[1] Boisset, J. C., et al. (2018). "Mapping the physical network of cellular interactions." Nat Methods.

A cell's function is influenced by the environment, or niche, in which it resides. Studies of niches usually require assumptions about the cell types present, which impedes the discovery of new cell types or interactions. Here we describe ProximID, an approach for building a cellular network based on physical cell interaction and single-cell mRNA sequencing, and show that it can be used to discover new preferential cellular interactions without prior knowledge of component cell types. ProximID found specific interactions between megakaryocytes and mature neutrophils and between plasma cells and myeloblasts and/or promyelocytes (precursors of neutrophils) in mouse bone marrow, and it identified a Tac1(+) enteroendocrine cell-Lgr5(+) stem cell interaction in small intestine crypts. This strategy can be used to discover new niches or preferential interactions in a variety of organs.

[2] Svensson, V., et al. (2018). "SpatialDE: identification of spatially variable genes." Nat Methods **15**(5): 343-346.

Technological advances have made it possible to measure spatially resolved gene expression at high throughput. However, methods to analyze these data are not established. Here we describe SpatialDE, a statistical test to identify genes with spatial patterns of expression variation from multiplexed imaging or spatial RNA-sequencing data. SpatialDE also implements 'automatic expression histology', a spatial gene-clustering approach that enables expression-based tissue histology.

[3] Kiselev, V. Y., et al. (2018). "scmap: projection of single-cell RNA-seq data across data sets." Nat Methods **15**(5): 359-362.

Single-cell RNA-seq (scRNA-seq) allows researchers to define cell types on the basis of unsupervised clustering of the transcriptome. However, differences in experimental methods and computational analyses make it challenging to compare data across experiments. Here we present scmap (<http://bioconductor.org/packages/scmap>; web version at <http://www.sanger.ac.uk/science/tools/scmap>), a method for projecting cells from an scRNA-seq data set onto cell types or individual cells from other experiments.

[4] Soneson, C. and M. D. Robinson (2018). "Bias, robustness and scalability in single-cell differential expression analysis." Nat Methods **15**(4): 255-261.

Many methods have been used to determine differential gene expression from single-cell RNA (scRNA)-seq data. We evaluated 36 approaches using experimental and synthetic data and found considerable differences in the number and characteristics of the genes that are called differentially expressed. Prefiltering of lowly expressed genes has important effects, particularly for some of the methods developed for bulk RNA-seq data analysis. However, we found that bulk RNA-seq analysis methods do not generally perform worse than those developed specifically for scRNA-seq. We also present conquer, a repository of consistently processed, analysis-ready public scRNA-seq data sets that is aimed at simplifying method evaluation and reanalysis of published results. Each data set provides abundance estimates for both genes and transcripts, as well as quality control and exploratory analysis reports.

[5] Alpert, A., et al. (2018). "Alignment of single-cell trajectories to compare cellular expression dynamics." Nat Methods **15**(4): 267-270.

Single-cell RNA sequencing and high-dimensional cytometry can be used to generate detailed trajectories of dynamic biological processes such as differentiation or development. Here we present cellAlign, a quantitative framework for comparing expression dynamics within and between single-cell trajectories. By applying cellAlign to mouse and human embryonic developmental trajectories, we systematically delineate differences in the temporal regulation of gene expression programs that would otherwise be masked.

[6] Sharma, A., et al. (2018). "Dissecting the sources of gene expression variation in a pan-cancer analysis identifies novel regulatory mutations." Nucleic Acids Res **46**(9): 4370-4381.

Although the catalog of cancer-associated mutations in protein-coding regions is nearly complete for all major cancer types, an assessment of regulatory changes in cancer genomes and their clinical significance remain largely preliminary. Adopting bottom-up approach, we quantify the effects of different sources of gene expression variation in a cohort of 3899 samples from 10 cancer types. We find that copy number alterations, epigenetic changes, transcription factors and microRNAs collectively explain, on average, only 31-38% and 18-26% expression variation for cancer-associated and other genes, respectively, and that among these factors copy number alteration has the highest effect. We show that the genes with systematic, large expression variation that could not be attributed to these factors are enriched for pathways related to cancer hallmarks. Integrating whole genome sequencing data and focusing on genes with systematic expression variation we identify novel, recurrent regulatory mutations affecting known cancer genes such as NKX2-1 and GRIN2D in multiple cancer types. Nonetheless, at a genome-wide scale proportions of gene expression variation attributed to recurrent point mutations appear to be modest so far, especially when compared to that attributed to copy number changes - a pattern different from that observed for other complex diseases and traits. We suspect that, owing to plasticity and redundancy in biological pathways, regulatory alterations show complex combinatorial patterns, modulating gene expression in cancer genomes at a finer scale.

[7] Kim, T., et al. (2018). "Octopus-toolkit: a workflow to automate mining of public epigenomic and transcriptomic next-generation sequencing data." Nucleic Acids Res **46**(9): e53.

Octopus-toolkit is a stand-alone application for retrieving and processing large sets of next-generation sequencing (NGS) data with a single step. Octopus-toolkit is an automated set-up-and-analysis pipeline utilizing the Aspera, SRA Toolkit, FastQC, Trimmomatic, HISAT2, STAR, Samtools, and HOMER applications. All the applications are installed on the user's computer when the program starts. Upon the installation, it can automatically retrieve original files of various epigenomic and transcriptomic data sets, including ChIP-seq, ATAC-seq, DNase-seq, MeDIP-seq, MNase-seq and RNA-seq, from the gene expression omnibus data repository. The downloaded files can then be sequentially processed to generate BAM and BigWig files, which are used for advanced analyses and visualization. Currently, it can process NGS data from popular model genomes such as, human (Homo sapiens), mouse (Mus musculus), dog (Canis lupus familiaris), plant (Arabidopsis thaliana), zebrafish (Danio rerio), fruit fly (Drosophila melanogaster), worm (Caenorhabditis elegans), and budding yeast (Saccharomyces cerevisiae) genomes. With the processed files from Octopus-toolkit, the meta-analysis of various data sets, motif searches for DNA-binding proteins, and the identification of differentially expressed genes and/or protein-binding sites can be easily conducted with few commands by users. Overall, Octopus-toolkit facilitates the systematic and integrative analysis of available epigenomic and transcriptomic NGS big data.

[8] Ellis, S. E., et al. (2018). "Improving the value of public RNA-seq expression data by phenotype prediction." Nucleic Acids Res **46**(9): e54.

Publicly available genomic data are a valuable resource for studying normal human variation and disease, but these data are often not well labeled or annotated. The lack of phenotype information for public genomic data severely limits their utility for addressing targeted biological questions. We develop an in silico phenotyping approach for predicting critical missing annotation directly from genomic measurements using well-annotated genomic and phenotypic data produced by consortia like TCGA and GTEx as training data. We apply in silico phenotyping to a set of 70 000 RNA-seq samples we recently processed on a common pipeline as part of the recount2 project. We use gene expression data to build and evaluate predictors for both biological phenotypes (sex, tissue, sample source) and experimental conditions (sequencing strategy). We demonstrate how these predictions can be used to study cross-sample properties of public genomic data, select genomic projects with specific characteristics, and perform downstream analyses using predicted phenotypes. The methods to perform phenotype prediction are available in the phenopredict R package and the predictions for recount2 are available from the recount R package. With data and phenotype information available for 70,000 human samples, expression data is available for use on a scale that was not previously feasible.

[9] Zambelli, F., et al. (2018). "RNentropy: an entropy-based tool for the detection of significant variation of gene expression across multiple RNA-Seq experiments." Nucleic Acids Res **46**(8): e46.

RNA sequencing (RNA-Seq) has become the experimental standard in transcriptome studies. While most of the bioinformatic pipelines for the analysis of RNA-Seq data and the identification of significant changes in transcript abundance are based on the comparison of two conditions, it is common practice to perform several experiments in parallel (e.g. from different individuals, developmental stages, tissues), for the identification of genes showing a significant variation of expression across all the conditions studied. In this work we present RNentropy, a methodology based on information theory devised for this task, which given expression estimates from any number of RNA-Seq samples and conditions identifies genes or transcripts with a significant variation of expression across all the conditions studied, together with the samples in which they are over- or under-expressed. To show the capabilities offered by our methodology, we applied it to different RNA-Seq datasets: 48 biological replicates of two different yeast conditions; samples extracted from six human tissues of three individuals; seven different mouse brain cell types; human liver samples from six individuals. Results, and their comparison to different state of the art bioinformatic methods, show that RNentropy can provide a quick and in depth analysis of significant changes in gene expression profiles over any number of conditions.

[10] Wagner, J., et al. (2018). "Metaviz: interactive statistical and visual analysis of metagenomic data." Nucleic Acids Res **46**(6): 2777-2787.

Large studies profiling microbial communities and their association with healthy or disease phenotypes are now commonplace. Processed data from many of these studies are publicly available but significant effort is required for users to effectively organize, explore and integrate it, limiting the utility of these rich data resources. Effective integrative and interactive visual and statistical tools to analyze many metagenomic samples can greatly increase the value of these data for researchers. We present Metaviz, a tool for interactive exploratory data analysis of annotated microbiome taxonomic community profiles derived from marker gene or whole metagenome shotgun sequencing. Metaviz is uniquely designed to address the challenge of browsing the hierarchical structure of metagenomic data features while rendering visualizations of data values that are dynamically updated in response to user navigation. We use Metaviz to provide the UMD Metagenome Browser web service, allowing users to browse and explore data for more than 7000 microbiomes from published studies. Users can also deploy Metaviz as a web service, or use it to analyze data through the metavizr package to interoperate with state-of-the-art analysis tools available through Bioconductor. Metaviz is free and open source with the code, documentation and tutorials publicly accessible.

[11] Yu, Y., et al. (2018). "XPAT: a toolkit to conduct cross-platform association studies with heterogeneous sequencing datasets." Nucleic Acids Res **46**(6): e32.

High-throughput sequencing data are increasingly being made available to the research community for secondary analyses, providing new opportunities for large-scale association studies. However, heterogeneity in target capture and sequencing technologies often introduce strong technological stratification biases that overwhelm subtle signals of association in studies of complex traits. Here, we introduce the Cross-Platform Association Toolkit, XPAT, which provides a suite of tools designed to support and conduct large-scale association studies with heterogeneous sequencing datasets. XPAT includes tools to support cross-platform aware variant calling, quality control filtering, gene-based association testing and rare variant effect size estimation. To evaluate the performance of XPAT, we conducted case-control association studies for three diseases, including 783 breast cancer cases, 272 ovarian cancer cases, 205 Crohn disease cases and 3507 shared controls (including 1722 females) using sequencing data from multiple sources. XPAT greatly reduced Type I error inflation in the case-control analyses, while replicating many previously identified disease-gene associations. We also show that association tests conducted with XPAT using cross-platform data have comparable performance to tests using matched platform data. XPAT enables new association studies that combine existing sequencing datasets to identify genetic loci associated with common diseases and other complex traits.

References

1. Boisset JC, Vivie J, Grun D, Muraro MJ, Lyubimova A, van Oudenaarden A: **Mapping the physical network of cellular interactions**. *Nat Methods* 2018.

2. Svensson V, Teichmann SA, Stegle O: **SpatialDE: identification of spatially variable genes**. *Nat Methods* 2018, **15**(5):343-346.

3. Kiselev VY, Yiu A, Hemberg M: **scmap: projection of single-cell RNA-seq data across data sets**. *Nat Methods* 2018, **15**(5):359-362.

4. Soneson C, Robinson MD: **Bias, robustness and scalability in single-cell differential expression analysis**. *Nat Methods* 2018, **15**(4):255-261.

5. Alpert A, Moore LS, Dubovik T, Shen-Orr SS: **Alignment of single-cell trajectories to compare cellular expression dynamics**. *Nat Methods* 2018, **15**(4):267-270.

6. Sharma A, Jiang C, De S: **Dissecting the sources of gene expression variation in a pan-cancer analysis identifies novel regulatory mutations**. *Nucleic Acids Res* 2018, **46**(9):4370-4381.

7. Kim T, Seo HD, Hennighausen L, Lee D, Kang K: **Octopus-toolkit: a workflow to automate mining of public epigenomic and transcriptomic next-generation sequencing data**. *Nucleic Acids Res* 2018, **46**(9):e53.

8. Ellis SE, Collado-Torres L, Jaffe A, Leek JT: **Improving the value of public RNA-seq expression data by phenotype prediction**. *Nucleic Acids Res* 2018, **46**(9):e54.

9. Zambelli F, Mastropasqua F, Picardi E, D'Erchia AM, Pesole G, Pavesi G: **RNentropy: an entropy-based tool for the detection of significant variation of gene expression across multiple RNA-Seq experiments**. *Nucleic Acids Res* 2018, **46**(8):e46.

10. Wagner J, Chelaru F, Kancherla J, Paulson JN, Zhang A, Felix V, Mahurkar A, Elmqvist N, Corrada Bravo H: **Metaviz: interactive statistical and visual analysis of metagenomic data**. *Nucleic Acids Res* 2018, **46**(6):2777-2787.

11. Yu Y, Hu H, Bohlender RJ, Hu F, Chen JS, Holt C, Fowler J, Guthery SL, Scheet P, Hildebrandt MAT *et al*: **XPAT: a toolkit to conduct cross-platform association studies with heterogeneous sequencing datasets**. *Nucleic Acids Res* 2018, **46**(6):e32.