Bioinformatics Advance Access published August 27, 2016

Bioinformatics, 2016, 1–9
doi: 10.1093/bioinformatics/btw510
Advance Access Publication Date: 2 August 2016
Original Paper

Systems biology

Predicting G protein-coupled receptor downstream signaling by tissue expression

Yun Hao[1] **and Nicholas P. Tatonetti**[1,*]

[1]Departments of Biomedical Informatics, Systems Biology, and Medicine, Columbia University, New York, NY, 10032, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on April 12, 2016; revised on July 12, 2016; accepted on July 28, 2016

Abstract

Motivation: G protein-coupled receptors (GPCRs) are central to how cells respond to their environment and a major class of pharmacological targets. However, comprehensive knowledge of which pathways are activated and deactivated by these essential sensors is largely unknown. To better understand the mechanism of GPCR signaling system, we integrated five independent genome-wide expression datasets, representing 275 human tissues and cell lines, with protein-protein interactions and functional pathway data.
Results: We found that tissue-specificity plays a crucial part in the function of GPCR signaling system. Only a few GPCRs are expressed in each tissue, which are coupled by different combinations of G-proteins or β-arrestins to trigger specific downstream pathways. Based on this finding, we predicted the downstream pathways of GPCR in human tissues and validated our results with L1000 knockdown data. In total, we identified 154,988 connections between 294 GPCRs and 690 pathways in 240 tissues and cell types.
Availability and Implementation: The source code and results supporting the conclusions of this article are available at http://tatonettilab.org/resources/GOTE/source_code/.
Contact: nick.tatonetti@columbia.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

G protein-coupled receptors (GPCRs) comprise the largest family of transmembrane signaling molecules—more than 800 distinct human proteins. Each member of this family shares in a common seven transmembrane (7TM) α-helical fold. As highly versatile membrane sensors, they regulate many physiological processes including immunization, sensation, behavioral and mood regulation, homeostasis modulation (Katritch *et al.*, 2012). GPCRs also emerge as crucial players in growth and metastasis of many tumors (Dorsam and Gutkind, 2007). Mutations in GPCRs have been linked to many human genetic diseases including retinitis pigmentosa (RP), hypo- and hyperthyroidism, nephrogenic diabetes insipidus, bleeding disorder and even carcinomas (Insel *et al.*, 2007; Schoneberg *et al.*, 2004).

GPCRs are activated by a heterogeneous set of endogenous ligands. These ligands may be light-sensitive compounds, odors, pheromones, hormones and neurotransmitters. Due to their accessibility on cellular membranes, their central role in cell communication, and the wide variety of potential functions, GPCRs are the targets for nearly one third of modern small molecule therapeutics (Hopkins and Groom, 2002).

GPCRs are transducers of extracellular stimuli to signal intracellular changes. In their inactive state, GPCRs are bound to a heterotrimeric G protein complex including three subunits, G_α, G_β and G_γ. The binding of an extracellular ligand will initiate a conformation change in GPCR, which then activates the bound G_α subunit. The activated G_α subunit dissociates from the complex by an exchange from GTP to GDP. The dissociated G_α and $G_{\beta\gamma}$ subunits binds to other intracellular proteins to trigger downstream signaling or metabolic pathways separately (Digby *et al.*, 2006). Meanwhile, G-protein-coupled receptor kinases (GRKs) are recruited to the ligand-bound GPCR and phosphorylate the receptor. Once phosphorylated, GPCR binds to β-arrestins, preventing further coupling of GPCR to G proteins. β-arrestins have a dual role, however, and can also interact with intracellular proteins inducing downstream effects (Metaye *et al.*, 2005). In the whole process, the signal of ligand is passed from GPCR to two types of transducer molecules: G protein and β-arrestin, contributing to G protein-dependent signaling and G protein-independent signaling, respectively.

GPCRs are usually expressed at low levels, with 1% of genes in the genome only accounting for 0.001–0.01% of expressed sequence tags (Fredriksson and Schiöth, 2005). What's more, expression levels of GPCRs vary dramatically by tissue. Regard *et al.* analyzed transcript levels of 353 GPCRs in 41 adult mouse tissues and found that GPCRs that are highly expressed in a given tissue usually exhibit important function to that tissue. For example, light-detecting opsins are highly and specifically expressed in eye, and dopamine, gamma-aminobutyric acid (GABA) receptors are highly expressed in central nervous system (Regard *et al.*, 2008). These findings suggest that GPCR signaling system consists of different expressed proteins depending on the tissue type and target different downstream pathways to fulfill tissue-specific functions.

Driven largely by their utility in drug development, the study of individual GPCR function and their downstream target pathways has become a research topic of great importance. With this information, many new avenues of research are available. For example, the action of pathways without any direct drug targets can be modulated through targeting the GPCRs that trigger them. In addition, the downstream pathways in the target tissue of a drug can be connected to cellular mechanism or on-target side effects of the drug while the downstream pathways in other tissues can be connected to the off-target side effects. However, the low expression level and difficulty in crystallization have made comprehensive analyses so challenging that most GPCRs still remain the '*terra incognita*', or unexplored territory, in functional genomics (Katritch *et al.*, 2012; Regard *et al.*, 2008; Tobin *et al.*, 2008). Importantly, GPCRs do not interact with intracellular proteins directly, but through the coupling of G protein or recruitment of β-arrestin. Studying the direct partners of GPCRs will not reveal their downstream effects. Research has shown that transducer isotypes can have distinct interacting partners and signaling roles. For example, $G_{\alpha s}$ subunit can activate cAMP-dependent pathway by stimulating the production of cAMP while $G_{\alpha i}$ inhibits the production (Birnbaumer, 2007). β-arrestin2 mediates dopaminergic synaptic transmission while β-arrestin3 actives ERK-1/2 pathway (Beaulieu *et al.*, 2005; Oakley *et al.*, 2000). Exactly which β-arrestins and G proteins associate with each GPCR and if this varies by tissue type remains unknown, further complicating the systematic identification of GPCR-targeted pathways.

We introduced a data-driven method GOTE, to systematically predict GPCR Downstream Pathways Signaling by Tissue Expression (Fig. 1). With many genome-wide expression profiling datasets of different human tissues or cell types available, we were able to identify the expressed proteins of GPCR signaling system in each tissue type, these include the GPCR proteins themselves as well as the transducer proteins (G-proteins and β-arrestins). We hypothesized that if the proteins of a pathway significantly interact with the transducer isotypes in a particular tissue, then the pathway is more likely to be the signaling outcome of the GPCRs that are highly expressed in that tissue. Based on this hypothesis, we designed a statistical test to evaluate the association between pathways and GPCRs in each tissue and applied it to five independent datasets (four from normal tissues, one from cancer cell lines). We tested the robustness of GOTE by comparing the similarity of results across the normal human tissue datasets. We found both concordance and discordance in the predicted pathways. The predicted pathways of some GPCRs such as *CASR, TBXA2R, CXCR4* and *GPR56* can be connected to their molecular function or cause of disease by our method. Due to the lack of true standard specifying the true relationship between GPCR and pathways, we used GPCR knockdown datasets to perform a weak validation, in which we observed whether the expression of a pathway is significantly changed after the knockdown of GPCR. We concluded with a discussion of the potential uses of these results, such as the study of complex human disease or molecular mechanism of drugs. All of our data and code are made publicly and freely available.

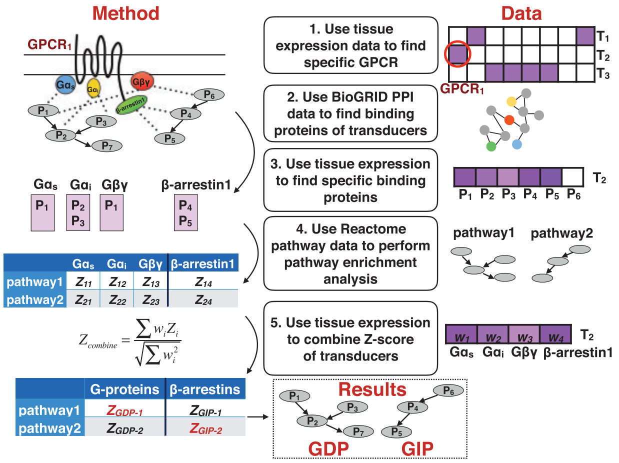

Fig. 1. Workflow and data used in GOTE. The left side shows the workflow of GOTE. The center gives a simple description of each step. The right side shows the data used in each step. In the first step we used tissue expression data to find specifically expressed GPCR in each tissue. Second, for each transducer (G-protein or β-arrestin), we obtained a list of binding proteins using the BioGRID PPI data. Third, the list of binding proteins is filtered by tissue expression, resulting in lists of tissue-specific binding proteins. Fourth, pathway enrichment analysis is performed based on the tissue-specific binding proteins of each transducer using Fisher's Exact Test. For each pathway, a Z-score is calculated for each transducer. Fifth, for each pathway, the Z-scores of all G-proteins are combined together using Stouffer's Method. The weight of each Z-score is in proportion to the tissue expression of each G-protein. The same analysis is repeated for the Z-scores of all β-arrestins. Finally, pathways with significant Z-scores are connected to GPCRs expressed in the same tissue as G-protein dependent pathways (those which are associated with G-proteins) or G-protein independent pathways (those which are associated with β-arrestins)

2 Materials and methods

2.1 Materials

2.1.1 Genes in GPCR signaling system

We downloaded 401 GPCRs and their class information from GtoPdb (http://www.guidetopharmacology.org/DATA/targets_and_families.csv) (Southan *et al.*, 2015), 53 G-proteins from gpDB (http://biophysics.biol.uoa.gr/gpDB/retrieve.jsp) (Theodoropoulou *et al.*, 2008). All GPCRs were classified into 6 main classes by GtoPdb: A, B, C, D (Adhesion), F (Frizzled) and O (Others). The gene symbol ID we used for 4 β-arrestins are SAG, ARRB1, ARRB2 and ARR3. The binding proteins of G-proteins and β-arrestins were identified through BioGRID Human PPI network release of 3.4 (Chatr-Aryamontri *et al.*, 2013).

2.1.2 Gene expression data

We chose five gene expression datasets deriving from three different platforms and covering both cancer cell line and normal human tissues (i) Human U133A Gene Atlas (referred as U133A in this paper) (Su *et al.*, 2004): a compendium of all transcript expression data in 84 human tissues running on the Affymetrix U133A microarray platform. This dataset was downloaded from BioGPS (Wu *et al.*, 2016). (ii) Human NCI60 Cell Lines (referred as NCI60 in this paper): a collection of all transcript expression data in 108 cancer cell lines running on the Affymetrix U133A microarray platform. This dataset was also downloaded from BioGPS (Wu *et al.*, 2016). (iii) HPM_RNA and (iv) HPM_PRT were both released from the Human Proteome Map (Kim *et al.*, 2014). They used Mass spectrometry to measure the levels of peptide sequences in 30 human tissues and mapped them to Human Refseq protein sequences (HPM_PRT) and corresponding Refseq genes to infer RNA level expression (HPM_RNA). (v) GTEx Analysis V6 RNA-seq data (referred as GTEx in this paper) (Melé *et al.*, 2015): a collection of transcriptome data in 53 human normal tissues running on the Illumina TrueSeq RNA sequencing platform (we used the 'Gene RPKM' file provided). We mapped all the array or gene IDs in five datasets to Uniprot protein ID (UniProt Consortium, 2014) and the average expression value was calculated if multiple arrays were mapped to a same Uniprot ID. Next, we converted all the expression value x to $\log_2(x+2)$ to adjust for 0 and extremely large values, so that all the converted values are no less than 1. We then normalized the expression of each gene by the median of all tissues, and took the log conversion again so that the final value will follow an approximate normal distribution across all the genes. Then, we calculated a Z-score based on the normal distribution to represent the level of differential expression of a gene in a tissue. Z-score was converted to P-value with the 'pnorm' function in R.

2.1.3 Reactome pathways

We used Reactome pathways as data source for pathways (Croft *et al.*, 2014). The Uniprot to pathway mapping file was downloaded from http://www.reactome.org/download/current/UniProt2Reactome.txt. In this paper, we considered 2223 pathways with size between 5 and 500.

2.2 Methods

2.2.1 Predicting downstream pathways of GPCRs

There are four parameters in GOTE: P-value threshold of GPCR t_1 and a threshold for GPCR specificity score t_2, P-value threshold of binding protein t_3, P-value threshold of enriched pathways t_4. We set t_1, t_3 and t_4, as 0.05 and t_2 as 0 in this paper.

First, we found the GPCRs that are specifically expressed in each tissue. To quantify the correlation between a GPCR and all the tissues in one dataset, we calculated a GPCR specificity score S_{gpcr} for a tissue defined as following:

$$T_{\text{gpcr}} = \begin{cases} 0(p_{\text{gpcr}} \geq t_1) \\ 1(p_{\text{gpcr}} < t_1) \end{cases} \quad S_{\text{gpcr}} = \frac{T_{\text{gpcr}}}{\sum_{\text{All Tissues}} T_{\text{gpcr}}}$$

where p_{gpcr} is the expression P-value of GPCR in that tissue. And those with S_{gpcr} greater than t_2 were considered as GPCRs specifically expressed in that tissue.

Next, we connected downstream pathways to specific GPCRs in that tissue through the binding proteins of transducers (G-protein or β-arrestin). For each transducer, we found its binding protein through BioGRID PPI network and those with expression P-value less than t_3 were selected as highly expressed binding proteins. We then assessed the correlation between binding proteins and each Reactome pathway by one side Fisher's Exact Test (Fisher, 1922). Then, we combined the Z-scores from every transducer by Stouffer's Z-score method to obtain a final Z-score for each pathway (Stouffer, 1949).

$$Z_{\text{combine}} = \frac{\sum w_i Z_i}{\sqrt{\sum w_i^2}}$$

Here, the Z-score of every transducer were combined with a weight w_i, referring to the expression of tranducer i. Thus, a transducer with high expression in the tissue will have more influence on the downstream pathway. Eventually, we used 'pnorm' function in R to transform Z-score into P-value. The pathways with P-value less than t_4 were considered as our predicted downstream pathways for those specific GPCR mentioned above. The pathways were classified into G-protein dependent pathways (deriving from the binding proteins of G-protein) and G-protein independent pathways (deriving from the binding proteins of β-arrestin).

2.2.2 Expression dataset from L1000 gene knockdown experiment

We downloaded the expression dataset of 448 737 gene knockdown experiments from lincscloud.org with perturbation type of 'trt_sh' (Duan *et al.*, 2014). In each experiment, a gene was knocked down in a particular cancer cell line and the expression of all the genes was measured before and after the knockdown. In total, 11 617 experiments were found to knock down a GPCR gene. Altogether, 145 unique GPCRs and 16 cell lines from NCI60 were included in L1000 gene knockdown dataset. Four levels of data are provided by lincscloud.org: raw, unprocessed flow cytometry data (level 1), Gene expression values (level 2), normalized expression value (level 3) and signatures with differentially expressed genes computed by robust z-scores for each profile relative to population control (level 4). We used level 4 data, Z-score of each gene representing the level of expression change after the gene knockdown. We then preprocessed the dataset by mapping array IDs to Uniprot IDs. The average Z-score was calculated if multiple array IDs were mapped to a same Uniprot ID.

2.2.3 Defining a weak reference standard for GPCR target pathways

Since the absolute value of Z-score represents the level of gene expression change after the knockdown, we quantified the expression change of a pathway by using Stouffer's method to combining all the Z-scores (absolute value) of genes in the pathway.

$$Z_{\text{pathway}} = \frac{\sum_{i=1}^{N} |Z_i|}{\sqrt{N}}$$

N represents the number of genes in a pathway. A pathway with Z_{pathway} greater than 4.08 (correspond to P-value < 2.2e−05, corrected for multiple hypothesis testing) is considered to experience significant change after the knockdown of a GPCR in a cell line. Such pathways are defined in our weak standard as a 'positive' and the rest, with no significant change, are defined as 'negative'. If GOTE identified a number of pathways as 'positive' for a GPCR g in a cell line c, we obtained the reference standard of 'positive' and 'negative' from the experiment with g knocked down in c. Then we calculated precision (TP/TP + FP), recall (TP/TP + FN) and specificity (TN/TN + FP). The performance of GOTE was compared to the following two methods: (i) HighExp: we created this method as a comparison to GOTE. GOTE connects the tissue-specific GPCRs to the pathways that are enriched by the tissue-specific binding proteins of transducers, while HighExp connects the tissue-specific GPCRs to all the tissue-specific pathways without considering the information from transducers. In each tissue, we found all the proteins with expression *P*-value less than t_3 and performed pathway enrichment analysis with Fisher's Exact Test based on these proteins. The pathways with *P*-value less than t_4 are defined as tissue-specific pathways. The pathways predicted by GOTE are usually a subset of the pathways predicted by HighExp. (ii) Random: we created this method to represent null distribution. In this method, we randomly assigned pathways to each GPCR without considering any other information.

3 Results

3.1 GPCR signaling system has different expressed proteins in distinct tissues

We found that the GPCR signaling system is highly tissue-specific across all five datasets. Each GPCR is highly expressed in 3–5% of all tissues in five datasets (Supplementary Figs S1–S5). We found that on average 7.84 ± 1.11 GPCRs in U133A, 8.83 ± 0.60 GPCRs in NCI60, 12.23 ± 2.00 GPCRs in HPM_RNA, 10.40 ± 2.00 GPCRs in HPM_PRT and 27.56 ± 3.00 GPCRs in GTex are highly expressed in each tissue when using *P*-value of 0.05 as a threshold for high expression (Supplementary Fig. S5A) and the remainder exhibit no or very low expression in both RNA (Supplementary Figs S1–S3 and S5) and protein level (Supplementary Fig. S4). We found tissue specificity for 57 signaling transducers, 53 G-proteins and 4 β-arrestins, as well (Supplementary Fig. S7–S11). On average, only 1–2 transducers are highly expressed in each tissue (Supplementary Fig. S6A) and each transducer is highly expressed in 3–5% of all the tissues (Supplementary Fig. S6B).

We then looked at the binding proteins of transducers. About 31 of 53 G-proteins and 4 of 4 beta-arrestins have protein–protein interaction data available from BioGRID. Among 31 G-proteins, 15 belong to α subunit, 5 belong to β subunit and 11 belong to γ subunit (Supplementary Fig. S12). The binding proteins of different transducer are enriched with different Reactome pathways (Supplementary Tables S1 and S2). For example, *GNA12* mediates the cell-cell junction and adherence ($P\text{-value} < 10^{-3}$, Supplementary Table S1) while *GNAI1* initiates the adenylate cyclase inhibitory pathway ($P\text{-value} < 10^{-4}$, Supplementary Table S1). For β-arrestins, isotype 3 is significantly associated with protein translation while isotype 4 interacts with NOTCH1 signaling process ($P\text{-value} < 10^{-2}$, Supplementary Table S2). We found that binding proteins were also tissue-specific. On average, only 3–5% binding proteins are highly expressed in each tissue and each binding protein is highly expressed in 3–5% of all the tissues (Supplementary Figs S13–S17).

3.2 Using tissue-specific expression to predict downstream pathways of GPCRs

A workflow of GOTE is shown in Figure 1 (for details, refer to Section 2). As a data-driven method, the goal of GOTE is to use expression data to find specific GPCRs and pathways in each tissue and connect them together. The specific pathways were found through enrichment analysis of proteins binding to each transducer (G-proteins and β-arrestins) of GPCR with more highly expressed transducers having more influence on the recruitment of downstream pathways.

3.3 Concordance and discordance of findings in five datasets

We used GOTE to predict downstream pathways of GPCRs in five expression datasets described in the section 2.1.2. General statistics of results were summarized in Table 1. Of all five datasets, GTEx has results for the largest number of GPCRs in both G-protein dependent pathways and G-protein independent pathways (237 and 215). With the exception of U133A dataset, we found more G-protein dependent pathways than G-protein independent pathways. On average, each GPCR were connected to pathways in 1–9 tissues. The standard deviations of all these statistics are large, suggesting the number of pathways and tissues varies by GPCR. We also compared the size of predicted pathways in five datasets (Supplementary Fig. S18). The median size of predicted pathways is relatively consistent among five datasets (distributed between 60and 80) and with no significant difference from the size distribution of all pathways from Reactome ($P\text{-value} > 0.05$, *t*-test).

Next, we compared the G-protein dependent and independent pathways predicted in each tissue. The results of HPM_PRT dataset is shown in Table 2 and the other datasets are shown in Supplementary Table S4. In HPM_PRT, we connected 277 G-protein dependent pathways and 167 G-protein independent pathways to 119 GPCRs. For simplicity, we classified all 30 tissues in HPM_PRT dataset into eight systems (Supplementary Table S3). All the systems except Immune have more G-protein dependent pathways than G-protein independent pathways. Reproductive system has the most G-protein dependent pathways while Immune system has the most G-protein independent pathways. The overlap between G-protein dependent and independent pathways in each system is not

Table 1. Summary of results in all five datasets

	U133A	NCI60	HPM_RNA	HPM_PRT	GTEx
General statistics of results for G protein dependent pathway (#:number)					
# GPCRs	118	112	129	119	237
# Pathways	293	305	217	277	238
# Tissues	66	98	26	27	50
# Connections	12484	15078	6957	5834	32665
number of G protein independent pathways per GPCR					
average	46.78	75.75	43.98	41.97	76.91
SD	42.61	65.68	30.9	40.3	58.96
Maximum	174	246	161	221	218
Minimum	1	2	1	1	2
number of tissues per GPCR					
average	4.92	7.79	2.43	2.36	5.97
SD	5.16	9.27	1.89	1.88	4.16
Maximum	24	40	9	9	18
Minimum	1	1	1	1	1
General statistics of results for G protein independent pathway					
# GPCRs	102	102	112	90	215
# Pathways	443	303	159	167	185
# Tissues	50	55	15	15	32
# Connections	32706	13149	4061	2743	29311
number of G protein independent pathways per GPCR					
average	124.18	78.91	30.09	27.31	62.17
SD	81.38	62.73	26.89	28.75	30.46
Maximum	356	258	104	110	135
Minimum	3	2	2	2	1
number of tissues per GPCR					
average	4.64	4.62	1.75	1.74	4.89
SD	4.81	5.12	1.09	1.08	3.54
Maximum	21	23	5	6	16
Minimum	1	1	1	1	1

significant by Fisher's Exact Test (*P*-value > 0.5). This is consistent in all five datasets (Supplementary Table S4). However, the G-protein dependent and independent pathways predicted across all systems are significantly overlapped with each other (*P*-value = 6.96e−06), suggesting the overlap happens between different systems. This is also consistent in all five datasets (Supplementary Table S4).

Many pathways were repeatedly predicted in more than one tissue (Supplementary Table S5). For G-protein independent pathways, the common pathways are consistently associated translation and cell cycle process across five datasets. For example, 'Eukaryotic Translation Elongation' is one of the top 10 most common pathways in all five datasets. In addition, many G-protein independent pathways in the cancer cell lines dataset-NCI60 are related to apoptosis, consistent with other studies (Ahn *et al.*, 2009; Revankar *et al.*, 2004). By comparison, the common G-protein dependent pathways represent a variety of biological process and vary by datasets. For example, neural pathways such as 'Transmission across Chemical Synapses', 'GABA receptor activation' are commonly predicted in HPM and GTEx dataset. Pathways about other signal transduction system such as 'Signaling by Hippo', 'Activated TLR4 signaling' and 'Nuclear events mediated by MAP kinases' are commonly predicted in U133A and NCI60 dataset. In addition, we also found G-protein dependent pathways participate in the regulation of ion channel through pathways such as 'Activation of G protein gated Potassium channels' and 'Inhibition of voltage gated Ca^{2+} channels via Gbeta/gamma subunits' in GTEx dataset.

Table 2. Summary of results for each system in HPM_PRT dataset

System	#GPCR	#GDP	#GIP	#Overlap
Endocrine	7	17	0	0
Digestive	51	71	9	1
Nervous	44	49	46	7
Cardiovascular	29	74	39	5
Respiratory	8	12	11	0
Reproductive	29	152	52	22
Urinary	15	19	0	0
Immune	25	60	86	15
Average	26	56.75	30.38	6.25
Total	119	277	167	90

#: number. GDP: G-protein dependent pathways. GIP: G-protein independent pathways. Overlap: between GDP and GIP.

3.4 Class C and D GPCRs are commonly connected to pathways in nervous system

We used bar plot to show the number of G-protein dependent (Fig. 2A) and independent (Fig. 2C) pathways predicted for each GPCR in HPM_PRT (The other four datasets are shown in Supplementary Fig. S19–S22). We chose HPM_PRT because the tissues in the dataset are diverse and cover eight various human anatomy systems. All the GPCRs were grouped by their class annotation from GtoPdb. Class A, B and F GPCRs have downstream pathways in various systems while class C and D show preference on particular systems. For example, class C GPCRs, including calcium-sensing,

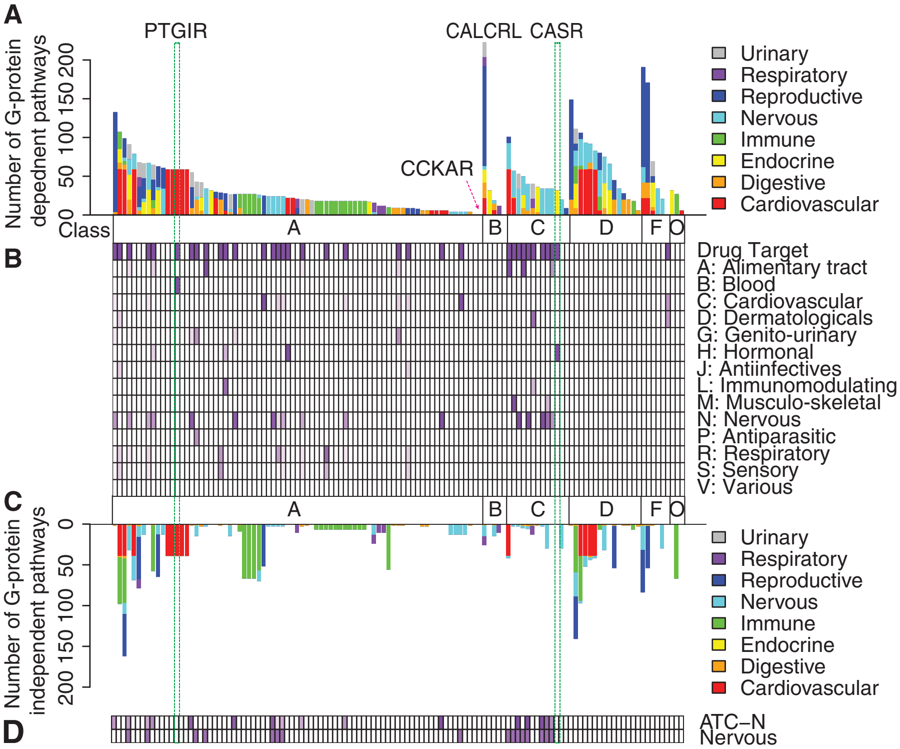

Fig. 2. Results of HPM_PRT dataset. (A,C) Bar plot showing the number of predicted G-protein dependent (A) and independent (C) pathways, respectively. The *x*-axis indicates different GPCRs, which are grouped by six families. Each bar with a unique color indicates the number of predicted pathways in a tissue type. (B) A heatmap shows a mapping between GPCR and the Anatomical Therapeutic Chemical (ATC) Classification System categories of the drugs known to bind them. The first row indicates whether or not the GPCR is a known drug target. Each cell colored from white to deep purple indicates the percentage of drugs that target the GPCRs that are classified into each category. (D) Correlation between ATC classification and the results of GOTE. The purple cell of top row shows the GPCR targeted by drugs belonging to the Nervous category of ATC classification. The purple cell of bottom row shows the GPCRs which are connected to nervous system by GOTE ($r = 0.543$, *P*-value = 1.76e−10

GABA and glutamate receptors are commonly connected to G-protein dependent pathways in nervous system. Class D (Adhesion) GPCRs are commonly connected to G-protein dependent pathways in nervous and cardiovascular system. Class D is also commonly connected to G-protein independent pathways in the immune system.

The number of predicted pathways varies greatly by GPCR. For example, *CALCRL*, a calcitonin-gene-related peptide receptor has 223 predicted G-protein dependent pathways from a variety of systems, the most among all GPCRs (Fig. 2A) while GPCRs such as *CCKAR*, a Cholecystokinin receptor that Mediates smooth muscle contraction of the gall bladder and stomach only have one predicted G-protein dependent pathway in the digestive system. G-protein dependent pathways appear in all eight systems while G-protein independent pathways mainly appear in immune, reproductive and cardiovascular system (Fig. 2A and C).

3.5 GPCR—pathway associations recapture known pharmacology

ATC is a pharmacological classification system based on the organ or system of action. Since GPCR are the targets of nearly one third of modern drugs, we compared the tissues and downstream pathways of GPCRs predicted by GOTE with the ATC classification of the drugs that target them (Supplementary Methods). For each GPCR, we calculated the proportion of drugs that belong to each ATC category (Fig. 2B).

Most drugs in ATC system target class A and C GPCRs. Many overlaps can be found when comparing the systems of a drug given by ATC and the systems connected to the GPCR by GOTE. For example, all three drugs targeting *PTGIR* act in 'Blood or blood forming organs (B)' according to ATC system, while in our results, *PTGIR* is connected platelets (cardiovascular system, Fig. 2B). In platelets, *PTGIR* is connected to 39 downstream pathways in such as 'Platelet activation, signaling and aggregation' (*P*-value = 0.0016, ranking No. 3 in G-protein independent pathways, Supplementary Table S6).

Extracellular calcium-sensing receptor (CASR) is the target of cinacalcet, a drug that acts as a calcimimetic to lower calcium level in blood (Shoback *et al.*, 2003). This drug is classified into 'systemic hormonal preparations (H)' by ATC and its target CASR is connected to endocrine system by GOTE since it is uniquely expressed in pancreas (the pancreas belongs to both endocrine and digestive system, we classified it into endocrine system). Thirty-two downstream pathways are connected to *CASR* in pancreas by GOTE. The most significant one is 'Depolarization of the Presynaptic Terminal Triggers the Opening of Calcium Channels' (*P*-value = 1.50e−06, ranking No.1 in G-protein dependent pathways, Supplementary Table S6).

Many GPCRs are the targets of drugs belonging to the 'Nervous (N)' class according to ATC. They are significantly overlapped with the GPCRs connected to the nervous system by GOTE (Fig. 2D). Pearson correlation coefficient between them is 0.543 (*P*-value = 1.76e−10). The Pearson correlation coefficient between GPCRs connected to digestive system by GOTE and GPCRs with drug classified into ATC-Alimentary tract (A) is 0.435 (*P*-value = 7.60e−07). The Pearson correlation coefficient between GPCRs connected to urinary, reproductive system by GOTE and GPCRs with drug classified into ATC Genito-urinary(G) is 0.424 (P-value = 1.57e-06).

3.6 The similarity of results depends on the similarity of cell types

We expect GOTE to predict similar pathways in similar tissues or cell types. To evaluate this hypothesis, we calculated the pairwise Jaccard similarity of predicted pathways between the same type of cell lines in NCI60. We paired together cell lines of the same and different cell types. The Jaccard similarity is shown as barplot in Figure 3A. On average, the G-protein dependent pathways predicted among similar cell types have a Jaccard similarity of 0.141 ± 0.028, higher than those predicted among different cell types 0.067 ± 0.006 (one side *t*-test *P*-value = 5.26e−06). The results are similar for G-protein independent pathways (0.141 ± 0.049 versus 0.063 ± 0.005, one side *t*-test *P*-value = 1.83e−03).

We also calculated pairwise Jaccard similarity of predicted pathways for all tissues in U133A dataset. The result is shown as heatmap in Figure 3B (G-protein independent pathways) and Supplementary Figure S23 (G-protein dependent pathways). Two clusters of tissues can be observed. Tissues in the same cluster have high Jaccard similarity to each other and belong to the same system, either neural (green cluster) or immune (red cluster).

3.7 Distinct datasets show consistency in predicted pathways

We calculated the pairwise Jaccard similarity of predicted pathways between the same tissue type in different datasets to test the robustness of GOTE against systematic error deriving from different platforms. The results were compared with a null distribution (Random) deriving from randomly assigning pathways to tissues (Supplementary Methods). The Jaccard similarity was shown as barplot in Figure 3C.

The comparison was conducted pairwise between four datasets containing normal tissues: U133A, HPM_RNA, HPM_PRT and GTEx. The results between GTEx and two HPM datasets were not shown because few overlapped tissues exist between them. Of all four datasets, HPM_RNA and HPM_PRT have the highest Jaccard similarity in predicted pathways. The average is 0.79 (95% CI: 0.66–0.92, *P*-value = 6.34e−11 compared to Random) for G-protein dependent pathways and 0.70 (95% CI: 0.50–0.87, *P*-value = 1.1e−05) for G-protein independent pathways. U133A and GTEx have high Jaccard similarity in predicted G-protein independent pathways. The average is 0.15 (95% CI: 0.10–0.20, *P*-value = 0.001 compared to Random). The similarity between U133A and two HPM datasets was not significant.

3.8 Predicted pathways go through expression change after GPCR knockdown experiment

We validated our GPCR-tissue-pathway predictions using an independent dataset from the L1000 experiment which measured the expression change of all genes after a GPCR was knocked down (Duan *et al.*, 2014). Using these expression data, we defined a weak reference standard that connects GPCRs to pathways by identifying significantly differentially expressed pathways in the knockdown experiment. For comparison, the performance of GOTE was compared to two other methods we created: HighExp and Random. HighExp looks at any pathway that is highly expressed in the tissue (not just the subset of pathways predicted for the GPCR by GOTE). Random represents the null distribution (Section 2.2.3).

We focused on the results of GOTE in NCI60 dataset since the cell lines used overlap with those used in the L1000 dataset. The precision, recall and specificity of results were calculated for three methods (Table 3). For G-protein dependent pathways, 92 connections between of GPCR and tissue can be evaluated (11% of all results). On average, the precision of GOTE is 0.22 ± 0.05, significantly outperforms two other methods (HighExp: 0.13 ± 0.03, Random: 0.08 ± 0.03). The recall of GOTE is

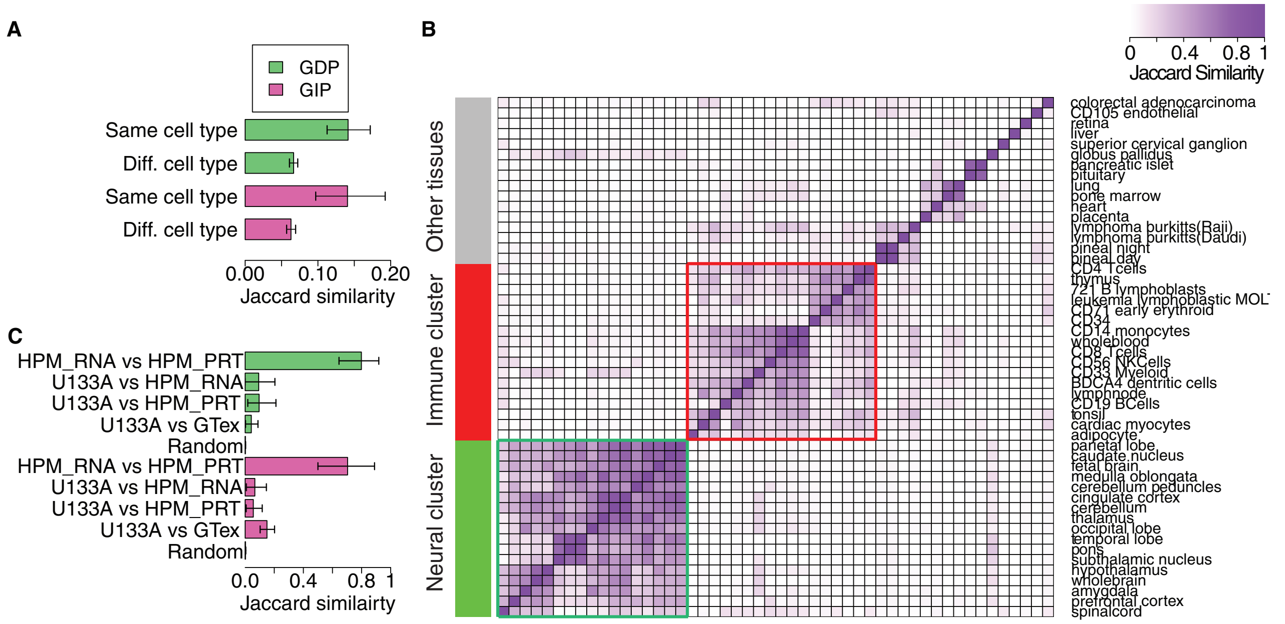

Fig. 3. Barplot of Jaccard similarity. GDP: G-protein dependent pathways. GIP: G-protein independent pathways. (A) Comparison of GOTE's results between different cell lines in NCI60 dataset. Each bar indicates the mean pairwise Jaccard similarity of cell lines belonging to same or different cell type. The error bar indicates 95% confidence interval of mean value calculated by bootstrap. For both G-protein dependent pathways (green) and G-protein independent pathways (pink), the Jaccard similarity between same cell types is significantly higher than Jaccard similarity between different cell types. (B) A heatmap showing the pairwise jaccard similariy of G-protein independent pathways among tissues in U133A dataset. Each column or row indicates a tissue. The color of each cell is proportion to the Jaccard similarity between the column and row. Two clusters of tissues are highlighted in the heatmap: green cluster of neural tissues at the bottom and red cluster of immune tissues in the middle. (C) Comparison of GOTE's results among four different datasets: U133A, HPM_RNA, HPM_PRT and GTEx. Each bar indicates the mean pairwise jaccard similarity of same tissue from two datasets. The error bar indicates 95% confidence interval of mean value calculated by bootstrap

Table 3. Validation of GOTE using L1000 GPCR knockdown data

	# positive pathways (95% CI)	Precision (95% CI)	Recall (95% CI)	Specificity (95% CI)
G-protein dependent pathways				
GOTE	173 (110–246)	0.22 (0.17–0.28)	0.03 (0.02–0.04)	0.994 (0.993–0.996)
HighExp		0.13 (0.1–0.17)	0.51 (0.45–0.57)	0.851 (0.848–0.855)
Random		0.08 (0.05–0.12)	0.007 (0.005–0.008)	0.994 (0.993–0.995)
G-protein independent pathways				
GOTE	223 (110–349)	0.28 (0.2–0.36)	0.10 (0.06–0.14)	0.989 (0.987–0.991)
HighExp		0.16 (0.1–0.22)	0.45 (0.37–0.52)	0.853 (0.85–0.857)
Random		0.1 (0.05–0.15)	0.015 (0.012–0.018)	0.986 (0.983–0.989)

CI: confidence interval.

The performance of three methods is compared: GOTE, HighExp and Random. Each column indicates the mean of precision: TP/(TP + FP), recall: TP/(TP + FN) and specificity: TN/(TN + FP). The number in bracket indicates 95% confidence interval of mean value calculated by bootstrap.

0.03 ± 0.01, outperforms Random (0.007 ± 0.001) but lower than HighExp (0.51 ± 0.06). The specificity of GOTE is 0.994 ± 0.001, outperforms HighExp (0.851 ± 0.003). For G-protein independent pathways, 38 connections between of GPCR and tissue can be evaluated (8% of all). The results have the same tendency as G-protein dependent pathways. On average, the precision, recall and specificity of GOTE is 0.28 ± 0.08, 0.10 ± 0.04 and 0.989 ± 0.002.

3.9 Increasing the cutoff of GPCR specificity contributes to high-confident results

We studied the influence of four parameters on the results of GOTE. These four parameters are as follows: (1) The *P*-value threshold for highly expressed GPCRs t_1 is to control the number of highly expressed GPCR in each tissue; (2) The GPCR specificity score threshold t_2 is to control the connections between GPCR and tissue. (3) The *P*-value threshold for highly expressed binding proteins of transducers t_3 is to control the number of binding proteins used for enrichment analysis; (4) The *P*-value threshold for enriched pathways t_4 is to control the number of enriched pathways. In a default setting, we use 0.05, 0, 0.05 and 0.05 for the four parameters to generate results. Here, the range of each parameter was expanded to a wider range to test influence on a benchmark statistics representing how much GOTE outperforms HighExp in precision.

The line graphs in Figure 4 show the benchmark statistics against each of four parameters. The P-value threshold of GPCR, binding protein and pathway do not have a strong influence on the accuracy of results (Fig. 4A, C and D). In contrast, different threshold of

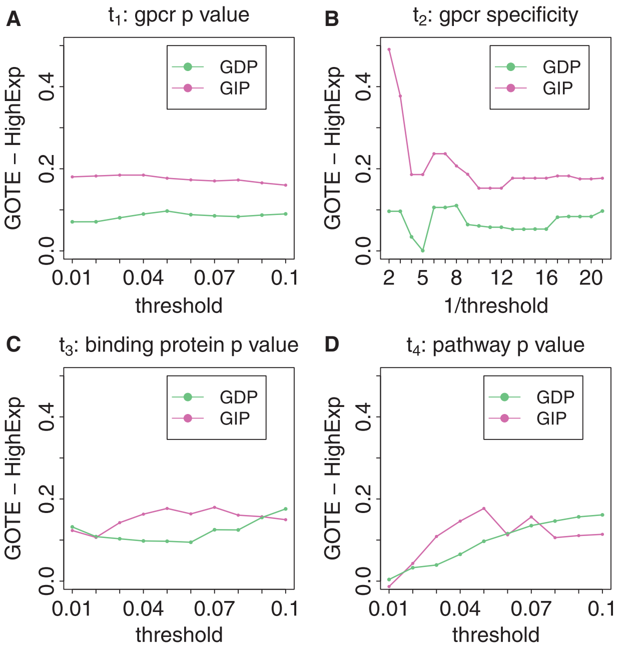

Fig. 4. Influence of four parameters on the prediction results of GOTE. GDP: G-protein dependent pathways. GIP: G-protein independent pathways. In each of four line graphs, the *x*-axis indicates the threshold (or 1/threshold in B) of one parameter. The *y*-axis indicates how much GOTE outperforms HighExp in precision when setting the parameter to the threshold of *x*-axis (the other three parameters as default setting). The four parameters are: (A) t_1: the *P*-value threshold for highly expressed GPCRs. (B) t_2: the GPCR specificity score threshold. (C) t_3: the *P*-value threshold for highly expressed binding proteins of transducers. (D) t_4: the *P*-value threshold for enriched pathways

GPCR specificity can lead to a dramatic change in the benchmark statistics (Fig. 4B). As the threshold for GPCR specificity score approaches to 1, GOTE significantly outperforms HighExp in precision.

4 Discussion

In this paper, we presented a data-driven method, GOTE, to systematically identify the downstream pathways of GPCRs in each tissue. We used gene expression datasets to find highly expressed binding proteins of G-proteins or β-arrestins in each tissue and performed pathway enrichment analysis to find enriched pathways. Then we connected the enriched pathways to specifically expressed GPCRs in the tissue as their downstream pathways. We found that the predicted pathways differ by tissue and GPCR while similar tissues have similar predicted pathways. With no true standard available, we created a 'weak' standard on our own with an independent dataset from L1000 gene knockdown experiment, to validate our GPCR-tissue-pathway predictions. In the validation, we tested the correlation between the knockdown of GPCR and the change of expression in downstream pathways. As co-expression is only a prerequisite of actual connection, this standard is less stringent and will unavoidably bring in some false positives. We controlled for this error by comparing the results to a null distribution Random and another reference method HighExp that connects highly expressed pathways to GPCRs in each tissue. GOTE outperforms both two methods in precision and specificity. GOTE has lower recall than HighExp because it chooses the tissue-specific pathways that are also enriched by the binding proteins of transducers. Thus GOTE is more stringent and is expected to have fewer predictions than HighExp.

As a data-driven method, our prediction results are dependent on the expression dataset used. Unfortunately, most current expression data are derived from different platforms. We tested the robustness of GOTE among four datasets with normal human tissue, each using a different technology. The similarity of predicted pathways between datasets is in alignment with the expression similarity (Supplementary Fig. S24), where the results are highly consistent between two mass spectrometry datasets and less consistent between these two mass spectrometry dataset and the microarray dataset. In addition, the five datasets do not employ a standard tissue naming scheme. This may contribute to a situation where a same name of tissue can refer to different types of cells in the tissue, which adds the difficulty of comparing predicted pathways between different datasets. Another limitation of GOTE is our current strategy of mapping GPCRs to their transducers. The relationship between GPCRs and transducers has not been well studied. Therefore, we used co-expression in the same tissue to connect GPCRs to transducers in GOTE. Consequently, this may bring in some false positive pairs of GPCRs and transducers since the correlation in expression does not necessarily mean the existence of actual connection between GPCRs and transducers. To mitigate these potential false positives, we provided a parameter in GOTE that controls GPCR tissue specificity. We showed that by setting a high threshold of this parameter, users can get more confident connections between GPCR and downstream pathways at the cost of fewer predictions.

We found our findings on many GPCRs agree with their molecular function, cause of disease, or the action of corresponding drug provided by other resources. For example, we found class C and D GPCRs are commonly connected to G-protein dependent pathways in nervous system. While class C GPCRs consists of calcium-sensing, GABA and glutamate receptors, which are synaptic receptors located primarily on neuronal cells, and class D GPCRs are reported to play an important role in brain development (Yona *et al.*, 2008). We also found the tissues connected to GPCR in our results correlate with the ATC classification of corresponding drug. GPCRs such as *CASR*, *PTGIR* are predicted with pathways in agreement with the molecular mechanism of their corresponding drugs (Supplementary Table S6). Some GPCRs have predicted pathways validated by other studies. For example, *CXCR4*, one of the co-receptors of HIV-1 in immunological cells participate in the activation of ion channels after the binding of gp120 (Lee *et al.*, 2003). We found this GPCR is connected to both HIV-related pathways (such as 'HIV Transcription Initiation') and ion channel-related pathways (such as 'Voltage gated Potassium channels') in CD8 Cells (Supplementary Table S7). Another GPCR *GPR56* is regulated by Blimp-1 in NK cells (Chang *et al.*, 2016), which is a transcription factor and modulates the MHC Class I antigen-processing and peptide-loading pathway (Doody *et al.*, 2007). In our results, this GPCR is connected to the 'Class I MHC mediated antigen processing & presentation' pathway in NK cells (Supplementary Table S8). Eventually, since GPCRs participate in controlling an extraordinary variety of physiological functions, mutations in GPCR can cause many complex diseases (Insel *et al.*, 2007; Schoneberg *et al.*, 2004). The pathways we predicted for these GPCRs may help us study the mechanism of these diseases. For example, mutations in *TBXA2R*, a thromboxane receptor (TP), will cause a bleeding disorder in human body (Hirata *et al.*, 1994). In our results, this GPCR is connected to blood clotting pathways such as 'Platelet activation, signaling and aggregation' (*P*-value = 1.62*e−03, ranking 3, Supplementary Table S9) in platelets. Meanwhile, we also found other interesting pathways such as

'Hyaluronan uptake and degradation' (*P*-value = 3.93*e−03, ranking 5, Supplementary Table S9), 'Hyaluronan metabolism' (*P*-value = 4.93*e−03, ranking 6, Supplementary Table S9). Hyaluronan plays an important role in skin wound healing events. Together with fibrin, it increases or stabilizes the volume and porosity of the clot and then serves as a physical support through which cells are trapped in the clot (Weigel *et al.*, 1986).

Funding

This work was supported by National Institutes of Health (R01GM107145) and the Herbert Irving Fellows award.

Conflict of Interest: none declared.

References

Ahn,S. *et al.* (2009) β-Arrestin-2 mediates anti-apoptotic signaling through regulation of BAD phosphorylation. *J. Biol. Chem.*, **284**, 8855–8865.

Beaulieu,J.M. *et al.* (2005) An Akt/β-arrestin 2/PP2A signaling complex mediates dopaminergic neurotransmission and behavior. *Cell*, **122**, 261–273.

Birnbaumer,L. (2007) Expansion of signal transduction by G proteins: the second 15 years or so: from 3 to 16 α subunits plus $\beta\gamma$ dimers. *Biochim. Biophys. Acta (BBA) Biomembr.*, **1768**, 772–793.

Chang,G.W. *et al.* (2016) The adhesion G protein-coupled receptor GPR56/ADGRG1 is an inhibitory receptor on human NK cells. *Cell Rep.*, **15**, 1757–1770.

Chatr-Aryamontri,A. *et al.* (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res.*, **41**, D816–D823.

Croft,D. *et al.* (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res.*, **42**, D472–D477.

Digby,G.J. *et al.* (2006) Some G protein heterotrimers physically dissociate in living cells. *Proc. Natl. Acad. Sci. USA*, **103**, 17789–17794.

Doody,G.M. *et al.* (2007) PRDM1/BLIMP-1 modulates IFN-γ-dependent control of the MHC class I antigen-processing and peptide-loading pathway. *J. Immunol.*, **179**, 7614–7623.

Dorsam,R.T. and Gutkind,J.S. (2007) G-protein-coupled receptors and cancer. *Nat. Rev. Cancer*, **7**, 79–94.

Duan,Q. *et al.* (2014) LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic Acids Res.*, **42**, W449–W460.

Fisher,R.A. (1922) On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. *J. R. Stat. Soc.*, **85**, 87–94.

Fredriksson,R. and Schiöth,H.B. (2005) The repertoire of G-protein–coupled receptors in fully sequenced genomes. *Mol. Pharmacol.*, **67**, 1414–1425.

Hirata,T. *et al.* (1994) Arg60 to Leu mutation of the human thromboxane A2 receptor in a dominantly inherited bleeding disorder. *J. Clin. Investig.*, **94**, 1662.

Hopkins,A.L. and Groom,C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.*, **1**, 727–730.

Insel,P.A. *et al.* (2007) Impact of GPCRs in clinical medicine: monogenic diseases, genetic variants and drug targets. *Biochim. Biophys. Acta*, **1768**, 994–1005.

Katritch,V. *et al.* (2012) Diversity and modularity of G protein-coupled receptor structures. *Trends Pharmacol. Sci.*, **33**, 17–27.

Kim,M.S. *et al.* (2014) A draft map of the human proteome. *Nature*, **509**, 575–581.

Lee,C. *et al.* (2003) Macrophage activation through CCR5-and CXCR4-mediated gp120-elicited signaling pathways. *J. Leukoc. Biol.*, **74**, 676–682.

Melé,M. *et al.* (2015) The human transcriptome across tissues and individuals. *Science*, **348**, 660–665.

Metaye,T. *et al.* (2005) Pathophysiological roles of G-protein-coupled receptor kinases. *Cell Signal*, **17**, 917–928.

Oakley,R.H. *et al.* (2000) Differential affinities of visual arrestin, βarrestin1, and βarrestin2 for G protein-coupled receptors delineate two major classes of receptors. *J. Biol. Chem.*, **275**, 17201–17210.

Regard,J.B. *et al.* (2008) Anatomical profiling of G protein-coupled receptor expression. *Cell*, **135**, 561–571.

Revankar,C.M. *et al.* (2004) Arrestins block G protein-coupled receptor-mediated apoptosis. *J. Biol. Chem.*, **279**, 24578–24584.

Schoneberg,T. *et al.* (2004) Mutant G-protein-coupled receptors as a cause of human diseases. *Pharmacol. Ther.*, **104**, 173–206.

Shoback,D.M. *et al.* (2003) The calcimimetic cinacalcet normalizes serum calcium in subjects with primary hyperparathyroidism. *J. Clin. Endocrinol. Metab.*, **88**, 5644–5649.

Southan,C. *et al.* (2015) The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. *Nucleic Acids Res.*, **44**, D1054–D1068.

Stouffer,S.A. (1949) *Adjustment during Army Life*. Princeton University Press, Princeton.

Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA*, **101**, 6062–6067.

Theodoropoulou,M.C. *et al.* (2008) gpDB: a database of GPCRs, G-proteins, effectors and their interactions. *Bioinformatics*, **24**, 1471–1472.

Tobin,A.B. *et al.* (2008) Location, location, location…site-specific GPCR phosphorylation offers a mechanism for cell-type-specific signalling. *Trends Pharmacol. Sci.*, **29**, 413–420.

UniProt Consortium,U. (2014) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.

Weigel,P.H. *et al.* (1986) A model for the role of hyaluronic acid and fibrin in the early events during the inflammatory response and wound healing. *J. Theor. Biol.*, **119**, 219–234.

Wu,C. *et al.* (2016) BioGPS: building your own mash-up of gene annotations and expression profiles. *Nucleic Acids Res.*, **44**, D313–D316.

Yona,S. *et al.* (2008) Adhesion-GPCRs: emerging roles for novel receptors. *Trends Biochem. Sci.*, **33**, 491–500.