GeneChip® Microarrays

Structure & Function of GeneChip Microarrays

Part I – Introduction and GeneChip Expression Microarrays

The **Human Genome Project** revealed much about our genetic sequence and discovered everyone's **DNA** to be 99.9% identical. However, even very small differences in DNA sequence can generate very big differences in **gene expression**, often having important biological implications. The same **gene** may be highly expressed in one person, but minimally or not expressed in another. This difference may be due to a small **mutation** or the influence of the environment that the organism lives in. For example, a person may begin to make more dark pigment in their skin by expressing the **melanin** gene due to an increase exposure to the sun. Every cell in our body contains the same **DNA sequence**. This is because each cell in the body of a multicellular organism came from the division of the original single cell that came from the fertilization of the reproductive cells from the parents. Despite this, every gene is not expressed in every cell. Some important genes such as those needed to extract energy from food molecules are active in all cells. While those genes producing a protein that protects the skin from harmful ultraviolet rays will only be expressed in skin cells.

Gene expression microarrays are powerful tools that help scientists study cellular genetics in a way never imagined before. Researchers study gene expression by measuring the amount of **RNA** copies a gene produces. When a gene is expressed it produces RNA which will help with the production of the final protein coded for by the gene. The gene expression microarray is a tool that tells them how much RNA a gene is making, if it's making any at all. An active gene producing a lot of RNA must be important for that specific cell type. Studying gene expression can help with understanding of cell function and lead to treatment of diseases in a much more specific manner than what is traditionally done today. Scientists use microarrays to try and find a link between a gene or group of genes and a disease, as well as to develop drugs to treat the disease and medical tests to diagnose the disease.

Understanding Disease

Gene expression arrays have been used to study virtually every type of disease. For simplicity's sake, however, let's use a hypothetical example to demonstrate a common way arrays might be used. Suppose that the public outcry over rude cell phone behavior – loud, inane, and inappropriate conversations in public places – reaches such a pitch that the government decides to fund investigative research looking for a cause and a treatment. Scientists hope to find some sort of genetic basis explaining the behavioral phone disorder and to then develop a drug that can treat it.

Genes and Disease

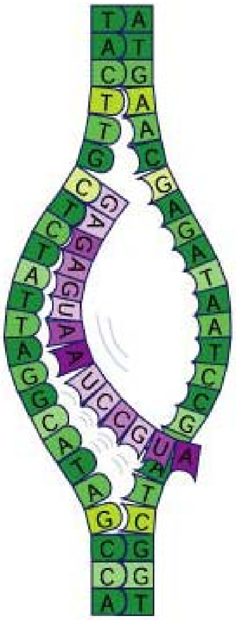
To find a treatment, researchers first have to find some sort of expression difference that inhibits the "polite" and "common sense" regions of the brain. Before the advent of microarrays, scientists would have searched previous academic research for genetic links to other forms of rude behavior, like talking in movies or standing in the express checkout lane with 25 items. They would have formed a hypothesis based on this previous research, looking for a genetic link to rude cell phone talkers. Scientists would test hundreds or even thousands of hypotheses in a process that was slow, expensive, tedious, and many times ended with inconclusive results. The revolutionary thing about microarrays is that they allow researchers to start without any hypothesis, or guesswork, at all. Arrays enable researchers to measure the expression of every gene in the entire human genome, even genes that have unknown functions. So by simply comparing the gene expression patterns of, say, 100 polite cell phone users and 100 rude cell phone users, scientists can quickly pinpoint differences in the patterns. If they see a pattern of genes consistently expressed or not expressed by rude cell phone talkers, that's where they start their research.

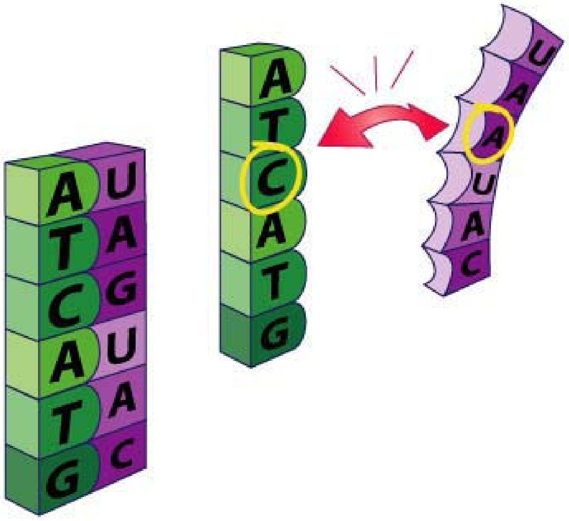
How the Array Works

While that may sound simple enough, it's been a long road to figure out just what genes are doing. It wasn't until 1963 that people even knew genes expressed different amounts of RNA. And it wasn't until 1977 that scientists had a practical tool to measure gene expression by a technique called "**Northern Blot Analysis**." The problem of course, is that they could only measure expression from one or a few genes at a time, a slow and tedious process. Now that we know there are approximately

30,000 genes, which are processed into hundreds of thousands of different variations, studying expression of the whole genome taking the one-gene-at-a-time approach, would be like draining the ocean using a cocktail straw. These arrays are built using the same type of technology that is used to manufacture semiconductor chips, but rather than etching miniature circuits, millions of strands of DNA are built vertically on a glass chip. That technology is combined with something called combinatorial chemistry, a chemical principle that in less than 100 steps, allows more strands of DNA than there are stars in the Milky Way to be built.

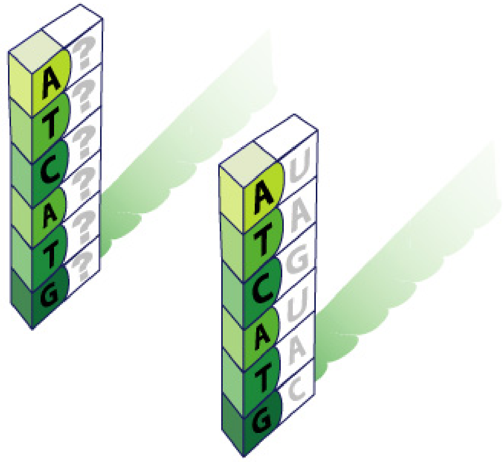
Natural Attraction. GeneChip microarrays use the natural chemical attraction between the DNA (on the array) and RNA target molecules (from the sample preparation) to determine the expression level -- how much RNA is being made -- of a given gene. There are only four molecules, or "**bases**", in every DNA chain: **adenine** (A), **guanine** (G), **thymine** (T) and **cytosine** (C). These four molecules partner: C partners with G and T partners with A. Pairing is a natural state for DNA and if you pulled the double helix apart, it would inevitably move back together, like two long chains of magnets that are attracted to each other. Like DNA, RNA is composed of 4 bases: A, G, C and **uracil** (U) instead of thymine. Uracil is related to its DNA equivalent, thymine, such that in RNA, C partners with G and U partners with A. However, unlike DNA, RNA is usually single stranded, meaning that it can easily bind to any other single stranded matching sequence, whether it's DNA or RNA. And the amazing thing is that these four bases are the basis the DNA code found not only in humans, but each and every organism on Earth. Flies, ducks, lizards, salmon, bacteria, and monkeys all have the same four base code in their nucleic acids.



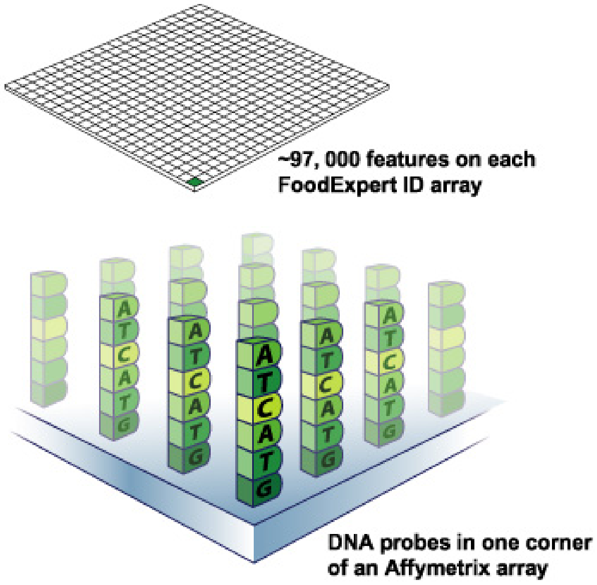


A Good Match Sticks, a Bad Match Doesn't. When a single strand of DNA (**ATCATG**) matches a strand of RNA (**UAGUAC**), the two strands are "**complementary**" and will stick to each other. However, if the bases aren't complementary, they won't fit together. An A won't pair with another A, another C or another G. Even a single base that doesn't match its partner (like a **C** and an **A** *as shown in the adjacent diagram*) could keep one single strand from sticking to another single strand.

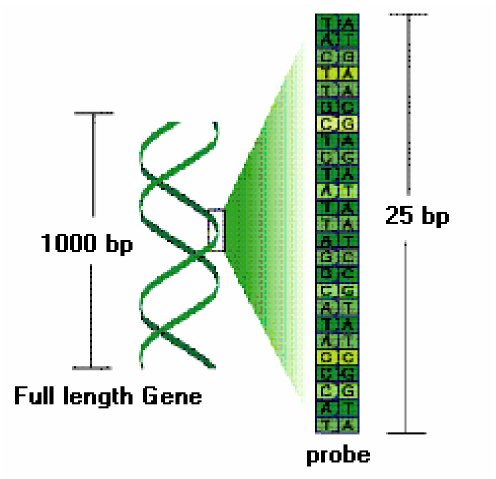
The Basic Principle. Microarrays use this base pairing attraction –known as **hybridization** – to help researchers identify what RNA sequences are present in a sample, and this then tells them what genes are being expressed by that organism and how much they are being expressed. Because C always pairs with G, and T always pairs with A, if you know that one side of the chain is ATCATG, you don't even need to see the other side — you know it is TAGTAC. Of course, if that second strand is RNA, the sequence will be UAGUAC.



From a Gene to a Probe. Microarrays can measure the expression of every known human gene. However, let's use a simple example that focuses on just one gene to illustrate how this technology works. The first step is to build a short DNA strand — a **probe** —onto the surface of a glass chip. This short strand is only 25 bases long and in our example represents a small section of a much larger gene, 1000's of bases long. Scientists compare the 25 base probe sequence to the rest of the human genome sequence to make sure it doesn't match any where else. This way, when an RNA molecule binds to the probe, researchers will know that the gene was expressed, because the only match possible is from that gene—no other sequence in the genome would match. The short probe on the array measures expression of the complete gene by sampling for a small section of the gene. It's like identifying the "*Star Spangled Banner*" without hearing the entire song, you might not need more than a few lines to name that tune. When RNA is prepared from a cell, there are hundreds of thousands of different sequences all mixed together. Measuring expression of a single gene in this mix, is like trying to play "name that tune" with hundreds of thousands of songs all playing at the same time. It's difficult and often times confusing to say the least. It's like listening for the National Anthem, with hundreds of thousands of other songs playing at the same time. The same holds true for DNA probes used to measure RNA expression levels. However, these microarrays are so accurate that they can detect even one specific RNA molecule out of 100,000 different RNAs that may be present in the mixed sample.


~97, 000 features on each FoodExpert ID array

DNA probes in one corner of an Affymetrix array



Features and probes

The surface of the Affymetrix array is like a 1.25 cm by 1.25 cm checkerboard. Each tiny square is 11 micrometers by 11 micrometers and holds one unique type of probe. These probes are built one molecule at a time, using the same type of manufacturing technology that is used to build computer semiconductors. The molecules are built one base at a time, one stacked on top of another, like checkers.

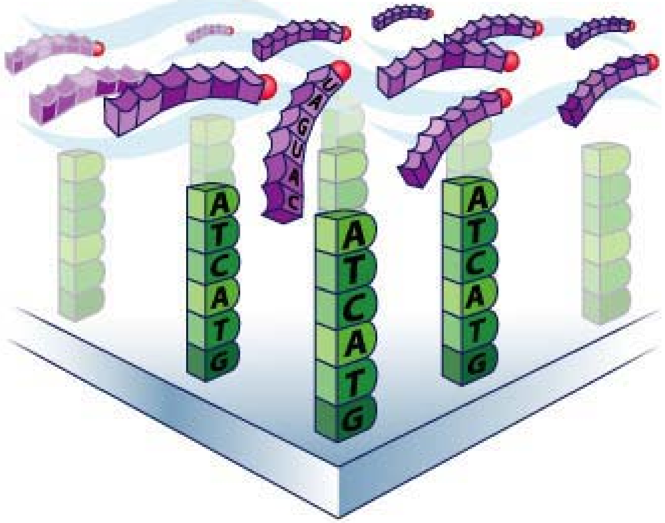
The most recent human genome array is about the size of a dime, with more than 1.3 million squares, called "**features**". It represents about 47,000 different RNAs— practically every known human gene that is expressed and eventually turned into a protein. Each feature on the array is about 11 microns across which is about one-fifth the width of a human hair! For our example, we are just going to look at one corner, or a single feature, of an imaginary array. Normally, each probe is 25 bases long, but for purposes of illustration here, let's abbreviate this standard probe length and say the probe is only a 6 base sequence, ATCATG.

Extract, Prepare, and Chop up the RNA

Now that we have a probe designed to measure expressed RNA, we have to extract the RNA from a biological sample, such as blood, tumor, or other body tissue. In a multi-step process, researchers extract RNA from the sample and make millions of copies through a process known as **PCR**. Copying the RNA allows it to be more easily detected on the array. At the same time the RNA is copied, molecules of a chemical called **biotin** (the small "cups" in the diagram to the right) are attached to each strand. These biotin molecules will act as a molecular glue for **fluorescent** molecules that will later be washed over the array. When researchers eventually scan the array with a laser, the fluorescent molecules will glow, showing where the sample RNA has stuck to the DNA probes on the array. Using another chemical process, we fragment the RNA chains up into millions of short pieces, all of which still have at least a few biotin molecules attached to them.

Wash Sample over the Array

The entire prepared RNA sample is **wash**ed over the array for 14 to 16 hours allowing the hybridization to occur. The number of molecules involved in this wash is staggering. There are millions of copies of each DNA probe (ACTATG) in every square on the chip, and there are also millions upon millions of pieces of tagged RNA from every expressed gene present in the sample. It's like the world's largest molecular "singles bar". All of the RNA strands from expressed genes are swimming around, looking for their perfect complement - molecular true love on a microarray. Sadly, most of them will not find a match on the array.



Sample RNA fragments (purple)
washed over DNA probe array (green)

A Committed Relationship

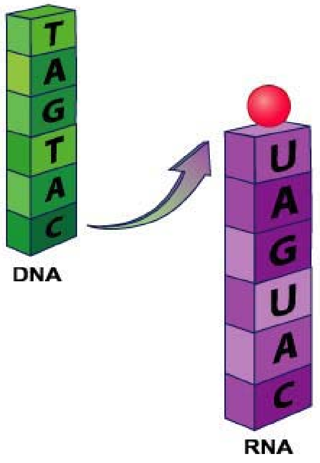
But somewhere, in the tagged sample of RNA washing over the array, a match will be made. If the sequence of bases in the sample RNA matches that of a DNA probe, then there will be a perfect match and the sample will stick to the probe.

Determining a Match

Let's assume that we have a match and that RNA in our sample bound to the probes built on the array. We then rinse the array, so that any RNA that didn't pair is washed away. The hybridized RNA is tagged with molecular glue (biotin); it's as if each hybridized square on the array has been coated with sticky glue.

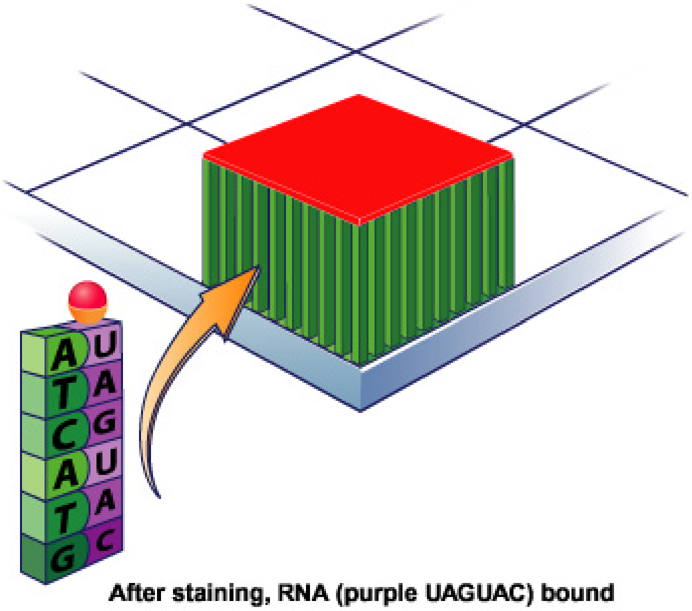
Making Glow in the Dark RNA

Because we can't see RNA, we can't directly figure out how much has stuck to the DNA probe on the array. Did only one strand attach? Or did 1,000,000 strands attach? To work around this problem, scientists make the RNA glow in the dark by using a fluorescent molecule that sticks to the biotin.

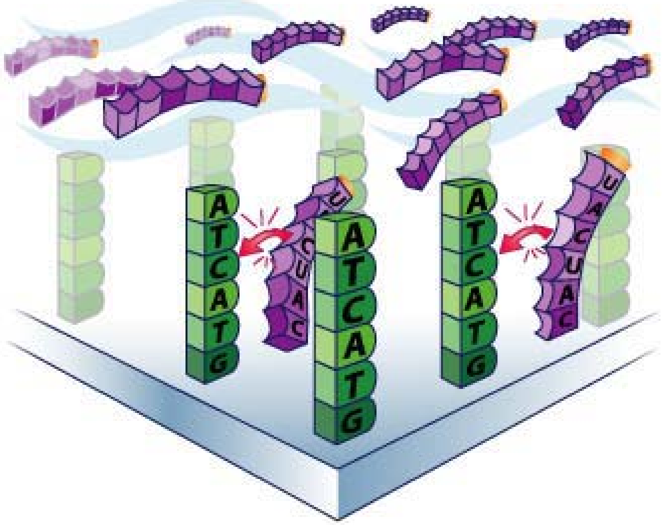


An Expressed Gene

Researchers wash the fluorescent stain over the array and the glow in the dark molecules (ball) stick to the biotin glue (small cup). It's like glitter painting in elementary school -- after pouring sparkle glitter all over the paper, you shake it off and the glitter only sticks to the places where there was glue. With microarrays, the fluorescent molecules are "shaken" away and



After staining, RNA (purple UAGUAC) bound to the DNA probe built on the array will fluoresce

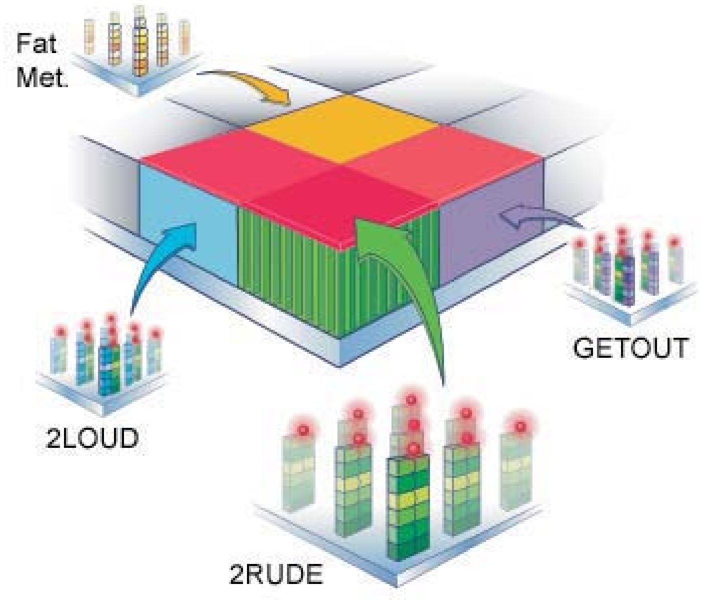


C does not stick to another C, so no match is made

the stain only sticks to those places on the array where RNA has bound. After all of this, researchers shine a laser light on the array, causing the stain to fluoresce or "glow". If a gene is highly expressed, many RNA molecules will stick to the probe, and the probe location will shine brightly when the laser hits it. If a gene was expressed at a lower level, less RNA will stick to the probe, and by comparison, that probe location will be much dimmer when it is hit with the laser.

No Match at All

If the sample RNA doesn't match it will be rejected by the probe on the array. Scientists will know that no match was made when they shine the laser on the probes and nothing glows.

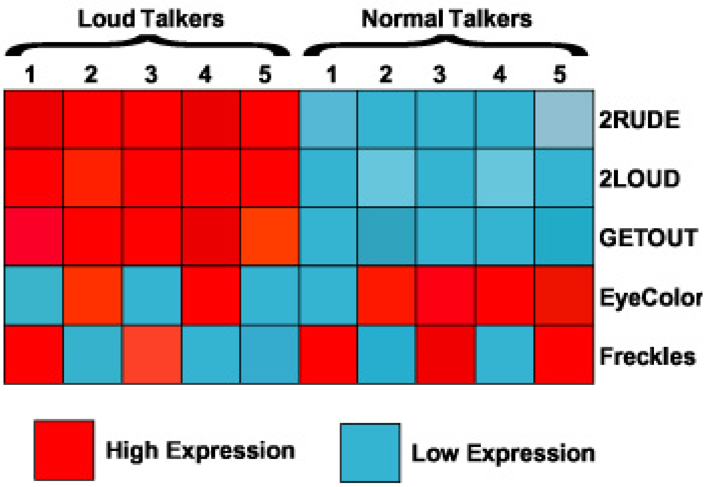


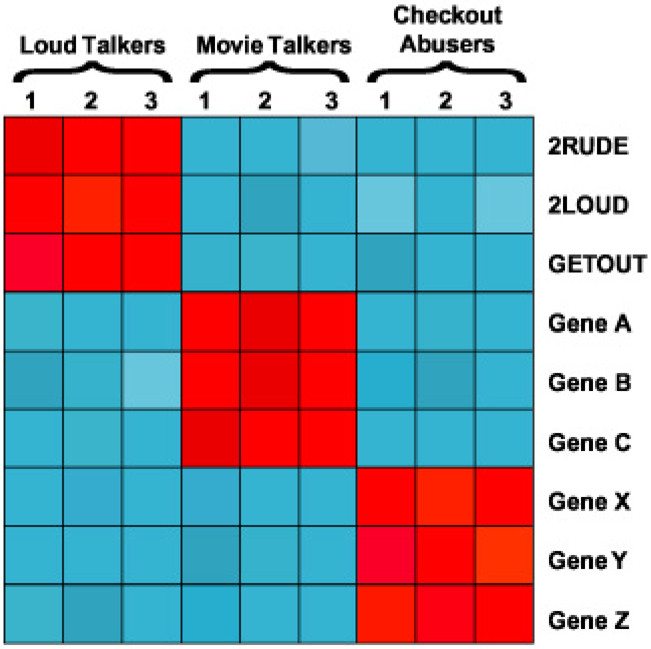
Measuring All Gene Expression at Once

So far we've been looking at expression from just one gene. And although GeneChip expression arrays can simultaneously measure tens of thousands of different genes, let's simplify it down to looking at expression from just four: Gene 1 (2RUDE) Gene 2 (2LOUD) Gene 3 (GETOUT) Gene 7 (Fat Met.) In this example, Gene1, Gene2, and Gene3 are expressed because fluorescent RNA has hybridized to teach of the probes. In their study, scientists find that these genes are only expressed by the loud speakers, and not at all by normal speakers. Because nothing is known about these genes other than their expression in rude people, scientists decide to call them 2RUDE, 2LOUD, and GETOUT – a RUDE pathway. Even though they aren't 100% sure what the genes do, they know they are consistently expressed by abusive cell phone talkers. Virtually everyone has these three genes, but the difference is that they are not equally expressed by everyone. In well-mannered individuals, the RUDE pathway sits idle and its genes are not transcribed into RNA. As a result, that RNA doesn't make proteins and those proteins don't drive a person to unconscionably rude behavior. The next step for the researchers is to use additional techniques showing that the proteins created by the RUDE genes function to suppress activity in the politeness and common sense regions of the human brain.

Comparing Gene Expression

The whole point of microarray gene expression analysis is to compare expression levels between two samples. In our example comparing loud talkers and normal-volume talkers, expression analysis found that 3 genes were more heavily expressed in rude people than in normal people. To represent this, researchers construct "**heat maps**", which are graphical displays that color code gene expression. Increased expression is color coded in dark grey while decreased expression is in light grey. The heat map to the right shows gene expression from 5 loud talkers and 5 normal talkers. It makes it clear that 2RUDE, 2LOUD and GETOUT are highly expressed in Loud Talkers, but not in Normal Talkers. Other genes expression levels, like those responsible for eye color or freckles, do not correlate with rude talking behavior.

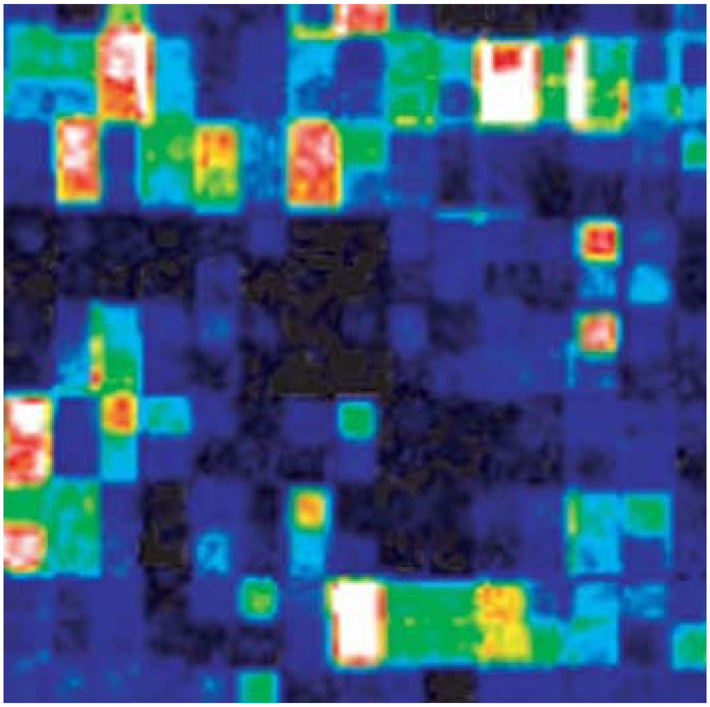




Classifying Disease

Having found genes responsible for rude behavior, our scientists explore if those same genes are responsible for other forms of rude behavior. They compare gene expression from 3 people who are loud cell phone talkers, 3 people with the rude behavior of talking at the movies, and 3 other people that rudely use the express checkout lane with 25 items. In this example, the researchers might find that expression of different genes can be used to classify the different types of rude behavior. In this case, expression of Genes A, B, C are markers for loud movie talkers, whereas expression of Genes X, Y, Z are markers for the checkout lane abusers. All these people suffer from seemingly identical rude behavioral disorders, but the genetics of each disease is quite different. By genetically classifying seemingly identical diseases, researchers can develop more targeted therapies to the distinct forms of otherwise indistinguishable disease.

An Actual Gene Expression Image



In reality, human expression arrays have over1.3 million different probes used to detect nearly 50,000 different RNA sequences. The result, when translated into a graphic by a computer, looks like a fuzzy television picture, but when viewed up close, looks just like an illuminated checker board. The fluorescence coming from each *checker square* or *probe location* tells researchers whether a gene is expressed or not. Some probes measure high concentrations of RNA (strong intensity, white and grey features) and some do not (weak intensity, light grey and black features). By analyzing images like this, scientists can measure how much of RNA was present in the sample. They can record that information, store it, and then compare it to another sample analyzed at a later date. For instance, if a scientist measures twice as much fluorescence for the 2RUDE gene in a second person, it's possible that person will be even more annoying than the first.

Drug Discovery

Ultimately, the goal of identifying a disease pathway, like RUDE, is to develop a treatment to counteract that pathway, restoring people to good health, and in this case, good behavior. But, before investing significant resources in drug development, researchers need to know that disrupting the RUDE pathway will actually disrupt disease—there's no point spending millions of dollars on a fool's errand. To validate a drug target, scientists use laboratory techniques to block a gene's function, giving them an early glimpse of how cells might respond to a drug blocking the 2RUDE, 2LOUD or GETOUT function. Microarrays are then used to understand this cellular response by examining gene expression across the whole genome. For instance, if scientists find that disrupting 2RUDE looks like it will have a more potent effect than disrupting 2LOUD or GETOUT, they might be inclined to develop an anti-2RUDE drug over the others.

From Chemical to Drug

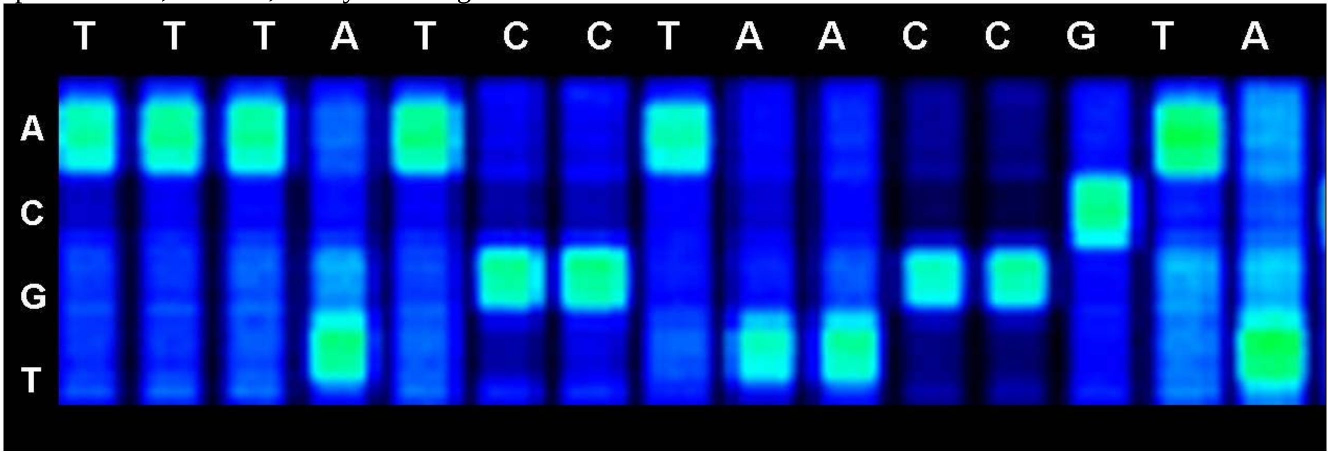
After identifying and validating a target, drug companies begin the long and expensive task of finding a chemical compound that can be developed into a successful drug. Using microarrays, scientists can screen libraries of compounds and determine how each compound affects cellular gene expression. Naturally, the compound they're looking for would inhibit RUDE pathway expression. But whole-genome expression analysis also lets scientists know about other genetic effects suggestive of a drug that would have far too many side-effects to be any good. For instance, if the changes in gene expression match those of a known toxin, like nerve gas, scientists would likely stay away from that compound, and concentrate on other molecules that don't look as risky. On the other hand, a compound may produce unexpected changes in gene expression that could be helpful for treating another disease, operating through a different molecular mechanism.

Treating People

Different people respond to drugs in different ways. Often, those responses are a result of the genetic differences between patients. Using microarrays, scientists can identify those genetic differences, and then use that information to predict which patients will respond favorably and which will not. For example, if a drug developed to suppress the RUDE pathway only worked in 20% of patients, it would be considered a dismal failure. However, by using gene expression profiles to pre-select those 20% of patients for whom treatment is likely to work, the pharmaceutical success rate would skyrocket. This type of *personalized medicine* not only leads to more successful drugs, but also avoids potential side-effects from a pointless treatment and can alert doctors to try a different, more effective type of therapy. In the case of RUDE cell phone behavior, perhaps a good scolding from mom might do!

Part II - GeneChip Resequencing Arrays

Let's now look at another type of array – the **Resequencing array**. This type of array has many potential uses, such as identifying DNA in a sample as coming from a certain organism or even a virus. It could be used to find out virus is involved in an outbreak or to determine which exact ***strain*** of the virus is present. This type of array could be used to monitor pathogens in the water or food supply, or to detect a whether a specific food is really "pure" without contamination from other foods. For example, it could be possible to make sure that the fish in tuna fish is really tuna and not a mixture of other fish. Similarly, very expensive sushi could be analyzed to make sure it was the real thing and not inferior fish. As long as there is DNA in the sample being analyzed, this array can be used! It is a powerful identification tool. The sample preparation for this array is different than the process outlined above for a gene expression array. With this array, it is the DNA which is isolated from the sample to be tested (not the mRNA). The DNA is fragmented using a **restriction enzyme** and is then amplified to create multiple copies of each fragment. The amplification is done through the process of PCR. Each fragment is labeled with Biotin, just as with the expression microarray, and then added to the chip. The chip is washed and stained for visualization purposes and then analyzed. Anywhere fluorescence shows up, the DNA fragments ('targets") have hybridized to the DNA probes in the features. The basic array structure is very similar to the gene expression array. The chip, feature, and probes follow the same basic concept. However, this array does not detect for the presence of mRNA in a sample by placing probes on the feature that represent unique segments of the gene. The probes that are used here are still made of segments of DNA 25 bases long, but the probes are used to basically sequence the DNA in the sample. A certain sequence then could be identified as a specific virus, bacteria, or any other organism.



For this array to work it is necessary to already have access to sequence information determined through prior research (that is why this is called "resequencing"). This way, when the DNA is sequenced it can be compared to known sequences for identification. Also, knowing the sequence helps build specific probes. This way, if you are screening for malaria, you only need to look at sequences that are specific for malaria strains allowing for quick identification. The actual resequencing array is very complicated, with each probe being slightly different than the one next to it. The main difference is that a set of four probes is used together to sequence one base. There are four versions of each probe to test to see if an A, G, C or T is found at a specific position on the DNA. In the 25 base long probe, the middle base (#13 – in bold below) is the variable nucleotide that is actually used to determine the actual base in that position of the target DNA (from the sample). Here is an example to give you a visual. Let's call this probe set W.

Probe W1 - ATCGGGTAAACTAAAGGCTACTGCCT
Probe W2 - ATCGGGTAAACTCAAGGCTACTGCCT
Probe W3 – ATCGGGTAAACTGAAGGCTACTGCCT
Probe W4 – ATCGGGTAAACTTAAGGCTACTGCCT

The arrays are built in rows of fours for the A, C, G and T possibilities. The next row of probes also contains 25 base long DNA segments (the probes) with the middle base being variable. It is actually exactly one base over from the W set (shown above) just before it. To help you out, here are the probes that would be found in the next row of 4 features on the array. Let's call this probe set X.

Probe X1 - TCGGGTAAACTGAAGGCTACTGCCTC
Probe X2 - TCGGGTAAACTGCAGGCTACTGCCTC
Probe X3 – TCGGGTAAACTGGAGGCTACTGCCTC
Probe X4 – TCGGGTAAACTGTAGGCTACTGCCTC

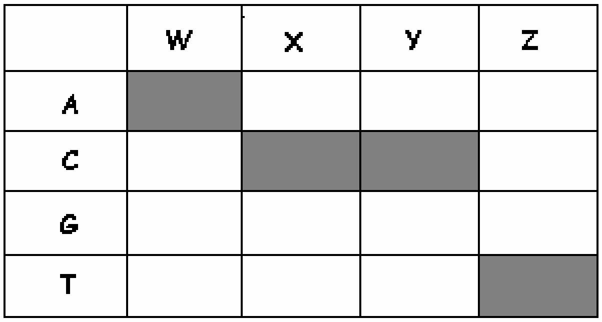
Notice, if you compare W1 to X1, they are basically shifted over one to the right and off by only two bases (the 13th base of each):

Probe W1 - ATCGGGTAAACTAAAGGCTACTGCCT
Probe X1 - TCGGGTAAACTGAAGGCTACTGCCTC

If we were to then continue with probe set Y, Z, etc. you could continue until you have a probe set of 4 for each base in the sequence of the DNA segment to be detected for. Here is a simplified diagram to help you visualize this. Notice it does not show all the probes, just a small section (enough for four bases). Normally, you would need to span the entire DNA. The numbers represent the base position that is being tested for by each probe:

```
DNA =================1 2 3 4===========================
                 Probe W ___1____
                    Probe X ___2____
                      Probe Y ___3____
                         Probe Z ___4____
```

The reading of the resequencing array is actually very easy. Even if you do not exactly understand why the target hybridized to one probe and not to the other and how this relates to the actual sequence, the reading is very simple. Here is an example using the situation above. "W" represents the four W probe set, and so forth. Going from left to right the sequence reads T G G A. You may be asking why, if the square that is fluorescing is in the A spot, is it read a T? Remember that T binds to A (and C to G). So, a fragment with a T at its' middle base should bind to a probe that has A as its' middle base. This is what is shown with Probe W1 above. In the first position, in which probe set W was used, the DNA target stuck to the W1 probes (with an A at the 13th base) on the feature. In the second position, which the X probe set was used to, there was a DNA target that stuck to the X2 probe, with a C at its' center position, but the actual sequence read is a G. And this continues on through the next two positions. Remember, that this is just four bases. Even the most basic viruses have thousands of base pairs in their genome. So, you could imagine why it is valuable why the array can have so many features (over 1.3 million).



	W	X	Y	Z
A	■			
C		■	■	
G				
T				■