.g., some cell-types have low abundance).
- PCA is very restrictive: each component is orthogonal.
- If we constrain the decomposition to have sparse and positive loadings we can recover some, but not all, variables of interest.



- We construct an example with 7 upstream factors; can we recover them?
- We can make this problem quite hard by making some upstream factors have low variance (very realistic: e.g., some cell-types have low abundance).
- PCA is very restrictive: each component is orthogonal.
- If we constrain the decomposition to have sparse and positive loadings we can recover some, but not all, variables of interest.



composition variables

05



These methods are data agnostic, they don't make use of gene identities!

- We construct an example with 7 upstream factors; can we recover them?
- We can make this problem quite hard by making some upstream factors have low variance (very realistic: e.g., some cell-types have low abundance).
- PCA is very restrictive: each component is orthogonal.
- If we constrain the decomposition to have sparse and positive loadings we can recover some, but not all, variables of interest.



► We want not just the most parsimonious but also the most <u>biologically</u> meaningful decomposition.



# PLIER: Pathway-Level Information ExtractoR

Idea: Make use of gene identities.



Prior knowledge matrix C is a binary geneset representation, where each column is a potentially co-regulated set of genes. Number of genesets is many times larger than k.

 Non-convex optimization problem is solved by block-coordinate minimization

- Non-convex optimization problem is solved by block-coordinate minimization
- All constants are set automatically

- Non-convex optimization problem is solved by block-coordinate minimization
- All constants are set automatically
- Running time depends on the size of the data and size of C

- Non-convex optimization problem is solved by block-coordinate minimization
- All constants are set automatically
- ► Running time depends on the size of the data and size of C
- We pre-compute the inverse of C and use it to find a set of active genesets in each iteration to be optimized with the elastic-net penalty

Recovering the pathway effects with PLIER



Recovering the pathway effects with PLIER





# Recovering the pathway effects with PLIER



- Performance across repeated simulations
- Recovering 30 pathway effects from a prior information database of 1000 pathways



#### How do we use PLIER?

Example on real human blood dataset (35 samples) with directly measured by Cytof



1

l	MOSERLE IFNA RESPONSE MIZUSHIMA AUTOPHAGOSOME_FORMATION GSE 19182. Ifng SVM Meutophis activated SVM Toolis regulatory (fregs) Hesca SPL/COSOME DE DE D	1. J	I Neurophi-Resing HANAI APOPTOSIS VIA_TRAIL_UP HANAI_CULISTER_7 VALX_AM_CULISTER_7 VALX_AM_CULISTER_7 VALX_AM_CULISTER_7 VALY_A
1	IMPS THTC: TYPE HISTONE ACETYL TRANSFERASE COMPLEX REACTOME RESPIRATORY LECTRON, TRANSPORT BICCARRA CDC42RAC PATHWAY KEGA ASTMA CL, I PROMOTER OPENING BICCARTA PROTEXSOME PATHWAY MPS 205 PROTEXSOME PATHWAY		I FLOTHO FEDATRIC ALL THERAPY RESPONSE_UP DEN INTERACT WITH LCAS MOOTHA TCA MARCHAR TO AN RESPONSE AND PAIL RARA FUSION MARCHAR FUSION RESPONSE AND PAIL RARA FUSION MARCHAR FUSION OF FAILING HEART UP MARALO HYPOXIA DN
1	REACTOME ACTIVATED AMPK_STIMULATES_FATTY_ACID_OXID MIPS_TAS_COMPLEX MIPS_TAS_TAS_TRESOUNT_CYTOPLASMIC   MIPS_TAS_TAS_TRESOUNT_CYTOPLASMIC   MIPS_TAS_TAS_TRESOUNT_CYTOPLASMIC   MIPS_TAS_TAS_TAS_TAS_TAS_TAS_TAS_TAS_TAS_TA	, <i>1</i>	GILMORE CORE NRB PATHWAY NAKAIMA EOSINOPHIL I MARTINEZ RESPONSE TO TRABECTEDIN VISALA RESPONSE TO TRABECTEDIN GRANDWALX IRF3 TARGETS DN KUMANDO RESPONSE TO NUTLIN 3A DN KUMANDO RESPONSE TO NUTLIN 3A DN BOSCO ALLERGEN NDUGED TRA ASSOCIATED MODULE KORKOLA UBHYCNIC, CARCINOLA VS SEMINOMA UP
1	TCELLAS TCELLAS TCELLAS TCELLAS MEGA2 DERDA DERDA LECAPENDER MEGA2 DERDA LECAPENDER MEGA2 DERDA		RICKIAAN DETASTASIS ON PHONG THE TARGETS UN VALX ANL CLUSTER & NAMICA AND CLUSTER & HAHTIGA ANTOSIS FINOSOBE CDA UP PHON HAPTICA ANTOSIS FINOSOBE CDA UP PHON HAPTICA DETABLICATION OF COMPACT AND ANTOSIS FIL CHANG SLISTER CYTOTOXICUTION BURTON ADPOCENESIS 12 DI DOOPD CONENSIS 12 DI
il T	NKcell-control Monocyte-Day0 co co to co Bcell-Memory_IgM - co co co co Bcell-naive		ZHENG FOXP3 TARGETS IN T LYMPHOCYTE_DN MAYBURD RESPONSE TO L663536 DN RAGHAVACHARI PLATELET SPECIFIC GENES DISTECHE_ESCAPED_FROM_X_INACTIVATION

1

1

				MOSERLE_IFNA_RESPONSE			10	Neutrophil-Resting
1				MIZUSHIMA_AUTOPHAGOSOME_FORMATION				HAMAI_APOPTOSIS_VIA_THAIL_UP
				GSE19182_Ifng				ERY3
				SVM Neutrophils		,		VALK_AML_CLUSTER_7
				SVM Mast cells activated				BOSCO INTERFERON INDUCED ANTIVIRAL MODULE
- I				SVM T cells regulatory (Tregs)		1		MARTINELLI_IMMATURE_NEUTROPHIL_UP
				KEGG SPLICEOSOME	1			SCHMIDT_POR_TARGETS_IN_LIMB_BUD_UP
				KEGG CHRONIC MYELOID LEUKEMIA		1		WANG NEOPLASTIC TRANSFORMATION BY COND1 MYC
				I MIPS TFTC TYPE HISTONE ACETYL TRANSFERASE COMPLEX				FLOTHO PEDIATRIC ALL THERAPY RESPONSE UP
		1		REACTOME RESPIRATORY FLECTRON TRANSPORT				DEN INTERACT WITH LCAS
				BIOCARTA CDC42RAC PATHWAY				MOOTHA TCA -
				KEGG ASTHMA				PARK TRETINOIN RESPONSE AND PML RARA FUSION
				REACTOME RNA POL L PROMOTER OPENING				CREIGHTON ENDOCRINE THERAPY RESISTANCE 2
1				BIOCARTA PROTEASOME PATHWAY			1	CHEN LVAD SUPPORT OF FAILING HEART UP
	1			MIPS 265 PROTEASOME				MANALO HYPOXIA DN
				REACTOME ACTIVATED AMPK STIMULATES FATTY ACID OXID		1		GILMORE CORE NEKB PATHWAY
		- I -		KEGG DNA REPLICATION		1		NAKAJIMA EOSINOPHIE
		1.1.1		MIPS FIE3 COMPLEX			1	MARTINEZ RESPONSE TO TRABECTEDIN
				MIPS 40S RIBOSOMAL SUBLINIT CYTOPLASMIC				VISALA RESPONSE TO HEAT SHOCK AND AGING DN
			1	KEGG BIROSOME			- I	GRANDVAUX IRF3 TARGETS DN
1				REACTOME GENERIC TRANSCRIPTION PATHWAY		1		KUMAMOTO RESPONSE TO NUTLIN 3A DN
11				REACTOME FORMATION OF ATP BY CHEMIOSMOTIC COUPLING		1		BOSCO ALLERGEN INDUCED TH2 ASSOCIATED MODULE
1.1				KEGG OXIDATIVE PHOSPHORYLATION	1			KORKOLA EMBRYONIC CARCINOMA VS. SEMINOMA LIP
1				KEGG LYROROME			1	SENGLIPTA ERNAL ANTICORRELATED
			1	TCELLA7		1		RICKMAN METASTASIS DN
				TOELLAR		1		PHONG THE TARGETS UP
	1.1			TOELLAG		1		VALK AML CILISTER 8
	-	1		TOELLAS	1			GUTIERREZ CHRONIC LYMPHOCYTIC LEUKEMIA DN
				TOELLAZ		1		HALTOLA MYCOCIE EUNCOIDES CD4 LID
				NEAT				PYEON HPV POSITIVE TUMORS TIP
				N/COAD				CHUNG RUSTER CYTOTOVICITY DN
				MEGAZ				DIDTON ADDOCENECIC 12
	1.1			DENDAL			1	LINI ADC TADOETC
				DENDAL		1		
				PlasmaGell-FromPBMG				ZEENO EOVRA TARGETO IN T LYMPHOCYTE DN
				NKCEII-CONITOI	1			MAYDUDD DECOONCE TO LOOSTO DN
				Monocyte-Dayu co co 4 o				DACUAVAOUADI DI ATTI ET ODEOIEIO OFNEO
				Bcell-Memory_IgM - o o o o o	· · · · · · · · · · · · · · · · · · ·			DIGTEOUE FOOTBED FDOW V INACTIVATION
				Bcell-naive	1			DISTECTE_ESCAPED_FROM_X_INACTIVATION

How do we know the pathways are real? we zero-out a random 1/5 of the genes for every pathway before optimization and check if we get them back in the loading.

1	1			MOSERLE IFNA RESPONSE III Neutrophil-Resting MIZUSHIMA, AUTOPHAGOSOME_FORMATION I HAMALAPOPTOSIS_VIA_TRAIL_UP	
				GSE19182 Ifing ERY3	
				SVM Neutrophils VALK_AML_CLUSTER_7	
	1			SVM Mast cells activated BOSCO INTERFERON INDUCED ANTIVIRAL MODULE	
				SVM T cells regulatory (Tregs) MARTINELLI_IMMATURE_NEUTROPHIL_UP	
				I KEGG SPLICEOSOME SCHMIDT POR TARGETS IN LIMB BUD UP	
	1			KEGG CHRONIC MYELOID LEUKEMIA WANG NEOPLASTIC TRANSFORMATION BY CONDI MYC	2
				I MIPS TETC TYPE HISTONE ACETYL TRANSFERASE COMPLEX FLOTHO PEDIATRIC ALL THERAPY RESPONSE UP	
			1	REACTOME RESPIRATORY FLECTRON TRANSPORT DEN INTERACT WITH LCAS	
				BIOCARTA CDC42RAC PATHWAY	
1				KEGG ASTHMA PARK TRETINOIN RESPONSE AND PML RARA FUSION	
				REACTIONE BNA POL L PROMOTER OPENING CREIGHTON ENDOCRINE THERAPY RESISTANCE 2	
1				BIOCARTA PROTEASOME PATHWAY	
	1			MIPS 265 PROTEASOME MANAEO HYPOXIA DN	
				REACTOME ACTIVATED AMPK STIMULATES FATTY ACID OXID GILMORE CORE NFKB PATHWAY	
			1	KEGG DNA BEPLICATION NAKAJIMA EQSINOPHIL	
		1		MIPS FIE3 COMPLEX MARTINEZ RESPONSE TO TRABECTEDIN	
				MIPS 405 BIBOSOMAL SUBUNIT CYTOPLASMIC VISALA RESPONSE TO HEAT SHOCK AND AGING DN	
				L KEGG BIROSOME GRANDVAUX IRF3 TARGETS DN	
1				REACTOME GENERIC TRANSCRIPTION PATHWAY	
11				BEACTOME FORMATION OF ATP BY CHEMIOSMOTIC COUPLING BOSCO ALLERGEN INDUCED TH2 ASSOCIATED MODUL	E
1				KEGG OXIDATIVE PHOSPHORYLATION KORKOLA EMBRYONIC CARCINOMA VS SEMINÓMA UP	
1				KEGG I YSOSOME SENGLIPTA EBNA1 ANTICORRELATED	
			1	TCFILTAZ	
				TCELLAG PHONG TIVE TARGETS UP	
	1			TCFI1A4 VALK AML CLUSTER 8	
		1		TCELLA2 GUTIERREZ CHRONIC LYMPHOCYTIC LEUKEMIA DN	
		1		TCELLA1 HAHTOLA MYCOSIS EURODIDES CD4 UP	
			1	NKA1 I PYEON HPV POSITIVE TUMORS UP	
		1		MEGA2 CHUNG BLISTER CYTOTOXICITY DN	
	1			MEGA1 BURTON ADIPOGENESIS 12	
	1.1			DENDA1	
		1.1		PlasmaCell_FromPBMC LL DCP2_BOUND_MBNA	
				NKcell-control ZHENG FOXP3 TARGETS IN T LYMPHOCYTE DN	
		1		Monocyte-David MAYBURD RESPONSE TO L663536 DN	
		1		Beell-Memory InM	
		1		Booling Disteche Escaped FROM X INACTIVATION	

- How do we know the pathways are real? we zero-out a random 1/5 of the genes for every pathway before optimization and check if we get them back in the loading.
- We can see many cell types.

MODENEMA AN DESIGNE FORMATION MODENEMA AND DESIGN VIA TRAIL LIP ENT SOM Neutronis SVM Neutronis			IGOEDIE IDII DEGDONGE				Manufacture bill Planetice a
decrement in modeline_commons decrement in the provided in the provid		N N	WUSERLE IFINA RESPONSE		1		HAMAI APOPTOSIS VIA TRAIL LIP
SVM Neurophis SVM Neurophis SVM Neurophis SVM Toels regulatory (res) FXMT cells regulatory (res) FXMT cells regulatory (res) FXMT cells regulatory (res) SXMT cells regulatory (res)		1 0	SEE10102 Has		1		ERV3
SVM Marchies scharted SVM Marchies Scharted SVM Marchies Scharted SVM Marchies Scharted SVM Marchies Scharter VEGG SPLIEGSSONE VEGG SPLIEGSSON		1 0	2/M Neutrophile			1	VALK AML CLUSTER 7
SVW Mad cells advated SVW Toells regulatory (Tregs) I KEGG SPUCEOSOME KEGG SPUCEOSOME K	1	1 3	2VM Meet cells activated			1	BOSCO INTERFERON INDUCED ANTIVIRAL MODULE
KEGG SPLICEOSOME STOLEN SCHMIDT POR TARGETS IN LINB BUD UP		3	SVM T cells activated		1		MARTINELLI IMMATURE NEUTROPHIL UP
KEGG CHRONIC MYELOID LEUKEMIA WANG NEOPLASTIC TRANSFORMATION BY CONDI MY		3	ZEGG EDLICEGEOME		1		SCHMIDT POR TARGETS IN LIMB BUD LIP
	1		CEGG_CURONIC_MVELOID_LEUKEMIA				WANG NEOPLASTIC TRANSFORMATION BY COND1 MYC
IMDE TETO TVE HISTONE ACETYL TRANSCERASE COMPLEY		IN IN	AIDS TETS TYPE LISTONE ACETYL TRANSEEDASE COMDLEY				ELOTHO PEDIATRIC ALL THERAPY RESPONSE UP
		1 0	ALL AND ALL AND ALL AND ALL AND	1			DEN INTERACT WITH LCAS
REACTOME RESPIRATORY ELECTION TRANSPORT		1 1	PLACTOME RESPIRATORY ELECTRON_TRANSPORT				I MOOTHA TCA
BACK TEETINON RESPONSE AND PAIL BARA FUSION	- ' I	L L L L L L L L L L L L L L L L L L L	COO ACTUMA		1		PARK TRETINOIN RESPONSE AND PMI RARA EUSION
DEADTONE DNA DOL I DDOMOTED ODENING CREGHTON ENDOCENE THERAPY RESISTANCE 2	1	D	CEGG ASTRIMA		·····	1	CREIGHTON ENDOCRINE THERAPY RESISTANCE 2
PROVIDE THAT DON'T A COLLEGE AND A COLLEGE A		n 0					CHEN LVAD SUPPORT OF FAILING HEART UP
MOONTA PROTEACONE AND A MANAGE MANAGEMENT OF THE REAL		0			1		MANALO HYPOYIA DN
DEACTOME ACTIVATED AMOV STIMULATES EATTY ACID OVID		N D	VIPS 205 PROTEKSOWE			1	GILMORE CORE NEKR PATHWAY
		1 1	COO DNA DEDITION			L	NAKA IMA EOSINOPHI
MIDE DE2 COMINY	1	1 1	AIDS EISS COMDLEY				MARTINEZ RESPONSE TO TRABECTEDIN
MIDS BIR DEPOSITION STORE STORE AND AGING DN	1	M	AIDS ARE DIROCOMAL CURUNIT OVTODI AGMIC	1			VISALA RESPONSE TO HEAT SHOCK AND AGING DN
READ BECOME SUBJECT OF PRIME	1		CECC DIROCOME				GRANDVALIX IRES TARGETS DN
REACTOME GENERIC TRANSCRIPTION PATHWAY KIMAMOTO RESPONSE TO NITLIN 3A DN	1	R	REACTOME GENERIC TRANSCRIPTION PATHWAY		1		KUMAMOTO RESPONSE TO NUTLIN 3A DN
REACTOME FORMATION OF ATP. BY CHEMIOSMOTIC COUPLING BOSCO ALL'ERGEN INDICED TH2 ASSOCIATED MODIL	L. L.	R	REACTOME FORMATION OF ATP BY CHEMIOSMOTIC COUPLING				BOSCO ALLERGEN INDUCED TH2 ASSOCIATED MODULE
KORKOLA EMBRYONIC CARCINOMA VS. SEMINOMA UP	1	K	(EGG OXIDATIVE PHOSPHORY ATION		1		KORKOLA EMBRYONIC CARCINOMA VS SEMINOMA LIP
KEGG LYSOSOME	1	K	(EGG_LYSOSOME			1	SENGUPTA EBNA1 ANTICORRELATED
I TCELLAZ		I T	ICELLA7		1		RICKMAN METASTASIS DN
TCELLAS PHONG TNE TARGETS UP	1	Ť.	ICELLA6		1		PHONG TNF TARGETS UP
TCELLA4 VALK AML CLUSTER 8		Ť	ICELLA4		1		VALK AML CLUSTER 8
GUTIERREZ CHRONIC LYMPHOCYTIC LEUKEMIA DN		Ť.	ICELLA2	1			GUTIERREZ CHRONIC LYMPHOCYTIC LEUKEMIA DN
TCELLA1 HAHTOLA MYCOSIS FÜNGOIDES CD4 UP		1 1	ICELLA1			1	HAHTOLA MYCOSIS FUNGOIDES CD4 UP
NKA1 PYEON HPV POSITIVE TUMORS UP		Ň	NKA1	1			PYEON HPV POSITIVE TUMORS UP
MEGA2 CHUNG BLISTER CYTOTOXICITY DN		I M	//EGA2				CHUNG BLISTER CYTOTOXICITY DN
MEGA1 BURTON ADIPOGENESIS 12	1	M	//EGA1	1			BURTON ADIPOGENESIS 12
DENDA1 LIN APC TARGETS		D	DENDA1				LIN APC TARGETS
PlasmaCell-FromPBMC LI DCP2 BOUND MRNA		P P	PlasmaCell-FromPBMC				LI DCP2 BOUND MRNA
NKcell-control ZFIENG_FOXP3_TARGETS_IN_T_LYMPHOCYTE_DN		N	VKcell-control				ZHENG_FOXP3_TARGETS_IN_T_LYMPHOCYTE_DN
Monocyte-Day0 go go yt ty MAYBURD_RESPONSE_TO_L663536_DN		M	Monocyte-Day0 m m m				MAYBURD_RESPONSE_TO_L663536_DN
Bcell-Memory IgM - 0 0 0 0 RAGHAVACHARI_PLATELET_SPECIFIC_GENES	1	B	Bcell-Memory IgM - Öööööö	1.			RAGHAVACHARI_PLATELET_SPECIFIC_GENES
Bcell-naïve DISTECHE_ESCAPED_FROM_X_INACTIVATION		I B	Bcell-naïve				DISTECHE_ESCAPED_FROM_X_INACTIVATION

- How do we know the pathways are real? we zero-out a random 1/5 of the genes for every pathway before optimization and check if we get them back in the loading.
- We can see many cell types.
  - ► 3 kinds of CD8 T cells.

	1	MOSERI E JENA RESPONSE			10	Neutrophil_Resting
1		MIZUSHIMA AUTOPHAGOSOME FORMATION				HAMAI APOPTOSIS VIA TRAIL UP
		GSE19182 Ifno	1			ERY3
	1	SVM Neutrophils				VALK AML CLUSTER 7
	1	SVM Mast cells activated		1		BOSCO INTERFERON INDUCED ANTIVIRAL MODULE
1		SVM T cells regulatory (Tregs)		1		MARTINELLI IMMATURE NEUTROPHIL UP
		I KEGG SPLICEOSOME				SCHMIDT POR TARGETS IN LIMB BUD UP
		KEGG CHRONIC MYELOID LEUKEMIA				WANG NEOPLASTIC TRANSFORMATION BY COND1 MYC
		MIPS TFTC TYPE HISTONE ACETYL TRANSFERASE COMPLEX				FLOTHO_PEDIATRIC_ALL_THERAPY_RESPONSE_UP
	1	REACTOME RESPIRATORY ELECTRON TRANSPORT				DEN_INTERACT_WITH_LCA5
1		BIOCARTA_CDC42RAC_PATRWAY				MOOTHA_TCA
1		KEGG_ASTHMA				PARK_TRETINOIN_RESPONSE_AND_PML_RARA_FUSION
		REACTOME_RNA_POL_I_PROMOTER_OPENING			. I.,	CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_2
		BIOCARTA_PROTEASOME_PATHWAY				CHEN_LVAD_SUPPORT_OF_FAILING_HEART_UP
		MIPS 26S PROTEASOME				MANALO HYPOXIA DN
		REACTOME_ACTIVATED_AMPK_STIMULATES_FATTY_ACID_OXID				GILMORE_CORE_NERB_PATHWAY
		KEGG_DNA_REPLICATION				NAKAJIMA_EUSINUPHIL
		MIPS_EIF3_COMPLEX	1			MARTINEZ RESPONSE TO TRABECTEDIN
1		MIPS_40S_HIBOSOMAL_SUBUNIT_CYTOPLASMIC				CRANDVALLY IDES TAD/CETS DAL
		REGG HIBUSUME				KUMAMOTO DECRONICE TO NUTLIN 24 DN
		REACTOME GENERIC TRANSCRIPTION PATRWAT		1		PORCO ALLEDGEN INDUCED THE ARCOCIATED MODULE
		KEGO OVIDATIVE DUCCEVED AT DI CHEMIUSMOTIC COUPLING	a			KORKOLA EMPRYONIC CARCINOMA VE CEMINOMA UR
		KEGG UXIDATIVE PHOSPHORITLATION				CENCLIDTA EDNAL ANTICODDELATED
		TCELLAZ		1		RICKMAN METASTASIS DN
1		TCELLAS		1		PHONG THE TARGETS UP
	1	TCELLAA		1		VALK AML CILISTER 8
	1	TCELLA2	1			GUTIERREZ CHRONIC LYMPHOCYTIC LEUKEMIA DN
		TCELLA1		1		HAHTOLA MYCOSIS FUNGOIDES CD4 UP
		NKA1	1			PYEON HPV POSITIVE TUMORS UP
		MEGA2				CHUNG BLISTER CYTOTOXICITY DN
		MEGA1				BURTON ADIPOGENESIS 12
	1	DENDA1			1	LIN APC TARGETS
		PlasmaCell-FromPBMC				LI DCP2 BOUND MRNA
		NKcell-control		1		ZHENG_FOXP3_TARGETS_IN_T_LYMPHOCYTE_DN
		Monocyte-Day0 co co co co				MAYBURD_RESPONSE_TO_L663536_DN
		Bcell-Memory_IgM - o o o o o	1.			RAGHAVACHARI_PLATELET_SPECIFIC_GENES
		Bcell-naïve				DISTECHE_ESCAPED_FROM_X_INACTIVATION

- How do we know the pathways are real? we zero-out a random 1/5 of the genes for every pathway before optimization and check if we get them back in the loading.
- We can see many cell types.
  - ► 3 kinds of CD8 T cells.
  - Naive and memory B-cells.



- How do we know the pathways are real? we zero-out a random 1/5 of the genes for every pathway before optimization and check if we get them back in the loading.
- We can see many cell types.
  - ► 3 kinds of CD8 T cells.
  - Naive and memory B-cells.
  - Very high frequency cell-types have multiple LVs.

# How do we use PLIER?

PLIER latent variables can be plugged into any downstream analysis that would normally be done at the gene level–for example eQTLs.

LV id	LV name	snps	cis-Gene(s)	corrected p-value
44	Mega/platelet 1	rs1354034	ARHGEF3	< 1.45e-10
133	Mega/platelet 2	rs1354034	ARHGEF3	0.01547
120	Histones	rs1354034	ARHGEF3	0.01889
97	Zinc fingers, pseudogenes	rs1471738	SENP7	< 1.45e-10
56	PLAGL1 associated, myeloid	rs9321957	PLAGL1	3.6e-05
42*	IKZF1 associated, myeloid	rs10251980	IKZF1	< 1.45e-10
17	NEK6 associated, myeloid	rs16927294	NEK6	0.00360
67	Neutrophils	rs13289095	PKN3,SET,ZDHHC12	0.01888
55*	NFE2 associated, erythrocyte	rs35979828	NFE2	< 1.45e-10
21	Interferon-gamma	rs3184504	SH2B3	5.9e-05
40	NFKB/TNF	rs12100841	PPP2R3C	0.00204
16	Myeloid/ILC	rs1138358	BCL2A1,MTHFS,ST20	0.00025

Interferon-gamma LV21 uses 3 pathways:

- REACTOME\_INTERFERON\_GAMMA\_SIGNALING
- GSE19182 Ifng
- SANA\_RESPONSE\_TO\_IFNG\_UP

# How do we use PLIER?

PLIER latent variables can be plugged into any downstream analysis that would normally be done at the gene level–for example eQTLs.

LV id	LV name	snps	cis-Gene(s)	corrected p-value
44	Mega/platelet 1	rs1354034	ARHGEF3	< 1.45e-10
133	Mega/platelet 2	rs1354034	ARHGEF3	0.01547
120	Histones	rs1354034	ARHGEF3	0.01889
97	Zinc fingers, pseudogenes	rs1471738	SENP7	< 1.45e-10
56	PLAGL1 associated, myeloid	rs9321957	PLAGL1	3.6e-05
42*	IKZF1 associated, myeloid	rs10251980	IKZF1	< 1.45e-10
17	NEK6 associated, myeloid	rs16927294	NEK6	0.00360
67	Neutrophils	rs13289095	PKN3,SET,ZDHHC12	0.01888
55*	NFE2 associated, erythrocyte	rs35979828	NFE2	< 1.45e-10
21	Interferon-gamma	rs3184504	SH2B3	5.9e-05
40	NFKB/TNF	rs12100841	PPP2R3C	0.00204
16	Myeloid/ILC	rs1138358	BCL2A1,MTHFS,ST20	0.00025

Interferon-gamma LV21 uses 3 pathways:

- ► REACTOME\_INTERFERON\_GAMMA\_SIGNALING
- GSE19182 Ifng
- SANA\_RESPONSE\_TO\_IFNG\_UP



#### LV eQTLs pathway associations



LV eQTLs pathway associations



Platelet









LV133 (Mega/platelet LV **early**) genes are expression in megakaryocyte precursors.

LV44 (Mega/platelet LV late) genes are megakaryocyte specific.



LV133 (Mega/platelet LV **early**) genes are expression in megakaryocyte precursors.

LV44 (Mega/platelet LV late) genes are megakaryocyte specific.

 rs1354034 is known to be pleitropic: it affects both mean platelet volume (MPV) and platelet counts (PLT).

Furman-Niedziejko A. et al. Relationship between abdominal obesity, platelet blood count and mean platelet volume in patients with metabolic syndrome.

- rs1354034 is known to be pleitropic: it affects both mean platelet volume (MPV) and platelet counts (PLT).
- MPV and PLT are negatively correlated due to the tight control on total platelet volume.

Furman-Niedziejko A. et al. Relationship between abdominal obesity, platelet blood count and mean platelet volume in patients with metabolic syndrome.

- rs1354034 is known to be pleitropic: it affects both mean platelet volume (MPV) and platelet counts (PLT).
- MPV and PLT are negatively correlated due to the tight control on total platelet volume.



Furman-Niedziejko A. et al. Relationship between abdominal obesity, platelet blood count and mean platelet volume in patients with metabolic syndrome.

- rs1354034 is known to be pleitropic: it affects both mean platelet volume (MPV) and platelet counts (PLT).
- MPV and PLT are negatively correlated due to the tight control on total platelet volume.
- Hypothesis: LV133 (early) is associated with platelet number and LV44 (late) is associated with volume.



Furman-Niedziejko A. et al. Relationship between abdominal obesity, platelet blood count and mean platelet volume in patients with metabolic syndrome.

- rs1354034 is known to be pleitropic: it affects both mean platelet volume (MPV) and platelet counts (PLT).
- MPV and PLT are negatively correlated due to the tight control on total platelet volume.
- Hypothesis: LV133 (early) is associated with platelet number and LV44 (late) is associated with volume.
- We have data from large GWAS studies of blood count variables that show some loci regulate PLT and MPV independently. How are these associated with our latent variables?



Furman-Niedziejko A. et al. Relationship between abdominal obesity, platelet blood count and mean platelet volume in patients with metabolic syndrome.

- rs1354034 is known to be pleitropic: it affects both mean platelet volume (MPV) and platelet counts (PLT).
- MPV and PLT are negatively correlated due to the tight control on total platelet volume.
- Hypothesis: LV133 (early) is associated with platelet number and LV44 (late) is associated with volume.
- We have data from large GWAS studies of blood count variables that show some loci regulate PLT and MPV independently. How are these associated with our latent variables?



phenotype	reported SNP	Close gene	LV 133 p-value	LV44 p-value	proxy SNP
PLT	rs2911132	ERAP2	2.4417e-05	0.13817361	rs2549803
MPV	rs108/6550	COPZ1	0.69933	1.1847e-05	rs108/6550

Table: Raw p-values. 80 platelet related SNPs tested.

Furman-Niedziejko A. et al. Relationship between abdominal obesity, platelet blood count and mean platelet volume in patients with metabolic syndrome.

# PLIER models transfer across datasets

Two human whole blood datasets:

- DGN: RNAseq US cohort
- NESDA: Affy European cohort

# PLIER models transfer across datasets

Two human whole blood datasets:

- DGN: RNAseq US cohort
- ► NESDA: Affy European cohort

PLIER decompositions performed independently

### PLIER models transfer across datasets

Two human whole blood datasets:

- DGN: RNAseq US cohort
- NESDA: Affy European cohort

PLIER decompositions performed independently



#### Correlation with phenotypes is more consistent in LV space



#### Some fun results

- Dataset from a collaborator: melanoma RNAseq , immunotherapy reponse (8 progressors, 11 responders).
- Very similart to the published Hugo et al. dataset \* (13 progressors, 15 responders). How do they compare?



<sup>\*</sup> Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma

### Usoskin et al. dataset

scRNAseq of mouse sensory neurons.



 PLIER returns a set of latent variables that are both maximally independent from each other and maximally aligned with prior information.

 PLIER returns a set of latent variables that are both maximally independent from each other and maximally aligned with prior information.

 Minimally supervised method: selects relevant pathways and discards thousands of irrelevant ones.

 PLIER returns a set of latent variables that are both maximally independent from each other and maximally aligned with prior information.

 Minimally supervised method: selects relevant pathways and discards thousands of irrelevant ones.

 Additional output matrix U provides the mapping between pathways and LVs for quick interpretation.

 PLIER returns a set of latent variables that are both maximally independent from each other and maximally aligned with prior information.

 Minimally supervised method: selects relevant pathways and discards thousands of irrelevant ones.

 Additional output matrix U provides the mapping between pathways and LVs for quick interpretation.

 Pathway-level estimates can be used in any subsequent analysis yielding mechanistic hypotheses.

# Questions and future directions

 Group-level regularization on samples: not every LV exists in every sample.

#### Questions and future directions

 Group-level regularization on samples: not every LV exists in every sample.

► Looking for LVs that maximize objectives other than variance.

#### Questions and future directions

 Group-level regularization on samples: not every LV exists in every sample.

► Looking for LVs that maximize objectives other than variance.

When are positivity constraints on the loadings necessary?

# Acknowledgments

- University of Pittsburgh Computational and Systems Biology
  - Wayne Mao

- Icahn School of Medicine at Mount Sinai
  - Stuart Sealfon
  - Elena Zaslavsky
  - Boris Hartman
- Funding
  - 1R01 HG009299-01A1
  - U24 DK112331-01
  - U54 HG008540-03
  - 1R03 MH109009-01A1