What's holding up progress against cancer?

*Time is not on our side when it comes to cancer: early diagnosis is key!**

50+ years of research milestones

- Oncogene discovery, e.g., P53 tumor suppressor gene (1979)
- Brca1 and 2 discovered 30 years ago (breast cancer)
- Human genome mapped (2003)
- Large-scale genomic arrays commercially available for 20+ years (now > 500k probes/array)
- Too many others to list!

And yet…

Genomic testing can take weeks to months for results

Price per test is $100s to $1000s

Tests tend to be narrowly tailored

Cancers typically involve 100s of millions of mutated cells, competing and evolving—with not just one malfunctioning gene, but dysregulation of dozens to hundreds)

*once a cancer is metastatic, it is generally incurable.

Challenges with lung cancer

*Fairly common disease: New case rate for lung cancer was 49.0 per 100,000 men and women per year. The death rate was 32.4 per 100,000 men and women per year.**

Radiography used to detect abnormalities, however…

- FNAs (fine needle aspirations) are then performed to directly check nodal abnormalities, followed by histological exam
- Up to 25 percent of patients may have abnormalities found through lung imaging; however, most of these are not lung cancer.
- Antigen targets for early screening were identified over a decade ago (but test still unavailable)
  - Blood tests could identify patients at high risk for cancer 6-12 months before tumors visible via radiographic imaging

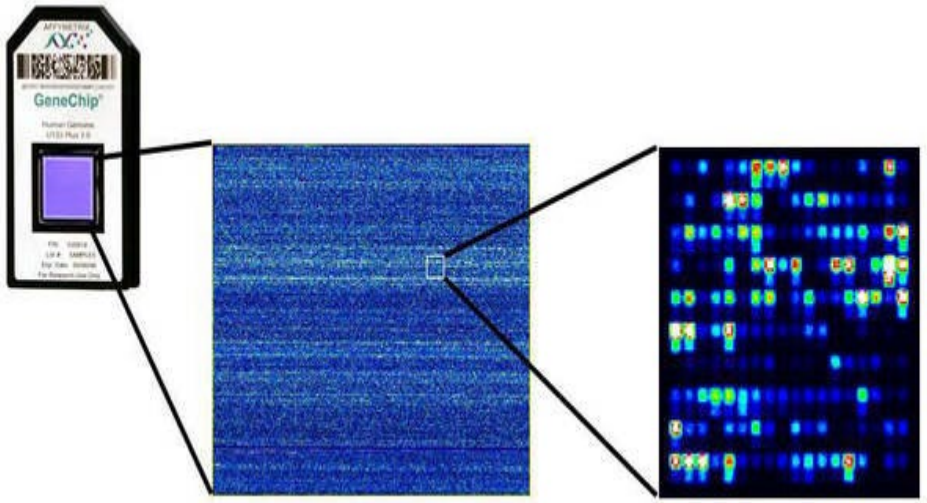
Clear need for better diagnostic tools

References: Fredhutch.org, private conversations with FHCC researchers, cancer.gov (NIH), https://abcnews.go.com/Health/Healthday/story?id=5809883&page=1

*Lung cancer is one of the deadliest cancers in the world

Paving the road from the lab to the office

Great technology for research, but…



Microarrays

- Huge number of sequences detected in parallel
- Costly (> $500/array)
- Requires sophisticated, expensive processing equipment
- Not well-suited for diagnostics

Assay development

…we need something simple

Microtiter plates

- Capacity is 24, 48 or more target genes
- Cheap ($3-4 each)
- Uses simpler, off-the-shelf equipment
- Can be deployed at point of care

Research has produced a treasure trove of genomic data

The expression database we chose to mine is just one out of many

TCGA - LUSC (Lung Cell Squamous Carcinoma) Expression Profiling by Array Dataset

- 56,907 transcripts
  - Genes and microRNA targets
- 551 patients (502 with lung cancer, 49 healthy)
  - Hard to collect as many true positive samples as this
- Genes named but not annotated

+

Alliance Genome's Homo sapiens gene annotation dataset

- 43,404 annotations
  - Brief descriptions of gene function (known or predicted)
  - All annotations automatically scraped from the literature
  - All annotations matched with genes in the LUSC database

The above datasets were merged to allow for careful selection of genes that might predict cancer

Data sources: https://portal.gdc.cancer.gov/ , https://www.alliancegenome.org/downloads#gene-descriptions

Finding the right needles in a haystack

How do we narrow down 57,000 measurements into a couple dozen?

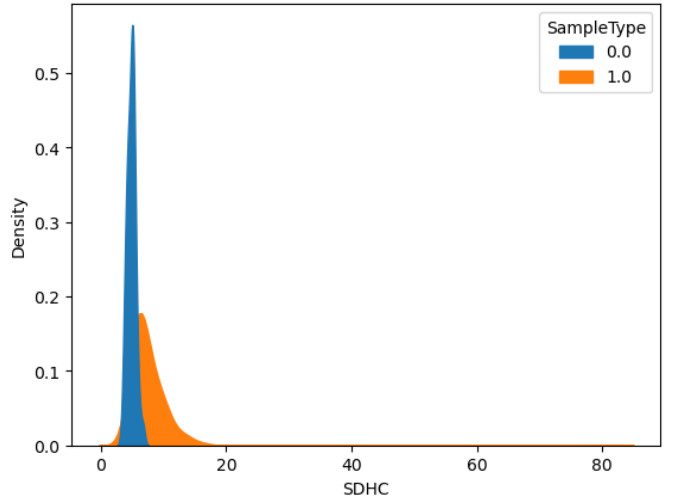
Key reasons for narrowing our feature search:	**<u>Selection criteria:</u>**	**<u>Rationale:</u>**
• Enable cheaper, faster processing	Select only those genes active in normal lung tissue	Improve accuracy, Increase test specificity (select lung-cell related genes)
	Reject known biomarkers of disease*	Avoid false positives
• Avoid overfitting (use fewer than 50 features in our predictor set)	Reject genes active across multiple (non-lung) cancers	Increase test specificity
	Avoid mRNAs	Ideally prefer genes that code for proteins, for later development of peripheral blood tests

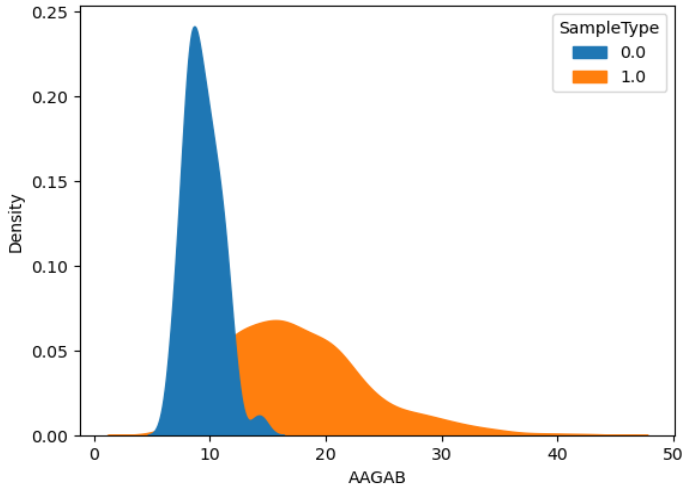
Lung cell oncogenes discovered elsewhere in the literature were not used; our model used a data-driven approach only

Except for normal sample activity, no expression data was used to make these feature selections

*a limited number of genes implicated in rare genetic or other unrelated conditions were allowed

The model employs a key discovery





At left we see the expression profiles of two genes, in normal (blue) and tumor (orange) tissue. These genes were randomly chosen and led to a surprising finding:

The variance in expression is strikingly different between the two tissue types

A literature search confirmed the validity of this observation*

To maximize model performance, we therefore decided to normalize all expression data by the standard deviation of gene expression in normal tissue samples

*PLoS Genet. 2011 Aug; 7(8): e1002207.Published online 2011 Aug 11. doi: 10.1371/journal.pgen.1002207PMCID: PMC3154954PMID: 21852951Variance of Gene Expression Identifies Altered Network Constraints in Neurological Disease

The final candidate gene list*

Not all genes are equally predictive; some may invite further investigation

	Gene Name	Importance
Upregulated genes	NKIRAS2	0.33
	SIM2	0.27
	COPA	0.17
	TMEM38B	0.16
	SNHG20	0.16
	DPPA2	0.16
	MSH5	0.11
	MAN2A1	0.08
	MAN1A2	0.07
	ZNF157	0.06
	DPPA4	0.05
	CIC	0.04
Downregulated genes	CFC1B	-0.02
	LAMA3	-0.03
	CCDC40	-0.06
	ATXN1L	-0.19
	NKIRAS1	-0.23
	STK40	-0.27
	TBX4	-0.55
	FENDRR	-0.60

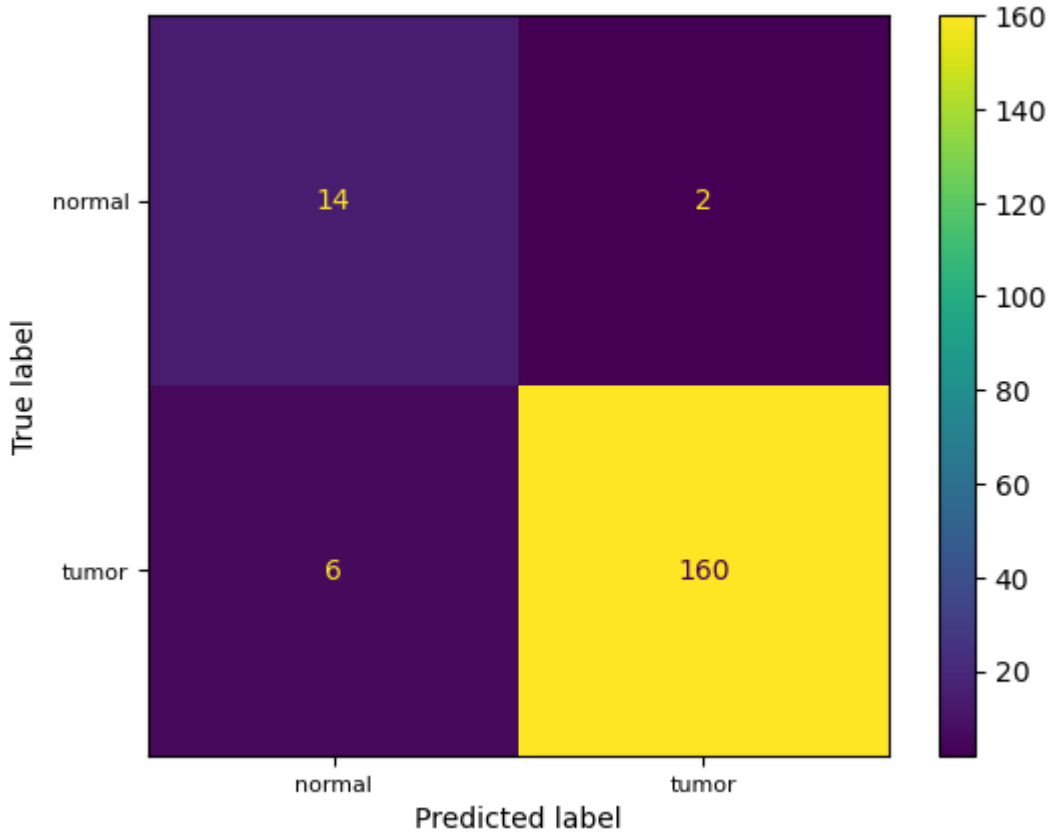
- 20 genes in the final set (ranked by feature importance)
- Compact set (small enough to use a 24-well titer plate with controls)
- Plenty of room for further experimentation/refinement of selections
- Genes can likely be added or removed based on additional criteria or through validation process

*Annotations and comments on gene function in Appendix B.

Initial model results were promising

*Results from unoptimized logistic regression model**

Sample	precision	recall	F1-score	support
Normal	0.7	0.88	0.78	16
Tumor	.99	.96	.98	166
accuracy			.96	182
Macro avg	.84	.92	.88	182
Weighted average	.96	.96	.96	182



Above results indicated we should proceed with model refinements and optimization

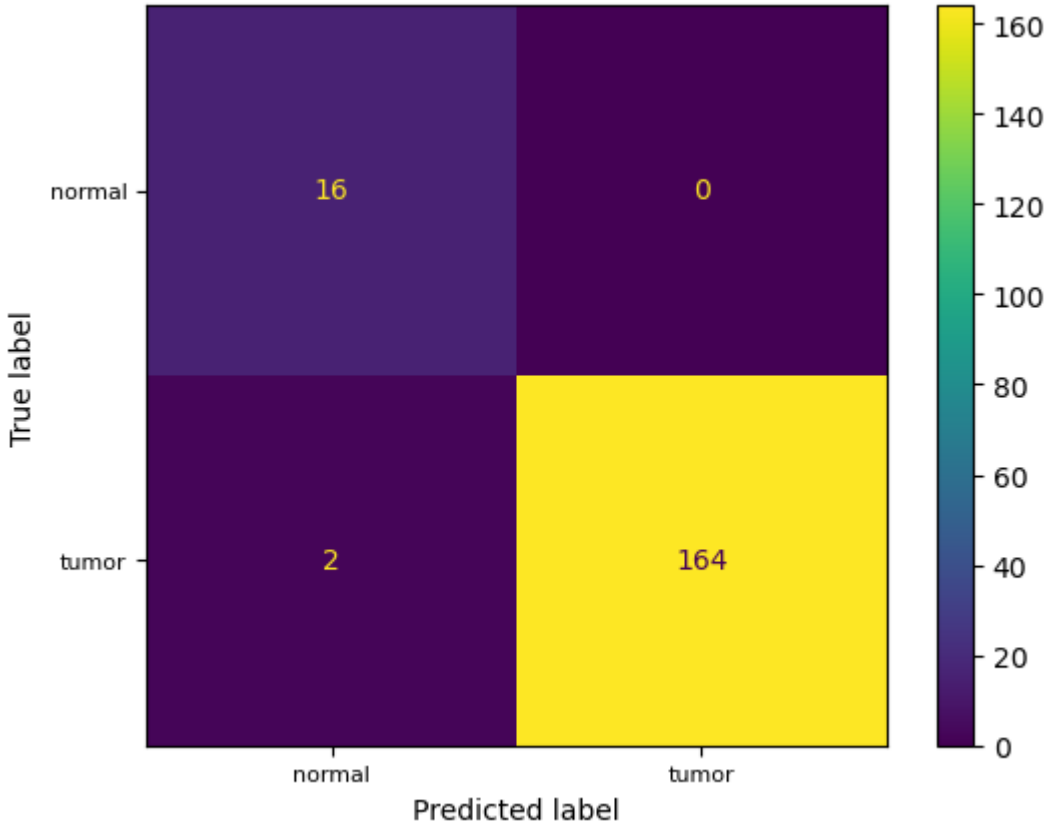
Goal: fewest false positives/false negatives possible

*Using a slightly larger gene list, and before adjusting to prevent data leakage

Excellent results obtained after optimizing model*

Model illustrates solid proof of concept

Sample	precision	recall	F1-score	support
Normal	0.89	1.0	0.94	16
Tumor	1.0	.99	.99	166
accuracy			.99	182
Macro avg	.94	.99	.97	182
Weighted average	.99	.99	.99	182



False positives/false negatives now very low

Model only includes 49 normal samples; would not expect to find additional improvement with other model types

*Optimal model parameters available with code notebook

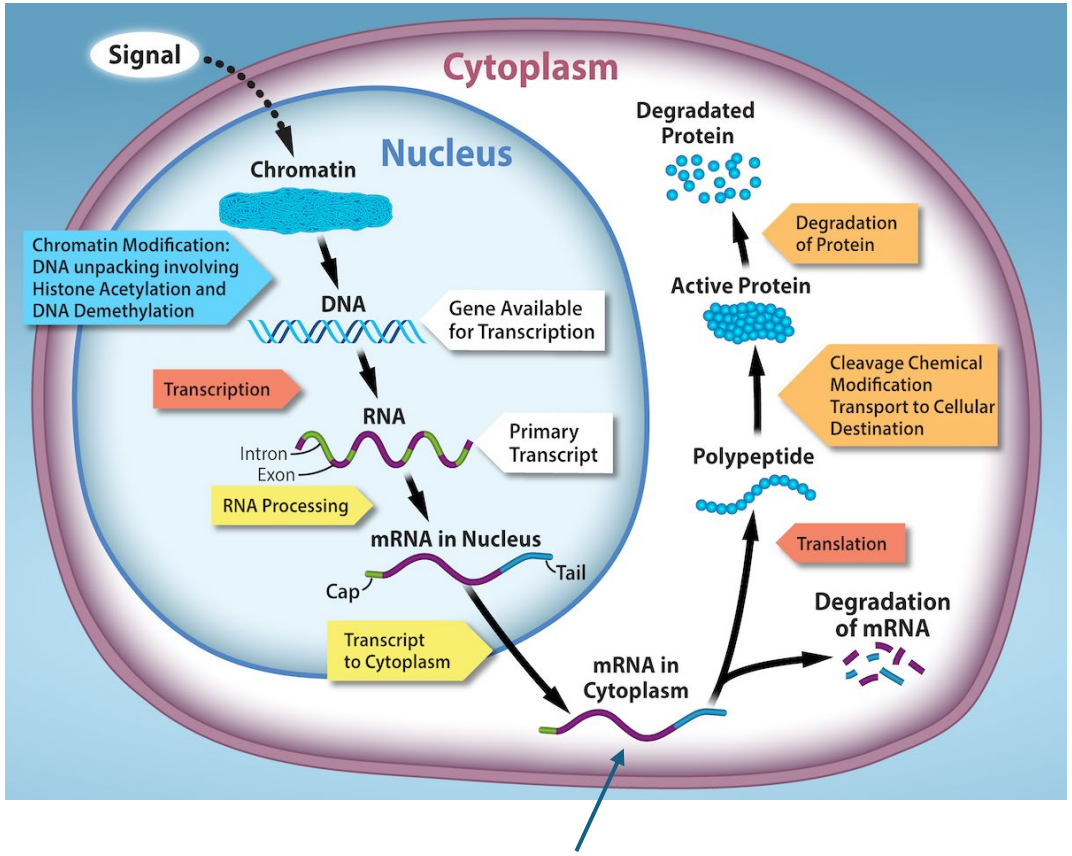
Yet many hurdles lie ahead

“You could make a lot of money off this”---not so fast!

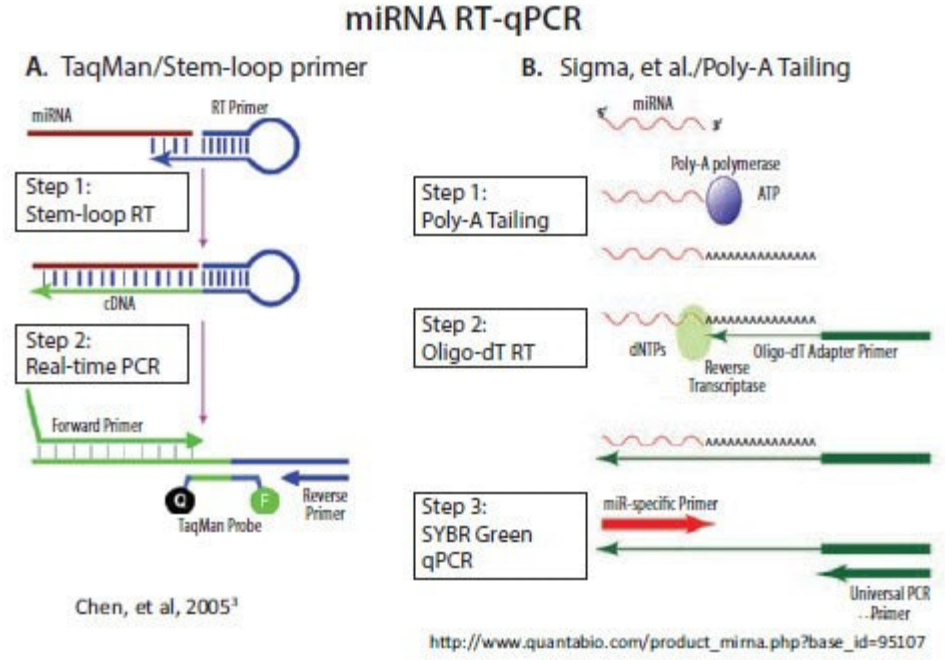
- Commercialization of diagnostic panels can take 15+ years
  - PCR panels for infectious disease developed circa 2006 for companion animals; similar products available for human use ~15 years later (MicroGenDX).
- Validation with large volume of real or archived (ATTC) samples costly, time-consuming (need a lot more than 49 normal samples)
- IP protection: US patent office limits patents on genetic sequences to 3/patent (hard to patent a panel of tests)
- Practitioners may have other test desiradata:
  - Inclusion of indicators for prognosis
  - Differentiation between types of lung cancer (small vs non-small cell)
  - Antigen vs gene expression test: use of peripheral blood vs FNA samples

Appendix A: Gene expression, PCR and microarrays

Biological processes and techniques/process for microarray measurement



Cells 'express' certain genes; mRNA for each is measured after sample processing
Expression is highly regulated!



Specific primers are used to amplify RNA from sample; these primers are bound to arrays, and target fluorescence measured after selective binding

Appendix B: Partial gene set annotations

Gene Name	Importance	Selected notes on biological function
NKIRAS2	0.33	Predicted to act upstream of or within several processes, including lung alveolus development; regulation of signal transduction; and surfactant homeostasis.
SIM2	0.27	Predicted to be involved in regulation of transcription by RNA polymerase II
COPA	0.17	Predicted to enable mRNA binding activity. Implicated in autoimmune interstitial lung, joint, and kidney disease.
TMEM38B	0.16	Predicted to enable potassium channel activity. Predicted to be involved in potassium ion transmembrane transport. Predicted to act upstream of or within several processes, including cellular response to caffeine; lung development; and regulation of cardiac muscle contraction by regulation of the release of sequestered calcium ion.
SNHG20	0.16	Predicted to be involved in cellular response to fatty acid; gene expression; and macrophage activation. Implicated in lung non-small cell carcinoma.
DPPA2	0.16	Predicted to act upstream of or within several processes, including epigenetic regulation of gene expression; lung-associated mesenchyme development; and positive regulation of stem cell proliferation.
MSH5	0.11	Predicted to enable double-stranded DNA binding activity. Implicated in lung non-small cell carcinoma; primary ovarian insufficiency 13; and spermatogenic failure.
MAN2A1	0.08	Predicted to enable alpha-mannosidase activity. Predicted to be involved in N-glycan processing. Predicted to act upstream of or within several processes, including lung alveolus development
MAN1A2	0.07	Predicted to enable mannosyl-oligosaccharide 1,2-alpha-mannosidase activity. Predicted to be involved in N-glycan processing. Predicted to act upstream of or within glycoprotein metabolic process; lung alveolus development; and respiratory gaseous exchange by respiratory system
ZNF157	0.06	Predicted to act upstream of or within lung alveolus development; mammary gland morphogenesis; and regulation of cell fate commitment.
DPPA4	0.05	Predicted to act upstream of or within lung-associated mesenchyme development.
CIC	0.04	Predicted to be involved in several processes, including learning or memory; regulation of DNA-templated transcription; and social behavior. Predicted to act upstream of or within lung alveolus development and negative regulation of transcription by RNA polymerase II
CFC1B	-0.02	Predicted to be involved in circulatory system development; nodal signaling pathway; and regionalization. Predicted to act upstream of or within several processes, including heart development; left lung morphogenesis; and spleen development.
LAMA3	-0.03	Predicted to enable integrin binding activity. Predicted to be an extracellular matrix structural constituent. Involved in endodermal cell differentiation. Implicated in junctional epidermolysis bullosa and lung small cell carcinoma.
CCDC40	-0.06	Involved in axonemal dynein complex assembly; determination of left/right symmetry; and lung development. Acts upstream of or within cilium assembly; flagellated sperm motility; and regulation of cilium beat frequency. Implicated in primary ciliary dyskinesia 15.
ATXN1L	-0.19	Predicted to be involved in learning or memory; regulation of DNA-templated transcription; and social behavior. Predicted to act upstream of or within several processes, including lung alveolus development; negative regulation of transcription by RNA polymerase II; and positive regulation of hematopoietic stem cell proliferation.
NKIRAS1	-0.23	Predicted to enable GTPase activating protein binding activity. Predicted to be involved in Ral protein signal transduction. Predicted to act upstream of or within lung alveolus development; regulation of tumor necrosis factor-mediated signaling pathway; and surfactant homeostasis.
STK40	-0.27	Predicted to enable ATP binding activity; protein serine kinase activity; and protein serine/threonine kinase activity. Predicted to act upstream of or within several processes, including glycogen metabolic process; lung development; and respiratory system process.
TBX4	-0.55	Involved in embryonic hindlimb morphogenesis; embryonic lung development; and skeletal system morphogenesis. Implicated in arthropathy and ischiocoxopodopatellar syndrome.
FENDRR	-0.60	Predicted to enable core promoter sequence-specific DNA binding activity. Predicted to be involved in in utero embryonic development and lung development. Predicted to act upstream of or within several processes, including chromatin organization; embryonic lung development; and lateral mesoderm development.