

# Linking Congenital Heart Disease to Cardiac Cell Types via Hierarchical Poisson Factorization

Zhenyu Zhang

## Abstract

*This study aims to explore the connection of congenital heart disease (CHD) to different cardiac cell types via single cell gene expression data and genetic mutation information. scRNA-seq data of PCW 6.5-7 human developing heart cells were retrieved from a previous study [1]. Major cardiac cell types were identified by cell markers using PCA and clustering analysis on scRNA-seq data. Hierarchical Poisson Factorization found 6 major hidden factors behind sequencing data that could lead to differentiation of 11 cell types. CHD risk genes were retrieved from OMIM and de novo mutations of CHD cases were retrieved from a previous study [2]. Majority CHD risk genes were expressed across more than one cell type. However, fibroblasts/smooth muscle cells and endothelial cells appear to have more risk genes significantly expressed. RAS signaling was enriched in endothelial cells and collagen synthesis was enriched in fibroblasts/smooth muscle cells. DNMs enrichment analysis of genes highly expressed in different cell types showed that CHD may not be dominantly associated with certain cardiac type. DNMs enrichment analysis showed 6 significant genes in CHD subtype LVO and CTD with multiple LoF DNMs. Three of them, FLT4, NOTCH1, and CHD7 predominantly expressed in endothelial cells.*

## Introduction

Congenital heart disease (CHD) is one leading cause of fatal birth defects. Previous studies have identified CHD risk genes and de novo mutations (DNMs) from comprehensive family studies.[2] However, connections between these risk genes/variants and different types of cardiac cells are still hidden due to lack of cell specific studies. Recently, the advancement of the single cell RNA sequencing technique has allowed large-scale cell-specific gene expression analysis. In this study, scRNA-seq data of human heart embryonic cells (post conception week 6.5-7) were used to find whether CHD is more associated with certain types of cardiac cells. We also studied risk genes with multiple loss of function (LoF) de novo mutations identified from previous studies in their ties to cardiac cell types.

PCA and clustering analysis are popular techniques to identify cell types along with expression analysis of cell specific marker genes.[3] Cell types can be easily visualized by some non-linear reduction method like UMAP. Conventionally, differential expression analysis was used to find potential cell specific markers. However, recent studies have shown that gene expression in a single cell is a stochastic process so that gene expression in different cells, even if they belong to the same type, are heterogeneous.[4] Sparsity in scRNA expression matrix further makes differential expression analysis unreliable. Therefore, a model can approximate the variational process might be superior in the identification of potentially cell-specific disease risk variants.

scHPF[5], a model based on hierarchical poisson factorization, appears to be appealing in this setting. Instead of directly running a single test across all cells, scHPF assumed there were  $k$  hidden factors that each gene or cell has a fixed budget to distribute over. A gene and a cell having similar factor distribution may indicate ties between the two. The top genes in each factor (ranked by factor scores) may have an indication on the distinct biological process behind corresponding cell type. Though the model cannot directly predict disease causing gene, by looking at factor score distribution of risk genes or de novo variants, we might have future research focus on the interactions of the genes and the corresponding cell types.

## Datasets

scRNA-seq data of human embryonic cardiac cells within 6.5–7 PCW were collected from a previous study of gene expression in developing human heart.[1] The whole dataset contains 3717 single cells, splitting into two separate experiments with average ~2900 genes and ~11000 unique transcripts count per cell. The accession number for the raw sequencing data reported in this paper is European Genome-phenome Archive (EGA): EGAS00001003996.

De novo mutation information in CHD patients were retrieved from a previous study of CHD probands recruited to the PGC and the Pediatric Heart Network (PHN) programs [2, Supplemental Dataset]. There were 2871 probands including 2,645 parent-offspring trios and 226 singletons. 1789 control trios were retrieved from a study of autism case; the controls consist of parent controls and unaffected siblings. [2, Supplemental Dataset]

## Methods

### *Identification of different cardiac cell types*

R package Seurat was used to perform clustering on the scRNA-seq dataset to explore identities of different cell types and similarities among them. Seurat is a popular package to perform QC and conventional analysis like PCA and clustering to scRNA-seq data. Seurat also has a strong method set to visualize dataset quality and its downstream analysis.

Quality control: Remove cells with >10% mitochondrial RNA count, cells with too few features (< 200 genes showing counts) and cells with too many features (>8000 genes showing counts). Gene expression is globally normalized across the whole dataset.

Clustering: PCA was used to reduce dimensionality. Number of top K dimensions were determined by straw plot and elbow plot of the PC numbers against variance. k nearest neighbour was then used for clustering the cell data. Clusters generated were visualized by UMAP for a better view of global structure since different cell types may share different similarities among them. For each cluster, marker genes were computed by Seurat findMarkers.

Gene signatures: Differential expression was performed to find markers for each cluster. Curated marker genes for major cardiac cell types (cardiomyocytes, fibroblast, epicardial cells etc.) were retrieved from previous studies using cellMarker[7], a manually curated cell marker resource. Gene expression of curated cell markers were done and visualized to identify cell types in each cluster. Top marker genes computed by Seurat were compared to curated marker genes and explored in PubMed literatures. At last, very similar cell types (>50% overlapped top 10 marker genes) were combined and a distinct cell type was assigned to each cell.

### *Identify latent factors by HPF*

Hierarchical poisson factorization (HPF) was performed on the scRNA-seq dataset by Python package scHPF to explore underlying latent factors that may link gene expression to cell types. The HPF model assumes each gene and cell can be vectorized by k factors following k independent gamma distributions, which are the prior distributions. Each gene and cell have a limited budget to distribute across the k factors. The model treats scRNA-seq expression count as an observation following a poisson distribution, which is the likelihood, being parameterized by the dot product of the gene and cell k-size vectors. Variational Bayesian method is used as an approximation to the posterior distribution. During the inference, the rate of k gamma distributions will be calculated for each gene and cell; therefore, the k-size vectors will be sampled from the inferred distributions.

Non-coding genes were filtered out from the expression matrix. To select the best K, there were two major criteria. The first one was minimum overlap between top genes of any two factors (<2 overlap in top 100 genes). The second one was to make each cell type have at least one unique enriched factor. Best K would be chosen whichever was stricter. Three trials were run with the best K value and the one trial with best loss (MSE) was recorded.

scHPF generated a k-factor score for each cell. Cells of the same assigned type were grouped together, and a mean cell score was calculated for each assigned cell type. A heatmap of mean cell scores visualized the relationship between cell types and hidden factors.

Curated marker genes expressed in each factor were visualized by heatmap to examine how well scHPF could differentiate cell types compared to conventional cluster analysis. Top 100 genes in each factor were conducted overrepresentation analysis by ReactoMe[8]. Biological functions or processes from the overrepresentation analysis were related to the corresponding cell type. Cell types that were clustered close to each other, such as atrial cardiomyocytes, and ventricular cardiomyocytes, were particularly compared with each other by distinctive factors.

### *Known CHD risk genes expression in each factor*

Combined 253 CHD risk genes were curated 212 genes from OMIM and human homologs of 61 genes from a mice CHD screen study (Supplemental Dataset).[9] Since each gene and cell has the same budget distributed across k factors, the k-factor score is a good measurement of preference against each factor for both a gene and a cell. A heatmap with genes on the row and factors on the column visualized the overall expression of genes in each factor. Similar was done to genes with de novo mutations in cases. Genes particularly enriched in certain factors were linked to cell types that dominate that factor. Further, these genes were performed overrepresentation analysis to detect potential biological processes that are disease related. Top biological processes were recorded (by adjusted p-value) for each factor.

### *Case control study of DNMs enrichment*

Case control comparison was between DNMs of 2645 case trios and that of 1789 control trios. This was to see whether cases were more enriched with DNMs than controls. The enrichment analysis was carried out by R package “denovolyzeR” [10]. The overall enrichment was calculated by comparing the observed number of DNMs across each variant class (syn, mis, lof:[del/ins+stop/start loss], prot:[lof+mis]) to expected under the null mutation model. The expected number of DNMs was calculated by taking the sum of each variant class specific probability multiplied by the number of probands in the study, multiplied by two. Then the Poisson test was used by “denovolyzeR” to calculate enrichment of observed DNMs.

Similar test was performed to the top 500 genes (<10 overlap) of k factors from the scHPF execution. This was to see whether there is any factor with DNMs enriched in top ranked genes. Doing this we can link the DNMs enrichment to the corresponding cell type.

### *Connect risk variants in CHD subtypes to cardiac cell types*

Out of 2645 cases, there were 5 major subtypes of on the basis of the major cardiac lesion: conotruncal defects (CTD, N=872), D-transposition of the great arteries (D-TGA, N=251), heterotaxy (HTX, N=272), left ventricular outflow tract obstruction (LVO, N=797), or Other (N=679). Using denovolyzeR, top enriched genes with multi-hits of DNMs were calculated for each subtype. By checking gene scores from scHPF, the latent factors enriched by the top genes generated from denovolyzeR were found and the result was visualized by a gene-factor score heatmap.

## **Results**

### *PCA and clustering*

Majority cells had less than 10% mitochondrial RNA. Number of features with counts was rallied around 2500-3000. Only ~10% total cells were excluded from the following analysis.

Principal component analysis was performed on normalized data. Straw Plot (Sup Fig 1) shows variation dropped sharply after 5 components and Elbow plot (Sup Fig 2) shows an elbow between component 5 to component 10. Using the first 5 components for clustering, 11 different clusters were identified. (Sup Fig 3) The following differential expression analysis (DEA) found markers suggesting the major cell type of each cluster. Cluster 0 and cluster 1 are close to each other under UMAP visualization. ACTA2, which was found as a top-ten gene for cluster 0 by DEA, is a marker gene of cardiac fibroblast [11]. The most significant gene in cluster 1 is OGN, which is a marker gene of smooth muscle cells to differentiate them from fibroblasts [12]. Cluster 2 and 4 have similar marker genes found by DAE showing they are capillary endothelial cells. EMCN is the most significant gene for cluster 2 and TMEM100 for cluster 4. EMCN is the marker gene for venous endothelial cells and TMEM100 is the marker gene for artery endothelial cells [13]. Cluster 3 and 5 are enriched with cardiomyocytes marker MYL6 and TNNT2. However, MYL2 [14] enrichment in cluster 3 indicates ventricular cardiomyocytes. ASCL [15] was enriched in cluster 9, which is heart neural crest cells. Using the curated cellMarkers, cluster 6, 7, 8 and 10 were identified as epicardial cells, erythrocytes, immune cells, and endothelial cells. Erythrocytes and immune cells were dumped for the rest of the analysis since they are not a major part of cardiac functions. Enrichment of part of curated cell markers were visualized through a heatmap. (Sup Fig 4) UMAP visualization with cell types remained (excluding immune cells and erythrocytes) was presented (Sup Fig 5)

### *scHPF execution*

Overlap of top ranked genes in each factor was rare until the number of factors used surpassed the number of identified cell types; however, not all cells had a noticeable unique factor enriched in by heatmap. In fact, even if K had increased over 10, much larger than the number of identified cell types, cell types that were closed to each other like atrial cardiomyocytes and ventricular cardiomyocytes have similar factor distribution shown by heatmap (Sup Fig 6). Also, we could see there were factors not enriched by any cell type identified by PCA clustering. However, different cell types appear to be distinct from each other in UMAP visualization, suggesting scHPF was able to detect different cell types by embedding cells into lower dimensions (Sup Fig 7)

### *scHPF factors and cell types*

The two explicit criteria of choosing K:

1. For top 500 genes of each factor, we want any of two factors have less than 5% of overlapping genes.

2. For cell scores, since each cell have a same budget to distribute over k factors, we want average cells can use 80 percent of its budget within 2 factors. This means average cells have a major preference to one factor; otherwise it is hard to associate a cell type with a factor.

When k is too large, these 2 criterions cannot be met. In our execution, k had to be smaller than 9. Here we used 6 because when we run  $k > 6$ , similar cell types like ventricular myocytes and atrial myocytes were still clustered in the same factor almost. These similar cell types are very hard to separate if we stucked with our 2 criterions. Running high k, we would have factors not favored by any cell types. In these factors, top genes were genes like ribosomal proteins which express in most cell types. Because we wanted to link risk genes to cell types, those genes were not in our interest.  $K = 6$  was the smallest k we found that each set of similar cell types had distinct preferred factor. For convenience of following analysis, we used  $k=6$ .

After comparing heatmaps of cell type versus factors of different K values,  $K = 6$  was selected for the best value to run final trials. A heatmap of cell mean scores against factors was generated. (Fig 1)

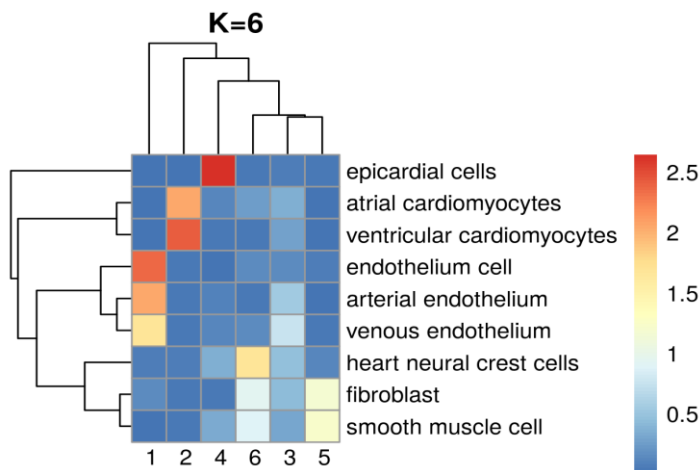


Fig 1.  $K=6$ , mean score heatmap of scHPF factors versus cell types (scale: (0, 2.8))

Cell types and their major/minor enriched factors were summarized in the supplement table (Sup Tab 1). A heatmap was generated for scores of curated marker cells versus each factor. (Sup Fig 8) This suggests that scHPF generated factor is a good low dimension representation of gene expressions.

#### *Expression of CHD risk genes and DNMs in scHPF factors*

Known CHD risk genes and DNMs expression scores in all 6 scHPF factors were generated and included in two separate pdf files (Supplemental File 1 and Supplemental File 2). Although each factor had noticeable numbers of genes distributing over half of the score budget in, factor 1 and factor 5 had the most such genes. Factor 1 is dominated by endothelial cells and factor 5 is dominated by smooth muscle cells and fibroblasts. Similar situations happened in expression of genes with case DNMs. Overrepresentation analysis found significant biological processes using genes differentially expressed in factor 1 and factor 5 (Sup Tab 2).

#### *Case control study of DNMs enrichment*

Table 1a. DNMs enrichment analysis for case trios (N=2645)

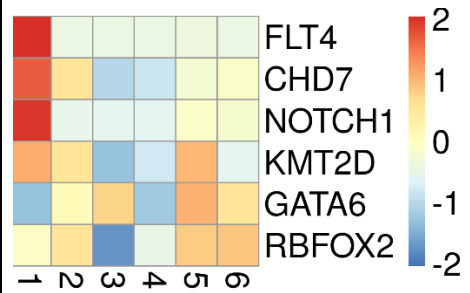
class	observed	expected	enrichment	pValue
syn	699	741.7	0.942	0.945
mis	1913	1666	1.15	1.78E-09
lof	373	231.3	1.61	7.16E-18
prot	2286	1897.3	1.2	2.9E-18
all	2985	2638.9	1.13	2.19E-11

Table 1b. DNMs enrichment analysis for control trios (N=1789)

class	observed	expected	enrichment	pValue
syn	484	501.6	0.965	0.79
mis	1194	1126.8	1.06	0.0243
lof	149	156.4	0.952	0.734
prot	1343	1283.2	1.05	0.0499
all	1827	1784.9	1.02	0.162

Table 1 showed that case trios have significantly higher DNMs enrichment, especially in potential damaging LoF mutations, than control trios. Enrichment analysis on top 500 genes (ranked by factor scores) of each factor showed that only factor 2 has some significant enrichment of overall mutations (Sup Tab 3). Further examining subtypes of CHD, we found 4 genes for LVO and 3 genes for CTD that had multi hits of LoF DNMs and were shown significance in enrichment analysis (pValue < 0.01) (Table 3). No multi-DNM hit genes were found in D-TGA and HTX subtypes.

Table 2 multi LoF DNMs enriched genes in CHD subtypes Fig 2 scHPF factor scores of enriched genes

CHD subtype	Genes (Multi LOF hits)						
LVO	KMT2D, RBFOX2, CHD7, NOTCH1						
CTD	CHD7, GATA6, FLT4						
D-TGA	NA						
HTX	NA						

CHD7 both appeared in LVO and CTD as a multi DNMs hit gene. Out of these 6 genes, FLT4 and NOTCH1 were significantly dominant of factor 1, and CHD7 was also but had a little preference to factor 2 (Fig 2).

## Discussion

### Biologically interpretable latent factors of scHPF

There exist many machine learning methods that aim to find lower dimensional representations of sparse high dimensional data. However, one advantage of scHPF is that the latent factors (lower dimension) generated are interpretable. Since each gene had the same budget to distribute over each factor, it is easy to rank top genes in each factor without any normalization. Putting the top genes in overrepresentation analysis, enriched biological functions were found in each factor. For example, striated muscle contraction in factor 2 (Sup Fig 6b) is a forceful contraction that happens more in ventricular cardiomyocytes [16]. Activation of HOX genes during differentiation is a biological process dominantly overrepresented in factor 4 (Sup Fig 6b) which is dominantly enriched by smooth muscle cells. There is literature evidence that the HOX gene is involved in differentiation of stem cells into smooth muscle cells [17]. Factor 7 (Sup Fig 6) top ranked genes overrepresents collagen formation, which is an important process in fibroblast cells [18].

### CHD risk genes with cardiac cell types

From the heatmap of risk gene expression in each scHPF factor, we find that factor 1 and 5 are predominantly preferred by more risk genes than other factors [Sup File 1]. With an overrepresentation analysis on CHD risk genes preferring factor 1, we find most of the significant processes are related to RAS signaling pathway. There is previous study shown that RAS/MAPK pathway malfunction is an important cause of CHD [19]; however, endothelial cells (dominantly prefers factor 1) (Fig 1) was rarely connected with this pathway in studying of CHD before. Same analysis with factor 5, we found collagen formation and assembly along with two other extracellular matrix processes related to collagen metabolism were enriched. Factor 5 was dominantly preferred by fibroblasts

and smooth muscle cells (Fig 1). Studies have shown that imbalance of collagen metabolism leads to myocardial fibrosis which is associated with various types of congenital and pediatric heart disease [20] [21].

#### *LoF DNMs with cardiac cell types*

Case control study (Table 1) shows that LoF mutations are more enriched in the CHD cases than the controls. Top 500 genes in each factor rarely overlap with each other, showing that they are distinctively expressed in their corresponding cell types. Doing DNMs enrichment analysis in these top genes, we do not find any of the factors have top ranked genes significantly enriched with potential LoF DNMs (pValue < 0.05) (Sup Table 3). This suggests there is no evidence that specific cell type is more associated with CHD than other cell types. Genes carrying high risk variants may express in different cell types and involve complex functional interactions across the heart. Instead of figuring out whether certain cell type is more associated with CHD, we use enrichment analysis to find highly risky CHD genes. We define genes with LoF pValue < 0.05 and multiple mutation hits as highly risky. We try to find such genes for each major subtype of CHD (LVO, CTD, D-TGA, and HTX). 4 genes were found to be highly risky in LVO and 3 genes in CTD. Conventionally, a heatmap of gene expression in clusters is a way to find the cell types that the genes mainly express. However, gene expression in single cell is a stochastic process and many genes were not expressed regularly in cells; in this case, genes like FLT4 and CHD7 are not very well identified in a heatmap. However, using scHPF factors for genes, we can find the predominantly preferred factors and the underlying cell types. Here, we found NOTCH1 and FLT4 are predominantly expressed in endothelial cells and CHD7 are mainly expressed in endothelial cells. Other 3 genes do not have preference to specific scHPF factors. NOTCH1 is said to be required for proper development of cardiac outflow tract associated with LVO [22][23]. FLT4 is a known gene associated with tetralogy of Fallot, which is a subtype of CTD while the underlying mechanism is not clear [24]. CHD7 is known to be a key chromatin remodeler in cardiovascular development, related to CHARGE Syndrome causing heart defects [25]. However, what role of endothelium cell and its related cell functions in this disease is still not yet understood.

#### **Conclusion**

Combining whole exome sequencing and single cell sequencing technique, we can connect genetic variants in CHD cases with certain cardiac cell type. With scHPF, we can implicitly quantify such connections and better visualize the relationship between risk genes and cell types. The de novo mutations enrichment analysis showed no evidence of certain cell type more associated with congenital heart disease. However, we find three highly risk genes NOTCH1, FLT4 and CHD7 associated with two CHD subtypes. The three genes are highly expressed in endothelial cells, indicating a future research direction of how some cases of CHD are developed.

#### **Reference**

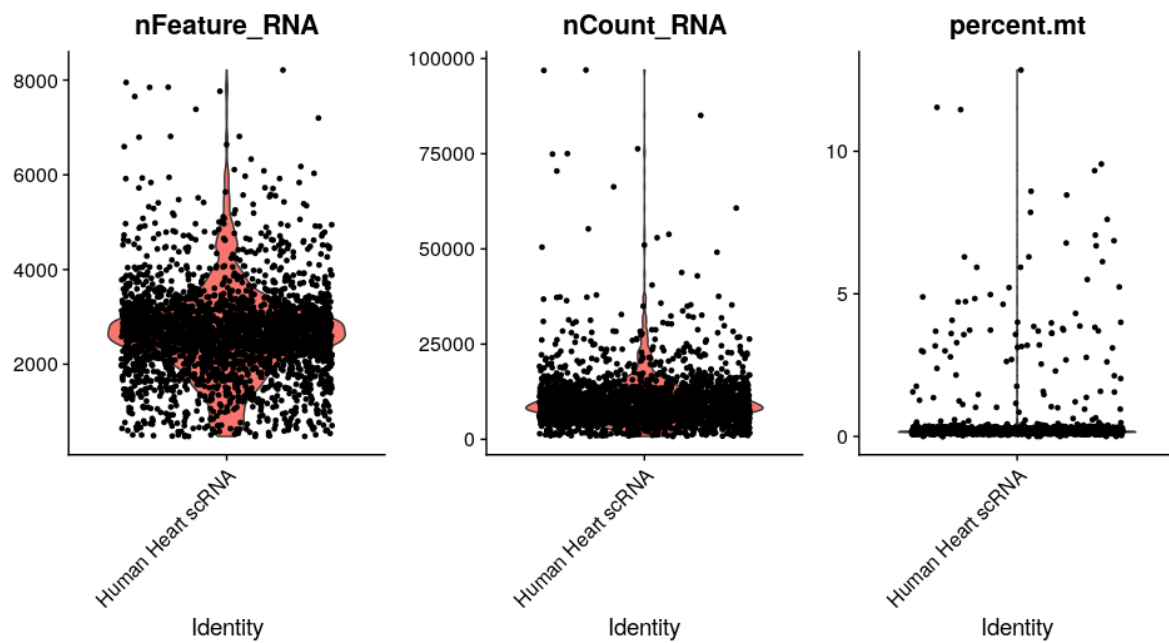
1. Asp, Michaela, et al. "A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing Human Heart." *Cell*, vol. 179, no. 7, 2019, doi:10.1016/j.cell.2019.11.025.
2. Jin, Sheng Chih et al. "Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands." *Nature genetics* vol. 49,11 (2017): 1593-1601. doi:10.1038/ng.3970
3. Zeisel A, Muñoz-Manchado AB, Codeluppi S, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*. 2015;347(6226):1138-1142. doi:10.1126/science.aaa1934
4. Stegle, O., Teichmann, S. & Marioni, J. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* 16, 133–145 (2015). <https://doi.org/10.1038/nrg3833>
5. Levitin HM, Yuan J, Cheng YL, et al. *De novo* gene signature identification from single-cell RNA-seq with hierarchical Poisson factorization. *Mol Syst Biol*. 2019;15(2):e8557. Published 2019 Feb 22. doi:10.15252/msb.20188557
6. Arrington, Cammon B et al. "Exome analysis of a family with pleiotropic congenital heart disease." *Circulation. Cardiovascular genetics* vol. 5,2 (2012): 175-82. doi:10.1161/CIRCGENETICS.111.961797
7. Xinxin Zhang, Yujia Lan, Jinyuan Xu, Fei Quan, Erjie Zhao, Chunyu Deng, Tao Luo, Liwen Xu, Gaoming Liao, Min Yan, Yanyan Ping, Feng Li, Aiai Shi, Jing Bai, Tingting Zhao, Xia Li, Yun Xiao, CellMarker: a manually curated resource of cell markers in human and mouse, *Nucleic Acids Research*, Volume 47, Issue D1, 08 January 2019, Pages D721–D728, <https://doi.org/10.1093/nar/gky900>

8. Fabregat, A., Sidiropoulos, K., Viteri, G. *et al.* Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* 18, 142 (2017). <https://doi.org/10.1186/s12859-017-1559-2>
9. Li, Y., Klena, N., Gabriel, G. *et al.* Global genetic analysis in mice unveils central role for cilia in congenital heart disease. *Nature* 521, 520–524 (2015). <https://doi.org/10.1038/nature14269>
10. Ware JS, Samocha KE, Homsy J, Daly MJ. Interpreting de novo Variation in Human Disease Using denovolyzeR. *Curr Protoc Hum Genet.* 2015; 87:7 25 1–15. [PubMed: 26439716]
11. Ivey, M. J., & Tallquist, M. D. (2016). Defining the Cardiac Fibroblast. *Circulation journal : official journal of the Japanese Circulation Society*, 80(11), 2269–2276. <https://doi.org/10.1253/circj.CJ-16-1003>
12. Deckx S, Heymans S, Papageorgiou AP. The diverse functions of osteoglycin: a deceitful dwarf, or a master regulator of disease? *FASEB J.* 2016;30(8):2651-2661. doi:10.1096/fj.201500096R
13. dela Paz NG, D'Amore PA. Arterial versus venous endothelial cells. *Cell Tissue Res.* 2009;335(1):5-16. doi:10.1007/s00441-008-0706-5
14. Sheikh F, Lyon RC, Chen J. Functions of myosin light chain-2 (MYL2) in cardiac muscle and disease [published correction appears in *Gene.* 2015 Oct 15;571(1):151]. *Gene.* 2015;569(1):14-20. doi:10.1016/j.gene.2015.06.027
15. Memic F, Knoflach V, Sadler R, et al. Ascl1 Is Required for the Development of Specific Neuronal Subtypes in the Enteric Nervous System. *J Neurosci.* 2016;36(15):4339-4350. doi:10.1523/JNEUROSCI.0202-16.2016
16. Shadrin, I Y et al. “Striated muscle function, regeneration, and repair.” *Cellular and molecular life sciences : CMLS* vol. 73,22 (2016): 4175-4202. doi:10.1007/s00018-016-2285-z
17. Klein D, Benchellal M, Kleff V, Jakob HG, Ergün S. Hox genes are involved in vascular wall-resident multipotent stem cell differentiation into smooth muscle cells. *Sci Rep.* 2013;3:2178. Published 2013 Oct 22. doi:10.1038/srep02178
18. Narayanan, A S et al. “Collagen synthesis by human fibroblasts. Regulation by transforming growth factor-beta in the presence of other inflammatory mediators.” *The Biochemical journal* vol. 260,2 (1989): 463-9. doi:10.1042/bj2600463
19. Williams K, Carson J, Lo C. Genetics of Congenital Heart Disease. *Biomolecules.* 2019;9(12):879. Published 2019 Dec 16. doi:10.3390/biom9120879
20. Wu M, Cronin K, Crane JS. Biochemistry, Collagen Synthesis. [Updated 2020 May 4]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 Jan-. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK507709/>
21. Tian J, An X, Niu L. Myocardial fibrosis in congenital and pediatric heart disease. *Exp Ther Med.* 2017;13(5):1660-1664. doi:10.3892/etm.2017.4224
22. Koenig SN, Bosse K, Majumdar U, Bonachea EM, Radtke F, Garg V. Endothelial Notch1 Is Required for Proper Development of the Semilunar Valves and Cardiac Outflow Tract. *J Am Heart Assoc.* 2016;5(4):e003075. Published 2016 Apr 22. doi:10.1161/JAHA.115.003075
23. Riley MF, McBride KL, Cole SE. NOTCH1 missense alleles associated with left ventricular outflow tract defects exhibit impaired receptor processing and defective EMT. *Biochim Biophys Acta.* 2011;1812(1):121-129. doi:10.1016/j.bbadis.2010.10.002
24. Reuter MS, Jobling R, Chaturvedi RR, et al. Haploinsufficiency of vascular endothelial growth factor related signaling genes is associated with tetralogy of Fallot. *Genet Med.* 2019;21(4):1001-1007. doi:10.1038/s41436-018-0260-9

25. Payne S, Burney MJ, McCue K, et al. A critical role for the chromatin remodeller CHD7 in anterior mesoderm during cardiovascular development. *Dev Biol.* 2015;405(1):82-95. doi:10.1016/j.ydbio.2015.06.017

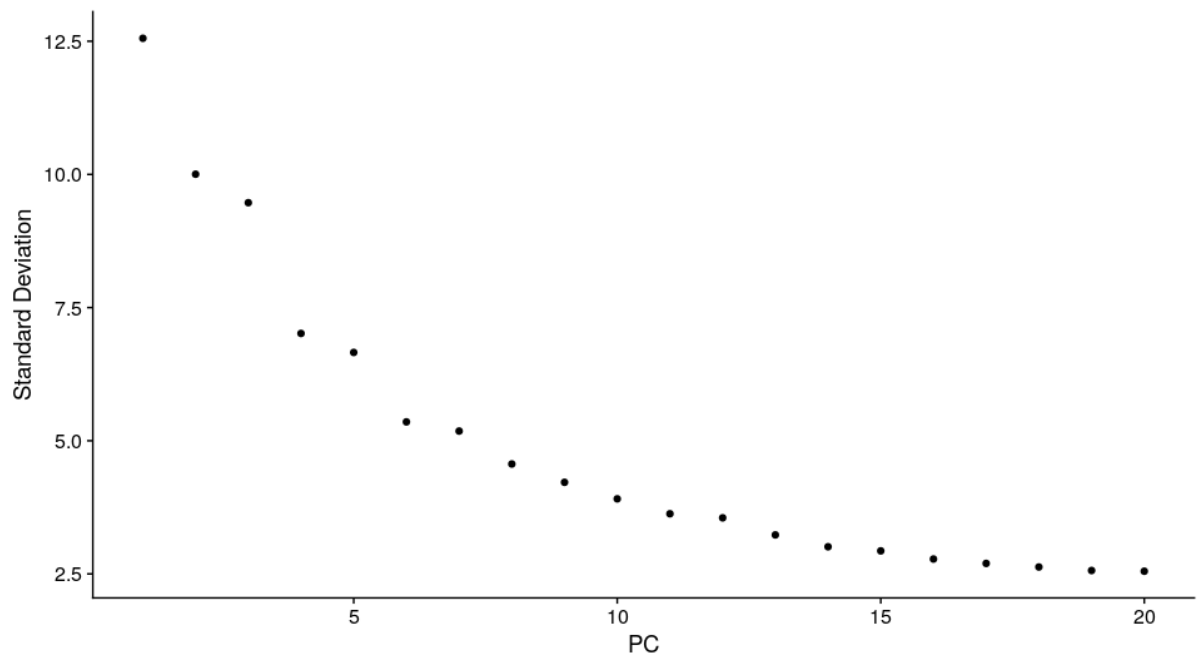
## Supplemental

**Fig 1** Quality control threshold for scRNA-seq data

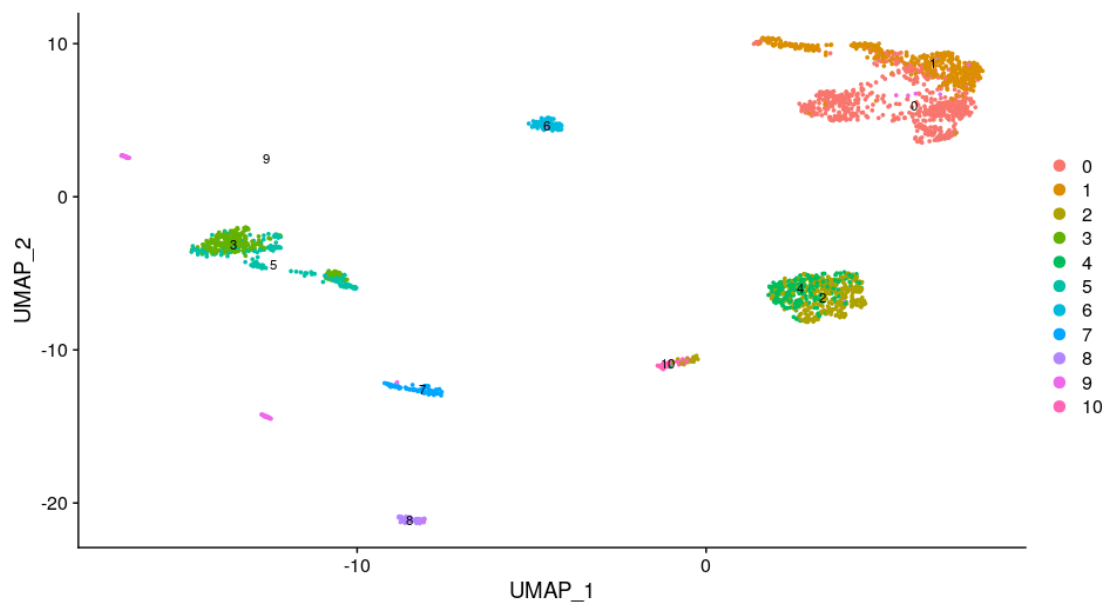


**Fig 2** Elbow plot of PCA analysis

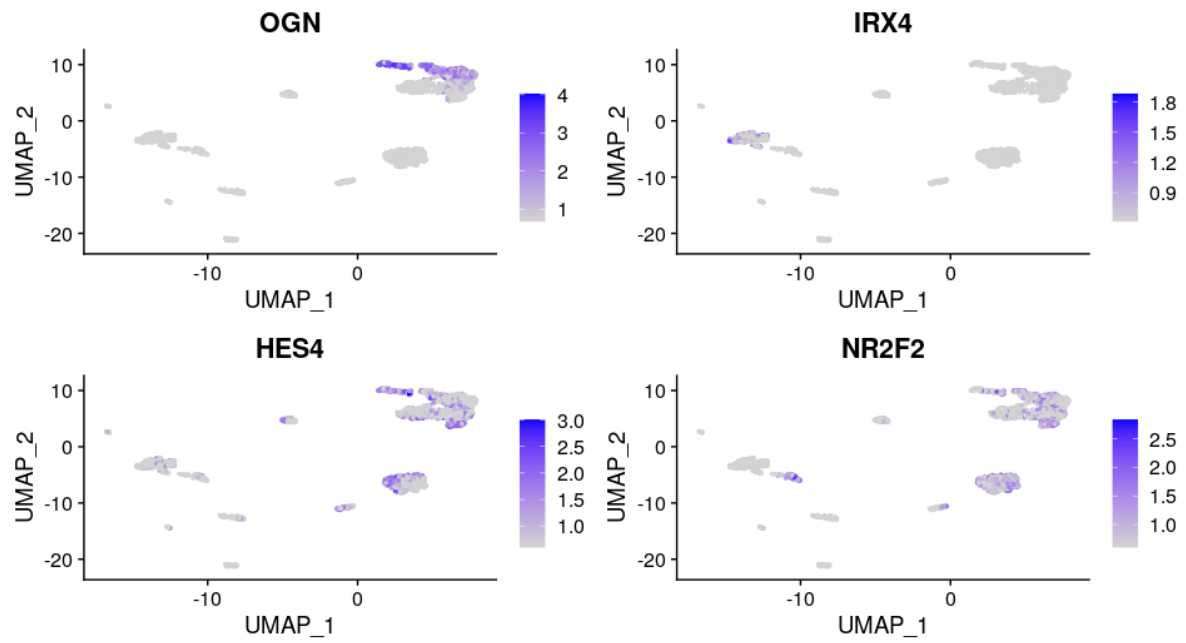




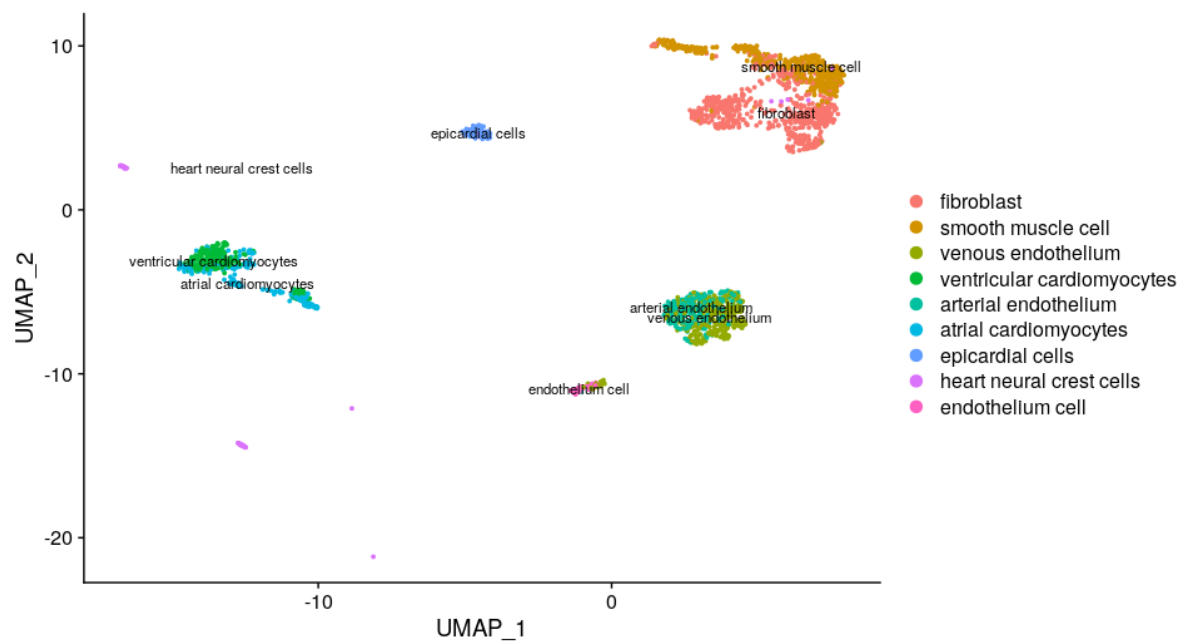
**Fig 3.** PCA clustering with cluster No.



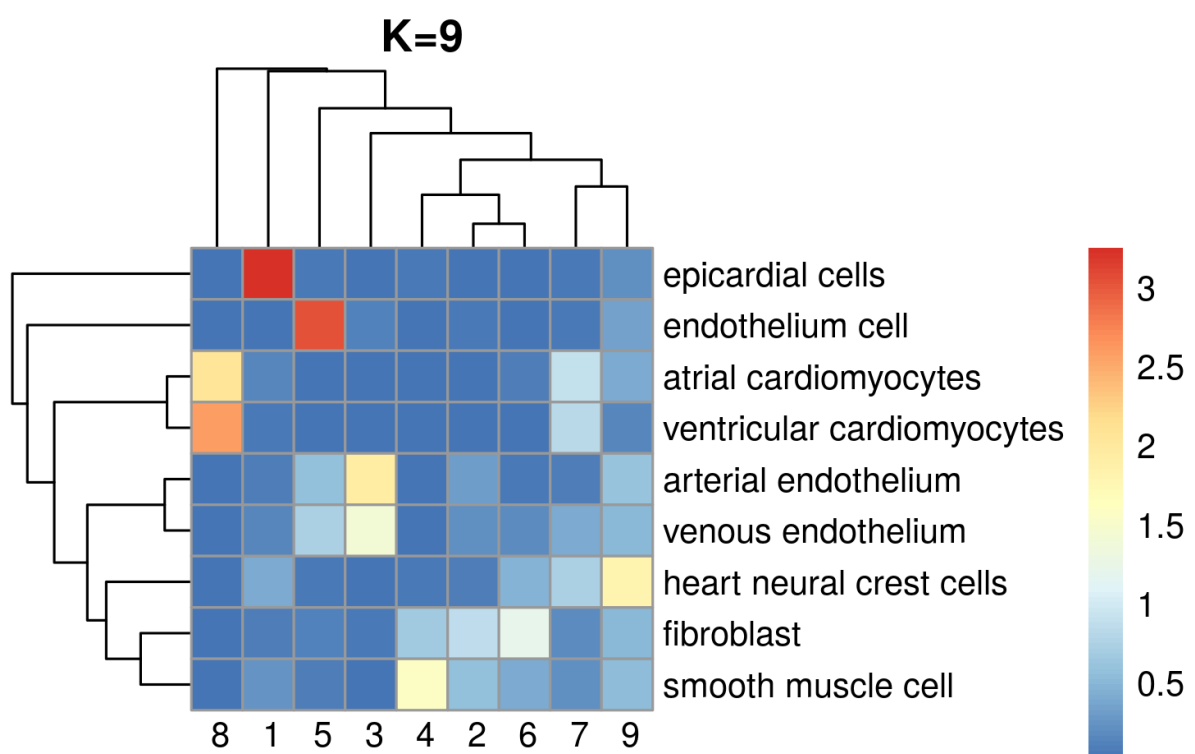
**Fig 4** a sample of curated cell marker expression visualized



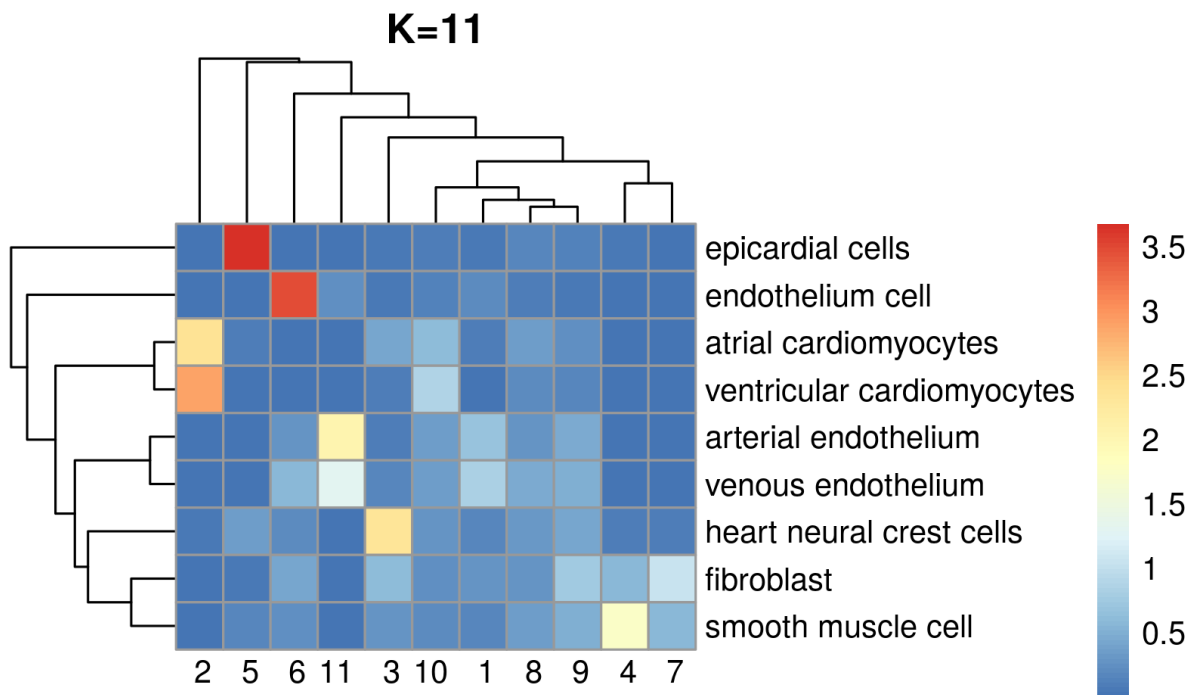
**Fig 5** PCA clustering with cell types labeled.



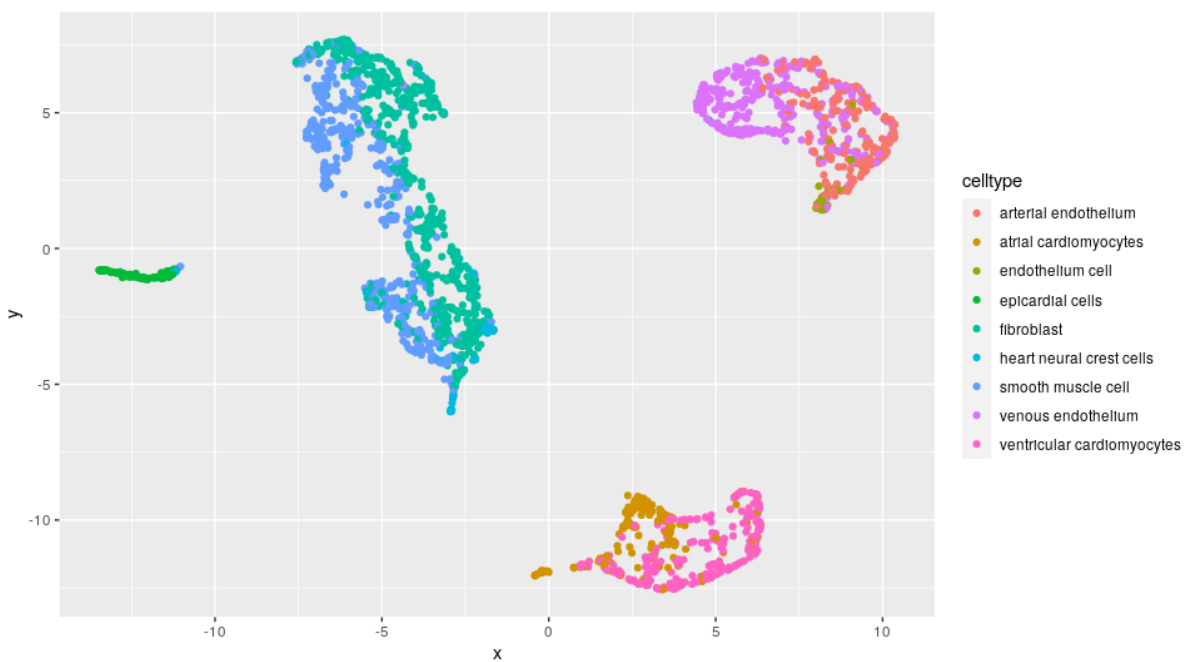
**Fig 6a** scHPF factor cell type heatmap (K=9)



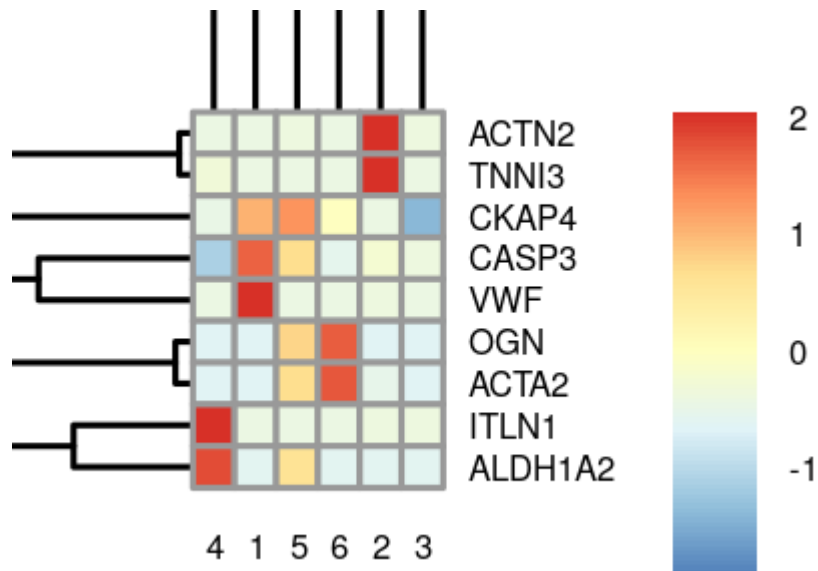
**Fig 6b** scHPF factor cell type heatmap (K=11)



**Fig 7** UMAP visualization of scHPF scores with assigned cell types



**Fig 8** score heatmap of curated cell markers in each scHPF factor



**Table. 1** cell types and enriched factor numbers

Cell type	Major factor	Minor factors
Epicardial cells	2	NA
Atrial cardiomyocytes	2	3, 6
Ventricular cardiomyocytes	2	3
Endothelium	1	NA
Artery endothelium	1	3
Venous endothelium	1	3
Heart crest neural cells	6	3,4

Smooth muscle cells	5	3,4,6
Fibroblasts	5	3,6

**Table 2a.** Factor 1 risk genes overrepresentation analysis

Pathway identifier	Pathway name	#Entities found	#Entities total	Entities ratio	Entities pValue	Entities FDR
R-HSA-5673000	RAF activation	6	44	0.002998705104614	2.07612010916236E-08	2.6188057991039E-06
R-HSA-6802953	RAS signaling downstream of NF1 loss-of-function variants	4	9	0.000613371498671	5.06786429488315E-08	2.6188057991039E-06
R-HSA-6802949	Signaling by RAS mutants	6	54	0.003680228992026	6.8899407179579E-08	2.6188057991039E-06
R-HSA-6802946	Signaling by moderate kinase	6	54	0.003680228992026	6.8899407179579E-08	2.6188057991039E-06

	activity BRAF mutants					
R-HSA-9649948	Signaling downstream of RAS mutants	6	54	0.003680228992026	6.8899407179579E-08	2.6188057991039E-06
R-HSA-5637812	Signaling by EGFRvIII in Cancer	5	27	0.001840114496013	7.45270910540441E-08	2.6188057991039E-06
R-HSA-5637810	Constitutive Signaling by EGFRvIII	5	27	0.001840114496013	7.45270910540441E-08	2.6188057991039E-06
R-HSA-6802955	Paradoxical activation of RAF signaling by kinase inactive BRAF	6	55	0.003748381380767	7.66938852425625E-08	2.6188057991039E-06
R-HSA-112412	SOS-mediated signalling	4	10	0.000681523887412	7.7023699973644E-08	2.6188057991039E-06

**Table 2b.** Factor 5 risk genes overrepresentation analysis

Pathway identifier	Pathway name	#Entities found	#Entities total	Entities ratio	Entities pValue	Entities FDR
R-HSA-3000170	Syndecan interactions	5	29	0.001976419273496	2.43768285890233E-07	5.87310277263553E-05
R-HSA-8874081	MET activates PTK2 signaling	5	32	0.002180876439719	3.9466549561773E-07	5.87310277263553E-05
R-HSA-1474290	Collagen formation	7	104	0.007087848429087	4.85380394432688E-07	5.87310277263553E-05
R-HSA-2022090	Assembly of collagen fibrils and other multimeric	6	67	0.004566210045662	6.51277160645947E-07	5.92662216187811E-05

	structures					
--	------------	--	--	--	--	--

**Table 3a.** Case DNMs enrichment in top 500 genes of each scHPF factors

Factor 1

class	observed	expected	enrichment	pValue
syn	23	21.8	1.05	0.43
mis	50	46.6	1.07	0.331
lof	10	5.9	1.7	0.0752
prot	60	52.5	1.14	0.167
all	83	74.4	1.12	0.172

Factor 2



class	observed	expected	enrichment	pValue
syn	36	29.5	1.22	0.138
mis	78	64.6	1.21	0.0576
lof	12	9	1.33	0.199
prot	90	73.6	1.22	0.0353
all	126	103.2	1.22	0.0161

Factor 3

class	observed	expected	enrichment	pValue
syn	13	15.3	0.85	0.756
mis	28	35.3	0.794	0.909
lof	5	4.8	1.05	0.52
prot	33	40.1	0.824	0.887
all	46	55.4	0.831	0.911

Factor 4

class	observed	expected	enrichment	pValue
syn	22	22	1	0.529
mis	58	48.3	1.2	0.0945
lof	10	6.5	1.53	0.125
prot	68	54.8	1.24	0.0467
all	90	76.8	1.17	0.0763

Factor 5

class	observed	expected	enrichment	pValue
syn	17	24.8	0.684	0.96
mis	61	54.2	1.13	0.194
lof	9	7.5	1.21	0.333
prot	70	61.7	1.14	0.159
all	87	86.5	1.01	0.493

Factor 6

class	observed	expected	enrichment	pValue
syn	21	21.3	0.987	0.554
mis	50	46	1.09	0.297
lof	10	5.9	1.71	0.074
prot	60	51.9	1.16	0.145
all	81	73.1	1.11	0.193

**Table 3b.** Control DNMs enrichment in top 500 genes of each scHPF factors

Factor 1

class	observed	expected	enrichment	pValue
syn	7	14.8	0.474	0.991
mis	30	31.6	0.951	0.633
lof	3	4	0.756	0.757
prot	33	35.5	0.929	0.686
all	40	50.3	0.795	0.94

Factor 2

class	observed	expected	enrichment	pValue
syn	13	20	0.651	0.961
mis	60	43.7	1.37	0.011
lof	5	6.1	0.819	0.728
prot	65	49.8	1.31	0.022
all	78	69.8	1.12	0.177

Factor 3

class	observed	expected	enrichment	pValue
syn	12	10.3	1.16	0.343
mis	22	23.9	0.922	0.676
lof	4	3.2	1.24	0.405
prot	26	27.1	0.959	0.609
all	38	37.4	1.01	0.485

Factor 4

class	observed	expected	enrichment	pValue
syn	14	14.9	0.941	0.626
mis	27	32.6	0.827	0.86
lof	6	4.4	1.36	0.283
prot	33	37.1	0.89	0.77
all	47	51.9	0.905	0.772

Factor 5

class	observed	expected	enrichment	pValue
syn	8	16.8	0.476	0.994
mis	28	36.7	0.764	0.94
lof	6	5	1.19	0.392
prot	34	41.7	0.815	0.901
all	42	58.5	0.718	0.99

Factor 6

class	observed	expected	enrichment	pValue
syn	13	14.4	0.903	0.679
mis	46	31.1	1.48	0.00737
lof	2	4	0.505	0.905
prot	48	35.1	1.37	0.0219
all	61	49.5	1.23	0.062