COLUMBIA UNIVERSITY

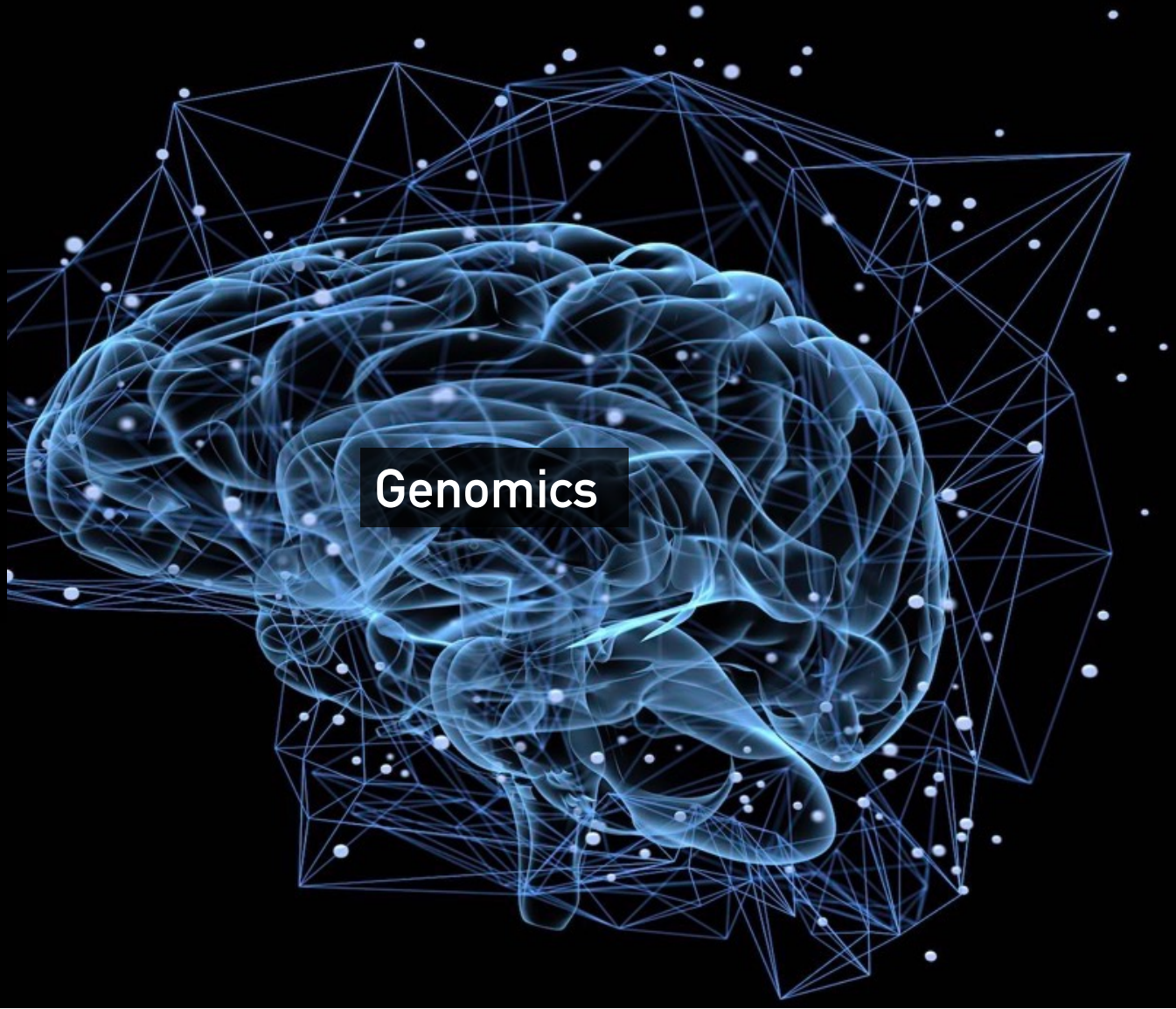
EECS 6895 Advanced Big Data and AI

Lecture 7: AI for Life Sciences II (Genomics)

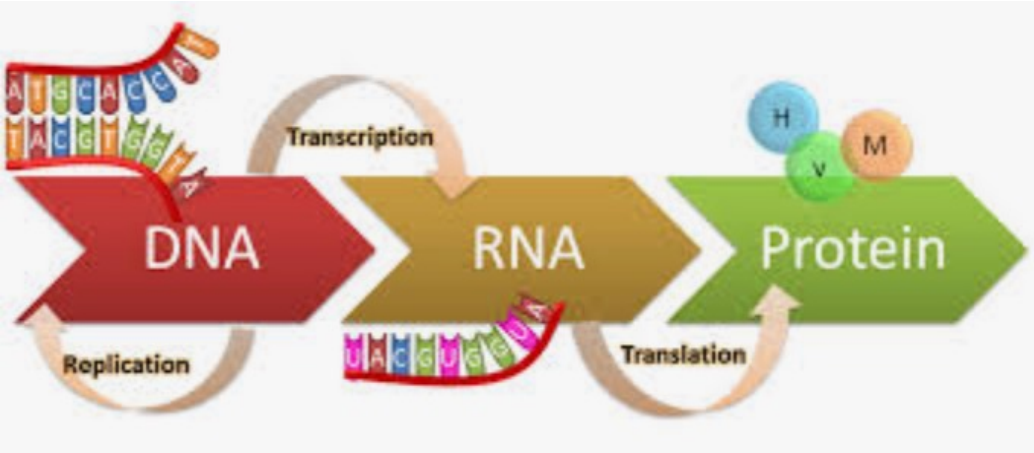
Prof. Ching-Yung Lin

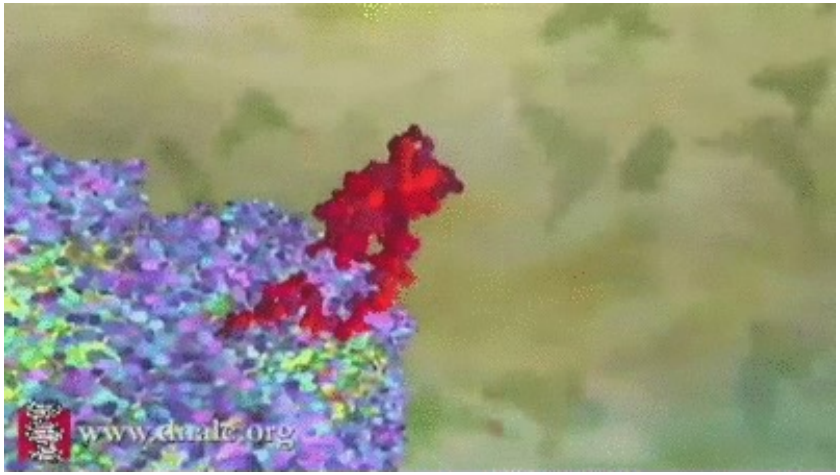
Columbia University

March 3rd , 2026

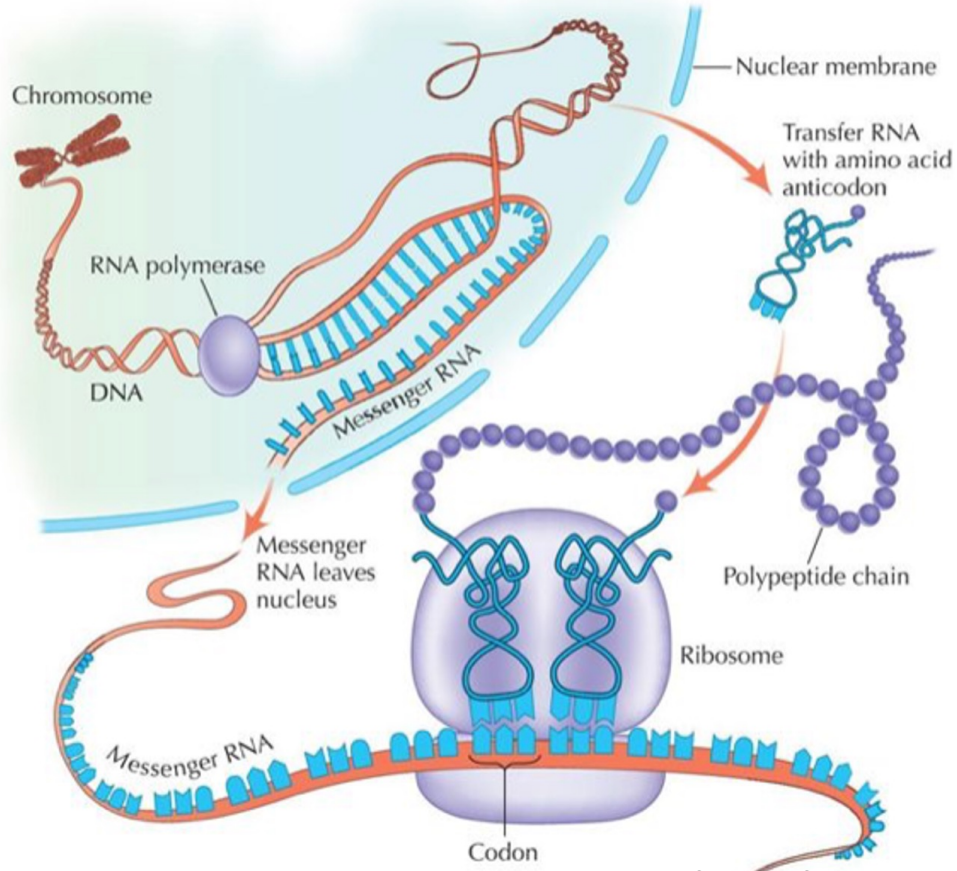

Genomics

Central Dogma of (Molecular) Biology





Forming Protein



Protein is the Foundation of All Life

A Protein == A Biological Machine Part

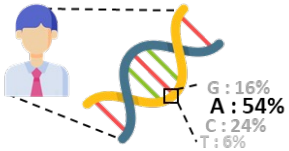
Graphen Health Services

-- Precision Health and Precision Medicine

General Consumers

1 *What am I?*

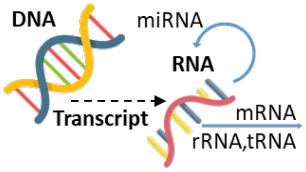
Born: Whole Genome Analysis



Use Whole Genome Sequencing (WGS) to understand the innate DNA genetic makeup, and analyze the association between individuals and disease risk through Single Nucleotide polymorphisms/variations (SNPs/SNVs) , providing an entry point for early diagnosis.

2 *How am I?*

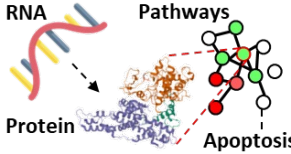
Now: Gene Expression Analysis



Epigenetic genetic behavior and regulations determine when and how extent genes affect us. Graphen further uses personal genetic behavior testing to understand current genetic behavior trends and analyze disease risks caused by acquired factors such as the environment. It also provides customers with existing disease states with the cause and course of the current disease.

3 *What will I be?*

Future: Pathways Dynamics

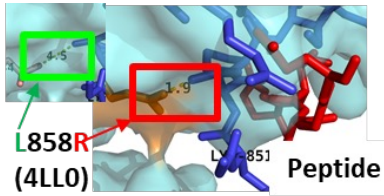


The dynamic expression of proteins in the biological signal transmission chain shows how genes affect body status and disease development. By combining genomic analysis with protein interaction networks, Graphen can instantly understand the correlation between key genes and disease processes.

Patients

4 *I want the most suitable drugs to me!!*

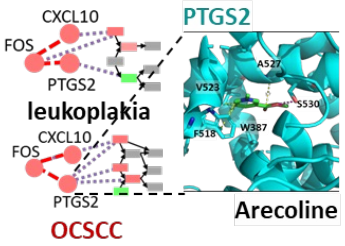
Gene/Protein-Mutation & Resistance Analysis



Genetic variation is inseparable from mutations and the occurrence of disease. Grphen understands how mutations in genes and proteins change the normal working of the body from the molecular structure level. In addition to elucidating the relationship between mutations and actual disease symptoms, Grphen also establishes the foundation for subsequent precision medicine and extends the development of dynamic disease classification and treatment strategies.

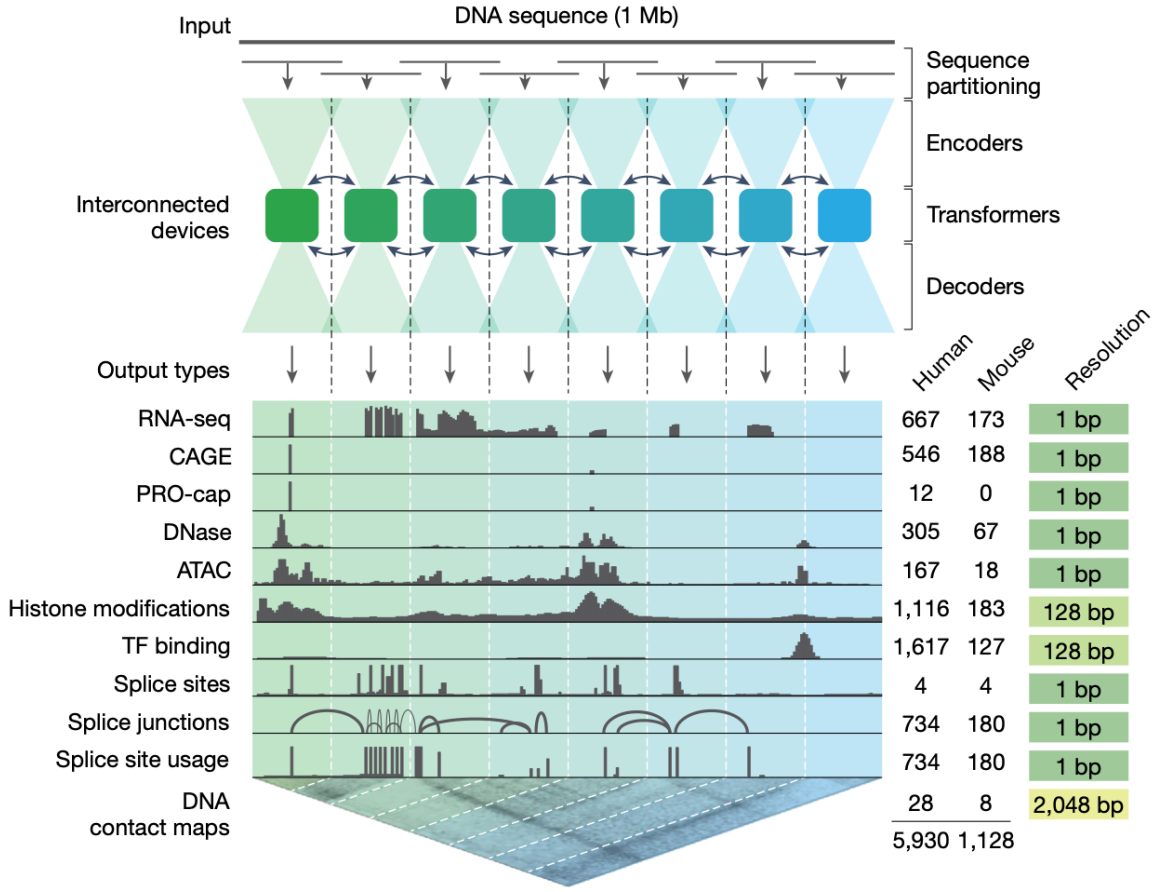
5 *Happening again?*

Prognosis Analysis

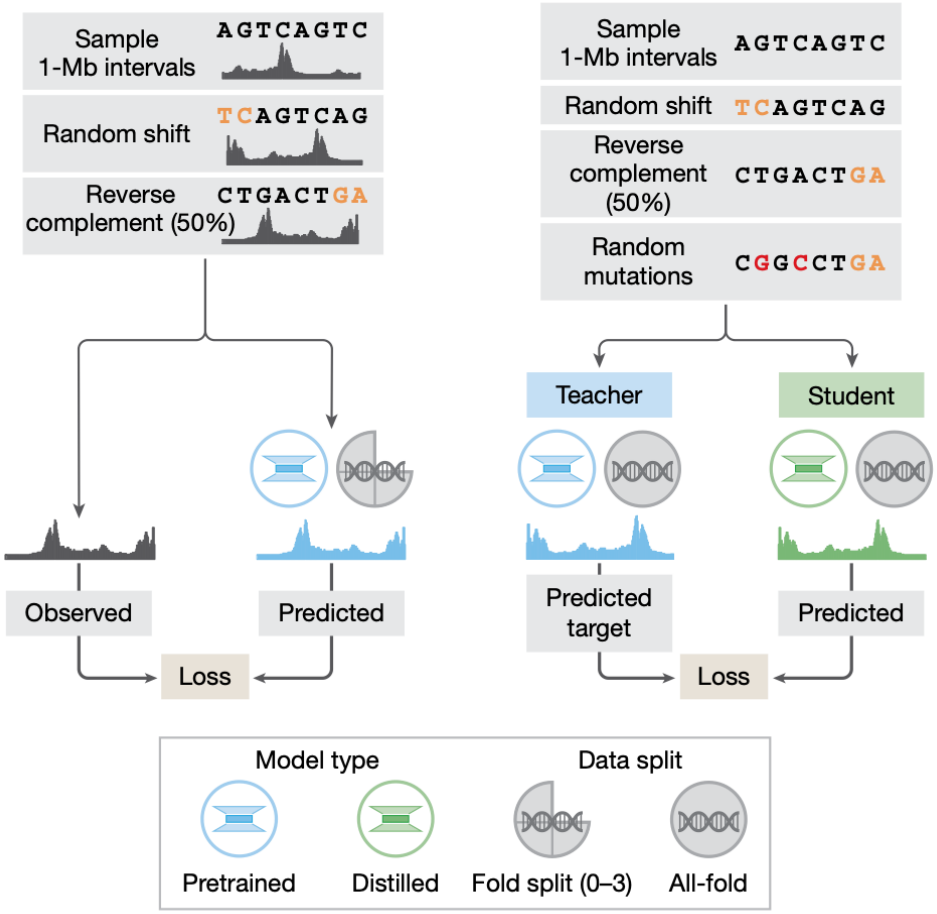


Graphen combines Genomics and Atoms analysis to develop multi-faceted disease classification services and predictions (gene & protein differences, biochemical pathway transmission chain patterns, structural mutations) in precision medicine, and provide analysis of prognosis prediction and monitoring strategies

AlphaGenome

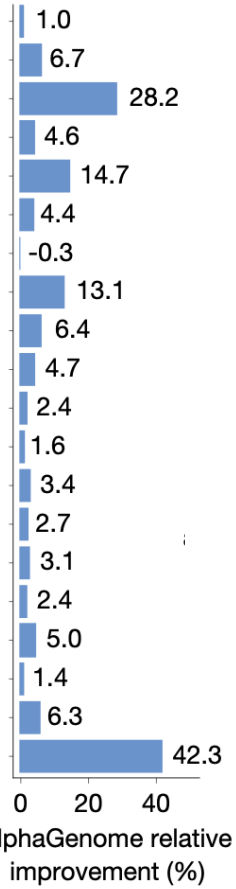


AlphaGenome



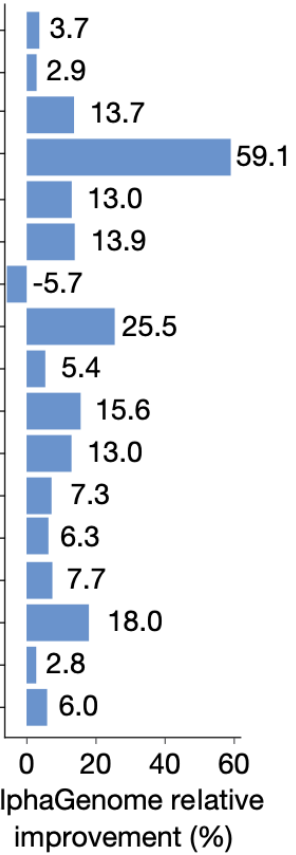
AlphaGenome

Modality	Evaluation	Metric	Resolution	Value	Comparison
Splicing	Splice site classification	auPRC	1 bp	0.79	DeltaSplice
	Splice site usage	Pearson *r*	1 bp	0.86	DeltaSplice
RNA expression	RNA-seq coverage	Pearson *r*	1 bp	0.59	Borzoi
			32 bp	0.78	Borzoi
	RNA-seq gene expr. LFC	Pearson *r*	Gene	0.57	Borzoi
	CAGE coverage	Pearson *r*	32 bp	0.74	Borzoi
			128 bp	0.71	Enformer
	Alternative poly(A)	Spearman *r*	Gene	0.87	Borzoi
DNA accessibility	DNase-seq coverage	Profile JSD	1 bp	0.51	ChromBPNet
		Pearson *r*	32 bp	0.86	Borzoi
			128 bp	0.87	Enformer
	ATAC-seq coverage	Profile JSD	1 bp	0.46	ChromBPNet
		Pearson *r*	32 bp	0.57	Borzoi
			128 bp	0.72	Enformer
Histone modifications	Histone ChIP-seq coverage	Pearson *r*	32 bp	0.69	Borzoi
			128 bp	0.71	Enformer
TF binding	TF ChIP-seq coverage	Pearson *r*	32 bp	0.55	Borzoi
			128 bp	0.58	Enformer
DNA contact maps	Orca contact maps	Pearson *r*	4,000 bp	0.79	Orca
	Orca contact maps cell type differences	Pearson *r*	4,000 bp	0.42	Orca



AlphaGenome

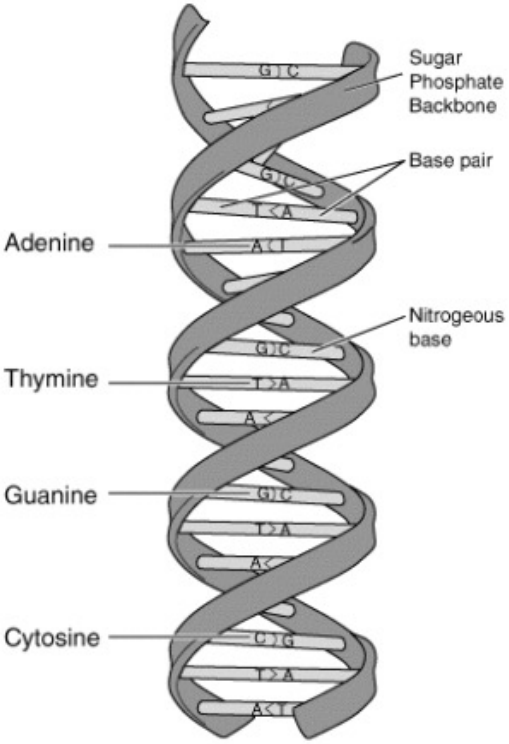
Modality	Evaluation	Type	Metric	Value	Comparison
Splicing	ClinVar splice site region	Causality	auPRC	0.57	Pangolin
	ClinVar non-coding		auPRC	0.66	Pangolin
	ClinVar missense		auPRC	0.18	DeltaSplice
	Splicing outlier (zero-shot)		auPRC	0.22	Pangolin
	Splicing outlier (supervised)		auPRC	0.28	AbSplice
	sQTL	Causality	auPRC	0.76	Pangolin
	MFASS		auPRC	0.51	Pangolin
RNA expression	eQTL	Direction	Spearman *r*	0.49	Borzoi
	eQTL (zero-shot)	Causality	auROC	0.71	Borzoi
	eQTL (supervised)	Causality	auROC	0.80	Borzoi
	ENCODE E2G (zero-shot)		auPRC	0.75	Borzoi
	paQTL		auPRC	0.63	Borzoi
DNA accessibility	CAGI5 MPRA	Causality	Pearson *r*	0.65	Borzoi
	ds/caQTL	Direction	Pearson *r*	0.70	ChromBPNet
	ds/caQTL	Causality	auPRC	0.52	Borzoi
TF binding	bQTL	Direction	Pearson *r*	0.55	Borzoi
	bQTL	Causality	auPRC	0.50	Borzoi



Genomics

- The molecules of DNA containing all the instructions to make the organisms ' working parts.
- If a cell is a computer, then its genome sequence is the software it executes.
- DNA is not just an abstract storage medium. It is a physical molecule that behaves in complicated ways.
- It also interacts with thousands of other molecules, all of which play important roles in maintaining, copying, directing, and carrying out the instructions contained in the DNA.
- The genome is a huge and complex machine made up of thousands of parts.
- We still have only a poor understanding of how most of those parts work, to say nothing of how they all com together as a working whole.
- Genetics treats DNA as abstract information. It looks at patterns of inheritance, or seek correlations across populations, to discover the connections between DNA sequences and physical traits.
- Genomics views the genome as a physical machine. It tries to understand the pieces that make up that machine and the ways they work together.

DNA, RNA and Proteins



- DNA is a polymer: a long chain of repeating units strung together.
- DNA has four units (called bases) that can appear: adenine, cytosine, guanine, and thymine, which are abbreviated as A, C, G, and T.
- Nearly all the information about how to make a living organism is ultimately encoded in the specific pattern of these four repeating units that make up its genome.
- **If DNA is the software, proteins are the most important hardware.**
- **Proteins are tiny machines that do almost all the work in a cell.**
- **Proteins are also polymers, made up of repeating units called amino acids.**
- **There are 20 amino acids. Some are large while others are small. Some have an electric charge while others do not. Some tend to attract water while others tend to repel it.**
- **When just the right set of amino acids is strung together in just the right order, it will spontaneously fold up into a 3D shape, to function as a machine.**

DNA, RNA and Proteins

- One of the main functions of DNA is to record the sequences of amino acids for an organism's proteins.
- It does this in a simple, straightforward way.
- Particular stretches of DNA directly correspond to particular proteins. Each sequence of three DNA bases (called a codon) corresponds to one amino acid. For example, the pattern AAA indicates the amino acid lysine, while the pattern GCC indicates the amino acid alanine.
- Going from DNA to protein involves another molecule, RNA, that serves as an intermediate representation to carry information from one part of the cell to another.
- RNA is yet another polymer and is chemically very similar to DNA. It too has four bases that can be chained together in arbitrary orders. To create a protein, the information must be copied twice.
- First the DNA sequence is transcribed into an equivalent RNA sequence, and then the RNA molecule is translated into a protein molecule.
- The RNA molecule that carries the information is called a messenger RNA, or mRNA for short.

Transcription Factors

- DNA tells us how proteins get made, but not when.
- A human cell has many thousands of different proteins it can make. Surely it doesn't just churn out copies of all of them, all the time?
- Clearly there must be some sort of regulatory mechanism to control which proteins get made when.
- In the conventional picture, this is done by special proteins called **transcription factors** (TFs).
- Each TF recognizes and binds to a particular DNA sequence. Depending on the particular TF and the location where it binds, it can either increase or decrease the rate at which nearby genes are transcribed.
- The job of DNA is to encode proteins. Stretches of DNA (called genes) code for proteins using a simple, well-defined code.
- DNA is converted to RNA, which serves only as an information carrier. The RNA is then converted into proteins, which do all the real work.
- The whole process is very elegant. And for many years, this picture was believed to be mostly correct. But, that reality is actually far messier and far more complicated.

Winding chromosomes -- Histones

- **DNA molecules** are called **chromosomes**.
- In bacteria, which have relatively small genomes, DNA exists as simple free-floating molecules.
- But eukaryotes (a group that includes amoebas, humans, and everything in between) have much larger genomes. To fit inside the cell, each chromosome must be packed into a very small space. This is accomplished by winding it around proteins called **histones**.
- But if all the DNA is tightly packed away, how can it be transcribed? The answer, of course, is that it can't. Before a gene can be transcribed, the stretch of DNA containing it first **must be unwound**.
- How does the cell know which DNA to unwind? The answer is still poorly understood. It is believed to involve various types of chemical modification to the histone molecules, and proteins that recognize particular modifications.
- Clearly there is a regulatory mechanism involved, but many of the details are still unknown.

Methylation of DNA

- DNA itself can be chemically modified through a process called **methylation**.
- The more highly a stretch of DNA is methylated, the less likely it is to be transcribed, so is another regulatory mechanism the cell can use to control the production of proteins.
- But how does it control which regions of DNA are methylated? This too is still poorly understood.
- In the previous section we said that a particular stretch of DNA corresponds to a particular protein.
- That is correct for bacteria, but in eukaryotes the situation is more complicated.
- After the DNA is transcribed into a messenger RNA, that RNA often is edited to remove sections and connect (or splice) the remaining parts (called **exons**) back together again.
- The RNA sequence that finally gets translated into a protein may be different from the original DNA sequence.
- In addition, many genes have multiple splice variants—different ways of removing sections to form the final sequence.
- This means **a single stretch of DNA can actually code for several different proteins**!

Different Types of RNAs

- In the conventional picture RNA is viewed as just an information carrier, but even from the early days of genomics, biologists knew that was not entirely correct.
- The job of translating mRNA to proteins is performed by ribosomes, **complicated molecular machines made partly of proteins and partly of RNA**.
- Another key role in translation is performed by molecules called transfer RNAs (or tRNAs for short).
- These are the molecules that define the "**genetic code**," recognizing patterns of three bases in mRNA and adding the correct amino acid to the growing protein.
- So, for over half a century we've known there were at least three kinds of RNA: mRNA, ribosomal RNA, and tRNA.

RNAs

Columbia University

Over the last few decades, many other types of RNA have been discovered. Here are some examples:

• Micro RNAs (miRNAs) are short pieces of RNA that bind to a messenger RNA and prevent it from being translated into proteins. This is a very important regulatory mechanism in some types of animals, especially mammals.

• Short interfering RNA (siRNA) is another type of RNA that binds to mRNA and prevents it from being translated. It's similar to miRNA, but siRNAs are double stranded (unlike miRNAs, which are single stranded), and some of the details of how they function are different.

• Ribozymes are RNA molecules that can act as enzymes to catalyze chemical reactions. Chemistry is the foundation of everything that happens in a living cell, so catalysts are vital to life. Usually this job is done by proteins, but we now know it sometimes is done by RNA.

• Riboswitches are RNA molecules that consist of two parts. One part acts as a messenger RNA, while the other part is capable of binding to a small molecule. When it binds, that can either enable or prevent translation of the mRNA. This is yet another regulatory mechanism by which protein production can be adjusted based on the concentration of particular small molecules in the cell.

- all these different types of RNA must be manufactured, and the DNA must contain instructions on how to make them.

- So, DNA is more than just a string of encoded protein sequences. It also contains RNA sequences, plus binding sites for transcription factors and other regulatory molecules, plus instructions for how messenger RNAs should be spliced, plus various chemical modifications that influence how it is wound around histones and which genes get transcribed.

- Now consider what happens after the ribosome finishes translating the mRNA into a protein. Some proteins can spontaneously fold into the correct 3D shape, but many others require help from other proteins called **chaperones.**

- It is also very common for proteins to need additional chemical modifications after they are translated.

- Then the finished protein must be transported to the correct location in the cell to do its job, and finally degraded when it is no longer needed.

- Each of these processes is controlled by additional regulatory mechanisms, and involves interactions with lots of other molecules.

Deep Learning for genomics research

- If this all sounds overwhelming, that's because it is! A living organism is far more complicated than any machine ever created by humans. The thought of trying to understand it should intimidate you!
- But this is also why machine learning is such a powerful tool.
- We have huge amounts of data, generated by a process that is both mind-bogglingly complex and poorly understood.
- We want to discover subtle patterns buried in the data. This is exactly the sort of problem that deep learning excels at!
- In fact, deep learning is uniquely well suited to the problem. Classical statistical techniques struggle to represent the complexity of the genome.
- They often are based around simplifying assumptions. For example, they look for linear relationships between variables, or they try to model a variable as depending on only a small number of other variables.
- But genomics involves complex nonlinear relationships between hundreds of variables: exactly the sort of relationship that can be effectively described by a deep neural network.

Transcription Factor Binding

- As an example of applying deep learning to genomics, let's consider the problem of predicting transcription factor binding.
- Recall that TFs are proteins that bind to DNA. When they bind, they influence the probability of nearby genes being transcribed into RNA.
- But how does a TF know where to bind?
- Like so much of genomics, this question has a simple answer followed by lots of complications.
- To a first approximation, every TF has a specific DNA sequence called its binding site motif that it binds to.
- Binding site motifs tend to be short, usually 10 bases or less.
- Wherever a TF's motif appears in the genome, the TF will bind to it.
- In practice, though, motifs are not completely specific. A TF may be able to bind to many similar but not identical sequences. Some bases within the motif may be more important than others.
- This is often modeled as a position weight matrix that specifies how much preference the TF has for each possible base at each position within the motif.

Transcription Factor Binding

- Of course, that assumes every position within the motif is independent, which is not always true. Sometimes even the length of a motif can vary.

- And although binding is primarily determined by the bases within the motif, the DNA to either side of it can also have some influence. And that's just considering the sequence!

- Other aspects of the DNA can also be important. **Many TFs are influenced by the physical shape of the DNA**, such as how tightly the double helix is twisted.

- If the DNA is methylated, that can influence TF binding. And remember that most DNA in eukaryotes is tightly packed away, wound around histones. TFs can only bind to the portions that have been unwound.

- Other molecules also play important roles. TFs often interact with other molecules, and those interactions can affect DNA binding.

- For example, a TF may bind to a second molecule to form a complex, and that complex then binds to a different DNA
- motif than the TF would on its own.

- Biologists have spent decades untangling these details and designing models for TF binding. Instead of doing that, let's see if we can use deep learning to learn a model directly from data.

Convolutional Model for TF Binding

- For this example, we will use experimental data on a particular transcription factor called JUND.
- An experiment was done to identify every place in the human genome where it binds.
- To keep things manageable, we only include the data from chromosome 22, one of the smallest human chromosomes.
- It is still over 50 million bases long, so that gives us a reasonable amount of data to work with.
- The full chromosome has been split up into short segments, each 101 bases long, and each segment has been labeled to indicate whether it does or does not include a site where JUND binds.
- We will try to train a model that predicts those labels based on the sequence of each segment.

Convolutional Model for TF Binding

- The sequences are represented with one-hot encoding. For each base we have four numbers, of which one is set to 1 and the others are set to 0. Which of the four numbers is set to 1 indicates whether the base is an A, C, G, or T.

- To process the data we will use a convolutional neural network. In fact, you will see the two models are remarkably similar to each other.

- This time we will use 1D convolutions, since we are dealing with 1D data (DNA sequences) instead of 2D data (images), but the basic components of the model will be the same: inputs, a series of convolutional layers, one or more dense layers to compute the output, and a cross entropy loss function.

Convolutional Model for TF Binding

- Let's start by creating a layer to define the inputs:

```
features = tf.keras.Input(shape=(101, 4))
```

- For each sample, we have a feature vector of size 101 (the number of bases) by 4 (the one-hot encoding of each base).

- Next, we create a stack of three convolutional layers, all with identical parameters:

```
import tensorflow.keras.layers as layers
prev = features
for i in range(3):
    prev = layers.Conv1D(filters=15, kernel_size=10,
                         activation=tf.nn.relu,
                         padding='same')(prev)
    prev = layers.Dropout(rate=0.5)(prev)
```

- We specify 10 for the width of the convolutional kernels, and that each layer should include 15 filters (that is, outputs).

- The first layer takes the raw features (four numbers per base) as input. It looks at spans of 10 consecutive bases, so 40 input values in total. For each span, it multiplies those 40 values by a convolutional kernel to produce 15 output values.

- The second layer again looks at spans of 10 bases, but this time the inputs are the 15 values computed by the first layer. It computes a new set of 15 values for each base, and so on.

- To prevent overfitting, we add a dropout layer after each convolutional layer. The dropout probability is set to 0.5, meaning that 50% of all output values are randomly set to 0.

Convolutional Model for TF Binding

- Next we use a dense layer to compute the output:

```
logits = layers.Dense(units=1)(layers.Flatten()(prev))
output = layers.Activation(tf.math.sigmoid)(logits)
keras_model = tf.keras.Model(inputs=features, outputs=[output, logits])
```

- We want the output to be between 0 and 1 so we can interpret it as the probability a particular sample contains a binding site. The dense layer can produce arbitrary values, not limited to any particular range. We therefore pass it through a logistic sigmoid function to compress it to the desired range.

- The input to this function is often referred to as logits. The name refers to the mathematical logit function, which is the inverse function of the logistic sigmoid.

- Finally, we create a KerasModel, telling it to use the cross entropy as the loss function:

```
model = dc.models.KerasModel(
    keras_model,
    loss=dc.models.losses.SigmoidCrossEntropy(),
    output_types=['prediction', 'loss'],
    batch_size=1000)
```

- Notice that for reasons of numerical stability, the cross entropy loss is computed from logits instead of the output of the sigmoid function. We indicate this by specifying 'loss' as the output type for the second output (it will be passed to the loss function in place of the output used for making predictions.

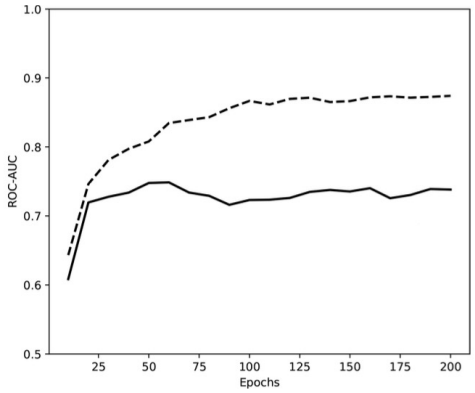
Convolutional Model for TF Binding

COLUMBIA UNIVERSITY

- Now we are ready to train and evaluate the model. We use ROC AUC as our evaluation metric. After every 10 epochs of training, we evaluate the model on both the training and validation sets:

```
train = dc.data.DiskDataset('train_dataset')
valid = dc.data.DiskDataset('valid_dataset')
metric = dc.metrics.Metric(dc.metrics.roc_auc_score)
for i in range(20):
    model.fit(train, nb_epoch=10)
    print(model.evaluate(train, [metric]))
    print(model.evaluate(valid, [metric]))
```

- The validation set performance peaks at about 0.75 after 50 epochs, then decreases slightly. The training set performance continues to increase, eventually leveling off at around 0.87. This tells us that training beyond 50 epochs just leads to overfitting, and we should halt training at that point:



Chromatin Accessibility

- The name chromatin refers to everything that makes up a chromosome: DNA, histones, and various other proteins and RNA molecules. Chromatin accessibility refers to how accessible each part of the chromosome is to outside molecules.

- When the DNA is tightly wound around histones, is becomes inaccessible to transcription factors and other molecules. They cannot reach it, and the DNA is effectively inactive. When it unwinds from the histones, it becomes accessible again and resumes its role as a central part of the cell's machinery.

- Chromatin accessibility is neither uniform nor static. It varies between types of cells and stages of a cell's life cycle. It can be affected by environmental conditions. It is one of the tools a cell uses to regulate the activity of its genome. Any gene can be turned off by packing away the area of the chromosome where it is located.

Chromatin Accessibility

COLUMBIA UNIVERSITY

- Accessibility also is constantly changing as DNA winds and unwinds in response to events in the cell. Instead of thinking of accessibility as a binary choice (accessible or inaccessible), it is better to think of it as a continuous variable (what fraction of the time each region is accessible).

- The data we analyzed in the last section came from experiments on a particular kind of cell called HepG2. The experiments identified locations in the genome where the transcription factor JUND was bound.

- The results were influenced by chromatin accessibility. If a particular region is almost always inaccessible in HepG2 cells, the experiment was very unlikely to find JUND bound there, even if the DNA sequence would otherwise be a perfect binding site. So, let's try incorporating accessibility into our model.

- First let's load some data on accessibility. We have it in a text file where each line corresponds to one sample from our dataset (a 101-base stretch of chromosome 22).

- A line contains the sample ID followed by a number that measures how accessible that region tends to be in HepG2 cells. We'll load it into a Python dictionary:

```
span_accessibility = {}
for line in open('accessibility.txt'):
  fields = line.split()
  span_accessibility[fields[0]] = float(fields[1])
```

Chromatin Accessibility

- Now to build the model. We will use almost exactly the same model as in the previous section with just two minor changes. First, we need a second feature input for the accessibility values. It has one number for each sample:

```
accessibility = tf.keras.Input(shape=(1,))
```

- Now we need to incorporate the accessibility value into the calculation. There are many ways we might do this. For the purposes of this example, we will use a particularly simple method. In the previous section, we flattened the output of the last convolution layer, then used it as the input to a dense layer that calculated the output.

```
logits = layers.Dense(units=1)(layers.Flatten()(prev))
```

- This time we will do the same thing, but also append the accessibility to the output of the convolution:

```
prev = layers.Concatenate()([layers.Flatten()(prev), accessibility])
logits = layers.Dense(units=1)(prev)
```

- That's all there is to the model! Now it's time to train it.

- At this point we run into a difficulty: our model has two different Input layers! Up until now, our models have had exactly one Input layer. We trained them by calling fit(dataset), which automatically connected the dataset's X field to the feature input. But that clearly can't work when the model has more than one set of features.

Chromatin Accessibility

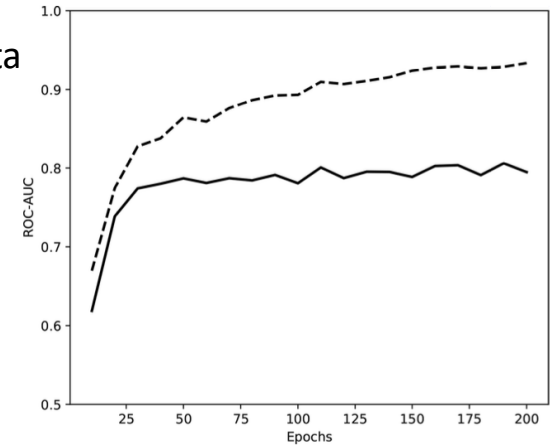
- This situation is handled by using a more advanced feature of DeepChem. Instead of passing a dataset to the model, we can write a Python generator function that iterates over batches.

- Each batch is represented by a tuple of the form (inputs, labels, weights), where each entry is a list of NumPy arrays to pass to the model or loss function:

```
def generate_batches(dataset, epochs):
  for epoch in range(epochs):
    for X, y, w, ids in dataset.iterbatches(batch_size=1000,
                                            pad_batches=True):
      yield ([X, np.array([span_accessibility[id] for id in ids])], [y], [w])
```

- Notice how the dataset takes care of iterating through batches for us. It provides the data for each batch, from which we can construct whatever inputs the model requires. Training and evaluation now proceed exactly as before. We use alternate forms of the methods that take a generator instead of a dataset:

```
for i in range(20):
    model.fit_generator(generate_batches(train, epochs=10))
    print(model.evaluate_generator(generate_batches(train, 1), [metric]))
    print(model.evaluate_generator(generate_batches(valid, 1), [metric]))
```



RNA Interference

- For our final example, let's turn to RNA. Much like DNA, this is a polymer composed of four repeating units called bases. In fact, three of the four bases are almost identical to their DNA versions, differing only in having one extra oxygen atom.
- The fourth base is a little more different. In place of thymine (T), RNA has a base called uracil (U). When a DNA sequence is transcribed into RNA, every T is replaced by a U.
- The bases G and C are complementary to each other, in the sense that they have a strong tendency to bond to each other. Likewise, the bases A and T (or U) are complementary. If you have two strands of DNA or RNA, and every base in one is complementary to the corresponding base in the other, the two strands will tend to stick together.
- This fact plays a key role in lots of biological processes, including both transcription and translation, as well as DNA replication when a cell is dividing.
- It also is central to something called RNA interference. This phenomenon was only discovered in the 1990s, and the discovery led to a Nobel Prize in 2006. A short piece of RNA whose sequence is complementary to part of a messenger RNA can bind to that mRNA.
- When this happens, it "silences" the mRNA and prevents it from being translated into a protein. The molecule that does the silencing is called a short interering RNA (siRNA).

RNA Interference

- Except there is much more to the process than that. RNA interference is a complex biological mechanism, not just a side effect of two isolated RNA strands happening to stick together.
- It begins with the siRNA binding to a collection of proteins called the RNA-induced silencing complex (RISC). The RISC uses the siRNA as a template to search out matching mRNAs in the cell and degrade them.
- This serves both as a mechanism for regulating gene expression and as a defense against viruses.
- It also is a powerful tool for biology and medicine. It lets you temporarily “turn off ” any gene you want. You can use it to treat a disease, or to study what happens when a gene is disabled. Just identify the mRNA you want to block, select any short segment of it, and create a siRNA molecule with the complementary sequence.
- Except that (of course!) it isn’t as simple as that. You can’t really just pick any segment of the mRNA at random, because (of course!) RNA molecules aren’t just abstract patterns of four letters. They are physical objects with distinct properties, and those properties depend on the sequence.
- Some RNA molecules are more stable than others. Some bind their complementary sequences more strongly than others. Some fold into shapes that make it harder for the RISC to bind them. This means that some siRNA sequences work better than others, and if you want to use RNA interference as a tool, you need to know how to select a good one!

RNA Interference

- Biologists have developed lots of heuristics for selecting siRNA sequences. They will say, for example, that the very first base should be either an A or G, that G and C bases should make up between 30% and 50% of the sequence, and so on.

- These heuristics are helpful, but let's see if we can do better using machine learning.

- We'll train our model using a library of 2,431 siRNA molecules, each 21 bases long.1 Every one of them has been tested experimentally and labeled with a value between 0 and 1, indicating how effective it is at silencing its target gene. Small values indicate ineffective molecules, while larger values indicate more effective ones. The model takes the sequence as input and tries to predict the effectiveness.

- Here is the code to build the model:

```
features = tf.keras.Input(shape=(21, 4))
prev = features
for i in range(2):
    prev = layers.Conv1D(filters=10, kernel_size=10,
                         activation=tf.nn.relu,
                         padding='same')(prev)
    prev = layers.Dropout(rate=0.3)(prev)
output = layers.Dense(units=1, activation=tf.math.sigmoid)(layers.Flatten()(prev))
keras_model = tf.keras.Model(inputs=features, outputs=output)
model = dc.models.KerasModel(
    keras_model,
    loss=dc.models.losses.L2Loss(),
    batch_size=1000)
```

RNA Interference

- This is very similar to the model we used for TF binding, with just a few differences. Because we are working with shorter sequences and training on less data, we have reduced the size of the model. There are only 2 convolutional layers, and 10 filters per layer instead of 15.

- We also use a different loss function. The model for TF binding was a classification model. Every label was either 0 or 1, and we tried to predict which of those two discrete values it was. But this one is a regression model. The labels are continuous numbers, and the model tries to match them as closely as possible. We therefore use the L2 distance as our loss function, which tries to minimize the mean squared difference between the true and predicted labels.

- Here is the code to train the model:

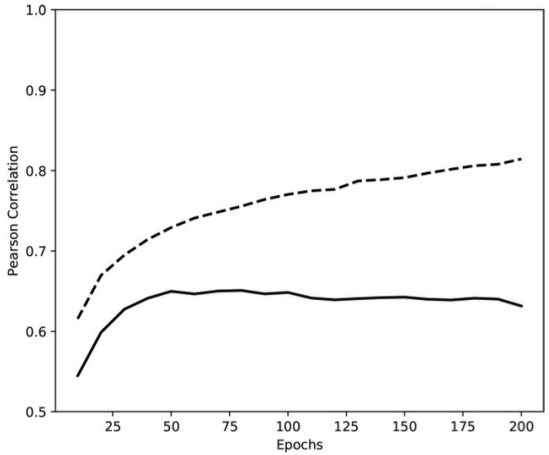
```
train = dc.data.DiskDataset('train_siRNA')
valid = dc.data.DiskDataset('valid_siRNA')
metric = dc.metrics.Metric(dc.metrics.pearsonr, mode='regression')
for i in range(20):
    model.fit(train, nb_epoch=10)

print(model.evaluate(train, [metric])['pearsonr'])
print(model.evaluate(valid, [metric])['pearsonr'])
```

RNA Interference

- For TF binding, we used ROC AUC as our evaluation metric, which measures how accurately the model can divide the data into two classes. That is appropriate for a classification problem, but it doesn't make sense for a regression problem, so instead we use the Pearson correlation coefficient. This is a number between –1 and 1, where 0 means the model provides no information at all and 1 means the model perfectly reproduces the experimental data.

- The validation set score peaks at 0.65 after 50 epochs. The training set score continues to increase, but since there is no further improvement to the validation set score this is just overfitting. Given the simplicity of the model and the limited amount of training data, a correlation coefficient of 0.65 is quite good. More complex models trained on larger datasets do slightly better, but this is already very respectable performance.



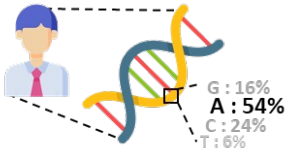
Graphen Health Services

-- Precision Health and Precision Medicine

General Consumers

1 *What am I?*

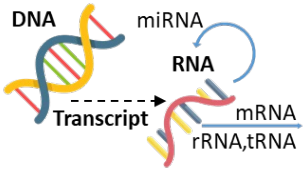
Born: Whole Genome Analysis



Use Whole Genome Sequencing (WGS) to understand the innate DNA genetic makeup, and analyze the association between individuals and disease risk through Single Nucleotide polymorphisms/variations (SNPs/SNVs) , providing an entry point for early diagnosis.

2 *How am I?*

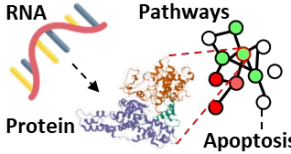
Now: Gene Expression Analysis



Epigenetic genetic behavior and regulations determine when and how extent genes affect us. Graphen further uses personal genetic behavior testing to understand current genetic behavior trends and analyze disease risks caused by acquired factors such as the environment. It also provides customers with existing disease states with the cause and course of the current disease.

3 *What will I be?*

Future: Pathways Dynamics

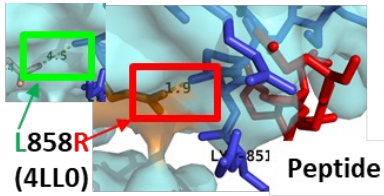


The dynamic expression of proteins in the biological signal transmission chain shows how genes affect body status and disease development. By combining genomic analysis with protein interaction networks, Graphen can instantly understand the correlation between key genes and disease processes.

Patients

4 *I want the most suitable drugs to me!!*

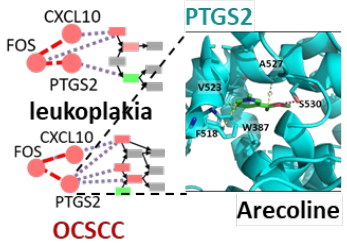
Gene/Protein-Mutation & Resistance Analysis



Genetic variation is inseparable from mutations and the occurrence of disease. Grphen understands how mutations in genes and proteins change the normal working of the body from the molecular structure level. In addition to elucidating the relationship between mutations and actual disease symptoms, Grphen also establishes the foundation for subsequent precision medicine and extends the development of dynamic disease classification and treatment strategies.

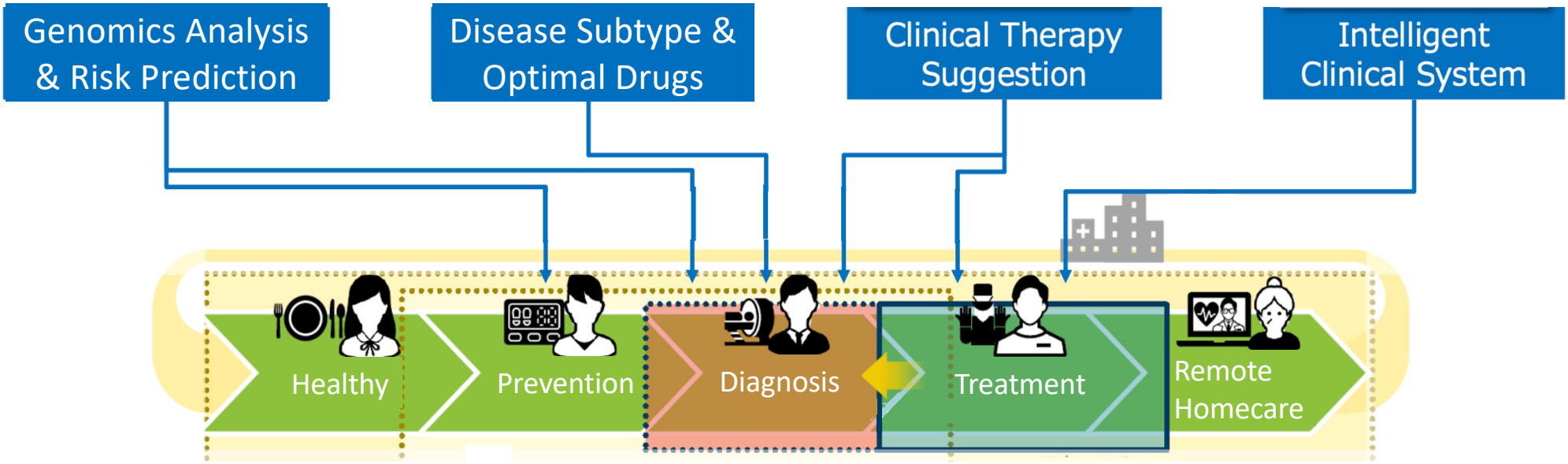
5 *Happening again?*

Prognosis Analysis



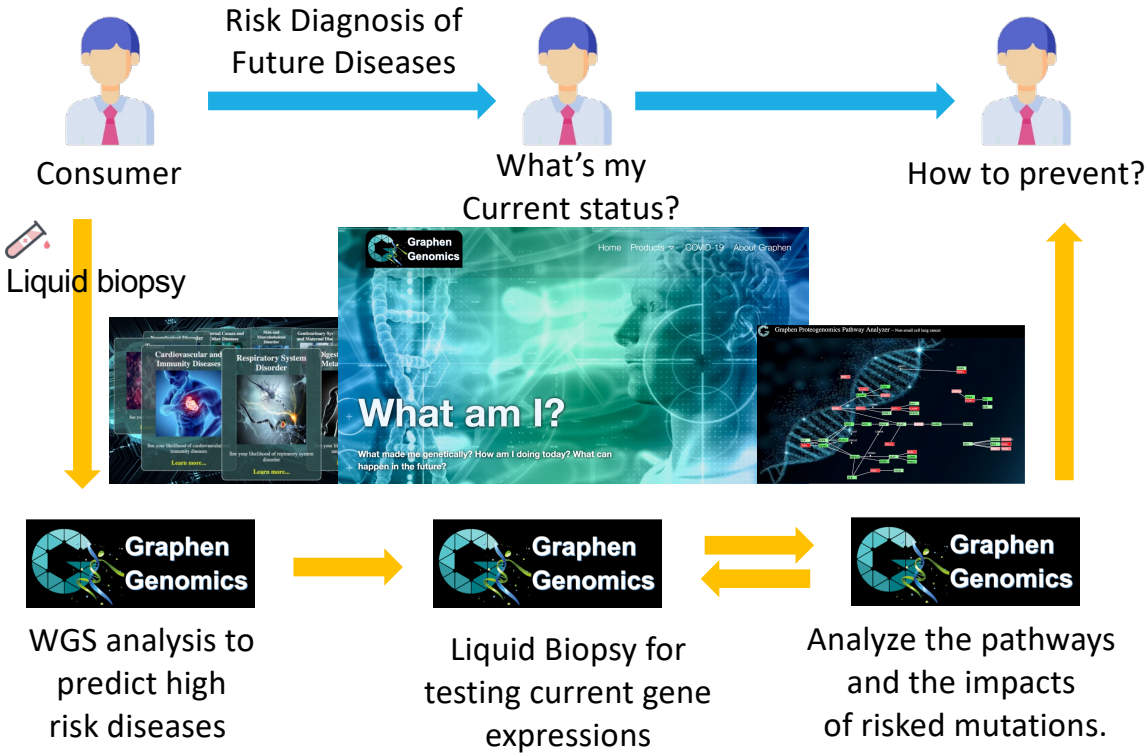
Graphen combines Genomics and Atoms analysis to develop multi-faceted disease classification services and predictions (gene & protein differences, biochemical pathway transmission chain patterns, structural mutations) in precision medicine, and provide analysis of prognosis prediction and monitoring strategies

Graphen Health helps health journey



Graphen Precision Health

Provide Predictive Health Monitoring Strategy for General Consumers



Service Goals

With Once-a-Lifetime Whole Genome Sequencing, a healthy or subhealthy person can understand his or her natural health risks.

With liquid biopsy of dynamic genome expression tests, we provide health risk monitoring for consumers.

Integrated Services

Understanding a patient's particular risk through WGS, and analyzing related-risky gene/targets. Further presenting the disease pathway and the influence of mutation. Periodically monitoring and giving practical suggestion of predictive health.

Graphen Precision Medicine

Based on the subtype of disease, and genome, providing treatment strategy to Patients.

Patient

Clinical pathologic diagnosis

Regular process

Therapeutic :
Chemotherapy/ Targeted therapy
Effect :
Survival ~13 months

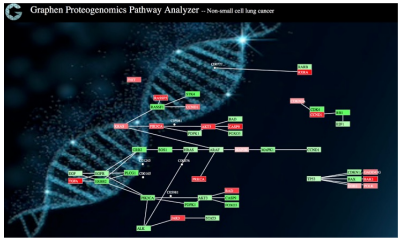
Not suitable

Change therapy

Painful
No progress

Precision Health process

Hospital

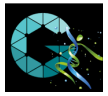


Liquid biopsy

Tissue

NGS company

Patient's gene sequencing



Graphen Precision Medicine

Analyze the gene behavior of your current disease status

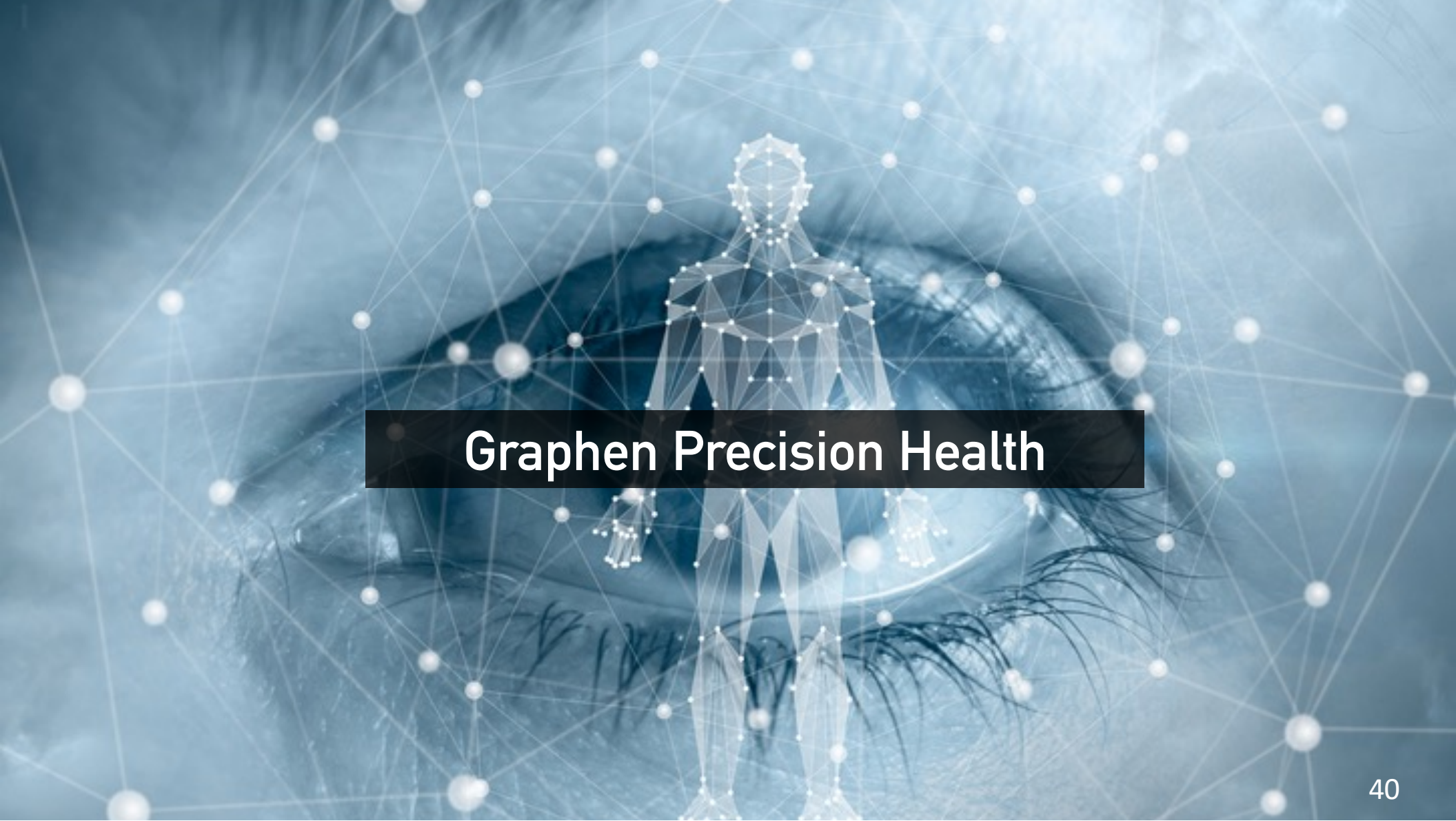
Provide disease status and clinical therapy information based on your gene behavior

Service Goal

Using cancer as example, through genome testing and expression, combining with WGS or WES results to provide the optimum drug suggestions. Considering drug-resistance, absorption, effectiveness, etc.

Integrated Service

After receiving Patient's genome sequencing data, using Graphen's dynamic pathway and mutation analysis to find the best protein target for the disease. Using drug-resistance to suggest to most appropriate treatment and drug usage information.

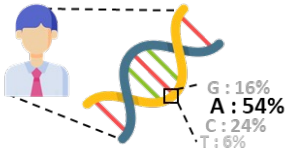

Graphen Precision Health

Graphen Precision Health

Biology Central Dogma from Gene to Protein Expression

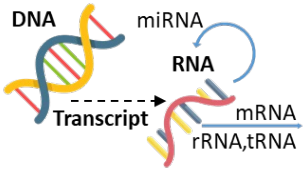
General Consumers

Born: Whole Genome Analysis



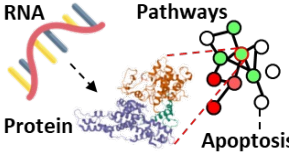
Use Whole Genome Sequencing (WGS) to understand the innate DNA genetic makeup, and analyze the association between individuals and disease risk through Single Nucleotide polymorphisms/variations (SNPs/SNVs) , providing an entry point for early diagnosis.

Now: Gene Expression Analysis



Epigenetic genetic behavior and regulations determine when and how extent genes affect us. Graphen further uses personal genetic behavior testing to understand current genetic behavior trends and analyze disease risks caused by acquired factors such as the environment. It also provides customers with existing disease states with the cause and course of the current disease.

Future: Pathways Dynamics

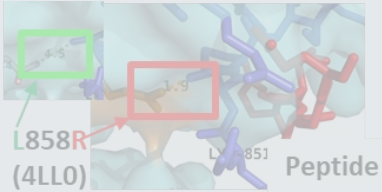


The dynamic expression of proteins in the biological signal transmission chain shows how genes affect body status and disease development. By combining genomic analysis with protein interaction networks, Graphen can instantly understand the correlation between key genes and disease processes.

Molecular-Level Precision Medicine Analysis

Patients

Gene/Protein-Mutation & Resistance Analysis

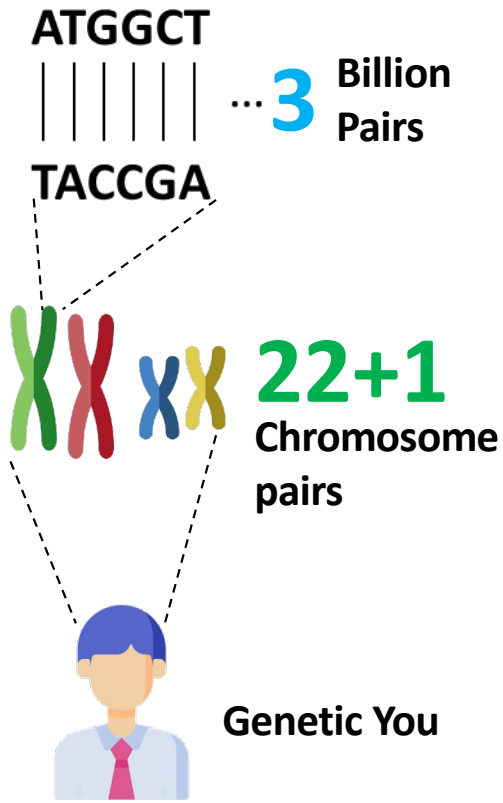


Genetic variation is inseparable from mutations and the occurrence of disease. Grphen understands how mutations in genes and proteins change the normal working of the body from the molecular structure level. In addition to elucidating the relationship between mutations and actual disease symptoms, Grphen also establishes the foundation for subsequent precision medicine and extends the development of dynamic disease classification and treatment strategies.

Prognosis Analysis

Graphen combines Genomics and Atoms analysis to develop multi-faceted disease classification services (gene & protein differences, biochemical pathway transmission chain patterns, structural mutations) in precision medicine, and provide analysis of drug resistance factors and information on today's most appropriate treatment strategies.

1. Whole Genome Sequencing Analysis



Genes dominate the physiological development of human life. We form our own unique genetic makeup by inheriting the genes of our parents. In addition to the acquired environment and other factors, humans can be said to be composed of up to 3 billion base pair "codes", and everyone's "codes" are slightly different.

In addition to making each person an independent individual, these small differences may also make us innately prone to certain diseases, which we used to call "physique" and can be understood through "family medical history." And now we can understand the potential risk of disease from a person's genetic information. Such as genetic diseases, diabetes, cardiovascular diseases, cancer, etc., by digitizing genetic information, we have the opportunity to understand the current process before the disease develops, and prevent and control it early, so as to regain the possibility of future health.

10 categories of 350+ diseases-a comprehensive understanding of your disease risk

With only one full genetic test, Graphen can analyze more than 350 disease risks and summarize your Digital Risk Profile in ten categories:

- Respiratory diseases
- Nervous system disease
- Digestive system diseases
- Cardiovascular and immune diseases
- Cancer, etc.

Let customers fully understand the possible future diseases, as well as the doubts about the past family medical history, and immediately take response measures and adjustment strategies.

Whole Genome Analysis Results

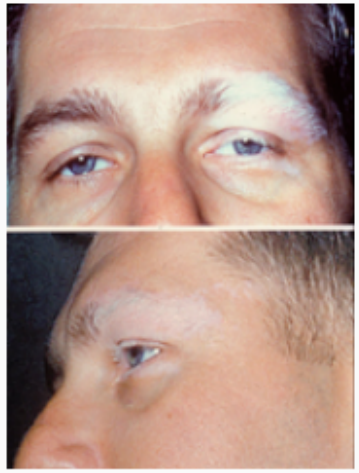
Display Language: English

Select Race: Any

Low Risk (0 - 30) | Mid Risk (30 - 60) | High Risk (60 - 100)

Potential Disease	潛在疾病	Category	疾病類別	Known Risky Alleles	Your Matched Alleles	Allele Match Ratio	Risk Score Total
age-related macular degeneration	年齡相關性黃斑變性	Eye, Ear and Mastoid disorder	眼耳鼻喉疾病	63	19	30%	56.57
exfoliation syndrome	剝脫綜合徵	Eye, Ear and Mastoid disorder	眼耳鼻喉疾病	6	1	17%	41.25
wet macular degeneration	濕性黃斑變性	Eye, Ear and Mastoid disorder	眼耳鼻喉疾病	17	4	24%	27.38
vogt-koyanagi-harada disease	沃格特-小柳-原田病	Eye, Ear and Mastoid disorder	眼耳鼻喉疾病	4	1	25%	20.45
central serous retinopathy	中央性漿液性視網膜病	Eye, Ear and Mastoid disorder	眼耳鼻喉疾病	2	2	100%	19.85
low tension glaucoma	低張力青光眼	Eye, Ear and Mastoid disorder	眼耳鼻喉疾病	2	2	100%	8.61
open-angle glaucoma	開角型青光眼	Eye, Ear and Mastoid disorder	眼耳鼻喉疾病	83	35	42%	8.45

Vogt–Koyanagi–Harada disease (VKH) is a multisystem disease of presumed autoimmune cause that affects pigmented tissues, which have melanin. The most significant manifestation is bilateral, diffuse uveitis, which affects the eyes. VKH may variably also involve the inner ear, with effects on hearing, the skin and the meninges of the central nervous system.



Close

birdshot chorioretinopathy	鳥狀脈絡膜視網膜病變	Eye, Ear and Mastoid disorder	眼耳鼻喉疾病	1	1	100%	6.33

Graphen Genomics

Risky Genome Mutations

Chrom.	Position	Your Genotype	In Gene	In Coding Area	Reference_SNP ID	Risk Allele	Your Risk Allele Match	Risk Allele Frequency	P-Value (mlog) ⇩	Potential Disease	潜在疾病	Category	疾病類別
6	32575658	GC			rs3021304-G	G	1	0.35	118	vogt-koyanagi-harada disease	沃格特-小柳-原田病	Eye, Ear and Mastoid disorder	眼耳鼻喉疾病

Graphen Genomics

Whole Genome Analysis by Risk Allele

Genome-wide association analysis of Vogt-Koyanagi-Harada syndrome identifies two new susceptibility loci at 1p31.2 and 10q21.3.

Hou S *et al.*

Nat Genet, 2014-08-10, *citation #*: 31

KNOW MORE..

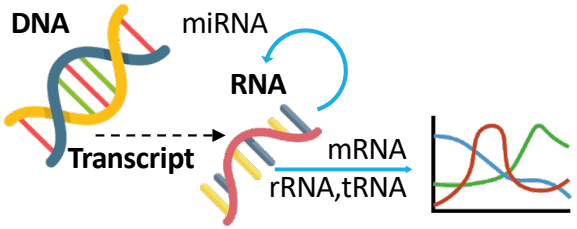
Genome-wide association analysis of Vogt-Koyanagi-Harada syndrome identifies two new susceptibility loci at 1p31.2 and 10q21.3.

To identify new genetic risk factors for Vogt-Koyanagi-Harada (VKH) syndrome, we conducted a genome-wide association study of 2,208,258 SNPs in 774 cases and 2,009 controls with follow-up in a collection of 415 cases and 2,006 controls and a further collection of 349 cases and 1,588 controls from a Han Chinese population. We identified three loci associated with VKH syndrome susceptibility (IL23R-C1orf141, rs117633859, P(combined) = 3.42 10(-21), odds ratio (OR) = 1.82; ADO-ZNF365-EGR2, rs442309, P(combined) = 2.97 10(-11), OR = 1.37; and HLA-DRB1/DQA1, rs3021304, P(combined) = 1.26 10(-118), OR = 2.97). The five non-HLA genes were all expressed in human iris tissue. IL23R was also expressed in the ciliary body, and EGR2 was expressed in the ciliary body and choroid. The risk G allele of rs117633859 in the promoter region of IL23R exhibited low transcriptional activation in a cell-based reporter assay and was associated with diminished IL23R mRNA expression in human peripheral blood mononuclear cells.

Read Paper

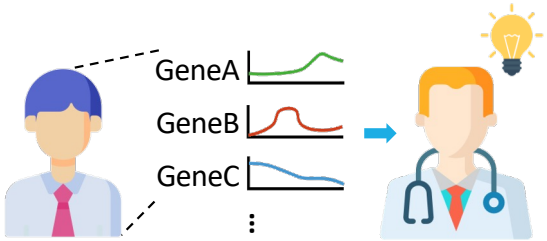
CLOSE

2. Graphen Gene Dynamic Detection Product Features



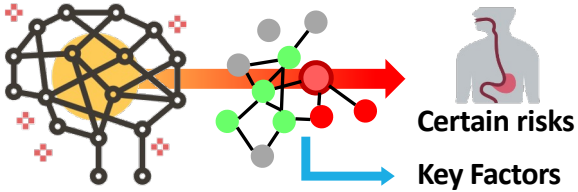
Genetic performance over time: Understand your genetic performance trends

Innate genetic composition contains possible future disease risks, but factors such as environmental factors, living habits, and bacterial virus infections cause changes in acquired gene expression over time, which are the main reasons for triggering diseases. Use genetic tracking to understand the trend of your genetic performance and grasp the safe distance between you and the disease.



Long-term tracking genetic resume: Your prevention and treatment strategy protection

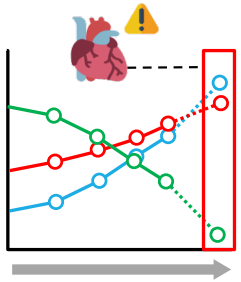
Differences in gene expression determine which lifestyle can reduce your risk of disease, and also affect the most appropriate treatment strategy when you are sick. Through long-term monitoring of your genetic performance trends, Graphen can provide tracking resumes for the first time to help doctors choose the best strategy for you, increase the quality of life and prevent the occurrence of diseases.



Advanced artificial intelligence technology: uncover the key factors and accurately prevent

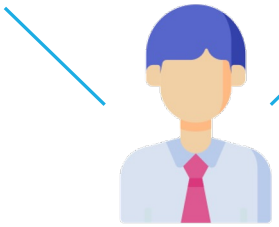
Graphen's AI technology analyzes the human's physiological transmission chain, gene and protein mutations, can judge how your gene performance causes specific disease risks, and design precise gene probes to find out key potential factors, see the best, and provide you with precise prevention strategies.

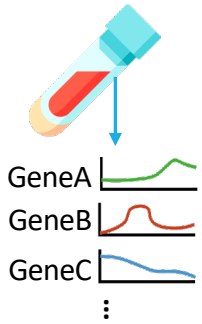
Graphen Gene Dynamic Detection Product Advantages



Pre-zero disease progression tracking

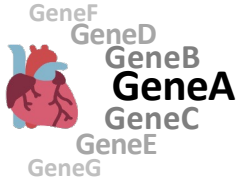
Continue to track changes in gene expression, allowing you to grasp the distance to the underlying disease, and accurate prevention in advance.





Traceable by blood draw

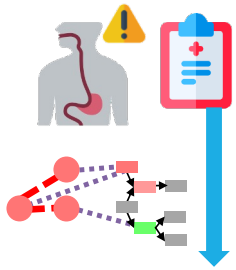
There is no need for complicated testing items, procedures and high costs, and blood draws can accurately track your health progress on a regular basis.



~100 plex

Up to one hundred kinds of target diversity detection

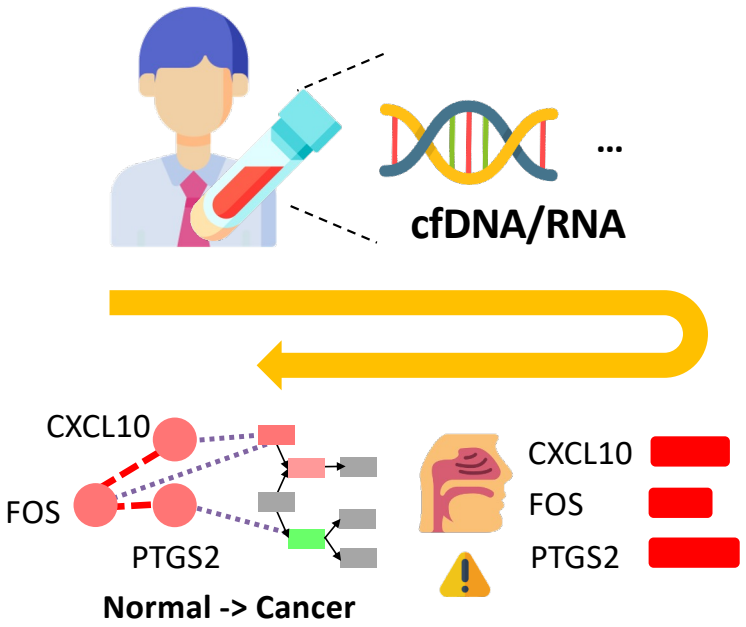
Up to one hundred targets can be evaluated in a single time, without repeated testing, and you can accurately understand your genetic performance trends.



Support health check follow-up evaluation

It can continue to track the risks of specific diseases after the health check, assisting individuals and corporate employees in the health management assessment.

Gene performance dynamics and liquid biopsy

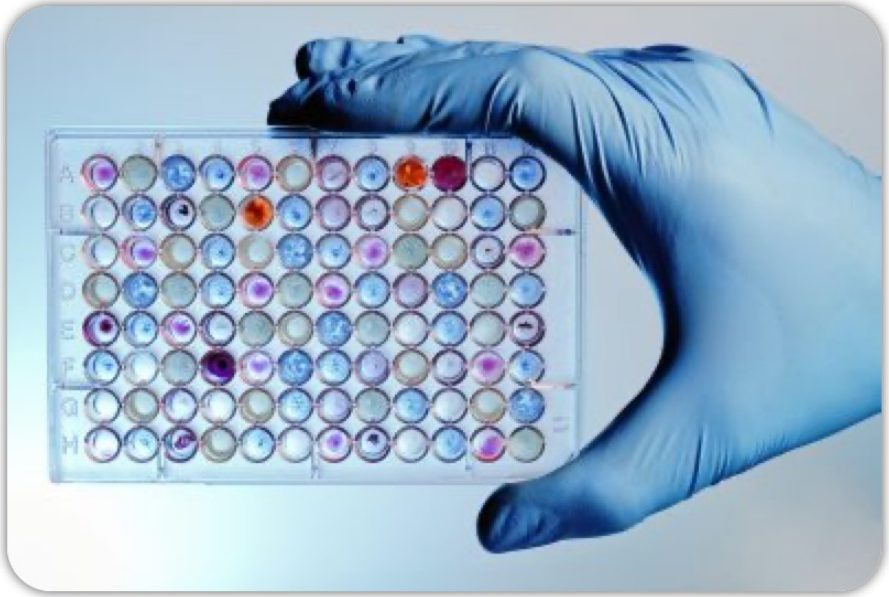


In order to adapt to factors such as living environment, daily habits, and coping with diseases and infections, the human body will adjust the way and degree of gene expression over time and physiological status to form gene dynamics. For example, when infected by viruses or bacteria, immune-related genes will be more active than usual to resist the invasion of foreign pathogens.

In terms of diseases, liquid biopsy captures circulating blood nucleic acid (cfDNA/RNA) by collecting blood. In a simple and less invasive way, analyze the subject's current gene dynamic behavior.

After the subject understands his innate genetic risk, through the dynamic behavior of gene expression, we can evaluate the distance between the subject and the disease status and take effective preventive measures.

Nearly a hundred target detections in a single test- accurately and quickly solve the needs of multiple diseases

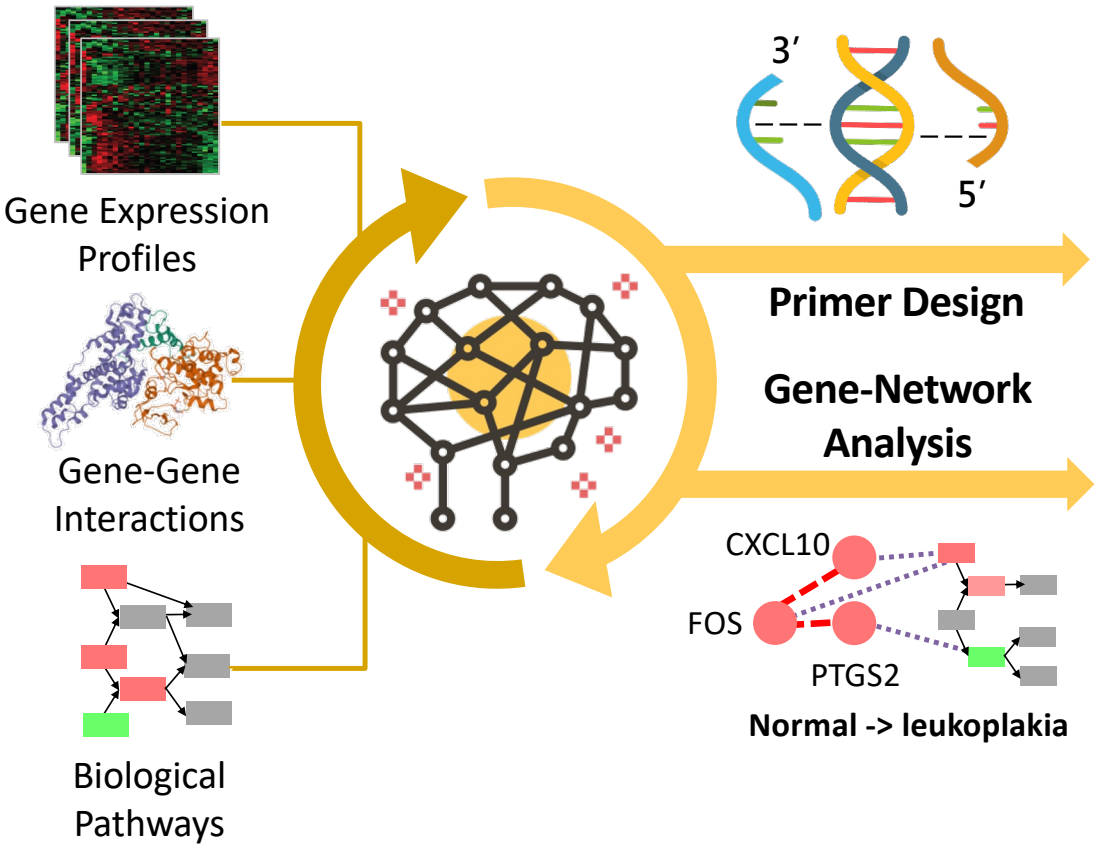


Graphen develops artificial intelligence and combines Multiplex PCR technology to break through the traditional PCR's limitation in the number of single target detections; a single detection can test up to nearly a hundred targets, expanding the depth and depth of gene mutations in disease risk analysis.

The breadth of multiple disease-related genes. With faster detection speed and lower detection cost, it meets the needs of modern personalized precision medicine.

Working with LiefOS

3. Graphen Health Risk Prediction Services: identify key risk factors and predict progress



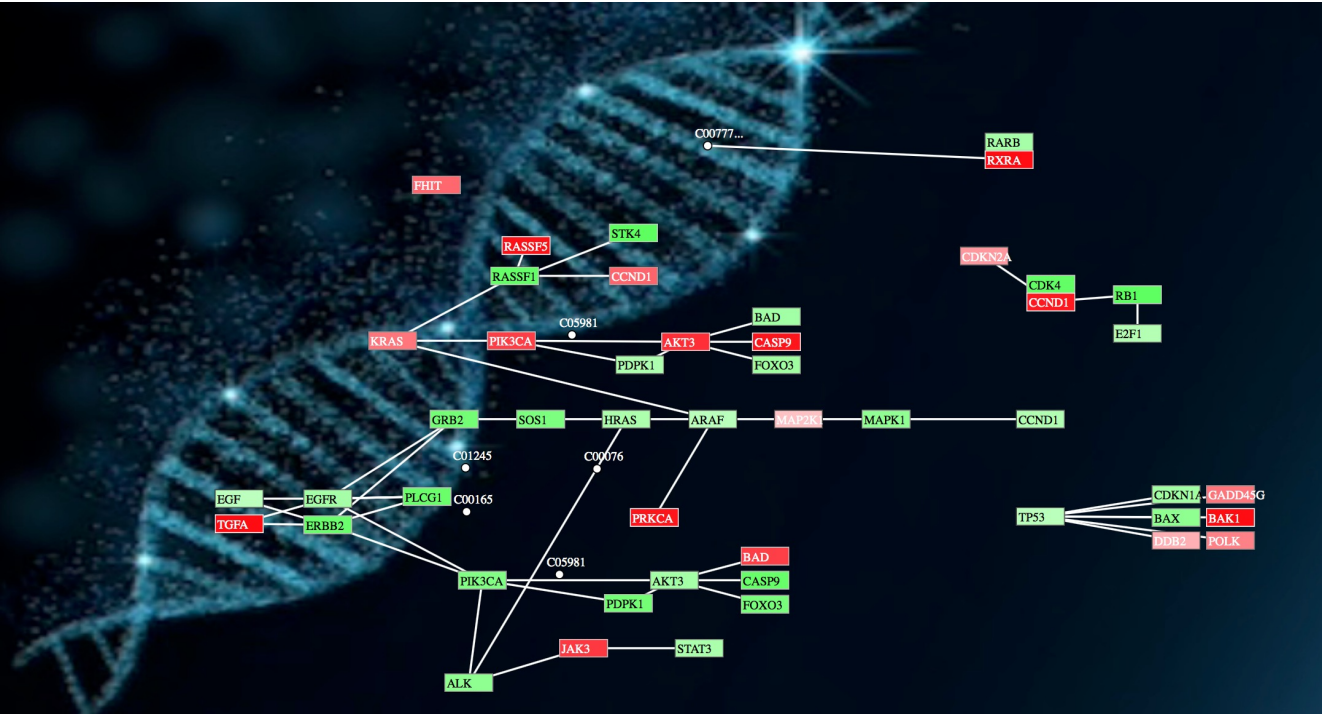
Graphen develops advanced artificial intelligence technology, applied to gene expression network analysis and primer design specially used for multi-target detection, to find out the key factors that cause your future disease risk.

Technology

- Network analysis/prediction of gene-biological delivery chain
- Gene/Translated Protein Mutation Analysis
- Gene target primer design/optimization
- Multi-target primer adhesion prediction

Graphen provides high-precision, low-false positive detection results, and at the same time to find out the key risk factors at multiple levels.

Gene dynamic performance and biological signal transduction-integrated gene interaction evaluation



Graphen integrates the Biological Signaling Pathways network formed by the interaction between genes into the evaluation process, linking current gene performance with future potential disease risks:

- Fold change rate of gene expression
- Upstream and downstream associated gene expression rate
- The delivery chain initiates the classification of biological functions

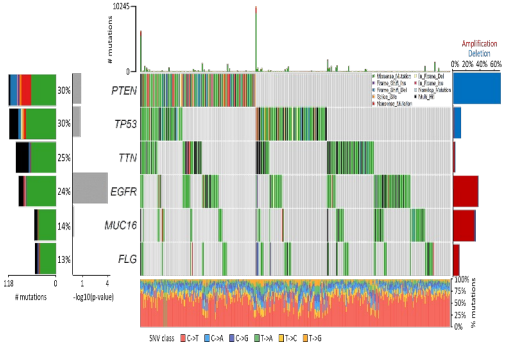
Graphen not only observes the change trend of individual genes, but also evaluates your health-disease progress based on the overall changes in your personal gene network, which is accurate and constructive.

Summary of Graphen Precision Health Services

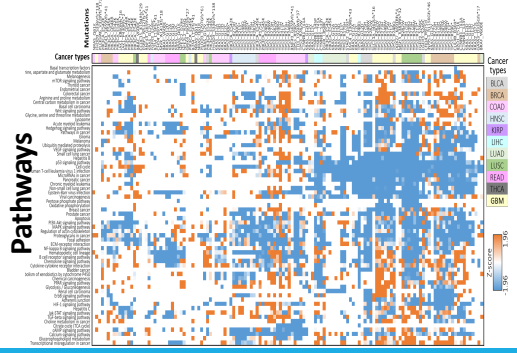
Columbia University

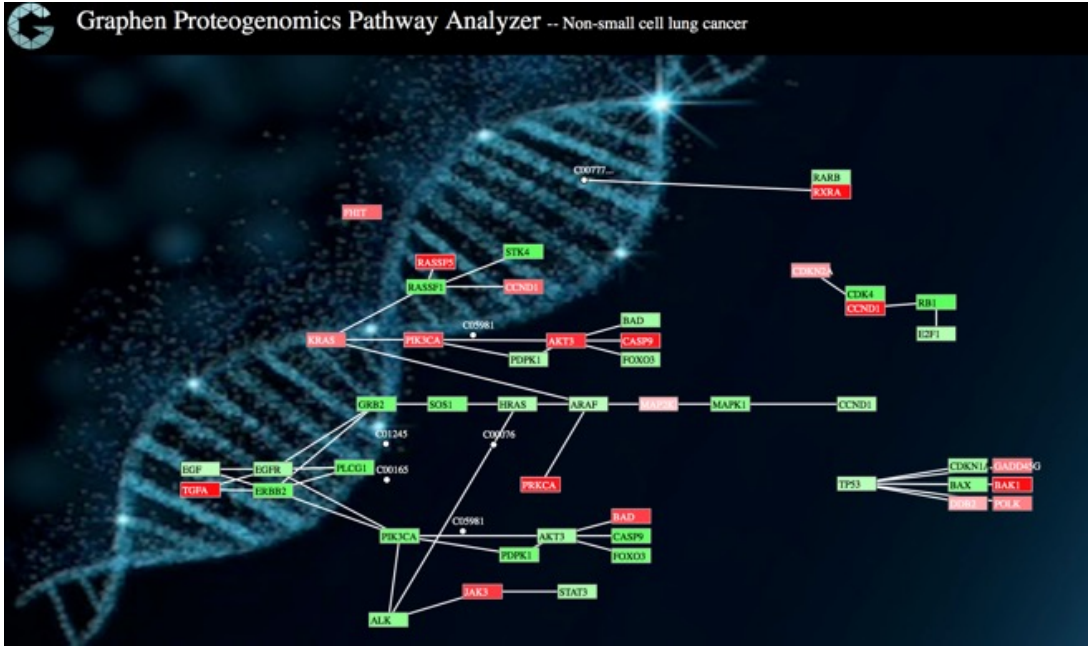
Graphen Precision Health

Mutation summary report



Pathway analysis





- Utilize **Genomics or Proteomics Testing** to extract key information of a person's 'Today' status.
- Utilize **Pathway Analytics** to predict future progress.
- Utilize Large-Scale **AI Article Reasoning** to gain **collective knowledge** of worldwide researchers and practitioners.

➢ Identify personal disease pathways

Graphen's biomarker identification technology published in *BMC Bioinformatics* (2022)

Research | Open Access | Published: 19 April 2022

CoMI: consensus mutual information for tissue-specific gene signatures

Sing-Han Huang

Present address: Graphen Inc., New York, NY, 10110, USA

BMC Part of Springer Nature

BMC Bioinformatics