Genome-wide Perturbations in Systems Biology

Introduction

In this course we have seen a number of techniques that help us to describe aspects of the state of a biological system. What is the sequence of a gene, how actively is a gene transcribed, what are the concentrations of key proteins, what other proteins do they interact with and how big is the flux of a metabolite?

Many of these techniques have been extended to the "omics" level. That is to say that it is possible to determine the sequence, concentration etc. of not one but of all - or at least a substantial portion of all - components of a system in parallel. However, the ability to describe a system's state is only one step on the way to understanding a system's function. A truly thorough understanding of a biological system includes the creation of a model that can predict how a system's state will change in response to a perturbation.

Traditionally the flexibility and throughput of the tools for introducing perturbations lag behind those of the tools for monitoring the state of the system. Hence, early perturbation studies focused on the disturbance of just a few carefully selected components of the system (e.g. the knock-out mutant for a key enzyme was compared with the wild-type organism). But, this situation is changing rapidly and tools for introducing precisely targeted perturbations on the genome-wide scale are starting to be used routinely.

This hand out describes two examples of this new generation of perturbation methods: RNAi is used to manipulate gene expression levels and CRISPR/Cas9 to manipulate DNA sequences. We will further discuss experimental strategies that allow these RNAi and CRISPR/Cas9 experiments to be performed on a genome-wide scale.

RNAi: from virus defense mechanism to research tool

RNAi stands for RNA-interference and describes a cellular process, during which short non-coding RNA molecules recognize other RNA molecules, for example messenger RNAs that code for proteins, and trigger their degradation.

The cellular processes underlying RNAi have their evolutionary origin in an ancient defense mechanism that evolved to protect organisms against double-stranded RNA viruses. Most eukaryotic organisms have retained the molecular machinery for these RNAi processes, *S. cerevisiae* being an exception. From these origins RNAi has evolved into a complex system that uses partially overlapping molecular pathways to achieve a number of different biological tasks.

The RNAi response that is best understood, is the response to so-called siRNAs (short-interfering RNAs). This is also the one used in the biotechnological applications discussed below. These siRNAs are very short (approx. 20bp) double-stranded RNA molecules that guide the RISC complex (RNA induced silencing complex) to recognize, degrade and thereby suppress the function of other RNA molecules (e.g. protein-coding messenger RNAs). What makes siRNA-based RNAi so attractive from a biotechnological standpoint is that the recognition between the siRNA and the targeted RNA is based on straightforward Watson-Crick base-pair complementarity. So in order to suppress the expression of a gene, the experimenter only needs to know the gene's sequence. Then he will be able to design a siRNA that specifically targets and degrades the RNA molecule transcribed from this gene. The applicability of this process was very quickly recognized: biotechnological uses of the RNAi mechanism sprang up within a year after the discovery of the basic biological process in 1999. RNAi-based technologies have been extremely popular ever since.

Off-target effects are a common problem in siRNA-based RNAi biotechnology applications

As mentioned above, siRNA-based RNAi is only one of a number of RNAi mechanisms, through which regulatory RNA molecules can modulate the activity of other RNA molecules. In these other RNAi mechanisms however, the molecular principles that steer the recognition between the regulatory RNA and the target RNA is much less straightforward than in the siRNA case. Particularly complicating is the fact that in these other RNAi mechanisms one regulatory RNA may modulate the activity of multiple target RNAs.

It is assumed, that many of the off-target effects (i.e. cases where an siRNA also modulates the activity of RNAs that are not complementary to the siRNA) can be accounted for by siRNAs inadvertently triggering one of these other types of RNAi responses.

Another possibility for off-target effects arises from the extreme abundance of artificial siRNAs that can overwhelm the cell's RNAi machinery so that the cell's

natural RNAi-based regulatory mechanisms are disrupted. Finally, it has been shown that some siRNA precursor molecules can trigger an inflammatory response.

Continual improvements in experimental protocols, siRNA delivery mechanisms and bioinformatics tools for the design of the optimal siRNA sequences have reduced the likelihood undesired secondary effects, but appropriate control experiments remain essential for all siRNA experiments.

RNAi reduces rather than eliminates targeted RNAs

Even optimally designed siRNAs merely reduce the expression of a gene rather than completely suppressing it. Reducing the concentration of the targeted RNA by 50-90% percent is considered typical. For this reason, RNAi experiments are colloquially referred to as "knockdown" experiments in order to distinguish them from classical knockout out experiments, in which the gene itself is disrupted so that the production of RNA ceases completely. However, the incomplete suppression of gene expression in RNAi experiments also bears advantages as it opens up the possibility of "titrating" the level of a targeted RNA by adjusting the concentration of the siRNA. Hence we can observe the way in which a system's response changes as a function of the targeted RNA's concentration. For the same reason RNAi experiments are reversible and the timing of the knockdown can be controlled. Unfortunately, the kinetics of RNAi mediated responses are relatively slow. Effects take many hours or even days to manifest and reverse themselves.

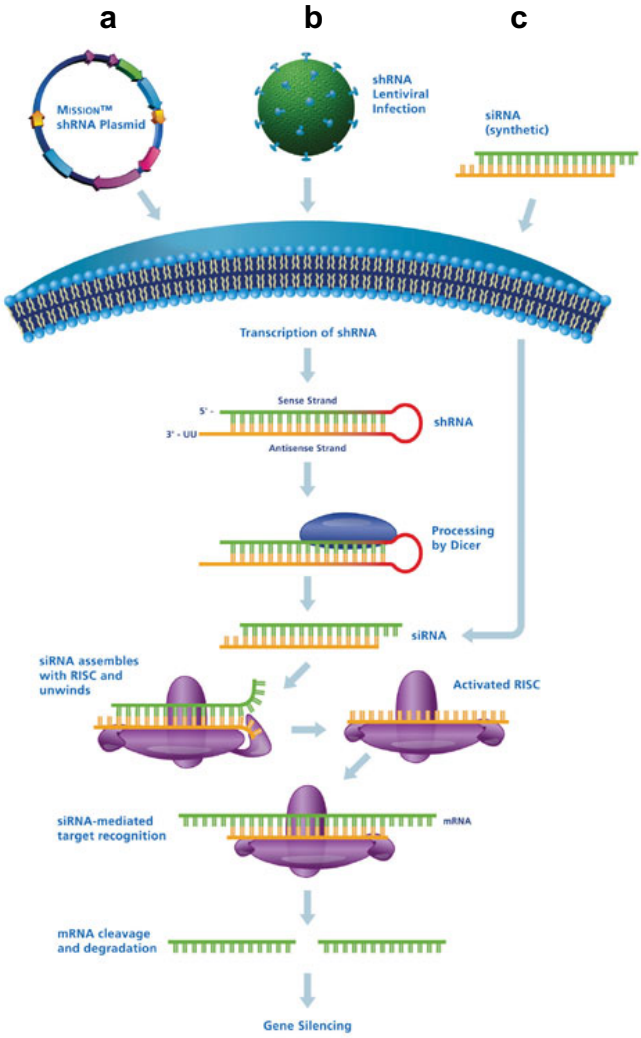

Figure 1 *Outline of the essential steps of the RNAi mechanism. The small interfering RNA (siRNA) sequences can be introduced either (**a**) via a plasmid, from which a precursor of the siRNA is transcribed, (**b**) via a lentivirus that permanently integrates the gene for the siRNA precursor into the cell's genome or (**c**) by introducing the siRNA directly. The siRNA is then bound to the RISC complex which removes one of the strands of the siRNA and uses the remaining strand to recognize the targeted RNA via base pairing. The resulting RNA dimer is cleaved and the targeted RNA is degraded.*

How to introduce siRNAs into living cells

One of the key advantages of the RNAi technology is that most eukaryotic cells naturally possess the cellular machinery to mount an RNAi response. All the experimenter needs to do is to provide siRNA molecules, or their precursors (e.g. short hairpin RNAs) inside the targeted cells. A number of alternative experimental strategies to introduce these RNAs have been developed and continue to be refined (Figure 1).

In Figure 1 you can see two basic strategies for introducing siRNAs into a cell. The first is to introduce synthetically generated siRNAs directly into the cell using reagents that transiently make the cellular membrane permeable. The second strategy is to introduce a piece of DNA (either inside a virus or in a plasmid) that codes for the siRNA precursor. This DNA is then either integrated permanently in the cell's genome, or it is passed from mother to daughter cells as an extra-chromosomal element. The production of the siRNA precursor then takes place inside the cell (*in situ*) where it is controlled by a specific promoter that was delivered on the same piece of DNA that codes for the siRNA.

siRNA precursors can be targeted to specific tissues and their production can be timed.

Particularly the virus-based methods for generating siRNAs *in situ* are continuously being improved, taking advantage of the full molecular biology toolbox. Figure 2 shows a still relatively simple sequence of a plasmid, into which a desired siRNA-coding sequence can be introduced using standard restriction-enzyme-based cloning techniques. These plasmids can be used to generate infectious lentiviral particles, which deliver them into eukaryotic cells. Inside the cells, these sequences are irreversibly integrated into the genome. In the example shown in Figure 2, the expression of the siRNA is controlled by the U6 promoter, which results in a continuous mid-level expression of the siRNA in all cells and tissues. Of course other promoters could be used: a tissue specific promoter for example, that transcribes the siRNA only in certain types of cells (e.g. liver cells) or an inducible promoter, which transcribes the siRNA only when a certain small molecule is introduced by the experimenter. The list of molecular biology tricks to target and control the expression of the siRNA-precursors inside the living organism is far too long to be listed here.

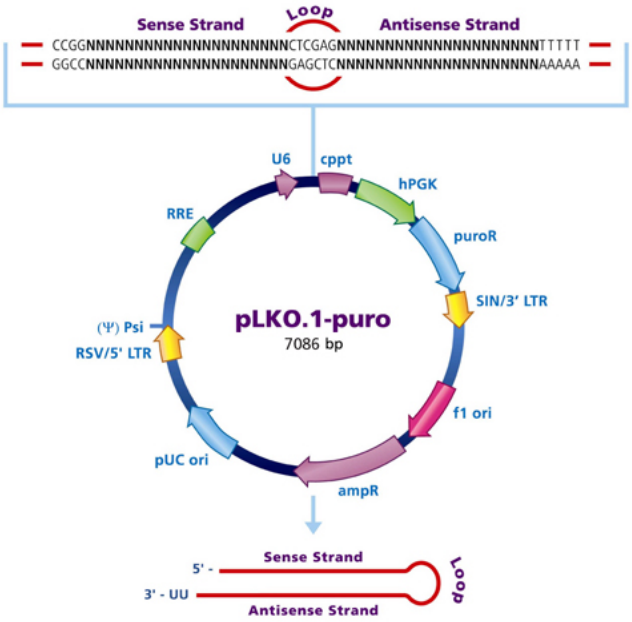

Figure 2 *Schematic map of the 7086bp long sequence of a lentivirus transfer vector that can be used to generate lentivirus particles, with which the DNA sequence coding for an siRNA and the promoter controlling its expression can be delivered and stably integrated into a host cell's genome. The elements named pUC ori, ampR and f1ori are elements that allow the vector to be propagated in E.coli cells for cloning. The RSV/5' LTR and SIN/3' LTR sequences allow the insertion of the intervening DNA sequence into a host cell's genome. The RRE sequence faciliates the packaging of the vector into the viral particle, the U6 promoter drives the expression of the siRNA sequence and the hPGK promoter ensures the production of the puroR puromycin resistance gene, which can be used to select those cells where integration of this DNA sequence into the genome was successful.*

CRISPR/Cas9 a new technology for RNA-directed DNA-editing

The origins of the CRISPR/Cas9 technology for genome editing lie in a bacterial self-defense mechanism that allows bacteria to protect themselves against recurring infections by viruses and transposons. The reason this rather arcane biological mechanism has made it to the cover pages of the world's newspapers is that its basic mechanism has been adapted to introduce precisely targeted cuts and sequence changes in eukaryotic genomes with great ease. Before the discovery of the CRISPR/Cas9 mechanism, introducing a genetic modification in a specific place in a eukaryotic genome was a labor-intensive and time-consuming trial-and-error process that required a great deal of expertise. Through CRISPR/Cas9 it has become a quick and straightforward

procedure that a trained technician can perform in a matter of weeks.

To learn more about the biological origins of the CRISPR/Cas9 system and how the natural mechanism was adapted for biotechnological uses, please have a look at the review article by Doudna & Charpentier provided on the Moodle page.

The reason the CRISPR/Cas9 system was such a breakthrough in genome editing is that the activity of the Cas9 protein's DNA cleavage activity can be directed to virtually any arbitrary location in a genome by a so-called single guide RNA (sgRNA), which identifies the targeted sequence by Watson-Crick base pairing. In this sense the CRISPR/Cas9 system is very similar to the RNAi technology. The important difference is however, that the CRISPR/Cas9 system acts on the DNA and not on the RNA level.

The basic function of the CRISPR/Cas9 system is relatively simple (Figure 3). Cas9 is a large multifunctional protein that binds a to an RNA molecule called a single guide RNA (sgRNA) and also binds to DNA. As the Cas9 protein slides along the DNA it pries open the DNA's double helix. As soon as it finds a sequence of DNA that can form perfect Watson-Crick base pairs with the 5' sequence of the bound sgRNA, the Cas9 protein clamps down on the DNA and uses its catalytic domains to cleave the DNA. The cell's own DNA repair mechanism then fixes this break in the DNA strand, but in doing so it introduces a number of nucleotides, thus disrupting the sequence of the cleaved gene so that the targeted gene loses its function (e.g. by introducing frame-shift mutations).

Unlike the RNAi mechanism, which occurs naturally in most eukaryotic cells, the Cas9 protein needs to be introduced into the targeted cells as well. This means that to disrupt the function of a gene in a cell's genome, one needs to introduce into the cell the gene for the Cas9 protein along with the gene for the sgRNA whose 5' end has been chosen to match the target gene. Luckily the expression of these components and their assembly into an active Cas9-sgRNA complex as well as the recognition and cleavage of the target sequence is indeed a very efficient process in most eukaryotic cells.

The mechanisms for transferring the Cas9 and sgRNA genes into the host cell are very similar to the plasmids and lentiviral mechanisms used for the RNAi technology. The same tools can be used to limit the expression to a specific tissue or to induce them at a particular time. Accordingly, the experimenter can select the tissue in which a gene is disrupted (or another DNA sequence is introduced), as well as the timing of the event.

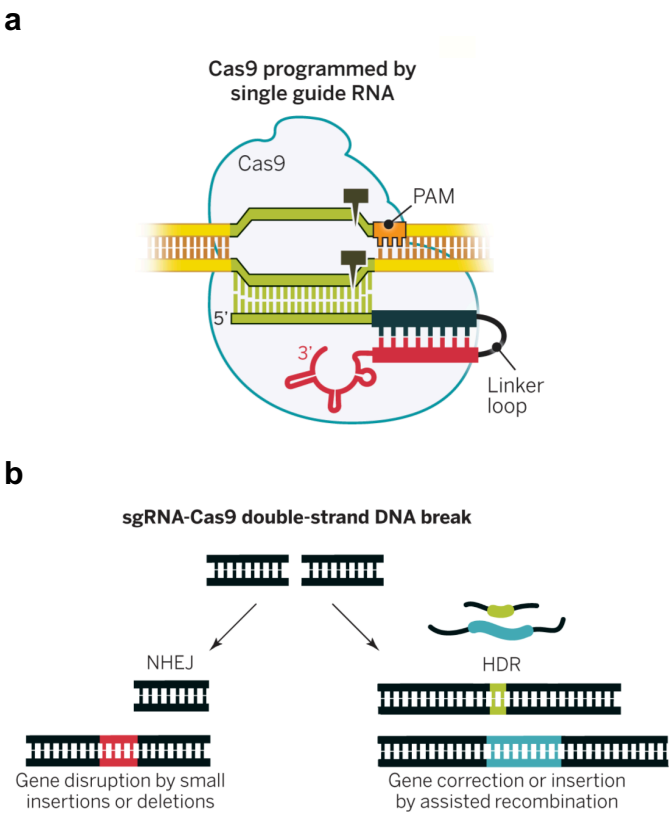

Figure 3 *The Cas9 protein (**a**) is an RNA-guided DNA restriction enzyme that cuts genomic DNA at sequences that can form Watson-Crick base pairs with the 5' portion of its bound sgRNA. The sgRNA is shown in three colors where the red and black part need to have a conserved sequence that facilitates binding of the sgRNA to the Cas9 protein but the green part can be freely chosen to recognize any DNA sequence. (**b**) The DNA double-strand break introduced by Cas9 is repaired by the cell's natural DNA repair system. Either the ends are joined non-homologously, which results in the insertion of a short stretch of random sequence and is likely to disrupt the function of the cut gene, or by homologous DNA-repair. The latter mechanism can be used to introduce new DNA sequences into the genome at the site of the cut. (adapted from Doudna & Charpentier Science 2014)*

Scaling up from single-gene to genome-wide perturbation experiments

In both RNAi and CRISPR/Cas9 technology, the targeted gene or RNA is specified by a linear RNA sequence, which recognizes its target via traditional Watson-Crick base pairing. Therefore, the only information that is needed to perturb (e.g. turn off) the function of a specific gene or RNA is the sequence of that gene or RNA. Because whole genome sequences are available for all relevant organisms, the task of designing the reagents

that will target each individual gene or RNA is reduced to a straightforward bioinformatics task that is easily automated. Generating the necessary reagents in the lab is almost as simple and comprises synthesizing the corresponding DNA sequences, which can then be introduced into a plasmid or lentivirus delivery systems with the help of standard molecular biology techniques. (At the end of this hand out you will find a description of how the many different DNA sequences needed for a genome-wide perturbation experiment are generated.)

Genome-wide perturbation experiments are experiments, in which all of an organism's genes or RNAs are targeted, but it is very important to note, that of course this is done one gene/RNA at a time. We generate a large number of cells or organisms where in each of them the function of a single gene or RNA is perturbed but in each cell it is a different gene or RNA. In principle it is conceivable to perform experiments, in which combinations of two or more perturbations are generated in each cell, but those types of experiments would be difficult to perform on a genome-wide scale because of the number of possible combinations being far to great for any highly automated process. In order to still examine perturbation-perturbation interactions, the current strategy is to perform genome-wide single perturbation experiments first and then to combine those perturbations that seem relevant because they change the behavior of the system in subsequent experiments.

Three types of genome-wide perturbation experiments: screens, selections and population shifts.

Most genome-wide perturbation experiments follow one of the following three strategies: screen, selection or population shift. The conceptually most straightforward approach is a **screen**. In a screen we create for each targeted gene (or gene segment) a separate pool of cells/organism and in each of these pools we perturb the function of one specific gene so that all cells in a pool are perturbed in the same way. We then observe each of these pools trying to find out in what way the perturbation changed the properties of the system. This means that we have to handle several tens of thousands of cell culture pools (one for each targeted gene or gene segment) and have to perform a phenotypic analysis of each one. This process is very resource intensive and is only practically feasible if the growth of the cell cultures can be automated (see for example the methods used in a yeast synthetic lethality screen) and if the phenotype of interest can be detected easily (e.g. fluorescent signal produced by a reporter gene). For certain model organisms this type of screening approach is certainly not suitable. A genome-wide perturbation in mice for instance, where the phenotype would be the metabolic profile in a liver biopsy would be endlessly time consuming, impossible to automate - and it would be very difficult to grow and maintain tens of thousands of different mouse strains in the first place.

The second strategy for a genome-wide perturbation experiment is a **selection**. In a selection one large pool of cells/organisms is transformed with a library of DNA constructs (e.g. lentiviruses or plasmids) that contain all siRNAs (or sgRNAs) to be tested. The pool of cells and DNA constructs has to be sufficiently large so that we can be certain to obtain for each DNA construct at least one cell/organism that carries it. The resulting population of cells is then grown under conditions where only those cells that show the phenotype of interest are able to survive. (Finding such growth conditions that translate the phenotype into a survival advantage is the great challenge in designing successful selection experiments.) These surviving cells/organisms then form colonies that can be detected by eye. These colonies are isolated and the siRNA or sgRNA sequence that allowed the cells/organisms to survive can be determined by DNA sequencing, using primers directed at a constant region of the employed DNA construct.

The third strategy is a **population shift** experiment. In this type of experiment one generates a very large pool of cells/organisms in which each siRNA (or sgRNA) construct is represented multiple times. This cell pool is then split into a small number of sub-pools of which each is grown under different conditions. Depending on the gene that was targeted, these cells grow and divide either faster or more slowly under different growth conditions and will end up being under- or overrepresented in the different pools. In other words: the population of cells will shift, because cells in the possession of a construct causing them to be better adapted to the growth conditions to which they have been subjected will thrive. Meanwhile other cells, whose inserted construct is not advantageous to their growth under these conditions, will grow slowly or not at all. Next generation sequencing is then used to determine the relative abundance of the different siRNA- (or sgRNA) coding constructs under the different growth conditions. This then allows the researcher to determine, which perturbations favor and which disfavor growth under different conditions. In the lecture by Prof. Beat Christen you will hear about a transposon-based method to perform such population shift experiments.

RNAi and CRISPR/Cas9 experiments are typically performed on cell lines but *in vivo* experiments are becoming possible.

Genome-wide perturbation experiments performed with RNAi and CRISPR/Cas9 are typically performed in cell-cultures rather than in whole organisms. The reason for this is that cell culture experiments are technically much simpler. First, it is much easier to introduce RNAi or CRISPR/Cas9 reagents into cells living in culture. To generate whole organisms, in which the same gene is knocked down/or altered in every cell, the modification typically has to be done on the egg cell, which in mammals then needs to be re-implanted into the mother animal - a technically very demanding process.

Also, the number of organisms required for genome-wide perturbation experiments is very great. Both RNAi and CRISPR/Cas9 experiments usually target 3-4 locations in each gene and for each locus several organisms are generated (both to bolster statistical significance and to ensure that, when a library is used, most genes are inactivated in at least one organism). Hundreds of thousands of organisms would have to be generated, raised and phenotyped, meaning that the procedure would quickly become very resource intensive.

In what regards humans, whole-organism experiments are obviously excluded for ethical reasons. This is of course not an issue in cell culture-based experiments.

The main disadvantage of using cell-cultures is that although they can be chosen to resemble the conditions in the whole organism, they generally fail to reproduce the complex interactions and spatial arrangement of multiple cell types that are often essential in the biological process under study. Tissue invasion in cancer research, or the build-up of plaques in blood vessels, for example, are difficult to reproduce in cell culture.

These shortcomings of cell culture experiments are widely recognized and three main approaches are adopted to address them.

Techniques for whole-organism genome-wide perturbation experiments are being developed and continually improved. For one of the smallest multicellular model organisms the nematode worm *C. elegans,* for example, genome-wide knockdown experiments are relatively common. Even in mice, creative experimental designs are starting to make it possible to randomly perturb different genes in different cells inside a specific tissue of a mouse embryo. The growth and proliferation of these differently perturbed cells can then be observed in a population shift-type experiment, all in the context of an intact living organism.

The second approach is designed in two steps, the first in which a series of genome-wide cell culture perturbation experiments (each designed to resemble a different aspect of the biological process of interest) are used to identify a set of candidate genes, which can then be perturbed in a second step in whole-organism experiments.

The third approach is based on organoid cultures. Organoid cultures can be thought of as sophisticated cell cultures, in which precursor cells are induced to grow into miniature versions of the natural tissue with all of the essential cell types arranged in the way they are found in the natural tissue. Such organoid cultures have now been developed to mimic the complex architecture of a variety of natural organs such as gut, kidney, pancreas, liver etc. Organoid cultures thus correspond to an intermediate level between whole organism experimentation and traditional cell-culture research.

Generation of custom nucleotide libraries

For genome-wide applications RNAi und CRISPR/Cas9 technologies depend on large libraries of short DNA/RNA strands (called oligos), each of these oligos specifically targeting a different gene in the genome. This requires a minimum of >20'000 different types of oligo - each of them with a distinct and precisely defined sequence. Also, each gene is targeted not just once but with 3-5 different oligo sequences, the experiment thus requiring libraries of ~100'000 custom-designed oligos. For the creation of such a library, traditional methods that synthesize oligos one at a time would be time and cost intensive.

So large oligo libraries are generated using highly automated and parallelized methods, such as the photolithographic solid surface method shown in Figure 4. The oligos are then released from the surface and are either pooled or released in a controlled way so that each type of oligo can be collected in a separate well of a microtiter plate.

For commonly studied organisms (e.g. drosophila, rat, mouse and human) such libraries are now commercially available.

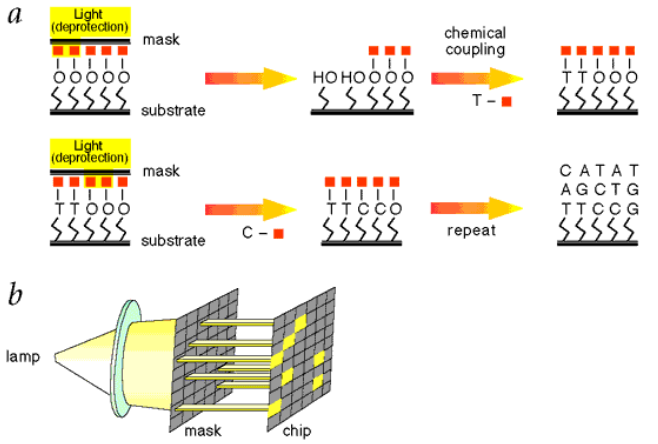

Figure 4 *a) Schematic representation of the principle behind the photolithographic method for custom oligonucleotide library synthesis. Light is used to de-protect (i.e. chemically activate) the ends of growing oligonucleotides in a well-defined area of a solid surface. Subsequent addition of a specific nucleotide (in this case a T) leads to incorporation of this nucleotide only in this activated area. By repeating the de-protection and chemical coupling steps with all four nucleotides (A,T,G,C) any desired nucleotide can be inserted at any given position, thus allowing any desired of nucleotide sequence to be assembled across the solid surface. b) The light activation of growing oligo chains in specific areas on the solid surface is achieved via a computer-controlled mask (equivalent to an LCD display) which allows the automation of the synthesizing process. (image source: Lipshutz et al. Nat. Genetics. 21 pp.20 (1999))*