Deliverable D5.4

Funding Scheme: THEME [ICT-2007.8.0] [FET Open]

Paving the Way for Future Emerging DNA-based Technologies: Computer-Aided Design and Manufacturing of DNA libraries

Grant Agreement number: 265505

Project acronym: CADMAD

Deliverable number: D5.4

Deliverable name: High level description of application libraries

Contractual Date[1] of Delivery to the CEC: M24
Actual Date of Delivery to the CEC: M25
Author(s)[2]: Sandra Meyer, Frank Edenhofer
Participant(s)[3]: UKB
Work Package: WP5
Security[4]: Pub
Nature[5]: R
Version[6]: 0.0
Total number of pages: 33 (including appendix)

1 As specified in Annex I

2 i.e. name of the person(s) responsible for the preparation of the document

3 Short name of partner(s) responsible for the deliverable

4 The Technical Annex of the project provides a list of deliverables to be submitted, with the following classification level:
Pub - Public document; No restrictions on access; may be given freely to any interested party or published openly on the web, provided the author and source are mentioned and the content is not altered.
Rest - Restricted circulation list (including Commission Project Officer). This circulation list will be designated in agreement with the source project. May not be given to persons or bodies not listed.
Int - Internal circulation within project (and Commission Project Officer). The deliverable cannot be disclosed to any third party outside the project.

5 **R (Report):** the deliverables consists in a document reporting the results of interest.
P (Prototype): the deliverable is actually consisting in a physical prototype, whose location and functionalities are described in the submitted document (however, the actual deliverable must be available for inspection and/or audit in the indicated place)
D (Demonstrator): the deliverable is a software program, a device or a physical set-up aimed to demonstrate a concept and described in the submitted document (however, the actual deliverable must be available for inspection and/or audit in the indicated place)
O (Other): the deliverable described in the submitted document can not be classified as one of the above (e.g. specification, tools, tests, etc.)

6 Two digits separated by a dot:
The first digit is 0 for draft, 1 for project approved document, 2 or more for further revisions (e.g. in case of non acceptance by the Commission) requiring explicit approval by the project itself;
The second digit is a number indicating minor changes to the document not requiring an explicit approval by the project.

Abstract

The goal of this work package is to provide a basis for generation of an interface between computer scientists and biologists, thereby bringing together developers of the DNA programming tools and potential end users of the system. This deliverable describes DNA libraries from different biological research fields designed by using the newly developed DNA library designer software DNALD and their potential biological applications.

Keywords[7]:

DNA Library design, DNALD, validation of CADMAD system

1. Introduction

a. Aim / Objectives

CADMAD aims at replacing conventional *de novo* DNA synthesis not only by high throughput computer-aided and automated DNA processing but also by exploiting DNA reuse on a large scale. The success of CADMAD technology will critically depend on a powerful and user-friendly interface and a high level of end user compliance. Work package 5 is intended to provide a profound basis for the generation of a reliable and robust interface between computer scientists, who develop the DNA processing tools, and biologists, as potential end users of the developed system. This deliverable deals with the high level description of application libraries designed by the end users of the system by using the newest version of the recently developed software DNALD. This deliverable is the first step to achieve our goal of testing and challenging the various subsystems of the CADMAD system. The libraries described here will, at a later stage of the project, be used to verify the breakthrough that computer-aided design and manufacturing can be effectively employed in DNA-based research and development.

b. State of the Art

At present the direct synthesis of genes is the most efficient way to generate functional genetic constructs. For this *de novo* synthesis activated monomers (protonated deoxyribonucleoside 3'-phosphoramidites) are sequentially added to a growing chain that is linked to an insoluble substrate (Letsinger and Mahadevan, 1965; Caruthers et al., 1987). In brief, the 3' phosphorus atom of the

[7] Keywords that would serve as search label for information retrieval

monomer is joined to the 5' oxygen atom of the growing chain, whereas a protecting group blocks the 5' -OH group of the monomer. In the next step the generated phosphite triester is oxidized to form a phosphotriester. Finally, the protecting group on the 5'-OH is removed, so that another monomer can be added to the growing chain in the next cycle. Single-stranded oligonucleotides with a length up to 100 monomers can be generated with this method. These oligonucleotides can then be enzymatically assembled to form complete synthetic genes.

In practice, several companies offer conventional gene synthesis (e.g. GenScript, Integrated DNA Technologies, Life Technologies and various others). Generally, they serve the end user with an internet-based order interface, in which the desired DNA-sequences can be uploaded by straightforward "copy and paste" action. Some companies (e.g. GenScript, Life Technologies and others) provide tools to further customize the DNA-synthesis to fulfil the needs of scientists such as codon optimization. Additionally, the DNA-sequence can be modified by removal of cryptic splice sites and RNA destabilizing sequence elements. These gene optimization steps appear to result in a maximized expression of the synthetic gene in the desired expression systems. The desired DNA-sequence is then synthesized and cloned in an output vector of choice before delivery to the end user. However, besides these obvious advantages also limitations become evident, that should be eliminated by using automated DNA processing. In this regard, the most important point is the time- and money-consumption of conventional gene synthesis. CADMAD aims at improving this point by reassembly and rearrangement of existing DNA fragments instead of conventional *de novo* DNA synthesis.

c. Innovation

To achieve the goal of automated DNA-processing a new textual DNA programming language and a new user interface as well as biochemistry and algorithms for computer-based approaches have to be developed. Afterwards, the output DNA has to be further analysed, verified and compared to conventionally synthesized DNA by applications by the end user. For this, potential end users from independent research fields involved in this work package drafted several DNA libraries using DNA library designer (DNALD) software developed in the course of this project. The diversity of the DNA libraries and their descriptions are presented in the results section of this report. The design and subsequent production of these DNA libraries should help to test and challenge the functionality and user friendliness of the CADMAD system.

2. Implementation

To analyze the functionality and user friendliness of the recently developed DNA programming language, six libraries have been drafted by the end users.

Number of library	Short name of partner	Title of library
1	UKB	Derivation of an Oct4 expression library to identify enhanced variants of Oct4 transcription factor
2	FMI	Sequence replacement library to identify determinants of CpG islands methylation states
3	UH	Glycosylation screening of the hRET-ECD
4	ETHZ	Combinatorial synthetic operon library
5	UNOTT	Investigating post-transcriptional regulation of *Pseudomonas aeruginosa* azurin by RsmA
6	UNOTT	Increase affinity of *Pseudomonas aeruginosa* PqsR quorum sensing signal receptor protein for N-oxide quinolones

The biological backgrounds and the specificities of these libraries, as well as the previous work performed and the objectives to be analyzed are described below. A comprehensive description of each library including the DNALD files can be found in appendix A.

3. Results

High-level description of designed application libraries

1. Derivation of an Oct4 expression library to identify enhanced variants of Oct4 transcription factor (UKB)

Oct4 transcription factor is involved in the self-renewal of undifferentiated pluripotent stem cells and plays a major role in changing the cellular fate of mammalian cells. For example, the reprogramming of somatic cell types into a multi- or pluripotent state involves ectopic activation of Oct4 (Takahashi and Yamanaka, 2006; Thier et al., 2012). Reprogramming can be achieved either by viral or protein

transduction or mRNA transfection of Oct4. As a transcription factor the protein consists of a DNA-binding domain, which enables sequence specific DNA binding, and two transactivation domains, that regulate gene expression by interacting with the transcriptional machinery (Figure 1).

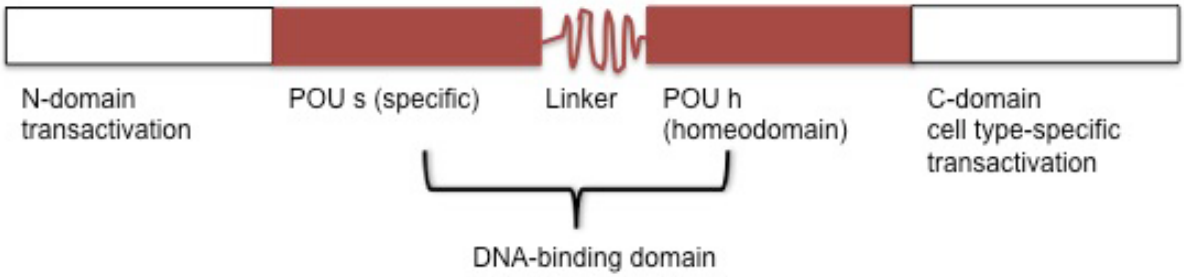

Figure 1: Schematic representation of domain structure of Oct4. Human Oct4-protein consists of 360 aa containing POU DNA-binding domain and two proline-rich regions. The POU domain has two independent subdomains: the N-terminal POU specific domain and a C-terminal hoomeodomain, which are connected by a linker amino acid stretch. Transactivation domains could be mapped to the N- and C-terminal proline-rich domains.

The designed library aims at identifying the domains crucial for reprogramming of cell fate of target cells and to generate enhanced versions of Oct4. Therefore, modified variants of Oct4 should be generated that lack one or more of their domains. Additionally, Oct4 variants are desired that are fused with different stabilizing peptides (SP1, SP2, SP3) and the potent transactivator domain VP16. The generated library products will later be used for (1) the expression of the fusion proteins in *E.coli* for subsequent protein transduction of mammalian cells, and (2) for viral transduction of mammalian cells.

2. Sequence replacement library to identify determinants of CpG islands methylation states (FMI)

DNA methylation at CpG dinucleotides is a heritable epigenetic modification associated with transcriptional silencing and is essential for mammalian development. Methylation levels are mainly controlled at the level of DNA sequence with little or no influence from neither chromatin nor the transcriptional environment. However, little is known about the identity of the regulatory sequences, which are able to tune DNA methylation states.

The two first years of the project were used to set up an experimental system allowing the profiling of methylation state acquisition from DNA libraries in a highly parallelized way (Figure 2). Using this system, thousands of fragments derived from the mouse genome were analyzed. This initial effort allowed (1) evaluation of the project feasibility and strategy limitations and (2) derivation of working

hypotheses on main mechanisms driving the acquisition of methylation profiles. The proposed library is intended to allow the experimental testing of such hypotheses.

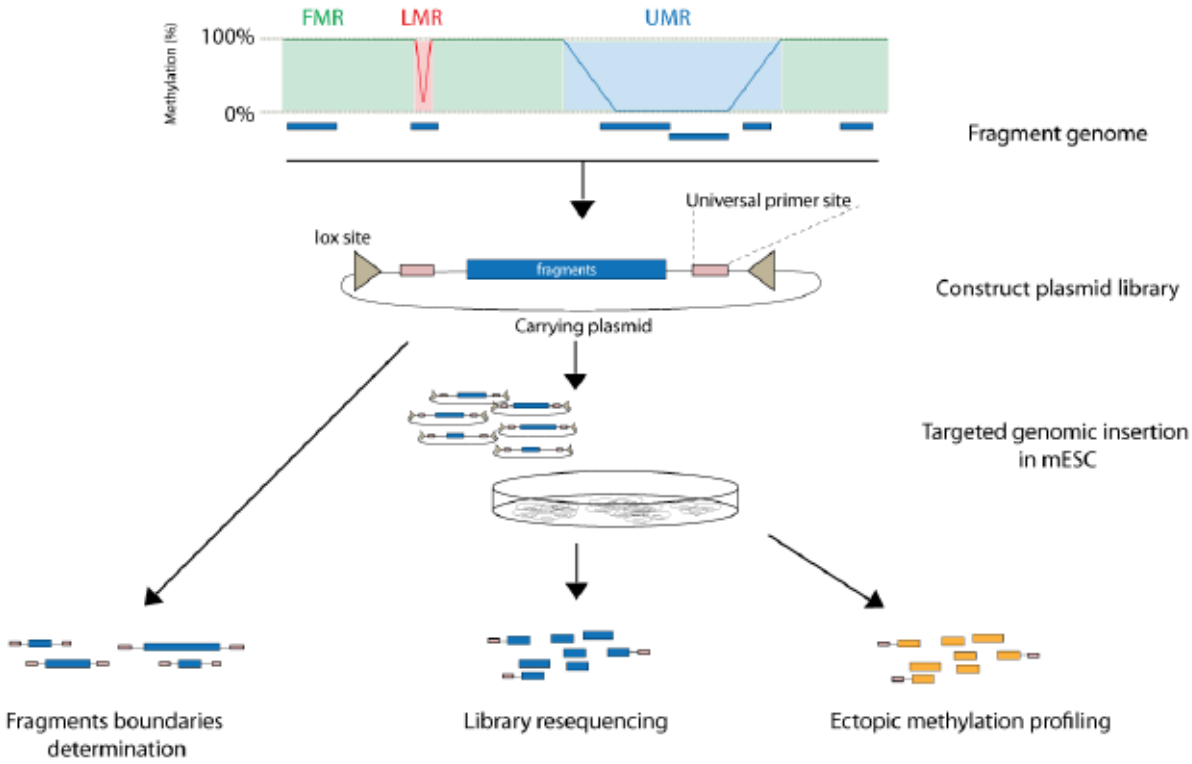

Figure 2: Summary of the experimental scheme used to test libraries. Synthesized libraries were cloned into a carrying vector for targeted, locus-specific insertion in mouse embryonic stem cells. Subsequently, the methylation status of inserted fragments was profiled by bisulphate sequencing. (FMR: fully methylated region, LMR: low methylated region, UMR: unmethylated region)

The initial work on the project has established the importance of CpG content of analyzed sequences in establishing various methylation states. In order to test the contribution of other sequence components independently, one needs to generate sequence variations within DNA fragments, while being conservative regarding the CG positions (Figure 3). Thereby, the contribution of different DNA determinants aside from CGs in the establishment of DNA methylation patterns will be obtained.

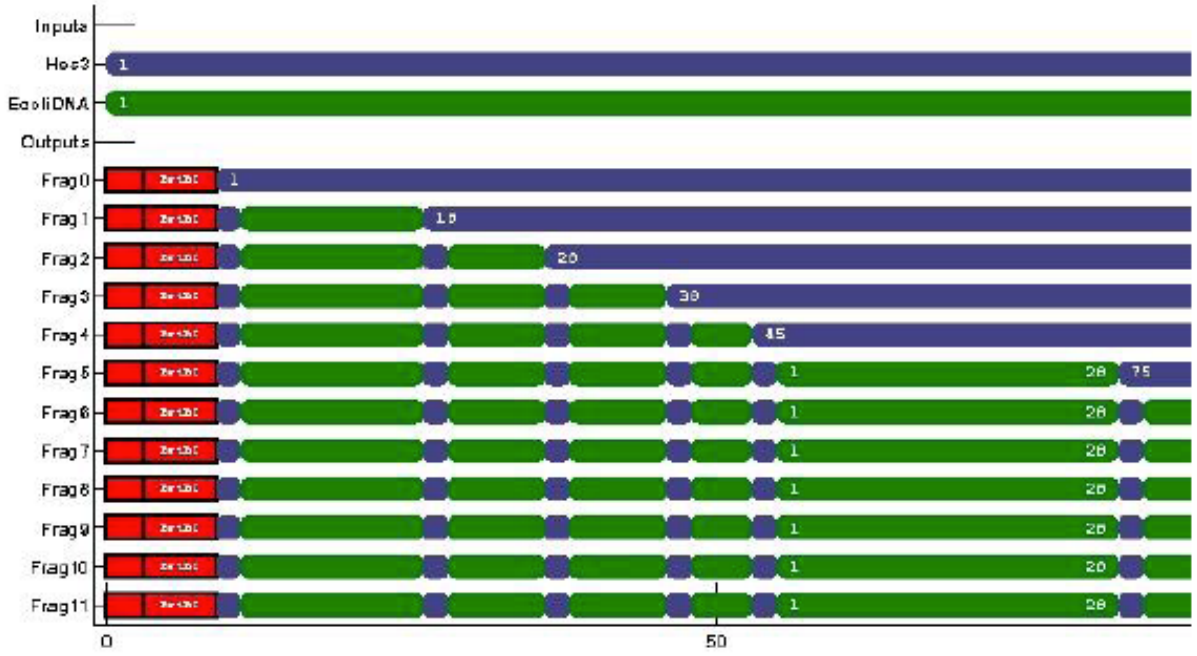

Figure 3: Sub-view of the library. Sequences flanked by CpGs are iteratively replaced from 5' to 3' *by E.coli* DNA.

3. Glycosylation screening of the hRET-ECD (UH)

Ret-oncogene receptor is a receptor tyrosine kinase which mutations are known to play a role in a variety of diseases. The extra-cellular domain (ECD) of human RET-oncogene receptor is highly post-transcriptionally modified. There is a belief that these modifications play a crucial role in protein folding, maturation and activity. hRET-ECD has twelve sites predicted to be N-linked glycosylated. However, only two sites (N151, and N394) have been shown to be N-linked glycosylated *in vivo* where N151 affects folding of the protein and N394 is associated with RADYS and HSCR1 diseases. There is no additional information about the other ten sites, and their significance in folding and in binding of extracellular substrates and later signaling remains vague.

The proposed combinatorial library consists of all possible combinations in which these sites of N-linked glycosylation would be mutated (N→Q) to prevent glycosylation. The resulting library will be screened *in vivo* for folding of the protein variants and for functional binding of substrates and further signaling utilizing reporter cell lines. The obtained results will allow to identify these sites in sequenced genomes of individuals and to predict the possible outcome, which may result in malfunction in the signaling pathway, and/or predisposition to certain diseases (e.g. RADYS, HSCR1), and/or even resistance to therapies (e.g. GDNF therapy for PD patients).

The size of the theoretical library, which consists of twelve positions to be mutated, is 4095 variants. This number is relatively high for a CADMAD test project. In order to reduce the size of the library a pre-screening will be performed. For this, single point mutations (only twelve variants) in hRET-ECD will be analyzes in CHO cells to exclude the sites, where Asn-to-Glu mutation causes the produced

protein to be degraded by the cellular quality control machinery. If a mutation prevents a protein from being expressed and delivered to the membrane, most probably any variant that contains this mutation will be misfolded and finally degraded, so there is no need to combine it with mutations in other sites. Herewith reducing the number of sites to be mutated to eight or nine, will allow generating a CADMAD test project with 255 or 511 variants, respectively.

Once the library is produced two types of screening will be performed: “folding” and “binding”. In folding assay CHO cells will be used in which the hRET-ECD library will be ligated to c-terminal part of hRet having GFP at the very end to express full-length hRET-GFP protein. Sub-cellular localization of the expressed fluorescent protein will report about the state of the protein. Localization in the plasma membrane shows normal folding, processing and translocation of the protein; ER or Golgi accumulation hints on problems associated with trafficking the protein to the cell surface; no fluorescence indicates non-stable protein, which is degraded by the cell machinery or not expressed on translational or transcriptional level. Binding assay will utilize a reporter cell line, which stably expresses GFRa1 co-receptor, which is needed for substrate binding and signaling. The library will be transiently transfected in frame of full-length hRet. Luciferase signal will report about the strength of the binding of soluble ligand and consequent signaling. In both assays the results will be semi-quantitative and one will be able to rank the mutations based on their contribution to the activity of the protein.

Despite a high biological outcome, the construction of the library cannot be implemented with the current version of DNALD, because the present release of the software does not allow the description of the library in an efficient way. However, the next release of the software will contain the required functions.

4. Combinatorial synthetic operon library (ETHZ)

A careful fine-tuning of enzyme levels is mandatory for efficient production of fine chemicals *in vivo* or *in vitro*. Enzyme concentrations can be tuned by varying expression levels for each gene of the pathway. To achieve above, each expression variant of each pathway gene has to be tested against all other possible expression variants of all other genes to identify the optimal situation. Therefore, combinatorial operon libraries are preferred because they could span the whole range of combinations for all regulatory elements of each gene.

The current objective in this regard is to optimize the carbon flow through glycolysis in order to increase dihydroxyacetone phosphate (DHAP) production in *E.coli* by fine-tuning of the relative and absolute amount of enzymes in the system.

The designed library is focused on testing the effect of promoter strength and ribosome binding site (RBS) strength in a synthetic polycystronic operon encoding the three main bottlenecks of DHAP biosynthesis (Bujara et al., 2011). Basically, the library consists of seven combinatorically recombined

modules: one for promoter (five strength options), three for RBS (five strength options) and three for enzyme-coding genes (one gene option per enzyme). Hence, a library of 625 variants will be generated.

In order to select the best variants from each series of variations a powerful analysis technique will be developed to rapidly perform detailed measurements for each pathway intermediate particularly focusing on how it is influenced by the variation of the expression of a specific gene. For this, enzyme activities analysis and DHAP quantification both based on cell-free extract assays on microplates coupled to automated kinetic measurements will be performed once the library is produced.

5. Investigating the post-transcriptional regulation of *Pseudomonas aeruginosa* azurin by RsmA (UNOTT)

Azurin production in *Pseudomonas aeruginosa* is positively controlled by the RNA-binding protein RsmA. This control is not exerted upstream of the ATG start codon (experimentally checked), and cannot be exerted downstream of the rho-independent transcriptional terminator. It could be possible that RsmA enhances the stability of the azurin mRNA by somehow binding to it, thus enhancing the quantity of protein that is translated from it. Rsm usually binds to mRNA at stem-loop structures having the following sequence: (U/A)CANGGANG(A/U). To bind AGGA or AGGGA have almost always been found on the single-stranded loops. The azurin ORF has three sites that could correspond to RsmA binding sites. Interestingly of the three AGGA sites found, two (the second and third) are located in the loops of potential short stem-loop structures, while in the remaining one the situation is slightly different. Thus seven parts can be defined, three of which could be altered to remove the AGGA sequences and the stem-loop structures, and then recombined to form new variants which may lose the positive regulation by RsmA. The parts to be conserved are 1, 3, 5 and 7. The parts to be altered are 2, 4 and 6. However, care must be taken to maintain the protein sequence, e.g. some alterations in part 2 may not be desirable because they would affect part 3.

The aim of the designed library is to confirm that RsmA binds to the azurin coding region and stabilizes the messenger RNA transcription as well as to investigate the contributions of each potential AGGA stem-loop to transcript stability.

The library variants are generated of parts 2, 4 and 6 with AGGA sequences and stem-loops removed while the translated sequence is preserved. Variants were produced using standard translation tools and custom algorithms to back-translate translations, filter for AGGA, detect and filter stem-loops. Outputs are 27 alternative azurin gene sequences with combinations of wildtype and two variants of each part.

The library output (azurin coding region with a C-terminal His-tag) will be cloned under control of the pBAD promoter making azurin synthesis arabinose inducible. Based on mutants and constructs already present in the lab a *P. aeruginosa* azurin mutant strain will be produced in which the level of

RsmA can be controlled by IPTG. Using the azurin expression construct and the *P. aeruginosa* ∆azu strain with controllable levels of RsmA the effect of different levels of RsmA on the level of azurin will be checked by western blot with antibodies against the azurin His-tag. If a positive correlation can be seen when azurin is expressed from the heterologous P_{BAD} promotor the wt. coding region will be replaced by an azurin variant where all three potential RsmA binding sites are changed. If this abolishes the positive correlation a variant will be tried where only one of the binding sites is changed. Depending on the outcome of these experiments more variants will eventually be tested.

6. Increase affinity of *Pseudomonas aeruginosa* PqsR quorum sensing signal receptor protein for N-oxide quinolones (UNOTT)

P. aeruginosa is a versatile bacterium that is able to grow in different environments such as soil, water, human and animal sewage, wounds and the lungs of cystic fibrosis (CF) patients. *P. aeruginosa* possesses three cell density dependent (quorum sensing) gene regulation systems: two *N*-acyl-homoserine lactone (AHL) based systems and a 2-alkyl-4-quinolone (AQ) based signalling system. CF sputum grown *Pseudomonas* show an enhanced production of the AQ quinolone signal. This may represent an important adaptive behaviour of *P. aeruginosa* in the lung (Hogardt and Heesemann, 2010). *P. aeruginosa* synthesize a number of quinolone compounds with different abilities to bind to and activate the quinolone quorum sensing transcriptional activator PqsR. Some, such as the N-oxide, bind poorly to PqsR (Fletcher et al., 2007). Data obtained by Palmer et al. indicate that growth of *P. aeruginosa* in CF sputum promotes increased synthesis of quinolones including high levels of 4-hydroxy-2-nonylquinoline *N*-oxide due to high levels of aromatic amino acids in the sputum (Palmer et al., 2007). PqsR has recently been crystallized (submitted). This led to the identification of key PqsR residues involved in binding respectively the alkyl chain and the bicyclic ring structure of the quinolone signal.

The aim of this library is the construction of a PqsR-based biosensor for detection of N-oxide quinolone compounds. Based on the recently submitted structure of PqsR key residues in the binding will be mutated to increase affinity for the *N*-oxide functional group of our target molecule. The hypothesis is that some combination of L207 (substituted for a polar, acidic or basic residue), its 1D neighbour L207 (substituted for an aliphatic, aromatic or polar residue) and the distant T265 substituted for an acidic amino acid residue) will yield additional information as to, and potentially increase, the binding of 4-hydroxy-2-nonylquinoline *N*-oxide (HHQNO) to PqsR.

The effects of changed key residues will be measured using a *P. aeruginosa pqsR* mutant with the *luxCDABE* reporter operon expressed from the PqsR dependent *pqsA* promotor is inserted into the chromosomes. Plasmids expressing mutated forms of PqsR will be introduced into the strain and the bioluminescence caused by expression of the lux reporter will be measured upon addition of N-oxide quinolone.

4. Conclusions

The libraries described above come from a variety of biological backgrounds and their products will be used for various scientific applications. Most of them could be successfully designed applying the recently developed DNALD software, which newest release includes many of the requirements and specifications that were defined by this work package in D5.3. These include the improvement of the graphical user interface (GUI) as well as a facilitation of sequence handling. The new version of DNALD now includes the possibility to zoom-in and zoom-out and change the length of fragment units in the GUI. It is now also possible to include comments in the DNALD file as well as to copy-and-paste sequences from other sources without the need to manually change their format. Additionally, one can now also define the reverse complements of sequences, and the possibilities to integrate mutations in sequences have been greatly enhanced. Despite this enhanced version of the software, there were still some challenges in designing the library from UH (Glycosylation screening of the hRET-ECD). However, these problems will be solved with the next version of the software from WP1, which will contain the required functions.

In the course of the design of the libraries presented here, the members of the work package have decided on parts of the libraries that will be shortly synthesized by conventional DNA synthesis. These fragments will later be used to compare the functionality of the DNA libraries generated by the CADMAD system to the ones generated by commercial conventional DNA synthesis. This will help to further validate the potency of the CADMAD system.

5. References

Bujara M, Schümmerli M, Pellaux R, Heinemann M, Panke S (2011) Optimization of a blueprint for in vitro glycolysis by metabolic real-time analysis. Nat Chem Biol 7(5):271-277

Caruthers MH, Barone AD, Beaucage SL, Dodds DR, Fisher EF, McBride LJ, Matteucci M, Stabinsky Z, Tang JY (1987) Chemical synthesis of deoxyoligonucleotides by the phosphoramidite method. Methods Enzymol 154: 287-313

Fletcher MP, Diggle SP, Crusz SA, Chhabra SR, Cámara M, Williams P (2007) A dual biosensor for 2-alkyl-4-quinolone quorum-sensing signal molecules. Environmental Microbiology *9*, 2683–2693.

Hogardt M Heesemann J (2010) Adaptation of Pseudomonas aeruginosa during persistence in the cystic fibrosis lung. International Journal of Medical Microbiology *300*, 557–562.

Letsinger RL, Mahadevan V (1965) Oligonucleotide synthesis on a polymer support. J Am Chem Soc 87: 3526-3527

Palmer KL, Aye LM, Whiteley M (2007) Nutritional cues control Pseudomonas aeruginosa multicellular behavior in Cystic Fibrosis sputum. Journal of Bacteriology *189*, 8079–8087

Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblasts by defined factors. Cell 126: 663-676

Thier M, Wörsdörfer P, Lakes YB, Gorris R, Herms S, Opitz T, Seiferling D, Quandel T, Hoffmann P, Nöthen MM, Brüstle O, Edenhofer F (2012) Direct conversion of fibroblasts into stably expandable neural stem cells. Cell Stem Cell 10(4): 473-479

6. Abbreviations

List all abbreviations used in the document arranged alphabetically.

AHL	N-acyl-homoserine lactone
AQ	2-alkyl-4-quinolone
CF	cystic fibrosis
DHAP	dihydroxyacetone phosphate
DNA	deoxyribonucleic acid
DNALD	DNA library designer
ECD	extracellular domain
E.coli	Escherichia coli
e.g.	for example (*exempli gratia*)
ER	endoplasmic reticulum
et al	and others (*et alii*)
ETHZ	Eidgenoessische Technische Hochschule Zuerich
FMI	Friedrich Miescher Institut
FMR	fully methylated region
GFP	Green fluorescent protein
GUI	Graphical user interface
LMR	low methylated regions
ORF	open reading frame
PD	Parkinson's disease
RBS	ribosome binding site

RNA	ribonucleic acid
UH	University of Helsinki
UKB	Universitaetsklinikum Bonn
UMR	Unmethylated region

Appendix A

Partner: UKB
Responsible person(s): Sandra Meyer, Frank Edenhofer
Email(s): smey@uni-bonn.de, f.edenhofer@uni-bonn.de

Library name (short): Cell Reprogramming
Library name (full title): Derivation of an Oct4 expression library to identify enhanced variants of Oct4 transcription factor.
Phase: Draft (February 2013)

Background: Oct4 transcription factor is involved in the self-renewal of undifferentiated pluripotent stem cells and plays a major role in changing the cellular fate of mammalian cells. For example, the reprogramming of somatic cell types into a multi- or pluripotent state involves ectopic activation of Oct4 (Takahashi and Yamanaka, 2006; Thier et al., 2012). Reprogramming can be achieved either by viral or protein transduction or mRNA transfection of Oct4. As a transcription factor the protein consists of a DNA-binding domain, which enables sequence specific DNA binding, and two transactivation domains, that regulate gene expression by interacting with the transcriptional machinery (Figure 1).

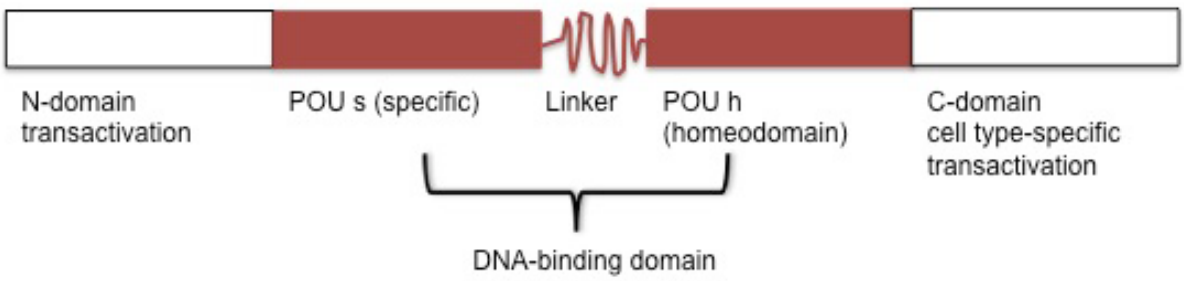

Figure 4: Schematic representation of domain structure of Oct4. Human Oct4-protein consists of 360 aa containing POU DNA-binding domain and two proline-rich regions. The POU domain has two independent subdomains: the N-terminal POU specific domain and a C-terminal hoomeodomain, which are connected by a linker amino acid stretch. Transactivation domains could be mapped to the N- and C-terminal proline-rich domains.

Aim: The designed library aims at identifying the domains crucial for reprogramming of cell fate of target cells and to generate enhanced versions of Oct4.

Library strategy: Modified variants of Oct4 should be generated that lack one or more of their domains. Additionally, Oct4 variants are desired that are fused with different stabilizing peptides (SP1, SP2, SP3) and the potent transactivator domain VP16.

Downstream processing planned: The generated library products will later be used for (1) the expression of the fusion proteins in *E.coli* for subsequent protein transduction of mammalian cells, and (2) for viral transduction of mammalian cells.

Keywords: Stem cell, reprogramming, protein transduction, heterologous expression

DNALD code

```
library Oct4 {

     inputs {
Oct4_N_terminal :=
"GCAGGTCATCTGGCAAGCGATTTTGCATTTAGTCCGCCTCCGGGTGGTGGCGGTGATGGTCCTGGTGGTCCGGA
ACCGGGTTGGGTTGATCCGCGTACCTGGCTGAGCTTTCAGGGTCCGCCAGGCGGTCCGGGTATTGGTCCAGGTGT
TGGTCCGGGTTCAGAAGTTTGGGGTATTCCGCCTTGTCCGCCACCGTATGAATTTTGTGGTGGTATGGCATATTG
TGGTCCGCAGGTTGGTGTGGGTCTGGTTCCGCAGGGTGGTCTGGAAACCAGCCAGCCGGAAGGTGAAGCCGGTGT
TGGTGTTGAAAGCAATAGTGATGGTGCATCACCGGAACCGTGTACCGTTACACCGGGTGCAGTTAAACTGGAAAA
AGAAAAACTGGAACAGAACCCGGAAGAAAGCCAG"
Oct4_POUspecific :=
"GATATTAAAGCACTGCAGAAAGAGCTGGAACAGTTTGCAAAACTGCTGAAACAGAAACGTATTACCCTGGGTTA
TACACAGGCAGATGTTGGTCTGACCCTGGGTGTTCTGTTTGGTAAAGTTTTTAGTCAGACCACGATTTGCCGTTT
TGAAGCCCTGCAGCTGAGCTTCAAAAATATGTGTAAACTGCGTCCGCTGCTGCAGAAATGGGTTGAAGAAGCAGA
T"
Oct4_linker := "AATAACGAAAACCTGCAAGAAATCTGTAAAGCCGAAACCCTGGTTCAGGCA"
Oct4_POUhomeodomain :=
"CGTAAACGTAAACGCACCAGCATTGAAAATCGTGTTCGTGGTAATCTGGAAAACCTGTTCCTGCAGTGTCCGAA
ACCGACCCTGCAGCAGATTAGCCATATTGCACAGCAGCTGGGTCTGGAAAAAGATGTTGTTCGTGTTTGGTTTTG
TAACCGTCGTCAGAAAGGTAAACGTAGC"
Oct4_C_terminal :=
"AGCAGCGATTATGCACAGCGTGAAGATTTTGAGGCAGCAGGTAGCCCGTTTAGCGGTGGTCCGGTTAGCTTTCC
GCTGGCACCGGGTCCGCATTTTGGTACACCGGGTTATGGTTCACCGCATTTTACAGCACTGTATAGCAGCGTTCC
GTTTCCGGAAGGTGAAGCATTTCCGCCTGTTAGCGTTACCACCCTGGGTAGCCCGATGCATAGCAAT"
SP1 :=
"ACCTACAAACTGATTCTGAATGGCAAAACCCTGAAAGGTGAAACCACCACCGAAGCAGTTGATGCAGCAACCGC
AGAAAAAGTCTTTAAACAGTATGCCAATGATAATGGCGTTGATGGTGAATGGACCTATGATGATGCAACCAAAAC
CTTTACCGTTACCGAA"
SP2 :=
"GAAGAAGCAAGCGTTACCAGCACCGAAGAAACCCTGACACCGGCACAAGAAGCAGCAGAAACCGAAGCAGCAAA
TAAAGCACGTAAAGAAGCAGAACTGGAAGCCGAAACCGCAGAACAA"
SP3 := "GAACGTAATAAAGAACGCAAAGAGGCCGAGCTGGAAGCAGAGACAGCAGAGCAG"
VP16 :=
"CTGGGTGATGGTGATAGTCCGGGTCCGGGTTTTACACCGCATGATAGCGCACCGTATGGTGCACTGGATATGGC
AGATTTTGAATTTGAACAGCATTTTACCGATGCCCTGGGCATTGATGAATATGGTGGC"

     }

NotI := "GCGGCCGC"
AvrII := "CCTAGG"
NheI := "GCTAGC"
EcoRI := "GAATTC"
BglII := "AGATCT"
BamHI := "GGATCC"
```

```
HindIII := "AAGCTT"
XhoI := "CTCGAG"
XmaI := "CCCGGG"
SalI := "GTCGAC"
Start_Codon := "ATG"
Linker := "ACCAGCGGTCTGGGTGGTGGTTCAGGTGGTGGTGGTAGCGGAGGTGGTGGCAGTGGT"


     outputs {

Oct4_WT_1 := NotI Start_Codon AvrII XhoI EcoRI Oct4_N_terminal BglII
Oct4_POUspecific BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII
Oct4_C_terminal SalI NheI NotI
Oct4_WT_2 := NotI Start_Codon AvrII XhoI XmaI EcoRI Oct4_N_terminal BglII
Oct4_POUspecific BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII
Oct4_C_terminal SalI NheI NotI

Oct4_SP1_1 := NotI Start_Codon AvrII XhoI SP1 EcoRI BglII Oct4_POUspecific
BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII Oct4_C_terminal SalI
NheI NotI
Oct4_SP1_2 := NotI Start_Codon AvrII XhoI SP1 EcoRI Oct4_N_terminal BglII
BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII Oct4_C-terminal SalI
NheI NotI
Oct4_SP1_3 := NotI Start_Codon AvrII XhoI SP1 EcoRI Oct4_N_terminal BglII
Oct4_POUspecific BamHI BamHI Oct4_POUhomeodomain HindIII Oct4_C_terminal
SalI NheI NotI
Oct4_SP1_4 := NotI Start_Codon AvrII XhoI SP1 EcoRI Oct4_N_terminal BglII
Oct4_POUspecific BamHI Oct4_linker BamHI HindIII Oct4_C_terminal SalI NheI
NotI
Oct4_SP1_5 := NotI Start_Codon AvrII.XhoI SP1 EcoRI Oct4_N_terminal BglII
Oct4_POUspecific BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII SalI
NheI NotI

Oct4_SP2_SP3_0 := NotI Start_Codon AvrII XhoI SP2 XhoI XmaI SP3 XmaI EcoRI
BglII Oct4_POUspecific BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII
Oct4_C_terminal SalI NheI NotI
Oct4_SP2_1 := NotI Start_Codon AvrII XhoI SP2 XhoI XmaI EcoRI
Oct4_N_terminal BglII Oct4_POUspecific BamHI Oct4_linker BamHI
Oct4_POUhomeodomain HindIII Oct4_C_terminal SalI NheI NotI
Oct4_SP2_2 := NotI Start_Codon AvrII XhoI SP2 XhoI XmaI EcoRI BglII
Oct4_POUspecific BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII
Oct4_C_terminal SalI NheI NotI
Oct4_SP2_3 := NotI Start_Codon AvrII XhoI SP2 XhoI XmaI EcoRI
Oct4_N_terminal BglII BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII
Oct4_C_terminal SalI NheI NotI
Oct4_SP2_4 := NotI Start_Codon AvrII XhoI SP2 XhoI XmaI EcoRI
```

Oct4_N_terminal BglII Oct4_POUspecific BamHI BamHI Oct4_POUhomeodomain HindIII Oct4_C_terminal SalI NheI NotI
Oct4_SP2_5 := NotI Start_Codon AvrII XhoI SP2 XhoI XmaI EcoRI Oct4_N_terminal BglII Oct4_POUspecific BamHI Oct4_linker BamHI HindIII Oct4_C_terminal SalI NheI NotI
Oct4_SP2_6 := NotI Start_Codon AvrII XhoI SP2 XhoI XmaI EcoRI Oct4_N_terminal BglII Oct4_POUspecific BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII SalI NheI NotI

Oct4_SP3_1 := NotI Start_Codon AvrII XhoI XmaI SP3 XmaI EcoRI Oct4_N_terminal BglII Oct4_POUspecific BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII Oct4_C_terminal SalI NheI NotI
Oct4_SP3_2 := NotI Start_Codon AvrII XhoI XmaI SP3 XmaI EcoRI BglII Oct4_POUspecific BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII Oct4_C_terminal SalI NheI NotI
Oct4_SP3_3 := NotI Start_Codon AvrII XhoI XmaI SP3 XmaI EcoRI Oct4_N_terminal BglII BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII Oct4_C_terminal SalI NheI NotI
Oct4_SP3_4 := NotI Start_Codon AvrII XhoI XmaI SP3 XmaI EcoRI Oct4_N_terminal BglII Oct4_POUspecific BamHI BamHI Oct4_POUhomeodomain HindIII Oct4_C_terminal SalI NheI NotI
Oct4_SP3_5 := NotI Start_Codon AvrII XhoI XmaI SP3 XmaI EcoRI Oct4_N_terminal BglII Oct4_POUspecific BamHI Oct4_linker BamHI HindIII Oct4_C_terminal SalI NheI NotI
Oct4_SP3_6 := NotI Start_Codon AvrII XhoI XmaI SP3 XmaI EcoRI Oct4_N_terminal BglII Oct4_POUspecific BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII SalI NheI NotI

Oct4_VP16_1_1 := NotI Start_Codon AvrII XhoI VP16 XhoI Linker BglII Oct4_POUspecific BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII Oct4_C_terminal SalI NheI NotI
Oct4_VP16_2_1 := NotI Start_Codon AvrII XhoI VP16 XhoI BglII Oct4_POUspecific BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII Oct4_C_terminal SalI NheI NotI
Oct4_VP16_3_1 := NotI Start_Codon AvrII XhoI VP16 XhoI Linker BglII BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII Oct4_C_terminal SalI NheI NotI
Oct4_VP16_4_1 := NotI Start_Codon AvrII XhoI VP16 XhoI Linker BglII Oct4_POUspecific BamHI BamHI Oct4_POUhomeodomain HindIII Oct4_C_terminal SalI NheI NotI
Oct4_VP16_5_1 := NotI Start_Codon AvrII XhoI VP16 XhoI Linker BglII Oct4_POUspecific BamHI Oct4_linker BamHI HindIII Oct4_C_terminal SalI NheI NotI
Oct4_VP16_6_1 := NotI Start_Codon AvrII XhoI VP16 XhoI Linker BglII Oct4_POUspecific BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII SalI NheI NotI

```
Oct4_VP16_7_1 := NotI Start_Codon AvrII XhoI XhoI Linker BglII
Oct4_POUspecific BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII
Oct4_C_terminal SalI NheI NotI
Oct4_VP16_8_1 := NotI Start_Codon AvrII XhoI XhoI BglII Oct4_POUspecific
BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII Oct4_C_terminal SalI
NheI NotI

Oct4_VP16_1_2 := NotI Start_Codon AvrII XhoI Oct4_N_terminal BglII
Oct4_POUspecific BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII Linker
XhoI VP16 SalI NheI NotI
Oct4_VP16_2_2 := NotI Start_Codon AvrII XhoI BglII Oct4_POUspecific BamHI
Oct4_linker BamHI Oct4_POUhomeodomain HindIII Linker XhoI VP16 SalI NheI
NotI
Oct4_VP16_3_2 := NotI Start_Codon AvrII XhoI Oct4_N_terminal BglII BamHI
Oct4_linker BamHI Oct4_POUhomeodomain HindIII Linker XhoI VP16 SalI NheI
NotI
Oct4_VP16_4_2 := NotI Start_Codon AvrII XhoI Oct4_N_terminal BglII
Oct4_POUspecific BamHI BamHI Oct4_POUhomeodomain HindIII Linker XhoI VP16
SalI NheI NotI
Oct4_VP16_5_2 := NotI Start_Codon AvrII XhoI Oct4_N_terminal BglII
Oct4_POUspecific BamHI Oct4_linker BamHI HindIII Linker XhoI VP16 SalI NheI
NotI
Oct4_VP16_6_2 := NotI Start_Codon AvrII XhoI Oct4_N_terminal BglII
Oct4_POUspecific BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII XhoI
VP16 SalI NheI NotI
Oct4_VP16_7_2 := NotI Start_Codon AvrII XhoI Oct4_N_terminal BglII
Oct4_POUspecific BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII Linker
XhoI SalI NheI NotI
Oct4_VP16_8_2 := NotI Start_Codon AvrII XhoI Oct4_N_terminal BglII
Oct4_POUspecific BamHI Oct4_linker BamHI Oct4_POUhomeodomain HindIII XhoI
SalI NheI NotI

     }
}
```

Partner: FMI
Responsible person(s): Juliane Schmidt, Arnaud Krebs, Dirk Schübeler
Email(s): juliane.schmidt@fmi.ch, arnaud.krebs@fmi.ch, dirk@fmi.ch

Library name (short): Methylation determinants
Library name (full title): Sequence replacement library to identify determinants of CpG islands methylation states.
Phase: Draft (February 2013)

Background: DNA methylation at CpG dinucleotides is a heritable epigenetic modification associated with transcriptional silencing and is essential for mammalian development. Methylation levels are mainly controlled at the level of DNA sequence with little or no influence from neither chromatin nor the transcriptional environment. However, little is known about the identity of the regulatory sequences which are able to tune DNA methylation states.

Status: The two first years were used to set up an experimental system allowing the profiling of methylation state acquisition from DNA libraries in a highly parallelized way (Figure 1). Using this system, thousands of fragments derived from the mouse genome were analyzed. This initial effort allowed (1) evaluation of the project feasibility and strategy limitations and (2) derivation of working hypotheses on main mechanisms driving the acquisition of methylation profiles. The proposed library synthesis projects are intended to allow the experimental testing of such hypotheses.

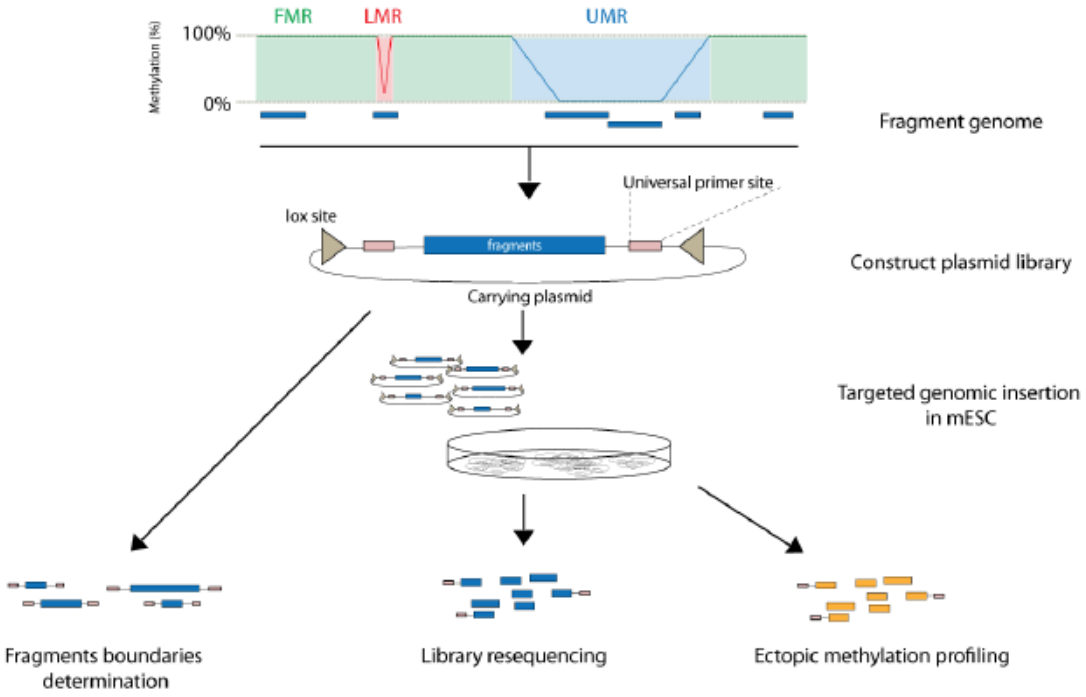

Figure 5: Summary of the experimental scheme used to test libraries. Synthetized libraries were cloned into a carrying vector for targeted, locus-specific insertion in mouse embryonic stem cells. Subsequently, the methylation status of inserted fragments was profiled by bisulphite sequencing. (FMR: fully methylated region, LMR: low-methylated region, UMR: unmethylated region)

Aim: Initial work on the project has established the importance of CpG content of analyzed sequences in establishing various methylation states. In order to test the contribution of other sequence components independently, one needs to generate sequence variations within DNA fragments, while being conservative regarding the CG positions (Figure 2). Thereby, the contribution of different DNA determinants aside from CGs in the establishment of DNA methylation patterns will be obtained

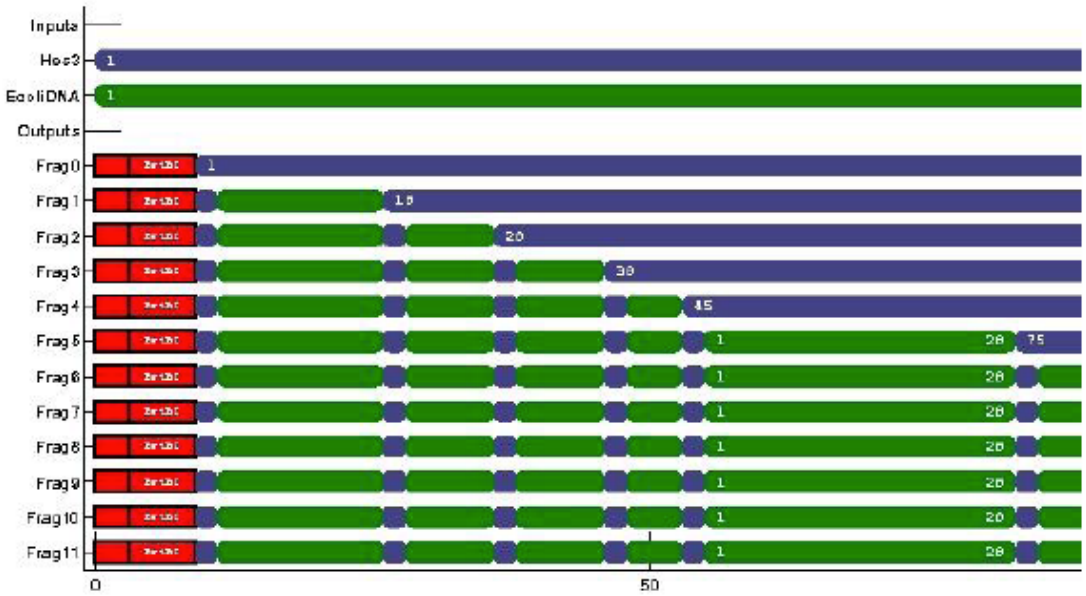

Figure 6: Sub-view of the library. Sequences flanked by CpGs are iteratively replaced from 5' to 3' by *E. coli* DNA.

Library strategy: Iterative replacement of the non-CpG DNA, composing the CpG island, by CpG-free prokaryotic DNA (*E. coli*).

Downstream processing planned: The library will be cloned into a recombination vector, electroporated in mammalian cells for genomic integration and tested for methylation.

Keywords: DNA methylation, epigenetics, CpG island

DNALD code

```
library Methylation_determinants {

     inputs {

Hes3 :=
"CGTCTATGGCTCTCAAACGCCTAGCAGCGTGGAGAATCGAGCCCCGCCCTCTCCCTTTCAGGCCCAGAGAAAGG
CGACCACAACTCTGGAACGTTCTCCTGCCGCTCCCCCACCCCCCGCCTTAGGGGGAGGGCCAGGTCTGGGTGGGT
TTAGGCCCCGGGCTGGTGATGTGACTGAAGCCCAGAGGCGAGCGCTGCCACGCGCCGGGTAGCTTTAAAGCGCTC
GGGCCAGTCGGCCCGGGAGGCCCCCTTGCATAGCTACGGCTCTGGGGCTGTGGCGGCGCGGTACAAAGGCGCCGC
GGGGCAGGCGGGCGGGAGGCACCGCACAAAAGGAGCAGCCGAGTGTTAAGGGCATTGTGCCTGGCGCACAAAGCG
GCGCCGCTCCGCCTTCTCCAGGCCAAACTCCCGGCGCCCCGCGGCGCACGGGCCTCCAGCTGCCGCACATCTGTA
GGAAGGCCGGGCGCGCGCCTCACCTGGGCG"
EcoliDNA :=
"AGAAAAAACAAATTTAATACAAAGGCTATTTGATAATGTTGAGTCTATATTTAATGAAGTACCTGTCAGCATTT
TAGTGAATGATATTTTTATGAATGATTTCTTTATGAAAAATCCTGAGATGATTTTGGTACTTCCCTCAGTTACTT
AAGAGTTATGAGGGTGAAAAGATTTATTTTGATAATTTAAAATATGATTTAAATGATAATGATAAGGAATCTAAT
AAAGAAATTTTGAAGAATCAACCAGATAAGTATCAAAGAAAAACTGAATAATGAATACAAACTTAGATTTAGAAT
GATGCAAACTATCTTGCAAT"

     }

linker := "CAT"
BstBI := "TTCGAA"
CpG := "CG"

     outputs {

Frag0 := linker BstBI Hes3 BstBI linker
Frag1 := Frag0[12:26=EcoliDNA[1:15]]
Frag2 := Frag1[29:36=EcoliDNA[1:8]]
Frag3 := Frag2[39:46=EcoliDNA[1:8]]
Frag4 := Frag3[49:53=EcoliDNA[1:5]]
Frag5 := Frag4[56:83=EcoliDNA[1:28]]
Frag6 := Frag5[86:100=EcoliDNA[1:15]]
Frag7 := Frag6[103:111=EcoliDNA[1:9]]
Frag8 := Frag7[114:126=EcoliDNA[1:13]]
Frag9 := Frag8[129:136=EcoliDNA[1:8]]
Frag10 := Frag9[139:168=EcoliDNA[1:30]]
Frag11 := Frag10[171:198=EcoliDNA[1:28]]
Frag12 := Frag11[201:202=EcoliDNA[1:2]]
Frag13 := Frag12[205:210=EcoliDNA[1:6]]
Frag14 := Frag13[215:215=EcoliDNA[1:1]]
Frag15 := Frag14[218:230=EcoliDNA[1:13]]
Frag16 := Frag15[233:234=EcoliDNA[1:2]]
Frag17 := Frag16[237:243=EcoliDNA[1:7]]
```

```
Frag18 := Frag17[246:248=EcoliDNA[1:3]]
Frag19 := Frag18[251:273=EcoliDNA[1:23]]
Frag20 := Frag19[276:290=EcoliDNA[1:15]]
Frag21 := Frag20[293:293=EcoliDNA[1:1]]
Frag22 := Frag21[298:306=EcoliDNA[1:9]]
Frag23 := Frag22[309:309=EcoliDNA[1:1]]
Frag24 := Frag23[314:320=EcoliDNA[1:7]]
Frag25 := Frag24[323:324=EcoliDNA[1:2]]
Frag26 := Frag25[327:334=EcoliDNA[1:8]]
Frag27 := Frag26[337:351=EcoliDNA[1:15]]
Frag28 := Frag27[354:376=EcoliDNA[1:23]]
Frag29 := Frag28[379:385=EcoliDNA[1:7]]
Frag30 := Frag29[388:388=EcoliDNA[1:1]]
Frag31 := Frag30[391:392=EcoliDNA[1:2]]
Frag32 := Frag31[397:398=EcoliDNA[1:2]]
Frag33 := Frag32[401:420=EcoliDNA[1:20]]
Frag34 := Frag33[423:423=EcoliDNA[1:1]]
Frag35 := Frag34[426:428=EcoliDNA[1:3]]
Frag36 := Frag35[433:433=EcoliDNA[1:1]]
Frag37 := Frag36[436:437=EcoliDNA[1:2]]
Frag38 := Frag37[440:452=EcoliDNA[1:13]]
Frag39 := Frag38[455:471=EcoliDNA[1:17]]
Frag40 := Frag39[474:475=EcoliDNA[1:2]]
Frag41 := Frag40[482:492=EcoliDNA[1:11]]

     }
}
```

Partner: UH
Responsible person(s): Konstantin Kogan, Adrian Goldman
Email(s): konstantin.kogan@helsinki.fi, adrian.goldman@helsinki.fi

Library name (short): RET
Library name (full title): Glycosylation screening of hRET-ECD
Phase: Draft (February 2013)

Background: Ret-oncogene receptor is a receptor tyrosine kinase which mutations are known to play a role in a variety of diseases. The extra-cellular domain (ECD) of human RET-oncogene receptor is highly post-transcriptionally modified. There is a belief that these modifications play a crucial role in protein folding, maturation and activity. hRET-ECD has twelve sites predicted to be N-linked glycosylated. However, only two sites (N151, and N394) have been shown to be N-linked glycosylated *in vivo* where N151 affects folding of the protein and N394 is associated with RADYS and HSCR1 diseases. There is no additional information about the other ten sites, and their significance in folding and in binding of extracellular substrates and later signaling remains vague.

Status: Progress has been made in the project and a RET-oncogene receptor (only extracellular part) from zebrafish source with minor alternations of the original DNA sequence has been successfully produced.

Aim: The aim of the library is to evaluate the role of glycosylations in hRET-ECD in protein folding, maturation, binding of extracellular substrates and subsequent signaling events.

Library strategy: We propose to build a combinatorial library of all possible combinations where these sites of N-linked glycosylation would be mutated (N->Q) to prevent glycosylation. The resulted library will be screened *in vivo* for folding of the protein variants and for functional binding of substrates and further signaling utilizing reporter cell lines. The obtained results will allow to identify these sites in sequenced genomes of individuals and to predict the possible outcome, which may result in malfunction in the signaling pathway, and/or predisposition to certain diseases (*e.g.* RADYS, HSCR1), and/or even resistance to therapies (*e.g.* GDNF protein therapy for PD patients).

Downstream processing planned: Once the library is produced we will perform two types of screenings: “folding” and “binding”. In folding assay we will use CHO cells where the library of hRET-ECD will be ligated to c-terminal part of hRET having GFP at the very end to express full length hRET-GFP protein. Sub-cellular localization of expressed fluorescent protein will report about the state of the protein. Plasma membranal localization shows normal folding, processing and translocation of the protein; ER or Golgi accumulation hints on problems associated with trafficking the protein to cell surface; and finally no fluorescence indicate non-stable protein, which is efficiently degraded by cell machinery or lack of expression on translational or transcriptional level. We will utilize the FACS to sort out the cells lacking fluorescence and confocal microscope to identify subcellular localization. Binding assay will utilized the reporter cell line which is stably express GFRa1 co-receptor needed for substrate binding and signaling, where we transiently transfect our library in a frame of full-length hRET. The Luciferase signal will report about the strength of the binding of soluble ligand and consequent signaling. In both cases the results will be semi-quantitative and we will be able to rank the mutations based on their contribution to the activity of the protein.

Keywords:
RET, oncogene, glycosylation

Partner: ETHZ
Responsible person(s): Gaspar Morgado, Sven Panke
Email(s): gaspar.morgado@bsse.ethz.ch; sven.panke@bsse.ethz.ch
Library name (short): Panke-CSOP
Library name (full title): Combinatorial synthetic operon library
Phase: Draft (February 2013)

Background: A careful fine tuning of enzyme levels is mandatory for efficient production of fine chemicals *in vivo* or *in vitro*. Enzyme concentrations can be tuned by varying expression levels for each gene of the pathway. To achieve above, each expression variant of each pathway gene has to be tested against all other possible expression variants of all other genes to identify the optimal situation. Therefore, combinatorial operon libraries are preferred because they could span the whole range of combinations for all regulatory elements of each gene.

Aim: Our current objective is to optimize the carbon flow through glycolysis in order to increase dihydroxyacetone phosphate production in *E. coli* by fine tuning of the relative and absolute amount of enzymes in the system.

Library strategy: Our approach will be focused on testing the effect of promoter strength and ribosome binding site (RBS) strength in a synthetic polycistronic operon encoding the 3 main bottlenecks of dihydroxyactone phosphate (DHAP) biosynthesis [1]. Basically, we want to combinatorically recombine seven modules: one for promoter (five strength options), three for RBS (five strength options) and three for enzyme-codifying genes (one gene-option per enzyme). Hence, we will generate a single library of 625 variants.

Downstream processing planned: In order to select the best variants from each series of variations we will develop a powerful analysis technique to rapidly perform detailed measurements for each pathway intermediate particularly focusing how it is influenced by the variation of the expression of a specific gene. We will perform enzyme activities analysis and DHAP quantification both based on cell-free extract assays on microplates coupled to automated kinetic measurements.

Keywords: Synthetic operon, gene fine tuning, cell-free synthesis, dihydroxyacetone phosphate.

[1] Bujara *et al. Nat. Chem. Biol.* **2011**. 7(5):271-7.

DNALD code

```
library Panke_combinatorial {

     inputs {
          fbaA :=
'ATGAGCAAAATCTTCGATTTCGTTAAACCGGGTGTTATTACCGGTGATGATGTTCAGAAAGTTTTTCAGGTTGC
CAAAGAAAACAATTTTGCACTGCCTGCCGTTAATTGTGTTGGCACCGATAGTATTAATGCAGTTCTGGAAACCGC
AGCCAAAGTTAAAGCACCGGTTATTGTTCAGTTTAGCAATGGTGGTGCAAGCTTTATTGCAGGTAAAGGTGTTAA
AAGTGATGTTCCGCAGGGTGCAGCAATTCTGGGTGCAATTAGCGGTGCACATCATGTTCATCAGATGGCAGAACA
TTATGGTGTTCCGGTTATTCTGCATACCGATCATTGTGCAAAAAAACTGCTGCCGTGGATTGATGGTCTGCTGGA
TGCCGGTGAAAAACATTTTGCCGCAACCGGTAAACCGCTGTTTAGCAGCCATATGATTGATCTGAGCGAAGAAAG
CCTGCAAGAAAACATTGAAATCTGCAGCAAATATCTGGAACGCATGAGTAAAATTGGTATGACCCTGGAAATTGA
ACTGGGTTGTACCGGTGGTGAAGAAGATGGCGTTGATAATAGCCATATGGATGCAAGCGCACTGTATACCCAGCC
GGAAGATGTTGATTATGCATATACCGAACTGTCCAAAATCAGTCCGCGTTTTACCATTGCAGCAAGCTTTGGTAA
TGTTCACGGTGTGTATAAACCTGGTAATGTGGTTCTGACCCCGACCATTCTGCGTGATAGCCAAGAATATGTTAG
CAAAAAACATAACCTGCCGCATAACAGCCTGAACTTTGTTTTTCATGGTGGTAGCGGTAGCACCGCACAAGAAAT
TAAAGATAGCGTTAGCTATGGCGTGGTGAAAATGAATATTGATACCGATACCCAGTGGGCAACCTGGGAAGGTGT
TCTGAATTATTACAAAGCCAATGAAGCATATCTGCAGGGTCAGCTGGGTAATCCGAAAGGTGAAGATCAGCCGAA
CAAAAAATACTATGATCCGCGTGTGTGGCTGCGTGCAGGTCAGACCAGCATGATTGCACGTCTGGAAAAAGCATT
TCAAGAACTGAATGCAATCGATGGCAGCTAA'
          ldh :=
'ATGGCAGATAAACAGCGCAAAAAAGTTATTCTGGTTGGTGATGGTGCAGTTGGTAGCAGCTATGCATTTGCACT
GGTTAATCAGGGTATTGCACAAGAACTGGGTATTGTGGACCTGTTCAAAGAAAAAACCCAGGGTGATGCCGAAGA
TCTGAGCCATGCACTGGCATTTACCAGCCCGAAAAAAATCTATAGCGCAGATTATAGTGATGCATCTGATGCAGA
TCTGGTTGTTCTGACCAGTGGTGCACCGCAGAAACCGGGTGAAACCCGTCTGGATCTGGTGGAAAAAAACCTGCG
TATTACCAAAGATGTGGTGACCAAAATTGTTGCCAGCGGTTTTAAAGGTATTTTTCTGGTTGCAGCAAACCCGGT
TGATATTCTGACCTATGCAACCTGGAAATTTAGCGGCTTTCCGAAAAATCGTGTTGTTGGTAGCGGCACCAGCCT
GGATACCGCACGTTTTCGTCAGGCACTGGCAGAAAAAGTTGATGTTGATGCACGTAGCATTCACGCCTATATTAT
GGGTGAACATGGTGATAGCGAATTTGCAGTTTGGAGCCATGCCAATGTTGCCGGTGTTAAACTGGAACAGTGGTT
TCAAGAAAACGATTATCTGAACGAAGCCGAAATCGTGGAACTGTTTGAAAGCGTTCGTGATGCAGCATATAGCAT
TATTGCAAAAAAAGGTGCCACGTTTTATGGTGTTGCAGTTGCCCTGGCACGTATTACAAAAGCAATTCTGGATGA
TGAACATGCAGTTCTGCCGGTTAGCGTTTTTCAGGATGGTCAGTATGGTGTTAGCGATTGTTATCTGGGTCAGCC
TGCAGTTGTTGGTGCAGAAGGTGTTGTTAATCCGATTCATATTCCGCTGAATGATGCGGAAATGCAGAAAATGGA
AGCAAGCGGTGCACAGCTGAAAGCAATTATTGATGAAGCATTTGCCAAAGAAGAATTTGCCAGCGCAGTGAAAAA
TTAA'
          glk :=
'ATGACCAAATATGCACTGGTTGGTGATGTTGGTGGCACCAATGCACGTCTGGCACTGTGTGATATTGCAAGCGG
TGAAATTAGCCAGGCAAAAACCTATAGCGGTCTGGATTATCCGAGCCTGGAAGCAGTTATTCGTGTTTATCTGGA
AGAACACAAAGTCGAAGTTAAAGATGGTTGTATTGCAATTGCCTGTCCGATTACCGGTGATTGGGTTGCAATGAC
CAATCATACCTGGGCATTTAGCATTGCCGAGATGAAAAAAAACCTGGGTTTTAGCCATCTGGAAATCATCAATGA
TTTTACCGCAGTTAGCATGGCAATTCCGATGCTGAAAAAAGAACATCTGATTCAGTTTGGTGGTGCAGAACCGGT
TGAAGGTAAACCGATTGCAGTTTATGGTGCAGGCACCGGTCTGGGTGTTGCACATCTGGTTCATGTTGATAAACG
TTGGGTTAGCCTGCCTGGTGAAGGTGGTCATGTGGATTTTGCACCGAATAGCGAAGAAGAAGCAATTATCCTGGA
AATTCTGCGTGCAGAAATTGGTCATGTTAGCGCAGAACGTGTTCTGAGCGGTCCGGGTCTGGTTAATCTGTATCG
TGCAATTGTTAAAGCCGATAATCGTCTGCCGGAAAATCTGAAACCGAAAGATATTACCGAACGTGCACTGGCAGA
TAGCTGTACCGATTGTCGTCGTGCACTGAGCCTGTTTTGTGTTATTATGGGTCGTTTTGGTGGTAATCTGGCCCT
```

```
GAATCTGGGCACCTTTGGTGGCGTTTTTATTGCCGGTGGTATTGTTCCGCGTTTTCTGGAATTTTTCAAAGCAAG
CGGTTTTCGTGCAGCCTTTGAAGATAAAGGTCGCTTCAAAGAATATGTGCATGATATTCCGGTGTACCTGATTGT
TCATGATAATCCTGGTCTGCTGGGTAGCGGTGCCCATCTGCGTCAGACCCTGGGTCATATTCTGTAA'
      }

      // Promoters are short DNA fragments and should be generated during Y
operations
      P1    :=
'ctggtttttccagcagacgacggagcaaaaactacccgtaggtgtagttggcgcaagcgtccgattagctcagg
ttttaagatg'
      P2    :=
'tttccagcagacgacggagcaaaaactacccgtaggtgtagttggcgcaagcgtccgattagctcaggttttaa
gatg'
      P3    :=
'agcagacgacggagcaaaaactacccgtaggtgtagttggcgcaagcgtccgattagctcaggttttaagatg'
      P4    :=
'cgacggagcaaaaactacccgtaggtgtagttggcgcaagcgtccgattagctcaggttttaagatg'
      P5    :=
'agcaaaaactacccgtaggtgtagttggcgcaagcgtccgattagctcaggttttaagatg'

      // RBSs are short DNA fragments and should be generated during Y
operations
      R1    := 'gacaaaaatctagaaataattttgtttaactttaagaaggagatatacaa'
      R2    := 'gggagctaacgagggcaaaaa'
      R3    := 'aataattttgtttaactttaagaaggagatatacat'
      R4    := 'atggtgttctccaatttttattaaattagtcgctacgagatttaagacgt'
      R5    := 'ctctaaaagcgcgctgaaacaagggcaggtttccctgccctgtgattttt'

      outputs {
            ^Library := (P1 + P2 + P3 + P4 + P5) (R1 + R2 + R3 + R4 + R5)
ldh (R1 + R2 + R3 + R4 + R5) fbaA (R1 + R2 + R3 + R4 + R5) glk
            // Combinatorial operon libray of 625 variants
      }

}
```

Partner: UNOTT
Responsible persons(s): Natalio Krasnogor, Jonathan Blakes, Birgit Koch
Email(s): {natalio.krasnogor, jonathan.blakes, birgit.koch}@nottingham.ac.uk

Library name (short): *Pseudomonas aeruginosa* azurin
Library name (full title): Investigating post-transcriptional regulation of *Pseudomonas aeruginosa* azurin by RsmA
Phase: draft (March 2013)

Background:
Azurin production in *Pseudomonas aeruginosa* is positively controlled by the RNA-binding protein RsmA. This control is not exerted upstream of the ATG start codon (experimentally checked), and cannot be exerted downstream of the rho-independent transcriptional terminator. It could be possible that RsmA enhances the stability of the azurin mRNA by somehow binding to it, thus enhancing the quantity of protein that is translated from it. RsmA usually binds to mRNA at stem-loop structures having the following sequence: (U/A)CANGGANG(A/U). To bind, AGGA or AGGGA have almost always been found on the single-stranded loops. The azurin ORF has three sites that could correspond to RsmA binding sites. Interestingly of the three AGGA sites found, two (the second and third) are located in the loops of potential short stem-loop structures, while in the remaining one the situation is slightly different. We can thus define 7 parts, 3 of which could be altered to remove the AGGA sequences and the stem-loop structures, and then recombined to form new variants which may lose the positive regulation by RsmA. The parts to be conserved are 1, 3, 5 and 7. The parts to be altered are 2, 4 and 6. However, care must be taken to maintain the protein sequence, e.g. some alterations in part 2 may not be desirable because they would affect part 3.

Aim:
Confirm that RsmA binds to the azurin coding region and stabilizes the messenger mRNA transcription. Investigate contributions of each potential AGGA stem-loop to transcript stability.

Library strategy:
Generated variants of parts 2, 4 and 6 with AGGA sequences and stem-loops removed while preserving translated sequence. Variants were produced using standard translation tools and custom algorithms to back-translate translations, filter for AGGA, detect and filter stem-loops. Outputs are 27 alternative azurin gene sequences with combinations of wildtype and 2 variants of each part.

Downstream processing planned:
The azurin coding region with a C-terminal His-tag will be cloned under control of the pBAD promoter making azurin synthesis arabinose inducible. Based on mutants and constructs already present in the lab. a *P. aeruginosa* azurin mutant strain will be produced in which the level of RsmA can be controlled by IPTG. Using the azurin expression construct and the *P. aeruginosa* Δ*azu* strain with controllable levels of RsmA we will by western blotting with antibodies against the azurin his-tag measure how different levels of RsmA affects the level of azurin. If we see a positive correlation also when azurin is expressed from the heterologous P_{BAD} promotor the wt. coding region will be replaced by an azurin variant where all three potential RsmA binding sites are chanced. If this abolishes the positive correlation we will try variant where the only one of the binding sites are changed. Depending on the outcome of these experiments more variants will eventually be tested.

Keywords (5):
Pseudomonas aeruginosa, azurin, post-transciptional regulation, RsmA, stem-loop

DNALD file:

```
library azurinLibrary {
inputs {
      azurin_PCR_product := '
AAGGT CCATGG
ATGCTACGTAAACTCGCTGCGGTATCCCTGCTGTCCCTGCTCAGTGCGCCACTGCTGGCTGCCGAGTGCTCGGTGGACAT
CCAGGGTAACGACCAGATGCAGTTCAACACCAATGCCATCACCGTCGACAAGAGCTGCAAGCAGTTCACCGTCAACCTGT
CCCACCCCGGCAACCTGCCGAAGAACGTCATGGGCCACAACTGGGTACTGAGCACCGCCGCCGACATGCAGGGCGTGGTC
ACCGACGGCATGGCTTCCGGCCTGGACAAGGATTACCTGAAGCCCGACGACAGCCGTGTCATCGCCCACACCAAGCTGAT
CGGCTCGGGCGAGAAGGACTCGGTGACCTTCGACGTCTCCAAGCTGAAGGAAGGCGAGCAGTACATGTTCTTCTGCACCT
TCCCGGGCCACTCCGCGCTGATGAAGGGCACCCTGACCCTGAAGTGA
GAATTC ACCTT
'}

      azurin      := azurin_PCR_product[12:458]

      // Parts such that no codons in 5'->3' Frame 1 are truncated
      Part_1 := azurin[1:258]
      Part_2 := azurin[259:273] is 'GGCCTGGACAAGGAT'
      Part_3 := azurin[274:327]
      Part_4 := azurin[328:342] is 'GGCGAGAAGGACTCG'
      Part_5 := azurin[343:360]
      Part_6 := azurin[361:375] is 'AAGCTGAAGGAAGGC'
      Part_7 := azurin[376:end] is azurin[376:447]

      // Parts 1, 3, 5, 7 are to be conserved.
      // Parts 2, 4, 6 are to be altered by removing RsmA binding sites with AGGA
and stem-loop structures.

      // 2/52 variants
      Part_2_altered := 'GGCCTGGACAAAGAC' + 'GGATTAGACAAAGAC'

      // 2/41 variants
      Part_4_altered := 'GGCGAGAAAGACAGC' + 'GGAGAGAAAGATAGC'

      // 2/70 variants
      Part_6_altered := 'AAGCTGAAAGAGGGC' + 'AAGTTAAAAGAGGGG'

      // control, reconstituted from new Part indices
      azurin_unaltered   := Part_1 Part_2 Part_3 Part_4 Part_5 Part_6 Part_7

      // targets with 'AGGA' and stem-loops in non-wt Parts 2, 4 and 6 removed
      azurin_altered     := Part_1 (Part_2 + Part_2_altered)
                                   Part_3 (Part_4 + Part_4_altered)
                                   Part_5 (Part_6 + Part_6_altered)
                                   Part_7

      NcoI  := 'CCATGG'
      _5_prime_end      := 'AAGGT' NcoI is azurin_PCR_product[1:11]

      His_tag     := 'CAT CAC CAT CAC CAT CAC'
      stop_codon  := 'TGA' is azurin[445:447]
      EcoRI := 'GAATTC'
      _3_prime_end      := His_tag stop_codon EcoRI 'ACCTT'

outputs {
      // 27 His-tagged variants with cloning sites
      azurin_library := _5_prime_end
                        (azurin_unaltered + azurin_altered) _3_prime_end } }
```

Partner: UNOTT
Responsible persons(s): Natalio Krasnogor, Jonathan Blakes, Birgit Koch
Email(s): {natalio.krasnogor, jonathan.blakes, birgit.koch}@nottingham.ac.uk

Library name (short): PqsR-based N-oxide quinolone biosensor
Library name (full title): Increase affinity of *Pseudomonas aeruginosa* PqsR quorum sensing signal receptor protein for N-oxide quinolones
Phase: March 2013

Background:
P. aeruginosa is a versatile bacterium that is able to grow in different environments such as soil, water, human and animal sewage, wounds and the lungs of cystic fibrosis (CF) patients. *P. aeruginosa* possesses three cell density dependent (quorum sensing) gene regulation systems: two *N*-acyl-homoserine lactone (AHL) based systems and a 2-alkyl-4-quinolone (AQ) based signalling system. CF sputum grown *Pseudomonas* show an enhance production of the AQ quinolone signal. This may represent an important adaptive behaviour of *P. aeruginosa* in the lung (Hogardt and Heesemann, 2010).*P. aeruginosa* synthesize a number of quinolone compounds with different abilities to bind to and activate the quinolone quorum sensing transcriptional activator PqsR. Some, such as the N-oxide, bind poorly to PqsR (Fletcher et al., 2007). Data obtained by Palmer et al. indicate that growth of *P. aeruginosa* in CF sputum promotes increased synthesis of quinolones including high levels of 4-hydroxy-2-nonylquinoline *N*-oxide due to high levels of aromatic amino acids in the sputum (Palmer et al., 2007). PqsR has recently been crystallized (submitted). This led to the identification of key PqsR residues involved in binding respectively the alkyl chain and the bicyclic ring structure of the quinolone signal.

Aim:
To construct a PqsR-based biosensor for detection of N-oxide quinolone compounds.

Library strategy:
Based on the recently submitted structure of PqsR we will mutate key residues in the binding to increase affinity for the *N*-oxide functional group of our target molecule. Our hypothesis is that some combination of L207 (substituted for a polar, acidic or basic residue), its 1D neighbour L207 (substituted for an aliphatic, aromatic or polar residue) and the distant T265 substituted for an acidic amino acid residue) will yield additional information as to, and potentially increase, the binding of 4-hydroxy-2-nonylquinoline *N*-oxide (HHQNO) to PqsR.

Downstream processing planned:
The effects of changes key residues will be measured using a *P. aeruginosa pqsR* mutant with the *luxCDABE* reporter operon expressed from the PqsR dependent *pqsA* promotor is inserted into the chromosomes. Plasmids expressing mutated forms of PqsR will be introduced into the strain and the bioluminescence caused by expression of the lux reporter will be measured upon addition of N-oxide quinolone.

Keywords (5):
Pseudomonas aeruginosa, cystic fibrosis, PqsR, quorum sensing, 4-hydroxy-2-nonylquinoline *N*-oxide

References:
Fletcher, M.P., Diggle, S.P., Crusz, S.A., Chhabra, S.R., Cámara, M., and Williams, P. (2007). A dual biosensor for 2-alkyl-4-quinolone quorum-sensing signal molecules. Environmental Microbiology *9*, 2683–2693.

Hogardt, M., and Heesemann, J. (2010). Adaptation of Pseudomonas aeruginosa during persistence in the cystic fibrosis lung. International Journal of Medical Microbiology *300*, 557–562.

Palmer, K.L., Aye, L.M., and Whiteley, M. (2007). Nutritional Cues Control Pseudomonas aeruginosa Multicellular Behavior in Cystic Fibrosis Sputum. Journal of Bacteriology *189*, 8079–8087

DNALD file:

```
library pqsr_mutagenesis {
inputs {

                  @db('http://www.pseudomonas.com/getAnnotation.do?locusID=PA1003')
      @genbank('http://www.ncbi.nlm.nih.gov/protein/15596200')
      @uniprot('http://www.uniprot.org/uniprot/Q9I4X0')
pqsr  :=
      aa'MPIHNLNHVNMFLQVIASGSISSAARILRKSHTAVSSAVSNLEIDLCVELVRRDGYKVEPTEQALRLIPYMRSL
LNYQQLIGDIAFNLNKGPRNLRVLLDTAIPPSFCDTVSSVLLDDFNMVSLIRTSPADSLATIKQDNAEIDIAITIDEELKISR
FNQCVLGYTKAFVVAHPQHPLCNASLHSIASLANYRQISLGSRSGQHSNLLRPVSDKVLFVENFDDMLRLVEAGVGWGIAPHY
FVEERLRNGTLAVLSELYEPGGIDTKVYCYYNTALESERSFLRFLESARQRLRELGRQRFDDAPAWQPSIVETAQRRSGPKAL
AYRQRAAPE*'
as '
ATGCCTATTCATAACCTGAATCACGTGAACATGTTCCTCCAGGTCATCGCCTCCGGTTCGATTTCCTCCGCTGCGCGGATCCT
GCGCAAGTCGCACACCGCGGTCAGCTCGGCGGTCAGCAACCTGGAAATCGACCTGTGCGTGGAGCTGGTCCGTCGGGACGGCT
ACAAGGTCGAACCCACCGAGCAGGCGCTTCGCCTGATCCCTTACATGCGCAGCCTGCTGAACTACCAGCAGCTGATCGGCGAC
ATCGCCTTCAATCTCAACAAGGGTCCGCGCAATCTCCGGGTGCTGCTGGACACCGCCATCCCGCCGTCGTTCTGCGATACGGT
GAGCAGCGTACTGCTCGACGATTTCAACATGGTCAGCCTGATACGCACCTCGCCCGCCGATAGCCTGGCGACGATCAAGCAGG
ACAACGCGGAAATCGATATCGCCATCACCATCGACGAGGAACTGAAGATCTCCCGCTTCAACCAGTGCGTGCTCGGCTACACC
AAGGCGTTCGTCGTCGCCCATCCGCAGCACCCGTTGTGCAATGCCTCCCTGCACAGCATCGCGAGCCTGGCCAATTACCGGCA
GATCAGCCTCGGCAGCCGCTCCGGGCAGCATTCGAACCTGCTGCGGCCGGTCAGCGACAAGGTGCTCTTCGTGGAAAACTTCG
ACGACATGCTGCGTCTGGTGGAAGCCGGCGTCGGATGGGGCATCGCGCCGCATTATTTCGTCGAGGAACGCCTGCGCAACGGT
ACCCTGGCAGTCCTCAGCGAACTCTACGAACCGGGCGGCATCGACACCAAGGTGTATTGCTACTACAACACCGCGCTGGAATC
CGAGCGCAGCTTCCTGCGCTTTCTCGAAAGCGCCCGCCAGCGCCTGCGCGAACTCGGCCGCCAGCGTTTCGACGATGCGCCGG
CCTGGCAACCGAGCATCGTCGAAACGGCGCAGCGCCGCTCAGGCCCGAAGGCGCTCGCGTACCGCCAGCGCGCCGCACCAGAG
TAG'
}

// residues of the active site from Figure 3. B
I149  :=    pqsr[149aa] is aa'I' as 'ATC'
A168  :=    pqsr[168aa] is aa'A'
I186  :=    pqsr[186aa] is aa'I'
Q194  :=    pqsr[194aa] is aa'Q'
L207  :=    pqsr[207aa] is aa'L'
L208  :=    pqsr[208aa] is aa'L'
V211  :=    pqsr[211aa] is aa'V'
F221  :=    pqsr[221aa] is aa'F'
I236  :=    pqsr[236aa] is aa'I'
Y258  :=    pqsr[258aa] is aa'Y'
I263  :=    pqsr[263aa] is aa'I'
T265  :=    pqsr[265aa] is aa'T'


      // asserting wildtype can be reconstituted from extracted intermediates
      pqsr_wt
      :=
      pqsr[1:148aa]
      I149 pqsr[150aa:167aa]
      A168 pqsr[169aa:185aa]
      I186 pqsr[187aa:193aa]
      Q194 pqsr[195aa:206aa]
      L207 pqsr[208aa:210aa]
      V211 pqsr[212aa:220aa]
      F221 pqsr[222aa:235aa]
      I236 pqsr[237aa:257aa]
      Y258 pqsr[259aa:262aa]
      I263 pqsr[264aa]
      T265 pqsr[266aa:end]
```

```
      is pqsr

      EcoRI             := 'GAATTC'
      _5_prime_end      := 'AAGGT' EcoRI

      His_tag           := 'CAT CAC CAT CAC CAT CAC'
      stop_codon        := 'TAG' is pqsr[333aa]
      SacI              := 'GAGCTC' is reverse(complement('GAGCTC'))
      _3_prime_end      := His_tag stop_codon SacI 'ACCTT'

      // aliphatic AAs
      A     := aa'A' as 'GCA'
      G     := aa'G' as 'GGT'
      I     := aa'I' as 'ATT'
      L     := aa'L' as 'CTG'
      V     := aa'V' as 'GTT'

      // polar AAs
      C     := aa'C' as 'TGT'
      M     := aa'M' as 'ATG'
      S     := aa'S' as 'AGC'
      T     := aa'T' as 'ACC'

      // cyclic AAs
      P     := aa'P' as 'CCG'

      // aromatic AAs
      F := aa'F' as 'TTT'
      W := aa'W' as 'TGG'
      Y := aa'Y' as 'TAT'

      // basic AAs
      H := aa'H' as 'CAT'
      K := aa'K' as 'AAA'
      R := aa'R' as 'CGT'

      // acidic AAs
      E := aa'E' as 'GAA'
      Q := aa'Q' as 'CAG'
      D := aa'D' as 'GAT'
      N := aa'N' as 'AAT'

      aliphatic_AAs     := A + G + I + L + V
      polar_AAs         := C + M + S + T
      cyclic_AAs        := P
      aromatic_AAs      := F + W + Y
      basic_AAs         := H + K + R
      acidic_AAs        := E + Q + D + N

      X :=  aliphatic_AAs + polar_AAs + cyclic_AAs +
            aromatic_AAs + basic_AAs + acidic_AAs

outputs {
      pqsr_lib    :=    // 115 variants
                        _5_prime_end
                        pqsr[1:148aa]
                        I149 pqsr[150aa:167aa]
                        A168 pqsr[169aa:185aa]
                        I186 pqsr[187aa:193aa]
                        Q194 pqsr[195aa:206aa]
```

```
                    (
                          (L207 (aliphatic_AAs + aromatic_AAs + polar_AAs))
                    +     ((acidic_AAs + polar_AAs + basic_AAs) L208)
                    )
                          pqsr[209aa:210aa]
                          V211 pqsr[212aa:220aa]
                          F221 pqsr[222aa:235aa]
                          I236 pqsr[237aa:257aa]
                          Y258 pqsr[259aa:262aa] I263 pqsr[264aa]
                    (T265 + acidic_AAs)
                          pqsr[266aa:end]
                          _3_prime_end
      }
}
```