Puzzling with the pathways of life

Maastricht University

Inaugurele rede

Prof. dr. ir. Chris Evelo

Faculty of Health, Medicine and Life Sciences

Puzzling with the pathways of life

Puzzling with the pathways of life

Chris Evelo

Inaugural lecture at Maastricht University.

Read March 6, 16.30 Aula Mindersbroedersberg, Maastricht

Geachte decaan, geachte leden van de begeleidingscommissie, dear professors, dear members of the bioinformatics and systems biology community, dear colleagues, dear students, beste familie en vrienden, I am really happy and honored that you are here today to listen to my inaugural lecture.

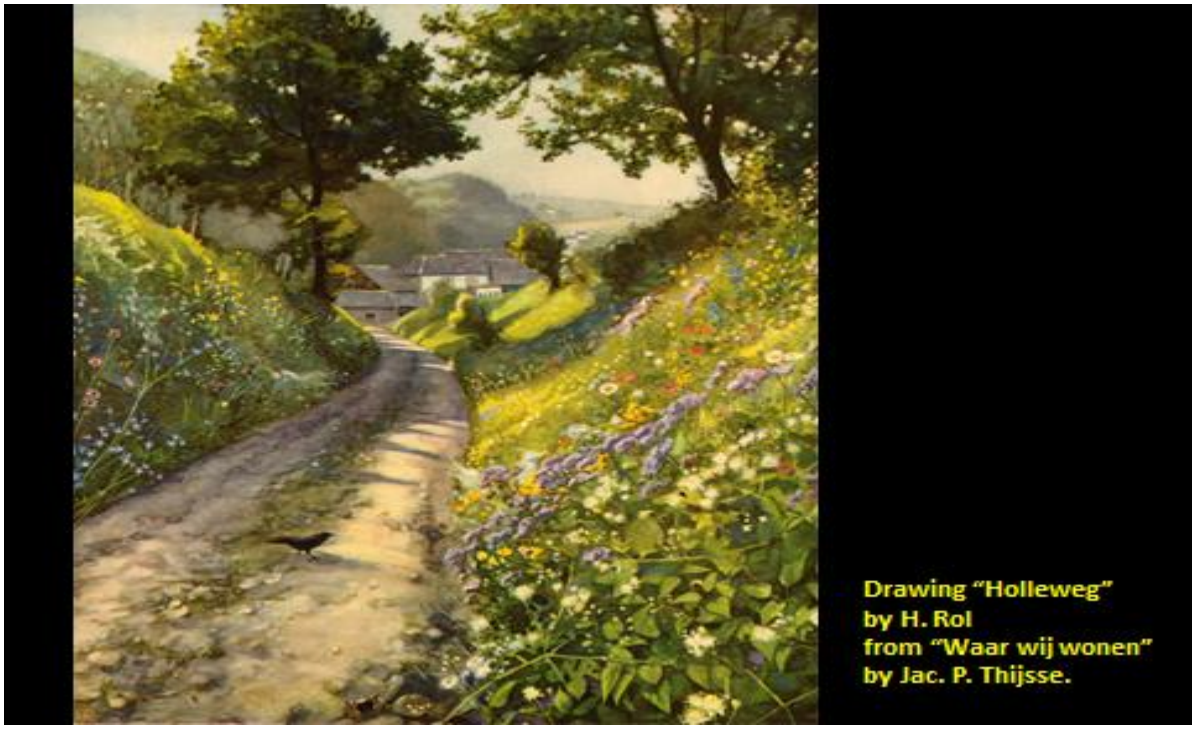
Een eerste zin in twee talen? De Universiteit Maastricht is er trots op de meest internationale universiteit van het land te zijn en natuurlijk is wetenschap zelf ook internationaal. Ik zal het hebben over het samenbrengen van kennis en daarmee ook over internationaal samenwerken. Een groot deel van onze studenten komt van buiten Nederland en veel van ons onderwijs is Engelstalig en dat geldt ook voor een groot deel van mijn eigen medewerkers. Veel internationale collega's hebben aangegeven niet hier te kunnen zijn, maar wel graag online te willen zien wat ik heb gezegd. In deze rede zal ik proberen duidelijk te maken dat wanneer onderzoekers verschillende talen, verschillende begrippen, verschillende standaarden gebruiken we niet moeten forceren ze hetzelfde te laten doen. Maar dat we moeten kijken naar verbanden, naar vertalingen. Een tweetalige oratie is echter wat veel gevraagd. Omdat wij buitenlandse collega's eigenlijk niet de kans geven Nederlands te leren doordat we ze altijd aanspreken in het Engels zal ik de rest van deze lezing dan ook in het Engels doen.

In my faculty, the faculty of health, medicine and life sciences, we study life itself, what it means to be healthy or to stay healthy, how people can be healed and even what life really is, how it all works. We focus on human biology and on real life issues. The southern part of Limburg unfortunately is the least healthy region in the Netherlands. We study consequences of the unhealthy habits that we have here and in the western world in general. Things like obesity, diabetes and

lung diseases, or on the positive side what kinds of nutrition or life styles are really healthy. This slide is from our website.

During our Open Days we ask prospective students research questions like “how do we prevent the jojo effect?”. Confronting? Yes.., but also touching upon real life. You may consider obesity a luxury problem, but it also kills people and it is extremely expensive for our health care system. They even paint questions on the pavement leading to the building where I work. In these days of reviving student engagement you could also ask “why don’t we allow or even encourage our students to paint research questions on the streets during the rest of the year?”

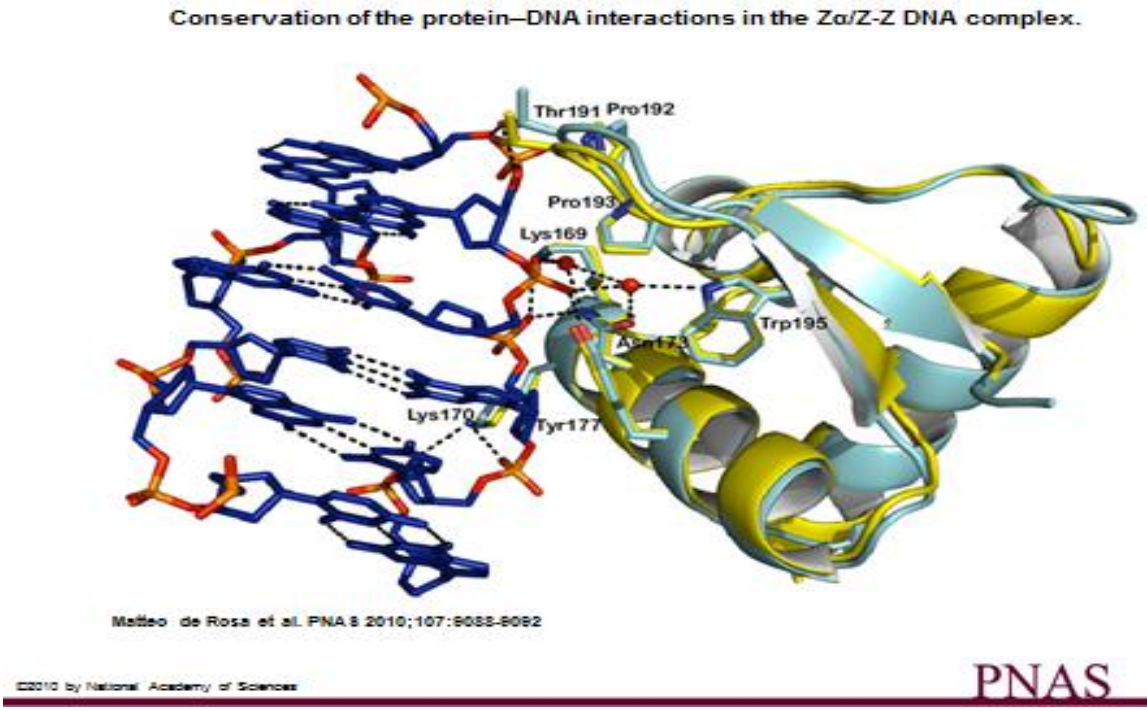
So, we study human biology. The problem is that biology is very complicated. That brings me to the first problem I want to address. We tend to underestimate how complicated biology really is. That already starts in high school; biology is considered one of the easy, not so technical, choices. That is probably related to the fact that we associate it with nice, beautiful things: flowers and butterflies. But look at this artist impression of a hollow road here in Limburg.



Granted, you have to be lucky to see so many flowers at any given time. But even at this level; can you really think for a moment that the interplay between those flowers, the trees, the birds and insects and all these smaller things would be simple to describe? Of course it isn't. The external complexity of a single organism or the beauty of many of them living together often makes us hold our breath. And we try to freeze the image in a picture. In that way we reduce the complexity. Without movement and without dynamic interactions it feels like we understand what we look at a little bit better. However, it is an inherent part of life that it changes. And we want to understand how it changes. How our hearts beat, how we fight an infection, how we become happy or obese, or both… Biology also is complex because every single individual is incomprehensibly complex. We describe the human body as an interplay of organs, which consist of cells, again consisting of organelles which are made up of larger molecules like DNA and protein, and small molecules like sugars, fat and vitamins.

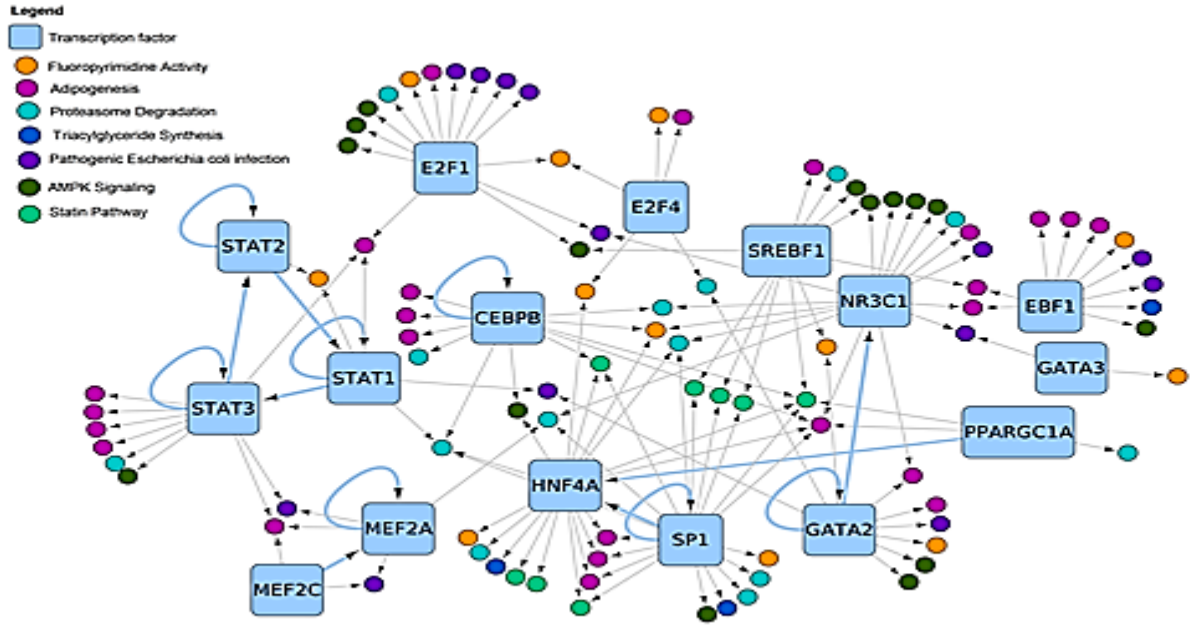
But if you are in high school and you are interested in studying complex things you will rather think about studying physics or so, so you can go into building rockets or computers. Biomedicine is considered less beta and people choose it for different reasons, very good reasons. Our students and staff members are driven by the idea that they want to heal people or want to keep people healthy. They

are good people, dedicated to a good cause. But, let's face it, not so many like to think about the chemistry that makes all these living things do what they do. In fact for many of us "chemical" doesn't even sound like it is about life. It sounds more like the stuff we make in smelly factories. And yet, the system below all those living things is of a chemical nature.



This picture shows some of the building blocks of life like a biochemist would look at them. On the left in red and blue you see a fragment of a DNA molecule and on the right you see a protein that interacts with that specific fragment.
For experts representations like this one are very relevant. They are important pieces in the puzzle of understanding life. If they look at a picture like this in detail they understand that if some of the amino acids in the protein are changed, the protein itself must change, because chemically and physically it just cannot keep the form needed to perform its normal function.

This is just one interaction between one protein and one piece of DNA.



Martina Kutmon, Chris T Evelo, Susan L Coort (2014) A network biology workflow to study transcriptomics data of the diabetic liver, BMC Genomics 15: 971

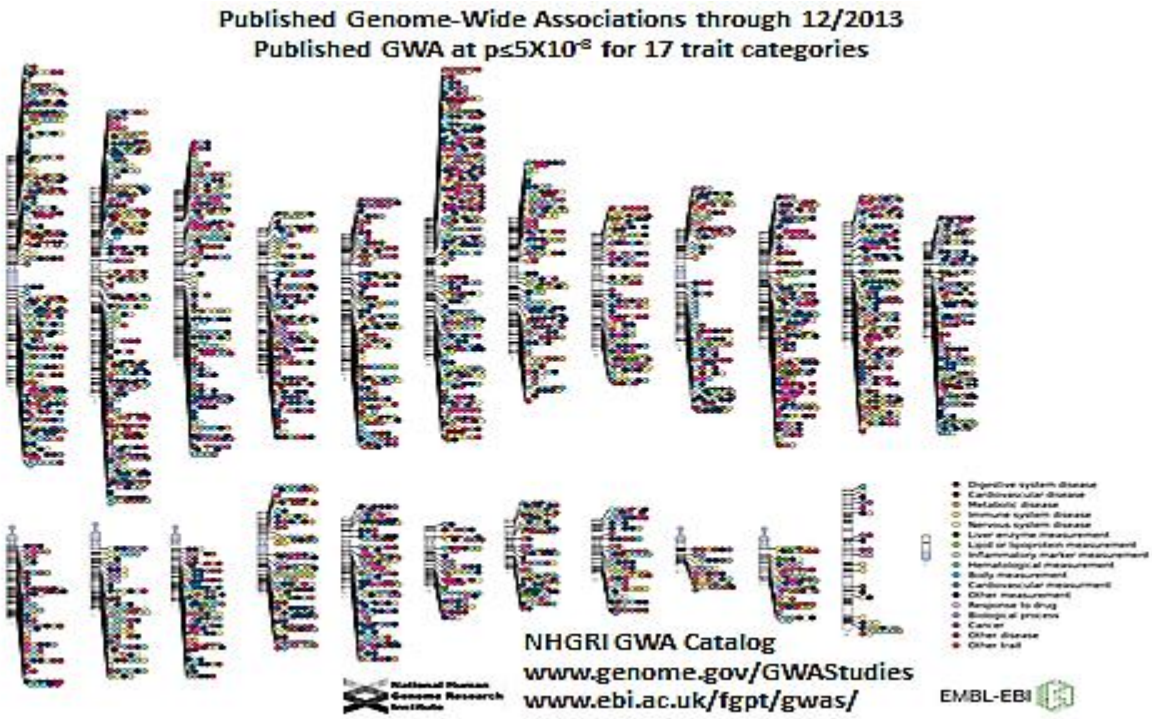
This picture comes from a research paper that we published just a month ago. Martina and Susan in my group reanalyzed data from studies in livers from type 2 diabetic patients. Every individual blue rectangle is a protein like the one you saw before. And every small circle is a gene that is influenced by such a protein. The colors indicate something of what we know about those genes, the processes they are involved in.

So here we are looking at a part of the complicated system involved in the development of diabetes. If you look at it a little more you'll see that these proteins in blue actually affect a lot of genes each. And many of these genes in the circles again are affected by more than one of the proteins. There even are blue lines indicating that they sometimes affect the genes that produce the other proteins indicated in the rectangles, sometimes even their own gene. Don't forget that to understand any one of these lines, these interactions in the network, you would have to look at the real chemistry of the protein, like we just did in the previous picture. What changes will make it fail? Only then can you start to understand why somebody that has a changed gene might develop diabetes earlier or later. And only then can you decide whether any personal treatment, with a drug for instance, might work for one person and not for another.

Real world health problems like diabetes do not just ask for disciplines that can deal with the inherent complexity, like graph

theory, neural network approaches or machine learning. They also ask for collaboration between disciplines. What often happens is that researchers see that a problem extends outside the boundaries of their own discipline and then they tend to stop at the border. What is needed is that people from different disciplines work together across these borders. They need to challenge each other. Especially in medicine we are not used to that. Sometimes things that we thought of as facts, as foundations of what we did and believed for years, start to shake when looked upon from another equally scientific perspective. You cannot say that the problem you work on every day needs to be studied on the systems level and then leave it to experts from for instance a more mathematical discipline to solve that. That is because these systems approaches too have limits and only the domain expert knows whether these limits are reached or not.

We often hope that modern research can find simple relationships between what is very deep in the system and what happens on the outside. We love simple relationships with simple consequences. This is about our own life after all, our own health. As a consequence we like to read about single genes found to be related with things as complex as autism. For that purpose we do a lot of so called genome wide association studies. Basically we ask the question “can we find places in the genome where people that have a different piece of DNA around there have a different trait, become more easily sick for instance?”.



This picture shows what GWAS results were known about a year ago for 17 different traits. It includes some that are really relevant for this not so healthy Limburg: cardiovascular diseases and metabolic diseases. It shows the whole genome, all 23 chromosomes. It must have cost billions in research money to get all that data. The good thing is of course that all that data is now available for further research. But is it also really worth that much? These variant-trait relationships do not show that a specific change in the genome causes the trait, they show that "something around there" might do that. So compared with the protein-gene puzzle pieces I described before it is more the hint "there should be a puzzle piece here somewhere, try to find it". And still… These are the things that make it into the newspapers.

That is because we think that if we understand such things we will be able to do something about it. But unfortunately many of these findings are a bit like saying you really need nuts and bolts to build an engine, so very true and so very useless. We do know that such conclusions are oversimplifications. We just like to believe in simple conclusions; it is easy to sell such an idea and it even makes actual products sell, superfoods for instance. Promising such outcomes even helps us to get research grants funded. Yet, we do know that life is based on a really complex system with extreme amounts of interactions and many possibilities to behave slightly different or alternatively many possibilities to respond to being pushed from the outside and remaining essentially unchanged. That made us redefine what we think health is. Health is not the absence of being sick. Health is the ability of that complex system to withstand many kinds of stress. You are not healthy if you stand upright, you are healthy if you remain standing upright if somebody pushes you. That complexity is why we talk about systems biology. If you feel that biology always was complex and systems biology is just a modern term for biology or maybe physiology, then yes, you are right. We like such terms though. Because for a while they help us indicate what we feel is important. But they are often overused and then become just fashion. When that happens such terms become nice to play "scientific term bingo" on social media during conferences. One way social media can help us to "stay real" in science.

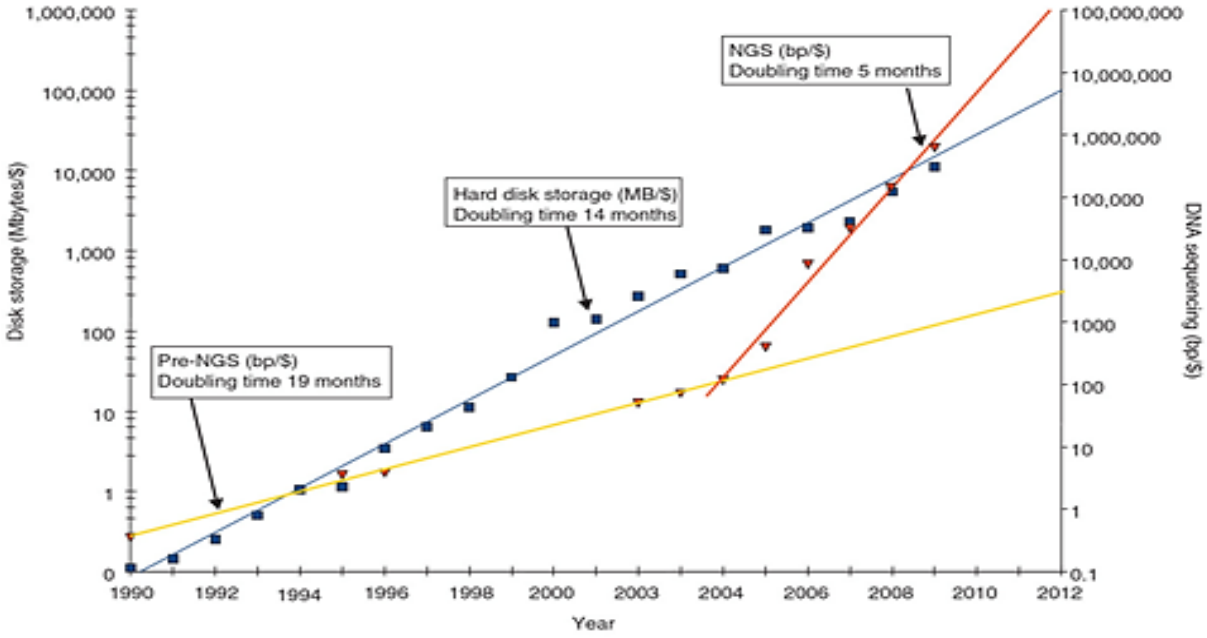
So how can we solve all that? What helps is that we get better at measuring things. Well to be honest, we are especially getting better

at measuring some things. We do not learn to measure that many new things. What we do learn is to measure the same things quicker often by using micro technology. We have more and more highly parallel measurement technologies that measure a lot, really a lot, of things in parallel. This is especially true for measurements related to genes and the expression of genes. We can easily and affordably sequence your entire DNA or compare all the genes expressed in a tumor with that in healthy tissue.

This is where another "bingo term" shows up: "big data". Measuring many things means we get lots of information. We need to organize that data in such a way that we can use it in combination with other things that we know or measured. That unfortunately sounds a lot simpler than it is.

We tend to think about "big data" problems as related to data size. And yes, we have such problems, even though computer storage keeps getting cheaper all the time. We can in fact store astronomical amounts of data. Two week ago I was at the European Southern Observatory head quarters in Munich for a European data infrastructure meeting. There they work with literally astronomical amounts of data, and they do very complex things with that. They also produce very beautiful pictures like this one.

But what struck me is that they told us that the growth rate of data that they produce is actually lower than the growth of the amount of computer storage you can buy for a certain amount of money. In other words astronomy does not have a data storage cost problem. They can keep storing all the information they collect and it even gets cheaper. Unfortunately you cannot say the same about biology. We do produce ever more data at such a pace that it gets more expensive to store it. This picture shows that.



From: Stein Genome Biology 2010 11:207 doi:10.1186/gb-2010-11-5-207

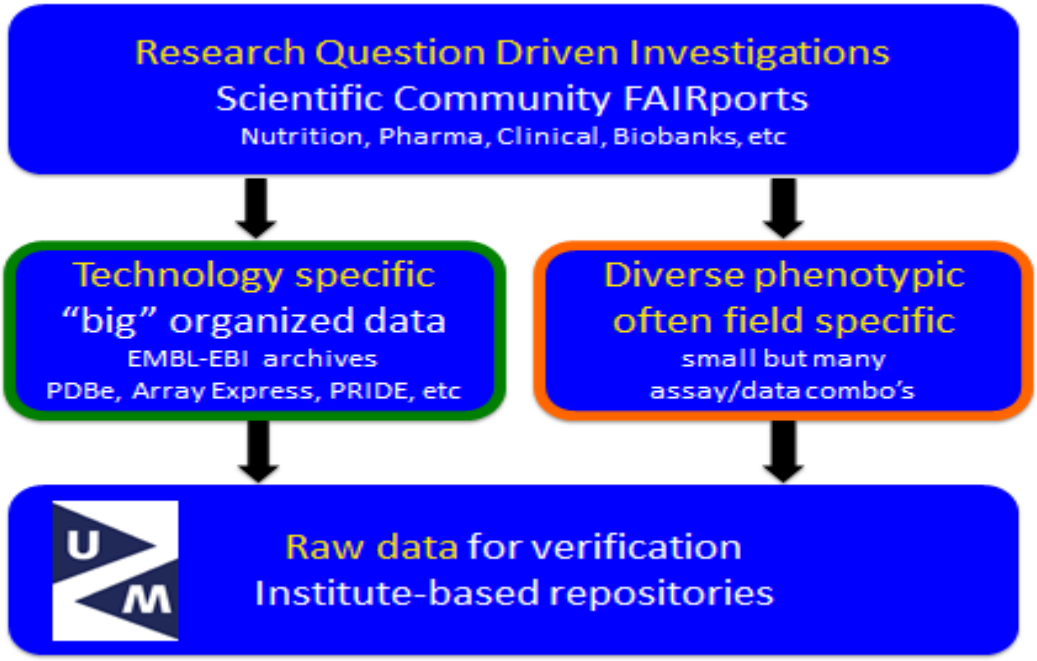
The blue line shows how much hard disc storage size you can buy for a given amount of money over time. The yellow line shows the pace at which we produced new DNA data just a few years ago. The important thing is that that line is less steep than the blue one. That means that we could store that increasing amount of data year after year and spend less, not more money on storage. For astronomy that still holds. Lucky astronomers. The problem is the red line, which is for the new DNA sequencing technologies. That line is steeper than the blue line. In other words we need to spend more money on storage every year. And while we can learn to store in more efficient ways that problem will not go away. We will have to learn to substantially reduce the amount of data we want to keep. In other words we need to learn to throw things away. In the medical field that is problematic. We are obliged to keep all the data that might be relevant to evaluate medical judgments. Since we actually cannot keep all raw data we need a better evaluation of what is and what

isn't needed for that and we need rules, laws even, that are in line with that.

Apart from these unsolved data volume problems there are other problems related to big biomedical data. To understand that it is helpful to think about an actual study. Let's take a modern one like we really perform them. Suppose we do a study where we want to find out about how our diet influences obesity. We want to understand why some people eat more, or why some people gain more weight even though they do not eat more. For that we of course collect information about what people really eat and about the actual weight changes. We probably collect a lot of other information too. To give just some examples that could be: blood pressure, response to challenges like drinking a glass of sugar water or information about other life style factors: "do they snack?", "do they exercise?", maybe even "are they happy?". Like I described we know that the system depends on the whole interplay of small molecules like protein and DNA and nutrients. Therefore, we often measure a lot of that as well. The expression of all the genes, thousands of metabolites or even the actual DNA sequence.
If we want to analyze all that data we need to organize it. What I want to address is how we should do that on a larger scale; for all biomedical studies in Europe for instance. There is an increasing notion that research results from government or EU funded research should be open. After all the research was paid for by the taxpayer. Why would the taxpayer pay for the same data again? But that assumes that we put the data somewhere where you can find it and from where you can really reuse it. Funders will in fact make that obligatory. Not just because that is efficient, not just because we can all do more if we can all use all the data that was produced before, but also because that is how science is supposed to work. You should be able to verify the findings that were published. And for that you need the data. That leads to better analysis and it also helps to prevent fraud. Unfortunately, that is needed too. There is an international movement to make data FAIR. Where FAIR is an acronym. F is for Findable. A is for Accessible, because if you know where it is but you don't get access it is a bit like a locked cookie jar. I is for Interoperable, because you want to be able to use it in combination with other data. Finally R is for Reusable. The latter for instance means that you need to know how exactly the data was

produced -- something we call data provenance -- and of course you also need a license that allows you to reuse it.
I spend a lot of my time on organizing interoperability. It means for instance that you need to know that things that are present in different resources actually are the same. Now it might seem to you that the easy way to do that would be to just name them the same. But unfortunately that is not always possible. That is in part because scientists are humans too, and where it would help if we all started to speak the same language that is just not going to happen. So even if working with one standard would be beneficial we need to be realistic and accept that some people use inches while others use common sense (ehhm centimeters). Apart from that different experts use different terms often for a reason. Or in one field of expertise you may need more granularity for some things and less for some others. Some researchers may just need white and black while others need fifty shades of gray. We need ways to get exact descriptions of what people mean using vocabularies and ontologies that are common in their own field. We often need to extend these vocabularies and ontologies. By the way, an ontology is a neat scientific way to describe something like a tree. An ontology used to describe an actual tree would know that the tree has branches and the branches have twigs and the twigs have leaves. That is not even all. Consider a leave that has fallen from the tree in autumn. Is that still a leave? For many purposes it is, but not for all. If you are looking at photosynthesis for instance you might not want to consider that fallen leave a leave. In this context we say science uses lenses to look at how we describe things, not just which things we consider equal, but also when we consider them equal and when not. We work with some computer science groups in the UK and at the VU to make these lenses really work in data science.

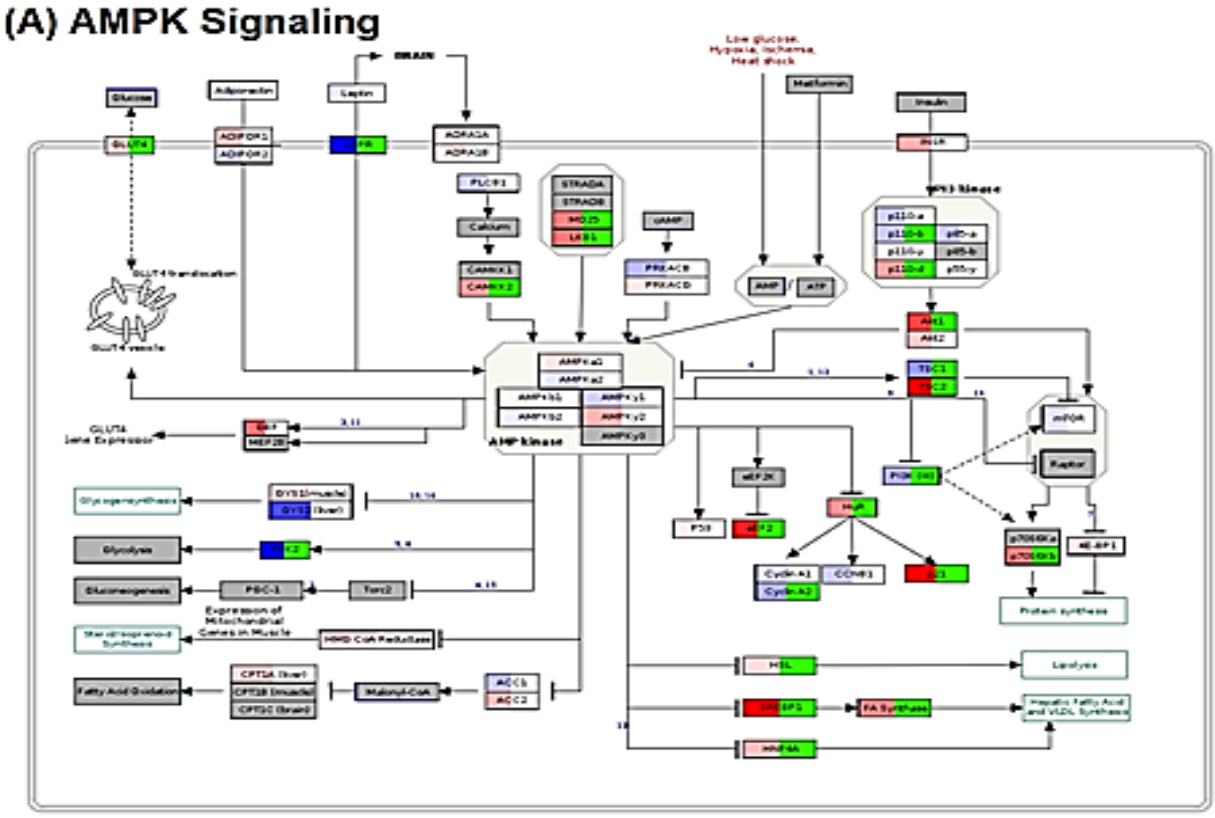
My colleague professor Mons, who is present here, focuses on the whole FAIR problem and encourages the development of an ecosystem of FAIRports. Building a useful FAIRport system will be whole lot easier however if we organize the data in a sensible way in the first place.



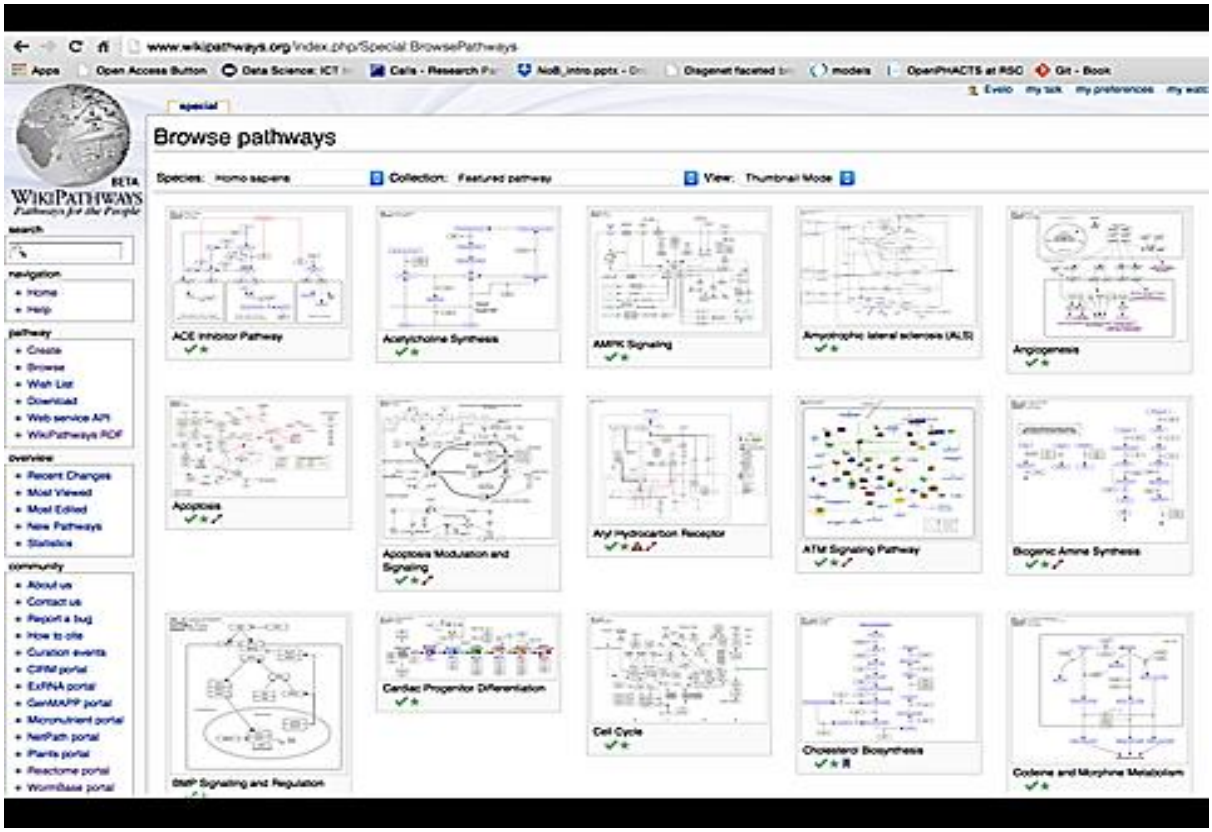
I created this diagram for the European data meeting at the Southern Observatory that I mentioned before. It is far less beautiful than the nebula picture from that observatory, but worth a look. First of all you should realize that all the arrows point downwards. At the top level are the FAIRports for the different domains. They have study capturing resources that are specific for a domain. The nutritional resource for instance knows that we are interested in diets and we typical perform cross over studies and do challenge tests. We are building such resources in international collaborations. They are, or should be, linked to the repositories on the left; the kind of resources that we typically use for big parallel datasets: they hold all the public genomics data, all the metabolite measurements, and so on. That box has a green border because to a large extend it is already there. The resources on the center right unfortunately are only in early stages of development. They are supposed to hold all "the other" things that researchers measure. Meaning they always need to contain a combination of what was measured, and how, and what the result was. The bottom part literally is closer to home. This is a repository like Maastricht University now develops to allow researchers to store and expose their own data. That is important and needed. But like I said, all the arrows in this diagram point downwards, and that is for a reason. If you want to expose your research data for reuse, and you for instance want to have it cited, then you need to make sure that the relevant parts are in the more international technology or research field specific resources at the higher levels. You need to

make sure that technology related data is described in the correct way, linked from these study descriptions at the top and put in the correct repository on the left. Finally all these other things you measure should go into domain specific resources on the right that are now being developed. That means Maastricht University does not just need a big local resource. What we do need too are trained data specialists that actually know about these other resources, especially the technology specific ones on the left, since these are used in multiple research fields. The researchers themselves should know about the resources specific for their own field. Unfortunately both these conditions currently are not met. Without that building your own data repositories might create more problems than it solves.

So far I have talked about the complexity of the biological system that underlies health and disease and described it as a large puzzle. I have indicated that what we understand and measure about the molecules and the interactions between those molecules can provide us the pieces of the puzzle. And I have described that we need to organize these pieces of information so that we can reuse them. The title of my lecture is “puzzling with the pathways of life”. So where are the pathways? Well let me show you one.



This is what biologists call a pathway. It shows an important regulatory process that in a liver cell activates a larger set of proteins. In fact you can see the liver cell itself. The double line around most of it is easily understood by biologists as the cell wall and the proteins that are drawn embedded in that wall thus are likely pumps that transfer things in or out the cell (like the glucose transporter on the top left), or receptors that bring a hormone signal to the inside (like the insulin receptor on the top right) or the opposite like the blue, green leptin exporter a bit right from the glucose transporter. Such a pathway thus holds a lot of information that is intuitively understood by an expert. This specific pathway was used in the diabetic liver study that I talked about when I showed you the network of blue regulating proteins affecting many genes. The green right half boxes show that the regulation of a gene was significantly different in the diabetic patients, the blue-red color shades to the left of that show whether that change was up or down. Pathways thus help you understand the process and assist in understanding the data. In fact they are a means to combine what we measured with what we already know. If you would work with that pathway on a computer screen you could also just click one of the many molecules is the pathway and get more information about it from various databases. For people from the 21st century that might not come as a big surprise. You are used to having pages with information showing up, and being clickable. But this is what the Internet was invented for originally. Bring all that information to the fingertips of the scientists. A lot of knowledge actually is available in different resources on the Internet. You can for instance easily and even fully automatically find information about genes, proteins and their relation with diseases. The GWAS resource that I showed before showing all the known variant-trait relationships is an example of that too. We created **WikiPathways**, in collaboration with another bioinformatics group at the University of San Francisco. WikiPathways a wiki, like Wikipedia. Everybody can contribute to it.

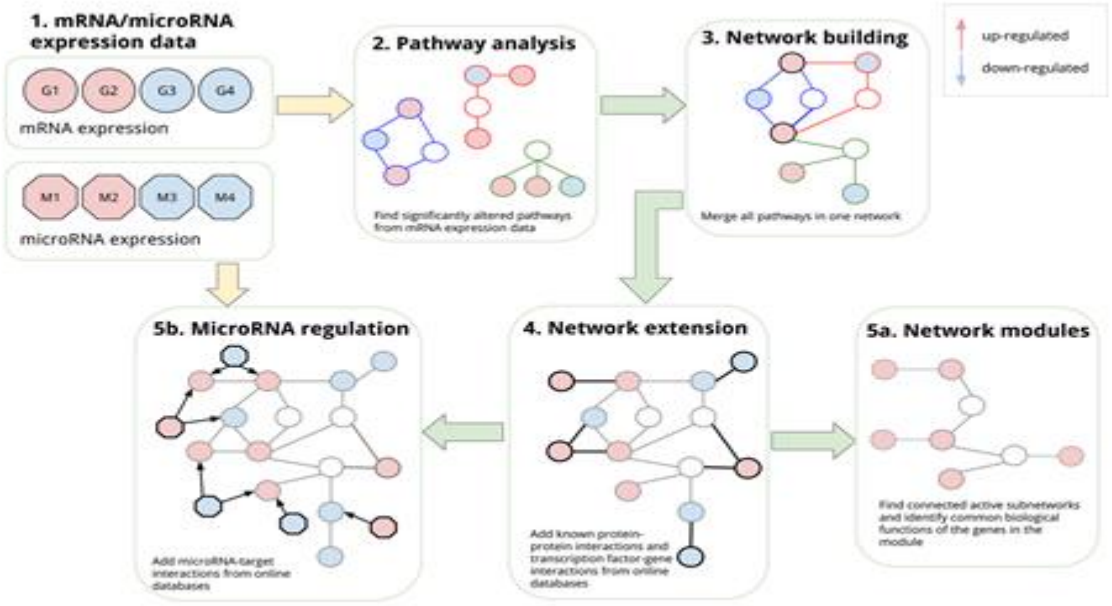


It has specific tools that make it easy to edit pathways and to use them to do the kind of analysis I just showed. We have around 2000 of such pathways and there are ten thousands of edits on these. So yes, we do get the information directly from the experts. Adding that information plus the evidence, for instance as references to the original scientific papers is very useful because it provides a context for research and because it helps solve the problem that all these things we know about are hidden in the literature and are not readily available in a computer readable or even in a comprehensive human readable form.

It is not just us or just WikiPathways. The scientific community as a whole has done enormous amounts of database curation, the monks work. I went to a Biocurator conference once and was surprised to see hundreds of people there alone that just read and evaluate papers and improve databases every day. Yet, when we talk about teaching our students academic skills about how to “find knowledge” we only teach them how to search the literature. Not where the best comprehensive resources are, how you search them, how you evaluate how good they are. We also do not teach them how to use automated programmatic methods to find that data. Today we all have smart phones with more computing power than what we used to bring people to the moon but we still only learn our students how

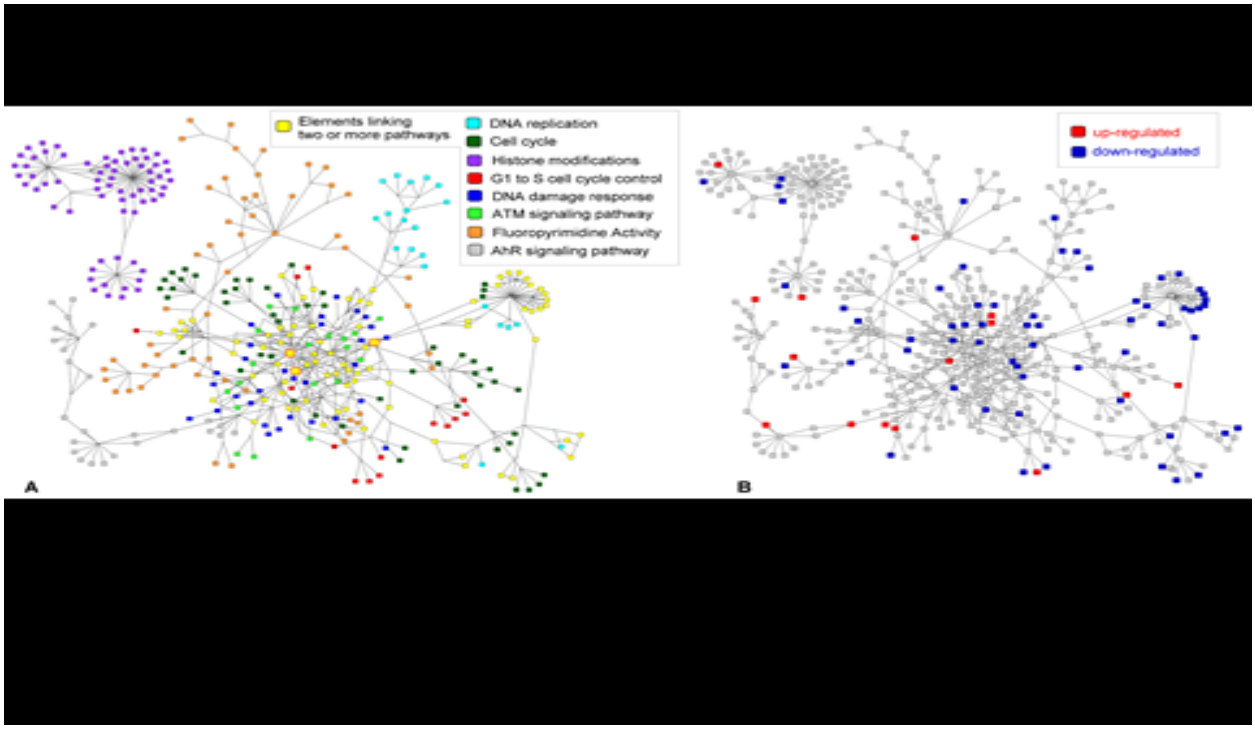
to produce new data and not how to experiment with things we already known.

Let me finish with another real world example that we very recently studied. As you may know Vitamin D is considered one of the more critical micronutrients, we are not really sure we always get enough of that especially at older age. That in it self leads to ideas about precision medicine. If older people really need more, we should come up with age dependent advices. Genetic variations influence how we convert Vitamin D in different forms. That is important for how much each of us needs individually.
We asked ourselves what Vitamin D really does on a systems level. My personal understanding is that it is a kind of spring hormone. When the sun starts to be more powerful in spring it lets you produce more Vitamin D. That Vitamin D then targets a lot of genes involved in a lot of processes that all make you more active. That maybe makes you want to do the fun things many animals do in spring. Vitamin D also plays an important role in prostate cancer and since there were some interesting datasets available we decided to look at these. This diagram shows what we did.

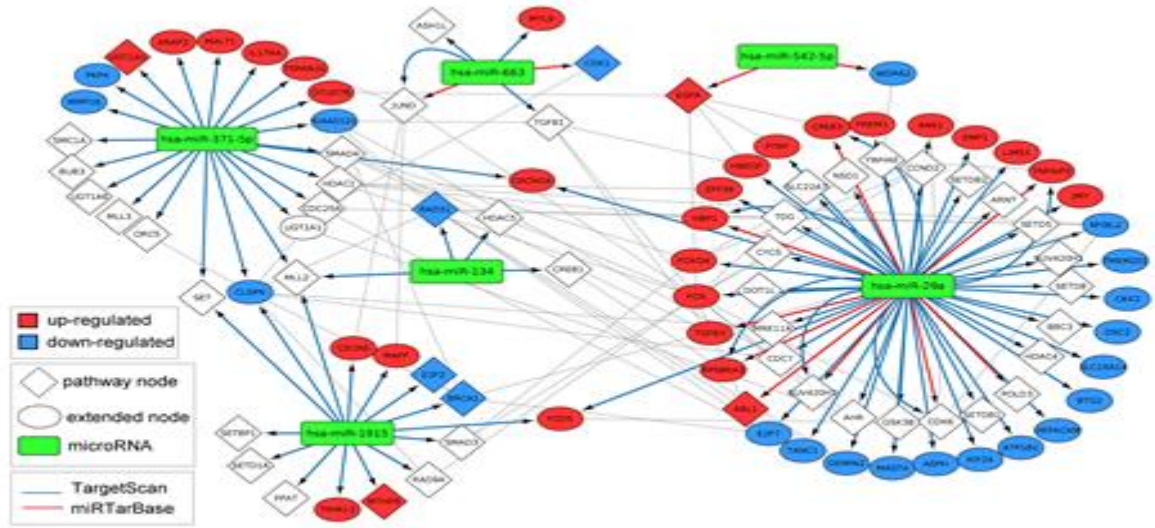


We took two datasets: one about expression of normal genes, the other about expression of small regulatory gene products. We did pathway analysis with the first set and found a number of cell cycle related and even a number of cancer related pathways that were changed in prostate cancer cells that were treated with Vitamin D. We built a large network out of these cell cycle related pathways

and… we extended that with neighboring proteins and genes from other resources. In that way we were able to find a very large fraction of the regulated genes in a single network. Finally we combined that with the miRNA data and looked for active parts of the network. The parts that we think might be the corner pieces of the puzzle.



The next picture gives an impression of the combined network, even before we extended that. Yes, it gets big. Vitamin D really does a lot of things. Note the number of blue, down regulated proteins to the very right. The left part shows that these actually appear in the overlap between cell cycle and DNA replication. That is a core aspect of cell division and thus really relevant for cancer.



The final slide shows some of what we think of as the corner pieces for this specific puzzle. Small gene products that influence genes in the pathways we found (the diamonds) or regulated genes known to be directly connected to these pathways (the ellipses). All these genes are affected by these central miRNAs. Of course such an end, never really is the end. We now need to evaluate this further, using yet other types of information and we need to check it with cancer specialists to see whether we can use this for therapeutic purposes already. It may also help us to better understand what Vitamin D does, and how much of that we really need.

I deliberately showed these four more complicated slides at the very end. I wanted you to understand what I mean by combining things. But I wanted you to have some idea of what we combine first. Professor Gert-Jan van Ommen director of the Dutch Biobanking initiative BBMRI-NL was once asked what the focus is of bioinformatics research in the Netherlands. He said: “we focus on everything”. Well… We focus on combining everything. Mostly we leave it to others to find the pieces. And actually, we even leave building the larger pieces, the pathways mainly to others too. What we try to do is combine all that to help solve real biological puzzles.

There are quite a few challenges that remain. We for instance want to get directionality in pathways right. So we can find drugs that hit things that hit things that are being hit by current drugs. Or we can find combinations of drugs that will work better than current single

treatments. We also want to integrate all that ongoing genetic variation work in our pathway approach more and we want to link to the modeling world where people calculate how fast processes work. On all these three things we currently have PhD students working. And all that work will play an important role in Maastricht's new Center for Systems Biology MaCSBio.

To do this work you have to work with many, many people from many different communities. That leads to a long list of people with whom I collaborate. People without whom this work would just not have been possible. Many of you were invited because you were important one way or another for the work I described or for the personal development that allowed me to do this, or just for the realization that life is fun and worth living, and for me that includes worth trying to understand it. I would need another 45 minutes to thank you all. So I won't. Even though some of you are very close to me. I do want to mention all the current and past members of my own group because they actually did the work that made this possible. I want to mention just two other people specifically. The late professor Bert Bijsterbosch from Wageningen University. During my own masters I tried to use irreversible thermodynamics approaches that I learned from Hans Westerhoff's early papers to describe why the lab research I had tried to do could never work. Professor Bijsterbosch's comment was: "you simplified the approach so far that it was no longer an irreversible process, but in doing so you actually proved your point." That remark made so early in my career always stayed with me. Make your puzzle as complicated as needed, but simplify it again to reach a solution. The other one I want to mention is professor Jos Smits. He was the one that said: "Chris you understand both about biology and about computers and we need a bioinformatics group. What do you think?" When I asked how long I could think about that he said: "we order tickets to fly to Palo Alto next Monday" so that's early enough. Jos, I never regretted the choice I made at that time. I do think this university and science in general would benefit if more people made such career changes. I also think universities should be more supportive of that.

I want to finish with this. I have shown that we currently develop a lot of infrastructure in the biomedical world that is badly needed for research. That is all done in temporarily funded projects. That infrastructure is then used to give away data for free because that is important. However, such infrastructure needs continuous support

and development. Now giving away things for free is not going to make you a lot of money. Since freely available knowledge and data is what is needed to bring science forward the consequence is that we need more long time support for that.

Ik heb gezegd.