Twitter Thread by KenOliveLab

KenOliveLab
@KenOliveLab

Extremely excited today to reveal the first of two great works (magnus opera?), just posted on bioarxiv, applying regulatory network analysis techniques to PDAC expression data to dissect the underlying biology of the disease. Tweetorial time! 1/

Eight years ago I launched an enduring collaboration with Andrea Califano @ColumbiaCancer to apply his regulatory network analysis techniques to #pancreaticcancer. So many amazing scientists have contributed to this work: @paslaise @hc_maurer @AlvaroCurielGa1 @ElyadaEla 2/

What is “regulatory network analysis”? At it’s heart, it’s a way of extracting more useful information from expression profiles. A fundamental flaw of differential gene expression (DGE) analysis is the assumption that each gene is independent. Biologist KNOW this is not true! 3/

DGE treats all ~25K detectable genes as SEPARATE variables, performs 25K T-tests, and then slaps on a multiple hypothesis correction to make the statisticians less dyspeptic. There is no consideration of the relationships between genes! 4/

Yet we KNOW genes are co-regulated in SETS by transcription factors and other REGULATORY FACTORS. Biologist: “p21 is a target of p53”. DGE: “Shhhh”. These relationships are completely ignored by expression analysis. THIS IS A TRAVESTY! 5/

Other limitations of expression analysis: 1) high variance for gene expression measurements; 2) high dimensionality (~25K!) means a big N in the denominator; 3) diff. expression does not suggest causality. 6/

Also, as reviewers love to note, RNA expression does not tell you anything about protein activity… or does it? You need an activity assay, right? What if I told you I could quantitatively measure enzyme activity using RNA expression data. Break out the pitchforks? 7/

Well, transcription factors are simply enzymes that alter the abundance of RNA transcripts for specific target genes. So RNAseq can quantify the products of the transcriptional enzymatic reaction, and therefore it can be used as an ASSAY to measure their PROTEIN ACTIVITY! 8/

To do this though, you need to know the targets for each regulatory factor. And that's the trick. Because if you use databases or services that collate this info from literature, you mix together info from lots of cell types and are limited to what others have already found. 9/

Enter the ARACNe algorithm, which uses information theory principles to computationally deduct target genes for regulatory factors, DE NOVO, using only gene expression profiles as input. The result is a global list of transcriptional relationships- a "regulatory network". 10/

With a regulatory network as a scaffold, one can transform an expression profile of ~25K genes to an ACTIVITY profile for perhaps 1500 regulatory factors. Benefits: 1) each activity measurement incorporates the expression of 100s of targets – way more precise/lower variance. 11/

2) notable dimensionality reduction; 3) incorporates relationships between genes; 4) most important, there is built-in mechanistic info. If you observe a difference in TF activity between two groups, they could be causal and you know their potential target gene effectors. 12/

To me, a critical piece here is the context specificity. The set of genes regulated by a TF in one cell type is hugely different than in another cell type! Finally, regulatory network analysis is extremely well validated by experiments dozens of papers. 13/

There's one more HUGE benefit to this approach: dealing with scRNAseq data. A prevailing challenge in the field is the "gene dropout problem" in which most genes in most cells of an scRNAseq profile have zero reads. Regulatory network analysis fixes this problem!!! 14/

Because the activity of each TF is calculated from hundreds of targets, you can determine the activity of protein that literally has ZERO reads in that cell. !!! This is transformative for the analysis of single cell expression data. 15/

OK, with your new 15-tweet education in regulatory network systems biology, here we go. In our new manuscript, we apply regulatory network analysis to the problem of molecular subtypes in PDAC. The nature of molecular subtypes in pancreatic cancer has long been controversial. 16/

Preface: I LOVE all of the published PDA molecular subtype papers. Each was well done and added important new contributions to the field. But they don't agree on details such as how many subtypes there are, or what genes distinguish them. 17/

The ISSUE is the TISSUE. PDAC has a huge amount of stroma intermixed within tumors. This makes it hard to extract information from the malignant epithelial cells. We dealt with this using laser capture microdissection (LCM-RNAseq) and

single cell expression analysis. 18/

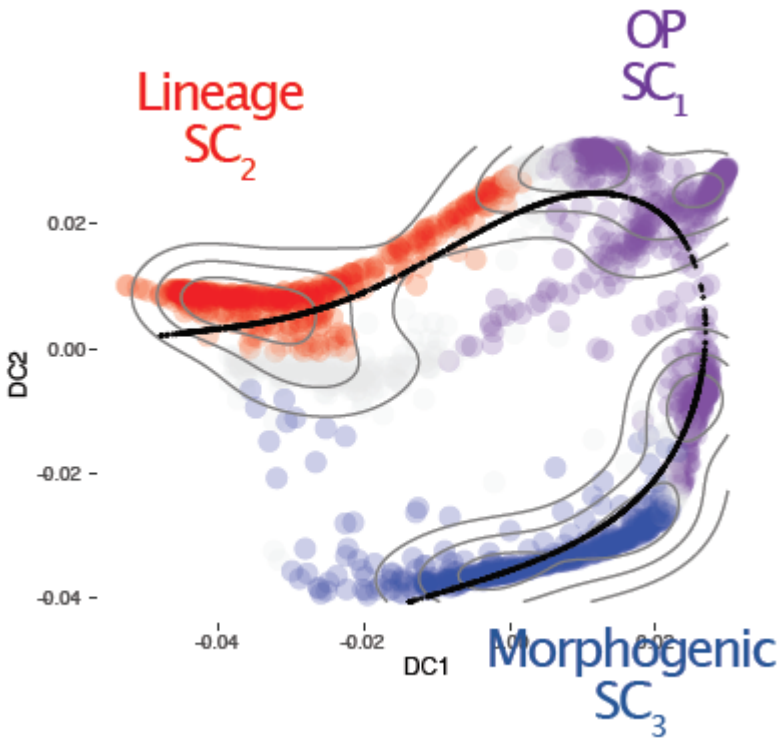
First, we performed LCM-RNAseq on a total of 200 human PDAC epithelial samples as well as 45 begin precursors (PanIN and IPMN). Data getting deposited ASAP to GEO! This was done on OCT samples and we have histopath on adjacent sections as well as outcomes. 19/

Then we applied regulatory network analysis and clustered based on ACTIVITY rather than gene expression, grouping tumors based on their GLOBAL REGULATORY STATE. We integrated multiple expression datasets with our LCM data using an algorithm called metaVIPER. 20/

Take Home Message: There are TWO epithelial subtypes at the bulk tissue level. The “Lineage” group shows elevated relative activity of GI transcription factors whereas the “Morphogenic” group has high activity of EMT genes and morphogen pathway transcription factors. 21/

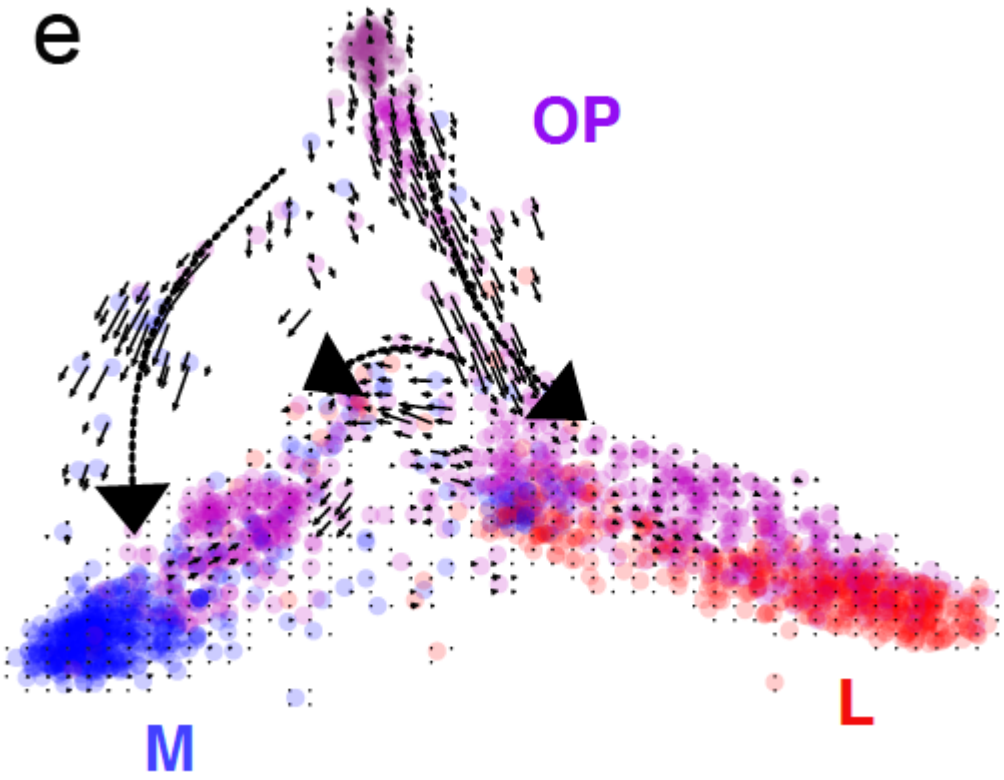
We confirm association between the less differentiated Morphogenic group and poor outcome. But there is only modest overlap with prior subtypes. Also, in the LCM data, tumors intermediate between the two states can be detected as a third cluster. So what’s going on there? 22/

To learn more, we applied the approach to PDAC scRNAseq data. However, instead of the two cell types we expected, we found three! Amazingly, the Lineage and Morphogenic states were almost perfectly recapitulated as two of the three cell types. 23/



BUT, the most common cell type in the tumor was completely invisible by bulk tissue analysis. May I introduce to you the Oncogenic Precursor (OP) cell! They are weird and awesome. They have high relative activity of EARLY GI transcription factors. 24/

OP cells are more WELL differentiated than the other cells, but may actually be the source of the other cell types. Yet they are proliferative and don't express typical stem/progenitor markers. This turns the concept of cancer stem cells on its head. Velocity analysis below. 25/



OP cells form about HALF the epithelial cells in most PDAC tumors. Because they are EVENLY represented in most tumors, they can't be detected by DIFFERENTIAL expression analysis. They are the Dark Matter of PDAC biology! 26/

So the two bulk subtypes emerge from the combination of a roughly constant amount of OP cells with a variable ratio of Lineage versus Morphogenic cells. Pretty cool, no? Also- those PDAC cell lines you use in 2D culture? They are a mixture of cellular subtypes! 27/

Finally, we validated these findings using two high throughput functional screening assays. In one, we knocked out thousands of transcription factors in six PDAC lines and showed that both general and subtype-specific dependencies. 28/

In the second screen, we showed that overexpressing one or more Lineage TFs could transform Morphogenic cells to a Lineage state. This affirms the regulatory dependencies of cellular subtypes in PDAC and hints at potential differences in therapeutic sensitivity. 29/

Important to acknowledge from @lustgartenfdn , @PancreasCenter , and the Columbia Irving Institute for Clinical Translational Research. Thank you so much for your support! 30/

Would love to hear critiques. What do you think? Do you believe in this master regulator stuff? Does this clarify transcriptional subtypes? Or just make a bigger mess? Looking forward to sharing our next manuscript on MRs of tumor progression very soon! 31/31

I forgot a key point. When you apply prior classifiers to single cells, a problem emerges: individual cells enrich for multiple subtypes. Or neither. This is biologically very confusing. Regulatory subtypes avoid this issue. 32/31