(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 October 2017 (19.10.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/178193 A1

(51) International Patent Classification:
*C12N 15/10* (2006.01) *C12Q 1/18* (2006.01)
*C07K 14/00* (2006.01) *C12N 15/63* (2006.01)
*C12Q 1/68* (2006.01)

(21) International Application Number:
PCT/EP2017/056454

(22) International Filing Date:
17 March 2017 (17.03.2017)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
16165307.6 14 April 2016 (14.04.2016) EP

(71) Applicant: **MAX-PLANCK-GESELLSCHAFT ZUR FÖRDERUNG DER WISSENSCHAFTEN E.V.** [DE/DE]; Hofgartenstraße 8, 80539 Munich (DE).

(72) Inventors: **TAUTZ, Diethard**; Seestraße 28, 24306 Plön (DE). **NEME, Rafik**; Rautenbergstrasse 55, 24306 Plön (DE). **AMADOR HIERRON, Cristina Isabel**; Lindevangs Alle 1, st. th., 2000 Frederiskberg (DK). **MC CONNELL, Ellen**; Rautenberg Straße 53, 24306 Plön (DE). **BHAVE, Devika**; Jungfernstieg 25, 24103 Kiel (DE).

(74) Agent: **VOSSIUS & PARTNER (NO 31)**; JÜRGEN MEIER, Siebertstraße 3, 81675 München (DE).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

**Published**:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

— *with sequence listing part of description (Rule 5.2(a))*

WO 2017/178193 A1

(54) Title: METHOD FOR THE IDENTIFICATION OF RANDOM POLYNUCLEOTIDE OR POLYPEPTIDE SEQUENCES WITH BIOLOGICAL ACTIVITY

(57) Abstract: The present invention relates to a method for identifying novel bioactive polynucleotides or polypeptides composed of random or almost random combinations of nucleotides or aminoacids, respectively. It may encompass the insertion of a library of polynucleotides with random or almost random nucleic acid sequences encoding for biomolecules such as RNAs and/or polypeptide chains in an expression vector, transformation of this expression vector library into suitable host cells, such that each cell carries one polynucleotide variant, and expression of the inserts of the respective expression vectors during the cultivation of the cells. Polynucleotides encoding for RNAs and polypeptides having biological activity are then identified through determining the changes in frequencies of individual polynucleotides in the pool of polynucleotide variants of the library of polynucleotides comprised by the host cells at two or more time points of cultivation. Polynucleotides that have changed in frequency during this phase are identified as polynucleotides encoding for biomolecules with biological activity. In particular, polynucleotides are considered as positively active when their frequency has increased and negatively active when their frequency has decreased between the time points of cultivation.

# Method for the identification of random polynucleotide or polypeptide sequences with biological activity

The present application claims the priority of the European patent application EP 16165307.6 filed on April 14, 2016 with the European Patent Office, the entire disclosure of which is hereby incorporated by reference.

The present invention relates to a method for identifying novel bioactive polynucleotides or polypeptides composed of random or almost random combinations of nucleotides or amino acids, respectively. It may encompass the insertion of a library of polynucleotides with random or almost random nucleic acid sequences encoding for biomolecules such as RNAs and/or polypeptide chains in an expression vector, transformation of this expression vector library into suitable host cells, such that each cell carries one polynucleotide variant, and expression of the inserts of the respective expression vectors during the cultivation of the cells. Polynucleotides encoding for RNAs and polypeptides having biological activity are then identified through determining the changes in frequencies of individual polynucleotides in the pool of polynucleotide variants of the library of polynucleotides comprised by the host cells at two or more time points of cultivation. Polynucleotides that have changed in frequency during this phase are identified as polynucleotides encoding for biomolecules with biological activity. In particular, polynucleotides are considered as positively active when their frequency has increased and negatively active when their frequency has decreased between the time points of cultivation at which the relative frequencies are determined.

Organismic life processes are based on the function and interaction of biomolecules like polypeptides (alternative terms used herein: proteins or peptides). Polypeptides consist of chains of amino acids and are encoded in the DNA and RNA of organisms and viruses. DNA and RNA consist of chains of nucleotides. There are 4 canonical nucleotides for RNA and DNA each, as well as 20 canonical amino acids that are used by organisms. DNA or RNA chains can consist of hundreds to millions of nucleotides,

polypeptide chains can consist of hundreds to thousands of amino acids. This implies that there are almost infinite combinatorial combinations for DNA, RNA and polypeptide sequences. However, while the sequence diversity of DNA, RNA and polypeptides is indeed very high, the structural diversity of folded polypeptides seems to be more limited. Comparative analysis of solved protein structures has revealed that only a few thousand stable folds may exist (Woolfson et al. 2015). On the other hand, it is known that many polypeptide chains are intrinsically disordered, i.e. do not form a stable structure or fold, but can still exert specific functions (Tompa et al. 2015).

Comparative genome analyses have shown that new polypeptide chains can arise *de novo* out of originally non-coding DNA during evolution (Schlötterer 2015). Comparative genome and transcriptome analysis have shown that there is a high evolutionary turnover of transcripts from non-coding DNA resulting in an expression and evolutionary testing of most of the genome (Neme and Tautz 2016). This underlines that many DNA sequences that are currently not yet known to be employed or are not employed for protein synthesis in nature may encode for functional peptides and/or RNAs. The identification of novel DNA sequences and/or corresponding encoded RNAs and/or encoded peptides is/are of high interest for pharmaceutical or diagnostic or other technical applications.

One way to identify novel DNA sequences and respective RNAs and/or respective peptides encoded thereof is to employ libraries of nucleic acids or peptides with random sequences, respectively. To identify those nucleic acids and/or peptides having a biological activity in these libraries, different screening methods have been applied.

For example, libraries of random peptides have been screened using the phage display technology (Omidfar & Daneshpour, 2015). The phage display technology allows identifying peptides which interact with a defined molecular structure (e.g. a protein or a chemical substance). Specifically, for phage display each individual random peptide variant is synthesized as part of the capsid proteins of the phage. By creating a number of phages, wherein each phage expresses a different peptide variant, a library of phage clone variants is created. These phage clone variants are then bound to a desired target (usually a protein, but any molecular structure is

suitable) and all non-bound phages are removed. Bound phages are eluted and replicated in new host cells. By these means random peptides which interact with the target can be identified. Optimized binding variants can be obtained by performing multiple screening cycles. The procedure has been proven useful for drug discovery (Omidfar & Daneshpour, 2015). However, the phage display method is limited to identify peptides that bind to a certain structure. Identification of random peptides that confer biological activity such as cell growth advantage or disadvantage by different biological mechanisms cannot be identified. Systematic evolution of ligands by exponential enrichment (SELEX) is another procedure that makes use of initially random sequences and binding to specific target molecules to obtain polymers with high binding affinity (Bruno 2015). However, in this case, one uses short DNA and RNA molecules (so-called aptamers) directly, rather than the translation products (Darmostuk et al. 2015). Similar as for the phage display technology, the polynucleotide chains are bound to a desired target and non-binding chains are removed. The binding variants are eluted and amplified *in vitro*. Again, this can be done in multiple cycles and can include additional mutagenesis steps in further cycles. This technology has also lead to some diagnostic and therapeutic applications, but is not as broadly used as protein based methods (Bruno 2015). SELEX is limited to select for DNA and RNA molecules that bind to certain structures. Again, identification of DNAs or RNAs that confer biological activity by different biological mechanisms cannot be identified.

Stepanov V.G. and Fox G. have described a method for *in vivo* selection of functional mini-genes from a randomized DNA library expressing combinatorial peptides in *E. coli* (Stepanov V.G. and Fox G., 2007). Specifically, this study describes a method for identifying from a randomized mini-gene library, those mini-genes which encode for peptides that confer resistance to stress conditions being nearly lethal concentrations of $NiCl_2$, $AgNO_3$ or $K_2TeO_3$ in the culture medium. However, this method has a number of disadvantages. For instance, the method requires application of a strong selection pressure. As apparent from the experiments performed by Stepanov V.G. and Fox G., this selection pressure led to a very high rate of secondary genomic mutations. In turn, this resulted in a very high rate of false positive hits, because in many cases the stress resistance was not conferred by the selected mini-gene but rather by a totally independent gene mutation in the *E. coli* genome. Moreover, the method is limited to

identification of mini-genes having a positive effect under the particular stress conditions applied; i.e. it is limited to identifying mini-genes that respond to a specific selection regime. Mini-genes that have a more general effect on, for example, cell growth or have a negative effect, for example, on growth under certain conditions, cannot be identified with such a method. Similarly, due to the experimental setup the method selects for mini-genes which confer a particular strong advantage under the applied stress conditions. Therefore, mini-genes conferring significant but mild advantages cannot be identified. Importantly, the method performed by Stepanov and Fox resulted only in identification of one polynucleotide and peptide encoded thereby for one stress condition used.

US 2012/0165225 A1 and US 8916376 B2 describe novel functionalized biomolecules and methods for generating such biomolecules. The methods described therein have the same limitations as described above in the context of Stepanov V.G. and Fox G.. In particular, the novel functionalized biomolecules described therein are restricted to biomolecules, which confer resistance to stress conditions induced by metals in the culture medium.

Thus, the technical problem underlying the present invention is the provision of means and methods for identifying polynucleotides encoding for biomolecules with biological activities, said means and methods being improved with respect to any of the above mentioned limitations. In particular, the technical problem can be seen as the provision of means and methods for reliably identifying a broad spectrum of polynucleotides encoding for biomolecules with biological activities, e.g. not only those that respond to a specific selection regime. Further, an object of the present invention is the provision of fully novel or at least so far not known polynucleotides which encode biomolecules with biological activity like, for example, RNAs or peptides. Further, an object of the present invention is to allow multiple bioactive peptides or RNAs (i.e. multiple peptides or RNAs with biological activity) to be detected in a single experiment. A further object of the present invention is to allow differential screening of bioactive peptides responding to different cell states.

The technical problem is solved by provision of the embodiments characterized in the claims and herein below.

Accordingly, in a first aspect, the present invention relates to a method for identifying a polynucleotide encoding a biomolecule with biological activity, said method comprising:

a) cultivating a population of host cells capable of expressing a library of polynucleotides, wherein each of said polynucleotides comprises a random nucleic acid sequence; and

b) determining the frequencies of (individual) polynucleotides in said library comprised in said population of host cells at a first time point and at a subsequent second time point during cultivation, wherein the biomolecules encoded by the polynucleotides of said library are expressed in said population between said first and said second time point,

wherein a change in the frequency of a polynucleotide in said library between said first and said second time point determined according to step b) identifies a polynucleotide encoding a biomolecule with biological activity.

In other words, the present invention relates to a method for identifying a polynucleotide encoding a biomolecule with biological activity, said method comprising:

a) cultivating a population of host cells capable of expressing a library of polynucleotides, wherein each of said polynucleotides comprises a random nucleic acid sequence; and

b) determining the frequencies of (individual) polynucleotides in said library comprised in said population of host cells at a first time point and at least one subsequent further time point during cultivation, wherein the biomolecules encoded by the polynucleotides of said library are expressed in said population between said first and said further time point(s),

wherein a (non-random or significant) change in the frequency of a polynucleotide in said library of polynucleotides between said first and one or more of said further time point(s) determined according to step b) identifies a polynucleotide with biological activity.

The present invention also relates to a method for identifying a polynucleotide encoding a biomolecule with biological activity, said method comprising:

a) cultivating a population of host cells capable of expressing a library of polynucleotides in at least 2, preferably at least 5 or most preferably at least 10

replicates, wherein each of said polynucleotides comprises a random nucleic acid sequence; and

b) determining for each of said replicates the frequencies of (individual) polynucleotides in said library comprised in said population of host cells at a first time point and at a subsequent second time point during cultivation,

wherein the biomolecules encoded by the polynucleotides of said library are expressed in said population between said first and said second time point,

wherein a significant change in the frequency of a polynucleotide in said library between said first and said second time point determined according to step b) identifies a polynucleotide with biological activity, and wherein the significance of change is assessed by a statistic test.

In particular, a statistic test to be employed in accordance with the invention is envisaged to allow determining the significance of frequency changes under the experimental boundary condition that were applied. For example, a statistic test to be employed in accordance with the invention is one as explicitly described herein and in the appended examples. In principle, any statistic test known in the art to be applied for assessing the probability of an observation within a random distribution may be employed in context of the present invention. In particular, a statistical test that is known in the art to be used for determining or analysing expression differences between cell states (for example RNA expression differences) may, for example, be applied for determining the significance of (a) frequency change(s). Such tests are, for example, described in Chen et al. (2011), which compared several of such test statistics (e.g. Wald test, likelihood ratio test, Fisher's exact test, variance stabilized test, and conditional binomial test) on simulated data and suggested that the Wald-Test is most powerful when only technical replicates are available. Thus, for example, the likelihood ratio test, Fisher's exact test, variance stabilized test, conditional binomial test or the Wald test may be employed in context of the present invention to determine (statistical) significance of (a) frequency change(s). Love et al. (2014) have developed the DESeq2 package as a general tool for differential analysis of count data. In one embodiment this package may be used to determine (significant) changes in frequencies of polynucleotides. Ching et al. (2014) compare five differential expression analysis packages. They propose to do a power analysis that takes the data structure into account for finding the optimal test under given experimental

conditions. Thus, in one aspect the teaching of Ching et al. may be used to select an optimal test for determining (significant) changes in frequencies of polynucleotides.

Furthermore, the present invention provides a method for identifying a deoxyribonucleic acid (DNA) molecule encoding a biomolecule with biological activity, said method comprising:

a) cultivating a population of host cells capable of expressing a library of DNA molecules, wherein each (or essentially each, e.g. at least 90%, preferably at least 95%, most preferably 100%) of said DNA molecules comprises a random nucleic acid sequence; and

b) determining the frequencies of individual DNA molecules in said library comprised in said population of host cells at a first time point and at a subsequent second time point during cultivation, wherein the biomolecules encoded by the DNA molecules of said library of DNA molecules are expressed in said population between said first and said second time point,

wherein a change in the frequency of a DNA molecule in said library of DNA molecules between said first and said second time point determined according to step b) identifies a DNA molecule encoding a biomolecule with biological activity.

In the methods of the present invention step a) may also be as follows: cultivating a population of host cells comprising a library of polynucleotides, wherein each (or essentially each) of said polynucleotides comprises a random nucleic acid sequence.

Similarly, in the methods of the present invention step a) may also be as follows: cultivating a population of host cells, wherein said population of host cells comprises a library of polynucleotides and is capable of expressing said library of polynucleotides, wherein each (or essentially each) of said polynucleotides comprises a random nucleic acid sequence.

Exemplary, however non-limiting, embodiments of the present invention are described as follows. In the context of these embodiments, at least one, preferably more, more preferably all, of the itemized steps as provided herein below are employed (non-limiting examples of these embodiments are also illustrated in the appended examples):

1. Generation of a clone library containing (polynucleotides comprising) randomized nucleotide sequences in suitable expression vectors that allow these sequences to be expressed as RNAs and with the option that the RNAs can be translated into polypeptides (for example by providing an appropriate start codon and ribosome initiation site).
2. Transformation of the library into host cells in which the RNA expression from the expression vector can be activated such that the cloned insert becomes expressed in the cell. The transformation process is preferably set up in a way to ensure that any cell receives not more than one vector with one sequence variant (this can for example be achieved by an excess of non-transformed cells over transformed cells and is implicit in the term "cloning").
3. Optional amplification of the library under conditions where the expression vector does not express the insert sequence.
4. Growth of samples of the library under conditions where the expression of the insert from the expression vector is activated. The growth conditions should preferably not be restrictive, i.e. should provide all cells good conditions of growth and the same probability of replicating. Their relative replication rate should only be influenced by the expression vector they carry. The first time point at which the frequency of the clones is assessed may be measured a few generations (for example about 1, 2, 3, 4, 5, 6, 7, 8, 9 or 10) after the activation of the expression vector, i.e. the induction of the expression of the polynucleotides. Selecting the first time point after the induction has the advantage that growth conditions are fully comparable between the first and the second time point.
5. Setting up replicates under the same growth conditions. The number of replicates may depend on a statistical power analysis that estimates the probability to detect a significant change in frequency in dependence of the number of cell divisions surveyed, number of different clones in the library, the dispersion (variance) between replicates and/or depth of sequencing, and the like (see also point 6).
6. Sequencing of the inserts of the expression vectors (randomized nucleotide sequences), preferably by a parallel sequencing procedure. This is to determine the relative frequency of the clones in the original library at said first time point,

as well as at said second time point of the experiment (optionally also at further time points).

7. Identification of clones that have changed their relative frequency among the cells during the course of the experiment, i.e. between the first and the second time point. Three types of clones can be expected - type 1: clones that have raised in frequency are expected to express a polynucleotide or peptide that is beneficial for the growth of the cell; type 2: clones that became reduced in frequency are expected to express a polynucleotide or peptide that is detrimental for the growth of the cell; type 3: clones that have not significantly changed their frequency are expected to express a polynucleotide or peptide that is neutral for the growth of the cell. Type 1 and 2 are referred to in the context of the invention as providing "biological activity" of the respective biomolecule.
8. Determination of significance values for the increase or decrease of peptide frequencies. These significance assessments may be made more powerful by running parallel experimental setups (replicates; see also item 5). This allows to estimate the variance in the experiment and/or to set a cutoff for significant outliers each under the conditions of the specific experiment.
9. Optionally: Confirmation of the results by repeating the whole experiment under the same conditions. Clones that are found in at least two independent experiments to show the same direction and preferably also the same magnitude of frequency change may be considered as (highly) confirmed.

Any, 2, 3, 4, 5, 6, 7, 8 or all of the above items may be combined with another embodiment of the invention described herein elsewhere. In particular they may, for example, be combined with the above mentioned methods for identifying a polynucleotide.

The gist of the present invention is the provision of random, yet, functional polynucleotides and/or polypeptides encoded by random polynucleotides and methods for identifying these. Preferably, the polynucleotides/polypeptides are generated and obtained by in vitro (random) modification(s) of DNA, in particular DNA obtained by randomized (bio-)chemical synthesis. In other words, the present invention provides for means and methods which allows for the selection of encoded polypeptides (or other

biomolecules) with biological activity. These may be generated from and/or based on the herein defined randomized (bio-)chemical or "artificial" polynucleotide sequences. The present invention provides for novel and inventive selection methods for such (*in vitro* generated) randomized (bio-)chemical or "artificial" polynucleotides. The nucleic acids sequences are selected/identified as encoding novel biomolecules (e.g. polypeptides or RNA) with biological activity, in particular biomolecules affecting the survival, growth and/or proliferation of the cells. The present invention provides for the identification of (a) potential biological function(s) of biomolecules which are the expression product of novel (artificially generated) polynucleotides. The present invention, therefore, provides for fast and reliable screening methods for the encoded biological function/activity of biomolecules that are expressed in the herein provided cellular systems on the basis of artificially generated polynucleotides that are introduced in said cellular systems. Accordingly, the present invention provides for (a) screening system(s) for artificially generated polynucleotides whereby artificially generated polynucleotides are selected/identified which encode for biomolecules with a biological function. The present invention is, in particular, useful since it allows for fast and reliable screenings/identifications of even a plurality of polynucleotides (potentially encoding a biomolecule with biological function) in parallel and/or in a short period of time.

Therefore, the methods of the present invention allow for identifying a polynucleotide within a library of polynucleotides which encodes a biomolecule with biological activity. Furthermore, the methods according to the present invention allow for the identification of polynucleotides comprising and/or having a random nucleic acid sequence. The method according to the present invention can also be employed for identifying (a) biomolecule(s) (in particular an RNA or a polypeptide) encoded by the respective polynucleotide(s). The biological activity of polynucleotides in host cells can be analyzed by performing the methods according to the present invention in the host cells. Different host cells may be used for this purpose as described in more detail herein elsewhere. The identified polynucleotides are characterized in that they have a biological activity (for example growth enhancement or growth inhibition). When performing the method according to the present invention, translation of the nucleic acid sequence of the identified polynucleotide(s) into the respective sequence of the biomolecule(s) (e.g. RNA sequence(s) and/or peptide sequence(s)) subsequently to

the identification step may be applied. The present invention is also directed to a method for identifying a biomolecule (e.g. RNA and/or peptide) with biological activity comprising the steps of a method for identifying a polynucleotide with biological activity (as described herein elsewhere) and (step of) determination of the nucleic acid or the amino acid sequence of the biomolecule (e.g. RNA and/or peptide) encoded by the identified polynucleotide with biological activity, respectively.

Previously known screening methods for identifying polynucleotides or peptides having biological activity are limited to polynucleotides/peptides that bind a specific target structure or promote cell proliferation and/or survival under selective (toxic or nearly toxic) culturing conditions. By contrast, the method of the present invention allows a more general approach, i.e. approach which allows identifying polynucleotides and corresponding encoded biomolecules (e.g. RNAs or peptides) that interact with, for example, different cellular pathways, also including pathways which may negatively affect cell growth (proliferation and/or survival). The selected polynucleotides are not restricted to those encoding for biomolecules (e.g. RNA or peptides) that interact with predefined target structures. This is achieved by identifying the polynucleotides conferring biological activity to a respective cell clone expressing the same by the effect on cell growth and cell proliferation and/or cell survival, respectively. In other words, the method of the present invention makes use of the evolutionary principle of natural selection. This is based on differences in reproductive fitness in a common environment (Boero 2015). In particular, cells that have an advantage in proliferation will show an enhanced reproductive rate, while individuals that have a disadvantage will show a decreased reproductive rate. Even very small differences in reproductive rate can become effective, in particular when the growth continues for multiple generations in the same environment. For example, assuming only a 5% fitness difference between two cells harboring different biomolecules (peptides or RNAs), one would have a 1.2-fold difference in the number of individuals after 4 generations of uninhibited growth, 1.5-fold at 10% and 5.1-fold at 50% (see, for example, Table 1).

Small differences in frequencies can also occur by random fluctuations. Hence, it is preferred to estimate for each experimental setup the statistical power for the detection of non-random fluctuations, by, for example, using approaches discussed by Ching et al. (2014) for the power of detection of RNA expression differences (which can be

analyzed analogous to the change in frequencies of polynucleotides in the context of the present invention) between cell states. The power of detection can be increased by including more replicates, for example, either through deeper analysis of a given setup or preferably through parallel setups. For each experiment it is preferred to control the false discovery rate (FDR) for independent test statistics. FDR is the probability that a certain fold-change could also occur by random fluctuation (Benjamini and Hochberg, 1995). In particular, this FDR is envisaged to be lower than, or equal to, 50% to make meaningful statements. Preferably, it is set at about 10% or 5%. The experiments in the appended example were, for example, evaluated at an FDR of 5%.

Methods known in the art, such as the method described by Stepanov V.G. and Fox G. (*loc. cit.*), may also make use of differences in the proliferation rate and/or survival of cells. However, in contrast to the means and methods of the present invention, these methods require culturing the host cells for a time period after which the cells with improved growth are present in such a high frequency that basically each cell analyzed comprises a polynucleotide which confers a very specific growth advantage. In other words, the culturing time and/or conditions are selected in a manner that the cells not comprising a polynucleotide conferring a growth advantage are essentially depleted in the cell cultures (i.e. the cells having a growth advantage due to the introduced polynucleotide are present in much higher frequencies than host cells with the remaining polynucleotides). This is required because these methods comprise the determination of polynucleotides conferring growth advantage by simply determining the polynucleotides comprised in a subset of the host cells rather than all host cells after the selection process by culturing under selection/stress conditions. This experimental set up requires that essentially each of the host cells remaining after the selection process is a host cell being transformed with a polynucleotide with the respective biological activity that improves cell proliferation/survival. Otherwise, many of the detected polynucleotides would be false positives in such experimental setup. Due to this requirement methods known in the art such as the method described by V.G. and Fox G. (*loc. cit.*) have a bias to primarily identify polynucleotides that confer a particular strong growth advantage under the stress/selection condition used. Moreover, only polynucleotides conferring a growth advantage under the specific selective conditions used can be identified.

The methods of the present invention could surprisingly solve this limitation by comparing the frequencies of the individual polynucleotides of the library of polynucleotides within the library at at least two different time points during culturing, i.e. at a first and at a second time point. Moreover, the limitations of the prior art have been solved by not applying selection pressure when performing the methods of the present invention.

These features of the means and methods of the present invention contribute to several advantages. First, this setup allows not only for identification of polynucleotides which promote growth (proliferation and/or survival) of the respective host cells but also polynucleotides which inhibit the same. Second, the method of the present invention is more sensitive, i.e. it allows for identification of polynucleotides having biological activities that promote or inhibit cell growth (proliferation and/or survival) in different strength. This enormously increases the spectrum and/or diversity of the polynucleotides/biomolecules which can be identified. In particular, it also allows identifying many bioactive peptides or RNAs in parallel in a single experiment. By contrast, the experimental setups of the previously known methods have a strong bias for identifying the polynucleotides conferring the highest cell proliferation/survival advantage, as they typically involve only determining the selected clones at the end of the selection cycles, thus typically identifying only a single clone in a given experiment. Third, due to the setup of determining the frequencies of polynucleotides within the host cell population at at least a first and a second time point during culturing, a particular short time of culturing the host cells and/or a reduced number of cell division cycles (cycles in which the cell number doubles) can be reached when identifying polynucleotides with biological activity in accordance with the means and methods of the invention. For example, the method of the present invention can identify a polynucleotide which confers biological activity already after only 1, 2, 3 or preferably 4 cell division cycles (i.e. the time during which the amount of cells is doubled) of the population of host cells. The reduction of the culturing time and/or the number of cell division cycles as achieved by the methods according to the present invention reduces the chance that secondary mutations in the host cell genome occur that may promote cell proliferation and/or cell survival and, in turn, lead to the identification of false positive polynucleotides. Fourth, the determination of frequency changes of many biomolecules (in particular peptides or RNAs) or the polynucleotides encoding said biomolecules in a given cell population results in a "fingerprint" of changes that can be

compared to the "fingerprint" from the same cells that are (genetically of pharmacologically) manipulated during the growth. The identification of biomolecules (e.g. peptides or RNAs) or the polynucleotides encoding said biomolecules that are specifically changed in either condition allows inferring an involvement in the specific pathways that were targeted by the genetic or pharmacologic manipulation. Accordingly, in another aspect, the methods of the present invention may be employed for two identical host cell populations in parallel, wherein one of the host cell populations is treated with a chemical substance (e.g. a pharmaceutically active chemical substance) and the other remains untreated (optionally also including replicates for each condition). In this aspect, the frequencies of polynucleotides of the library in the host cell population are assessed at a first and a subsequent second time point for both host cell populations. Subsequently, the frequencies and/or changes in frequencies of individual polynucleotides are compared between the two differently treated host cell populations. In the context of this aspect, the polynucleotides of the library that show a (significant) difference in the frequencies and/or changes in frequencies between the two host cell populations are identified as polynucleotides with biological activity. In particular, these polynucleotides are polynucleotides with a biological activity which is related to the chemical substance added to one of the host cell populations. In other words these nucleotides have the biological activity of altering the effect of a chemical substance on the host cell population. Similarly, instead of taking two genetically identical host cell populations that are treated differently (e.g. one is treated with a chemical substance), two host cell populations that are genetically different, e.g. by a defined mutation (e.g. knock-in or knock-out of a gene, mutation of gene or mutation of a gene expression control element (e.g. promoter) etc.) in the host cell genome may be employed and compared. The identified polynucleotides showing a (significant) difference in the frequencies and/or changes in frequencies between the two genetically different host cell populations encode biomolecules (e.g. peptides or RNAs) with biological activity that genetically interact with the pathway in which the mutated gene is involved. “Genetically interacting” means in this context that both genes may be involved in overlapping and/or parallel genetic pathways. In particular, in the context of the present invention an approach similar to, for example, shRNA library screening procedures (Sims et al. 2011) may be employed, wherein the library of polynucleotides is employed instead of an shRNA library.

The growth advantage (or disadvantage) principle underlying the means and methods of the invention does not require a direct interaction between the cells. In principle, only growth under the same general conditions is required. The decisive test lies in comparing growth (e.g. growth rates) within a given time frame, in particular between 2 time points. Hence, the means and methods can, for example, also be applied to growth conditions where cells are attached to a surface, or where individual cells grow in separate containers. However, the same or very similar/essentially the same growth conditions should be provided when such an experimental design is chosen. Further, differential reproductions rates of the host cells must be allowed to be in principle possible. The procedure is expected to work particularly well when growth conditions are optimal, e.g. such that cells can replicate at a high rate.

Given the nearly endless number of possible combinations of nucleotides and amino acids in polynucleotide or peptide chains, it will be very difficult (and for higher chain lengths technically impossible) to cover all combinations in a single clone library. However, the means and methods of the invention allow that a multitude of different polynucleotide libraries that comprise different or partially different random polynucleotides can be tested to obtain bioactive peptides or polynucleotides. This can, for example, be achieved by performing the methods of the present invention at least two times with two different polynucleotide libraries. The different libraries may, for example, be based on different strategies for synthesizing random combinations, or combinations that bias nucleotide or amino acid composition in desired directions.

In the context of the invention, each, or essentially each (for example at least 70%, 80%, 90%, 95% or 98%), of the polynucleotides of the library may comprise a different random nucleic acid sequence. In other words, the diversity of the random nucleic acid sequences to be comprised in the polynucleotides of the library to be screened in accordance with this invention may be rather high. This means that as much as possible different random nucleic acid sequences may be present in the population of host cells to be cultivated in accordance with the invention. However, in principle, the occurrence of (some, for example, at least 2%, 5%, 10%, 20%, 30%, 40% or 50%) polynucleotides which comprise the same nucleic acid sequence is also tolerated. In other words, certain, several or all polynucleotides may occur several times (e.g. at least 2, at least 3, at least 4, at least 5, at least 8, at least 10, at least 20, at least 50 or

at least 100 times) within the analyzed sample of the library of polynucleotides with random or partially random nucleotides. This repetitive occurrence of polynucleotides in the library of polynucleotides has the advantage that the chance that such a polynucleotide is detected in the sequencing is higher. Moreover, the more often a respective polynucleotide is detected (i.e. the more counts are detected during sequencing) the more precise determination of frequency changes between two time points is possible. Moreover, in the embodiments in which statistics are employed, the power of the statistics is increased through higher occurrence of the detected polynucleotides.

In the context of the invention, it is explicitly envisaged that the nucleic acid sequence comprised in polynucleotides of the library is an exogenous nucleic acid sequence, i.e. not a nucleic acid sequence that originates from the host cell to be employed and not a nucleic acid sequence that is comprised in the host cell's genome (or any other known genome). This exogenous character of the nucleic acid sequence is also implied by terms like „random" or „randomized" nucleic acid sequence. In one aspect, the terms „random" or „randomized" nucleic acid sequence mean that the nucleic acid sequence is an artificial nucleic acid sequence and/or does not occur in nature (or at least has not been described as occurring in nature). Likewise, the terms mean that the nucleotide sequence encodes an artificial, "random" or "randomized" biomolecule (e.g. RNA or protein) and does not encode a biomolecule (e.g. RNA or protein) that occurs in nature (or that at least has not been described as occurring in nature). In a specific aspect, „random" or „randomized" nucleic acid sequence means that there is a random distribution of the 4 nucleotides A, C, T, G within the nucleic acid sequence to be employed in accordance with the invention. However, (slightly) turning away from a strict random distribution of the 4 nucleotides is also possible, as long as this does not result in a nucleic acid sequence which is anyway present in the host cell and is comprised in the host cell's genome, respectively. The random or nearly random combination of nucleotides creates an enormous combinatorial potential, in particular, when extended sequences are considered. For example, a random combination of 150 nucleotides as it is used in the supporting experiments creates $4^{150}$ sequence variants, which corresponds to a number larger than the estimated number of atoms in the universe. Hence, it is exceedingly unlikely and nearly impossible that an already known sequence will be among the randomly or nearly randomly synthesized ones. This

extremely low chance can e.g. be also expressed by the term "essentially random" or "partially random".

The library of polynucleotides employed in this context comprises or consists of (or essentially consists of) polynucleotides having a completely or partially random nucleic acid sequence. In particular, 70% or more, preferably 80% or more, preferably 90% or more, preferably 95% or more, preferably 98% or more, preferably and 99% or more and, most preferred, 100% of the polynucleotides of the library have a completely or partially random nucleic acid sequence. "Partially", in this context, may, for example, mean that 70% or more, preferably 80% or more, preferably 90% or more, preferably 95% or more, preferably 98% or more, or preferably 99% or more of the nucleic acid sequence are random. In principle, an employed library of polynucleotides may consist only of random or partially random polynucelotides. However, it is also envisaged to use a polynucleotide library that comprises (but does not consist of) random or partially random polynucleotides (polynucleotides comprising a random nucleic acid sequence). For example, a library that comprises or consists of random or partially (see definition above) random polynucleotides and one or more control polynucleotides (see definition further below) may be employed. In particular, a library of polynucleotides comprising a control polynucleotide at a frequency of up to 5%, or up to 10% or up to 20% or up to 30% or up to 40% or up to 50% of the total number of polnyucleotides in the library may be employed.

It is also envisaged that the cultivated population of host cells may optionally in addition to a library of polynucleotides, in which each of said polynucleotides comprises a random nucleic acid sequence, may also comprise one or more polynucleotides that do not have a random polynucleotide sequence (non-random polynucleotide) or are known to have or to not have a biological function. In other words, these polynucleotides may have a nucleic acid sequence known in the art (e.g. a control polynucleotide). The host cell population may further be capable of expressing these one or more non-random polynucleotides that have a predetermined polynucleotide sequence. In principle, however, also polynucleotides that are not expressed (e.g. empty vectors) may be used. The one or more polynucleotides that do not have a random polynucleotide sequence may, for example be introduced together (in the same pool) with the library of polynucleotides into the host cells. In principle,

also a library of polynucleotides that comprises and/or consist of random or partially random polynucleotides and one or more non-random polynucleotides (e.g. a control polynucleotide) may be employed. The one or more polynucleotides of known sequence may e.g. comprise or consist of one or more control polynucleotide(s).

In the context of the present invention (a) control polynucleotide(s) can, in principle, be any polynucleotide(s) of known nucleic acid sequence, i.e. a previously described nucleic acid sequence (e.g. a naturally occurring nucleic acid sequence) or even a random or partially random nucleic acid sequence that is known to encode for a biomolecule with biological activity or to have a biological activity on its own. In principle, a polynucleotide that does not have a random polynucleotide sequence or a control polynucleotide may also be a polynucleotide that is comprised in the library of (random or partially random) polynucleotides employed in the context of the present invention. This may for example include cases in which by chance a polynucleotide sequence that is known in the art or naturally occurring is part of the library. As mentioned elsewhere herein, it is highly unlikely that during synthesis of a polynucleotide library with random or partially random sequences as employed in the context invention a polynucleotide that is known in the art or naturally occurring is synthesized and, thus comprised in the library (in particular considering the length of the random sequences).

In particular, the one or more control polynucleotide(s) sequences may comprise or be (a) polynucleotide(s) that do(es) not encode for a biomolecule with biological activity (e.g. an empty vector that is not expressed or a nucleic acid encoding a biomolecule without a biological activity (e.g. a known artificial polynucleotide sequence without biological activity)). A control polynucleotide that does not encode for a biomolecule with biological activity is also referred to herein as “neutral control polynucleotide”. In such case, the frequency change of said control polynucleotide between the first and the second time point in the methods according to the present invention might be used as control that gives an indication or a measure for random fluctuations in change in frequency. If such control polynucleotide is employed, in particular a (significant) change in the frequency of a polynucleotide of the library may be identified if the change in frequency of a polynucleotide of the library is higher (e.g. at least 1.5-fold,

preferably at least 2-fold, preferably at least 3-fold, most preferably at least 5-fold higher) than the change in frequency of the control polynucleotide(s).

The one or more control polynucleotide(s) sequences may, in particular, also comprise or be (a) polynucleotide(s) that encode(s) for a biomolecule with biological activity (in particular a growth promoting or inhibiting biological activity). In such case the population of host cells preferably comprises cells capable of expressing the control polynucleotides. Furthermore, in such case the control polynucleotide may indicate an exemplary fold-change/change in frequency of a polynucleotide that promotes or inhibits cell growth. In principle, the observed fold-change/change in frequency for such a control polynucleotide encoding for a biomolecule with biological activity may also be used as cut-off to determine when a change in frequency exists. For example, a change in frequency of a polynucleotide of the library may in such an embodiment only be detected if the fold-change is at least as high or higher (e.g. at least 1.5-fold, preferably at least 2-fold, preferably at least 3-fold, most preferably at least 5-fold higher) than the fold-change of the control polynucleotide(s) that encode(s) for a biomolecule with biological activity.
Any embodiment disclosed herein with the exception of those defining the random or partially random character of the polynucleotides of the library of polynucleotides employed in the context of the present invention can be applied *mutatis mutandis* to the control polynucleotide(s) and/or polynucleotide(s) that do(es) not have a random polynucleotide sequence. This particularly applies to all embodiments describing in which form (e.g. in a vector) the polynucleotides are comprised in the population of host cells.

Accordingly, in one embodiment the present invention relates to a method for identifying a polynucleotide encoding a biomolecule with biological activity, said method comprising:

a) cultivating a population of host cells comprising or consisting of a first and a second subpopulation (group) of host cells, wherein said first subpopulation (group) of host cells is capable of expressing a library of polynucleotides, wherein each of said polynucleotides comprises a random nucleic acid sequence, and wherein said second subpopulation (group) of host cells comprises a control polynucleotide; and

b) determining the frequencies of individual polynucleotides of said library of polynucleotides comprised in said population of host cells at a first time point and at a subsequent second time point during cultivation and determining the frequency of said control polynucleotide comprised in said population of host cells at said first time point and at said subsequent second time point during cultivation, wherein the biomolecules encoded by the polynucleotides of said library of polynucleotides are expressed in said population between said first and said second time point,

wherein a (significant) change in the frequency of a polynucleotide in said library of polynucleotides between said first and said second time point determined according to step b) identifies a polynucleotide (of said library) encoding a biomolecule with biological activity.

Included is particularly also a method for identifying a polynucleotide encoding a biomolecule with biological activity, said method comprising:

a) cultivating a population of host cells comprising a first and a second subpopulation (group) of host cells, wherein said first subpopulation (group) of host cells is capable of expressing a library of polynucleotides, wherein each of said polynucleotides comprises a random nucleic acid sequence, and wherein said second subpopulation (group) of host cells comprises a control polynucleotide; and

b) determining the frequencies of individual polynucleotides of said library of polynucleotides comprised in said population of host cells at a first time point and at a subsequent second time point during cultivation and determining the frequency of said control polynucleotide comprised in said population of host cells at said first time point and at said subsequent second time point during cultivation, wherein the biomolecules encoded by the polynucleotides of said library are expressed in said population between said first and said second time point,

wherein a significant change in the frequency of a polynucleotide of said library of polynucleotides between said first and said second time point determined according to step b) identifies a polynucleotide (of said library) encoding a biomolecule with biological activity,

and wherein said significance is assessed by comparison with the change in frequency of said control polynucleotide between said first and said second time point determined according to step b).

In one aspect, the control polynucleotide may be a polynucleotide having a biological activity and a change in the frequency of a polynucleotide of said library of polynucleotides between said first and said second time point determined according to step b), which is at least as high as the change in frequency of said control polynucleotide between said first and said second time point determined according to step b), identifies a polynucleotide (of said library) encoding a biomolecule with biological activity. The control polynucleotide used in such embodiment may preferably be a polynucleotide which is known to have a biological activity which increases or decreases the time needed per cell division of the host cells/doubling of the host cell number by 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2, 3, 5, or 10-fold. In principle, selecting a control polynucleotide having a biological activity conferring a lower increase or decrease per cell division of the host cells/doubling of the host cell number is preferred.

In another aspect, the control polynucleotide may be a polynucleotide having no biological activity (e.g. no biological activity affecting host cell growth and or proliferation) and a change in the frequency of a polynucleotide of said library of polynucleotides between said first and said second time point determined according to step b), which is higher (e.g. at least 1.5-fold, preferably 2-fold, most preferably 5-fold) than the change in frequency of said control polynucleotide between said first and said second time point determined according to step b), identifies a polynucleotide (of said library) encoding a biomolecule with biological activity.

In another embodiment the present invention relates to a method for identifying a polynucleotide encoding a biomolecule with biological activity, said method comprising:

a) cultivating a population of host cells capable of expressing a library of polynucleotides, wherein said library of polynucleotides comprises (or consists of) random or almost random polynucleotides; and
b) determining the frequencies of (individual or all individual) polynucleotides of said library comprised in said population of host cells at a first time point and at a

subsequent second time point during cultivation, wherein the biomolecules encoded by the polynucleotides of said library are expressed in said population between said first and said second time point,

wherein a change in the frequency of a polynucleotide in said library between said first and said second time point determined according to step b) identifies a polynucleotide encoding a biomolecule with biological activity.

In particular, the present invention also relates to a method for identifying a random or partially random polynucleotide encoding a biomolecule with biological activity, said method comprising:

a) cultivating a population of host cells capable of expressing a library of polynucleotides, wherein said library of polynucleotides comprises (or consists of) random or almost random polynucleotides and one or more control polynucleotide(s); and

b) determining the frequencies of (individual or all individual) random or almost random polynucleotides and said one or more control polynucleotides of said library of polynucleotides comprised in said population of host cells at a first time point and at a subsequent second time point during cultivation, wherein the biomolecules encoded by the random or partially random polynucleotides of said library are expressed in said population between said first and said second time point,

wherein a change in the frequency of a random or partially random polynucleotide in said library between said first and said second time point determined according to step b) identifies a random or partially random polynucleotide encoding a biomolecule with biological activity.

In one aspect, the control polynucleotide may be a polynucleotide having a biological activity and a change in the frequency of a random or partially random polynucleotide of said library of polynucleotides between said first and said second time point determined according to step b), which is at least as high as the change in frequency of said control polynucleotide between said first and said second time point determined according to step b), identifies a random or partially random polynucleotide (of said library) encoding a biomolecule with biological activity. The control polynucleotide used in such embodiment may preferably be a polynucleotide which is known to have a

biological activity which increases or decreases the time needed per cell division of the host cells/doubling of the host cell number by 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2, 3, 5, or 10-fold. In principle, selecting a control polynucleotide having a biological activity conferring a lower increase or decrease per cell division of the host cells/doubling of the host cell number.

In another aspect, the control polynucleotide may be a polynucleotide having no biological activity (e.g. no biological activity affecting host cell growth and or proliferation) and a change in the frequency of a polynucleotide of said library of polynucleotides between said first and said second time point determined according to step b), which is higher (e.g. at least 1.5-fold, preferably 2-fold, most preferably 5-fold) than the change in frequency of said control polynucleotide between said first and said second time point determined according to step b), identifies a polynucleotide (of said library) encoding a biomolecule with biological activity.

In one aspect, the methods for identifying a random or partially random polynucleotide encoding a biomolecule with biological activity described herein may also be repeated with a selected subset of the identified polynucleotide(s) (and optionally a control polynucleotide). This may be done to re-confirm the properties of the identified polynucleotides and/or biomolecules.

The biomolecule with biological activity encoded by the identified polynucleotide may, in the context of the present invention, be an RNA or, preferably, a polypeptide (also referred to as peptide or protein herein). Similarly, also the biomolecules as encoded by the polynucleotides of the library of polynucleotides are RNAs or polypeptides (also referred to as peptides or proteins herein).

Due to the random or partially random nucleic acid sequences of the library encoding the biomolecules (e.g. polynucleotides), the biomolecules (e.g. RNAs or polypeptide) encoded thereby are also random or almost random combinations of, for example, ribonucleotides or amino acids (also including modified derivatives of ribonucleotides or amino acids), respectively. Likewise, also in this context, “almost random” or “partially random” may, for example, mean that 70% or more, preferably 80% or more, preferably 90% or more, preferably 95% or more, preferably 98% or more, preferably and 99% or more of the, for example, ribonucleotides or amino acids (sequences), respectively, are random. Thus, the method of the present invention allows for

identifying from a large pool of polynucleotides (with random or almost random nucleic acid sequence) novel polynucleotides which encode for biomolecules (e.g. RNAs and/or peptides) that have a biological activity. Thus, novel polynucleotides encoding biomolecules with biological activity and the novel biomolecules (e.g. RNAs and/or peptides) encoded by said polynucleotides can be identified. One advantage of the method of the present invention is that novel polynucleotides and biomolecules (e.g. RNAs and/or peptides), which are not naturally occurring (i.e. which are artificial), or which are not yet known to be employed in nature, can be identified. These novel polynucleotides and corresponding biomolecules (e.g. RNAs and/or peptides) identified by the method of the present invention provide totally new possibilities to modify cellular pathways or to develop new drugs.

As mentioned above, the biomolecule with biological activity encoded by the identified polynucleotide may, in the context of the present invention, be an RNA or a polypeptide. Accordingly, the methods of the present invention may further comprise a step to assess whether the biomolecule having biological activity is the RNA or the polypeptide encoded by a respective identified polynucleotide. An illustrative example how to assess whether the biological activity conferred by the expression of a polynucleotide is conferred by the encoded RNA or by the encoded polypeptide is presented in the appended Examples (see Example 4). Accordingly, to evaluate whether the RNA or the polypeptide encoded by the respective polynucleotide confers the biological activity to the host cell, one may, for example, create a variant of the identified polynucleotide that prevents the translation of an in-frame polypeptide encoded by the randomized sequence (also referred to as inactivated polynucleotide in the following) and compare the growth of a cell containing this mutant variant of the identified polynucleotide with a cell bearing a control polynucleotide (e.g. an empty vector) and/or a cell bearing the identified "wild-type" polynucleotide encoding for a biomolecule with biological activity. Preferably, the variant that prevents the translation of an in-frame polypeptide encoded by the randomized sequence is a variant that bears a premature stop codon before the start of the randomized part of the respective polynucleotide, preferably directly before the start of the randomized part of the respective polynucleotide. In principle, however every other mutation preventing expression of the polypeptide normally encoded by the randomized part of the polynucleotide would also be suitable, e.g. a frame shift mutation. If a growth

comparison is performed in comparison to cells bearing a control polynucleotide without a biological activity and no difference between the growth of the cells bearing the control polynucleotide and the inactivated polynucleotide can be observed, this means that the biological active biomolecule is the peptide encoded by the identified ("wild-type") polynucleotide. By contrast, if also the inactivated polynucleotide variant would still confer a growth advantage or disadvantage compared to the cells bearing the control polynucleotide, the biological active biomolecule would be the RNA encoded by the identified ("wild-type") polynucleotide. When a growth comparison is made with cells bearing the identified wild-type polynucleotide, observing a similar growth for the cells bearing the inactivated polynucleotide and the cells bearing the identified ("wild-type") polynucleotide would indicate that the biomolecule with biological activity is the RNA encoded by the identified ("wild-type") polynucleotide. By contrast, if the identified ("wild-type") polynucleotide leads to a loss of the growth advantage or disadvantage otherwise confirmed by the identified ("wild-type") polynucleotide, the biomolecule with biological activity is the polypeptide encoded by the identified ("wild-type") polynucleotide.

The growth comparison can in principle be performed by culturing each clone individually and comparing the growth rates, e.g. by determining the change in cell numbers over time. Preferably, the comparison may however be achieved by culturing the clones together and determining the frequencies of the respective polynucleotides (and/or control polynucleotides) at at least two time points. The frequencies may be determined by sequencing as described in the context of the present invention. In other words, the growth comparison may in principle be a repetition of the methods according to the present invention with the only exception that instead of the polynucleotide library only the respective inactivated polynucleotide and the control polynucleotide and/or the wild-type polynucleotide are provided. Alternatively, the frequencies may also be determined by any different method that allows distinguishing the different DNAs comprised and the subsequent quantification thereof. For example, a restriction digest that allows generating DNA fragments that can be distinguished in size and a subsequent quantification using a Bioanalyzer may be employed (see example 4 for further details).

The Examples and Figures of the present invention exemplarily illustrate how the method of the present invention may be performed. In particular, an *E. coli* host cell

population was used in this respect. Specifically, for the purpose of performing the first step of the present invention, a population of host cells comprising a library of polynucleotides was generated. Expression vectors were constructed (see example 1) and were subsequently employed to screen for novel polynucleotides that encode for biomolecules (e.g. peptides (or in principle also RNAs)) with biological activity.

For the purpose of constructing a population of host cells to be employed in accordance with (the first step of) the present invention, suitable host cells (e.g. *E. coli* cells) were transformed with the constructed library of polynucleotides (which, for example, may be comprised in expression vectors) and subsequently cultured (e.g. in LB-medium (which, may comprise an antibiotic selecting for transformants (e.g. ampicillin)). In a preferred embodiment, the culturing is performed under (otherwise) non-selective (more preferably optimal) growth conditions. In such a first growth phase, the cells may be pre-cultured in the absence of further compounds like, for example, IPTG in order to amplify the transformed cells without expression of the biomolecules (e.g. peptides (and/or RNAs)) encoded by the polynucleotides of the library (which may be comprised in the respective expression vectors), i.e. biomolecules (e.g. peptides) comprising random or partially/almost random amino acid sequences (or RNAs comprising random nucleic acid sequences). In a second growth phase, which may, for example, involve four overnight culturing cycles, the cells were cultured in the presence of further compounds like, for example, IPTG, e.g. conditions under which the biomolecules (e.g. peptides (and/or RNAs)) encoded by the polynucleotides are expressed. During this second growth phase, samples of the culture may be collected at different time points during culturing, as described further below. Finally, the frequency of different clones (at these different time points) was assessed by isolating the polynucleotides (which may be comprised in vectors/plasmids) from each of the collected samples and assessing the frequency of individual polynucleotides (which may be comprised in vectors/plasmids) in that sample by a nucleotide sequencing of the respective insert sequences. The expressed random or partially/almost random insert sequences that were either enriched or depleted during culturing were identified as polynucleotides with biological activity (encoding for biomolecules with biological activity). Without being bound by any theory, they promote or inhibit the proliferation of the respective host cell clone. In particular, the growth advantage or growth disadvantage is conferred by the biomolecule (e.g. RNA and peptide) expressed from the random polynucleotide sequence employed.

In principle, the random (or essentially random) nucleic sequence of the polynucleotides of the library to be employed in accordance with the invention may have an equal or an unequal representation of the four nucleotides A, C, G and T at each position. An equal representation is preferred in this respect, but a biased representation, wherein the representation may include one, two, three or all four different nucleotides at either position could help to direct the outcome to particular amino acid combinations or a reduction of the possibility of creating premature stop codons. Optimised methods for creating random (or nearly random) sequences have, for example, been explored in the context of creating phage display libraries (Omidfar et al. 2015). These methods may also be employed in the context of the present invention for generating polynucleotide libraries with random or partially random nucleic acid sequences.

In the context of the invention, the length of the random nucleic acid sequence is, in principle, not particularly limited. However, relatively short nucleic acid sequences are preferred in this respect. For example, a random nucleic acid sequence to be employed in the context of the invention may have a length of 18 to 300 nucleotides, preferably 36 to 250 nucleotides and more preferably 120 to 180 nucleotides. Similarly, a partially/almost random nucleic acid sequence may, for example, comprise one or more random nucleic acid sequence stretch(es)/fragment(s) having a length of 18 to 300 nucleotides, preferably 36 to 250 nucleotides and more preferably 120 to 180 nucleotides. Due to the high combinatorial number of different nucleic acid sequence variants for a random sequence or sequence stretch/fragment of that length it is very unlikely that an employed sequence exists in known genomes or is naturally occurring. Similarly, also random nucleic acid sequences with a length of 63 to 300 nucleotides may, for example be employed. If the biomolecule intended to be encoded by the nucleic acid sequences is a peptide, the number of nucleotides in the polynucleotide is preferably a multiple of 3 in order to ensure that the random sequence or sequence stretch encodes a corresponding random amino acid sequence or sequence stretch in frame.

In the context of the invention, the random sequence may be expressed not in frame or, preferably, in frame.

In the context of the present invention "culturing", "cultivating" or "cultivation" means preferably maintaining and/or proliferating the cells. Maintaining and/or proliferating the cells *in vitro* is particularly envisaged. In a most preferred embodiment the cells are proliferated, i.e. they undergo cell division/mitosis.

In the context of the methods of the present invention "expressing a library of polynucleotides" means that the polynucleotides of the library are transcribed into RNA (e.g. m-RNA) and/or that the peptide encoded by the polynucleotide is expressed. Preferably only one polynucleotide of the library is thereby expressed per host cell of the population of host cells or subpopulation thereof. The "one" particularly pertains to the sequence of the polynucleotide. In other words, the "one" means that several copies of a polynucleotide having identical sequences may be expressed per host cell. It is further not excluded and also envisaged that a host cell may also harbour more than one polynucleotide (sequence) of the library of polynucleotides. This may in particular be the case when it is of interest to test the interaction between different polynucleotides within a single cell. The copy number per host cell can be, for example, controlled by adapting the transformation conditions (e.g. the ratio between cells and DNA to be transformed). It is well known in the art how the transformation conditions can be adapted to control the copy number of polynucleotides of the library per host cell.

In principle, the biological activity of the biomolecule to be identified in accordance with the invention may be any biological activity. However, a biological activity which is growth-related is particularly preferred. The meaning of "growth-related" is clear to the skilled person and may, for example, refer to any activity which influences growth (for example, enhances or impairs growth) of cells (in particular of the host cells to be employed in accordance with the invention). Specifically, in accordance with the invention, the biological activity may be cell growth promoting or cell growth inhibiting activity and/or cell survival promoting or cell survival inhibiting activity and/or biological activity that promotes or inhibits cell division, proliferation and/or survival.

In the context of the methods of the present invention "frequency" or "frequencies" of polynucleotides means the amount, abundance or percentage relative to the total number of polynucleotides of the library comprised in the host cell population or a

representative sample thereof. For example, the frequency at a defined time point can be determined by using a representative sample of the population of host cells at the respective time point of the cultivation (e.g. said first or said second time point of the methods of the present in invention), isolating the respective polynucleotides from the sample of host cells (e.g. the plasmid containing said polynucleotide), amplifying the polynucleotides of the library (e.g. the polynucleotides within this vectors) using a PCR with, for example, about 25 cycles or less (e.g. using a primer binding to a position in a vector/plasmid upstream of the polynucleotide of the library and a primer binding to a position in a vector/plasmid downstream of the polynucleotide of the library, wherein said primers also include the primer binding sites for subsequent sequencing) (this may avoid PCR bias) and subjecting the amplicons to DNA sequencing (e.g. Illumina MiSeq sequencing). The PCR for amplification of the polynucleotides before DNA sequencing is preferably performed with a PCR set up that allows for maintenance of differences in amounts of different polynucleotides. In particular, the Illumina amplicon sequencing kit may be employed in the context of the present invention. The optimal sequencing depth should be assessed as part of the statistical power analysis described above.

The number of occurrence of each specific polynucleotide among the sequenced reads constitutes the absolute amount or abundance of the specific polynucleotide. To make this comparable between time points, it may preferably be normalized by the total number of reads obtained from the sequencing of each time point. Preferably, this is achieved by adjusting the sampled read number to the lowest available for two time points or a time series. This absolute amount or abundance of a given polynucleotide variant can also be expressed as a percentage relative to the total number of polynucleotides of the library, or a percentage relative to the total number of different polynucleotides that are statistically evaluated. Alternatively, and more preferably, in this case the frequency can also be the number/amount/abundance of polynucleotides of the library of polynucleotides (as for example shown in the appended examples). A frequency change is then recorded as the difference in the number/amount/abundance of a given polynucleotide between two time points (e.g. said first and said second time point).

In the context of the methods of the present invention determining said frequency is preferably performed by sequencing, e.g. following the general strategy that is also used for determining expression frequency differences in naturally occurring RNAs (see Oshlack et al. 2010 for a review). As illustrated in the appended examples sequencing can, for example, be carried out with an Illumina MiSeq sequencer, following the standard sequencing protocols. The procedure starts with taking a sample of the host cells from the total number of host cells in the experiment. The sample size (in other words the number of host cells comprised in the sample) needs to be adjusted to the number of expected different clones in the library, as determined during the library construction process. To obtain a representative sample, one needs to sample a cell number that is larger than the expected number of different clones in the sample, preferably at least 5-fold larger and most preferably at least 10-fold larger. The (vector) DNA is then extracted from the sample. The inserts representing the expressed random polynucleotides are then PCR amplified using specific primers flanking the inserts (polynucleotides of the library) and providing the primer sites for subsequent sequencing. The resulting sequencing reads are quality checked using the procedures recommended by the supplier of the instrument and the sequencing kit (e.g. Illumina) and a non-redundant library of unique sequence reads is produced from all available sequences. This library serves as the reference for counting the number of occurrences of a specific sequence in the full sequencing run. A representative example how determining can be carried out can also be found in the appended example.

A change in the frequency of a polynucleotide in the library to be employed in accordance with the invention may be an increase or a decrease in the counted number or frequency between two time points (e.g. the first and the second time point). In particular, a change in the frequency may be an (significant) increase in the number or frequency by at least 1.1-fold, preferably by at least 1.5-fold, preferably by at least 2-fold, preferably by at least 2.5-fold, preferably by at least 3-fold and even more preferably by at least 10-fold or a (significant) decrease in frequency by at least 1.1-fold, preferably by at least 1.5-fold, preferably by at least 2-fold, preferably by at least 2.5-fold, preferably by at least 3-fold and even more preferably by at least 10-fold between two time points (e.g. the first and the second time point). A change in frequency of a polynucleotide in the library may in particular be assessed by a fold-

change cut-off when a polynucleotide is at least detected 2, preferably 3, preferably 4, preferably 5, or preferably 10 times at the first and/or the second time point. If a polynucleotide is not identified at one of the time points the frequency is set to one in order to calculate a respective fold change value. In one aspect, the fold change values may also be expressed as corresponding $\log_2$ values.

Alternatively, and preferred, a change in the frequency of a polynucleotide in the library to be employed in accordance with the invention may be a (statistically) significant increase or a (statistically) significant decrease in the counted number or frequency between two time points (e.g. the first and the second time point). A variety of statistic test that are appropriate for such analysis are known in the art (for example: Student t-test, Wald test, likelihood ratio test, Fisher's exact test, variance stabilized test or conditional binomial test). The goal is to assess the probability of the occurrence of a measured value within a distribution of random values, whereby the distribution of the random values can be influenced by experimental parameters. Modelling the parameters of particular experimental combinations can lead to choosing the optimal test statistic. Since the experimental procedures of the current invention are very similar to testing RNA expression differences in cells based on high throughput sequencing procedures, it is possible to follow, for example, the teaching in Chen et al. (2011) and Love et al. (2014) for employing the best test statistic and the teaching in Ching et al. (2014) for doing a power analysis of the statistical test under any experimental condition. Using these statistical tests, the statistical significance of a given change (e.g. the change in frequency of a polynucleotide between said first and said second time point) is determined by employing a probability cutoff value, the p-value (often set at 0.05). However, since the present invention uses the same data (namely the overall sequence counts) for determining the p-values of changes in multiple clones, one would preferably correct for this multiple testing. The most broadly used approach for high throughput data is to control for the false discovery rate (FDR) using the procedure of Benjamini and Hochberg (1995). This allows estimating how many of the statistically significant values obtained in the multiple statistical testing of a dataset are likely to be wrong. For the statistical testing in the present invention a false discovery rate (FDR) of at least 50%, preferably 10%, and most preferably 5% is employed.

Using state of the art statistics, such as, for example, the Wald test as suggested by (Chen et al. 2011) and, optionally, being implemented in DESeq2 (Love et al. 2014) allows the detection of meaningful changes in frequency regardless of the size of the effect. For example, polynucleotides of the library with fold changes in the order of 1.1 relative to the initial conditions can be identified as changed in frequency, i.e. as increased or decreased in frequency, provided that the variance across replicates allows a precise distinction. For this reason, it can be preferred not to apply a simple fold change cut-off value, but rather retain those cases with significant frequency changes (e.g. at a FDR of 5%) regardless of the effect size (fold change value). In particular, using significant frequency changes (e.g. at a FDR of 5%) regardless of the effect size may have the advantage that more polynucleotides are identified. Moreover, also the rate of potentially false positively identified polynucleotides may be reduced.

The Wald test is a parametric statistical test named after the Hungarian statistician Abraham Wald who was the first to formally describe it (Wald 1943) and it is well known in the art. For example, the Wald test can be performed as follows: Significance of a change in the frequency of a polynucleotide in the library of polynucleotides between the first time point and the second time point can be identified by dividing the $\log_2$-fold change in said frequency of a polynucleotide by the standard error, and comparing the resulting z-statistic to a normal distribution to obtain a p-value. The determined p-values can optionally be further corrected for multiple testing by controlling for the false discovery rate (Benjamini and Hochberg 1995). Additionally, also simulations, e.g. as described in Ching et al. (2014), can be performed to obtain statistical cutoff values.

Polynucleotides encoding biomolecules (e.g. peptides and/or RNAs) with biological activity may in principle be identified by performing the method for identifying polynucleotides encoding for biomolecules with biological activity according to the present invention with only one sample of the population of host cells. The method for identifying polynucleotides encoding for biomolecules with biological activity according to the present invention may, however, also be performed with at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9 or at least 10 replicates of the population of host cells. Preferably, at least 5 replicates are used. Most preferably at least 10 replicates are used. A “replicate” or “replicates” means (a) separate

culture(s) of the same population of the host cells which is/are treated under the same or similar conditions. The methods of the present invention are typically performed in parallel. However, in principle, they can also be performed sequentially, if the same culturing conditions can be ensured. Steps a) and b) of the method described herein are in principle performed separately for each replicate. Step b) of the method, may, if performed by DNA sequencing, however, be performed in the same reaction. In such case barcode sequences (i.e. short predefined DNA sequences of 4 to 10 nucleotides, e.g. 4, preferably 5, preferably 6, preferably 7, or preferably 8 nucleotides) may be used to distinguish the replicate samples, as described elsewhere herein. Using replicates has the advantage that the method according to the present invention has an increased statistical power and sensitivity, i.e. that a broader spectrum of polynucleotides encoding for biomolecules with biological activity and encoded biomolecules can be identified. In particular, even biomolecules which only contribute to a slight change in growth behaviour may be identified. Another advantage is that the anyhow particular low risk of identifying a false positive polynucleotide is even further lowered. In particular, the identification of false positives that may potentially arise from (a) mutation(s) in the genome of the host cell(s) that confer(s) a growth advantage or disadvantage. Using replicates in the methods of the present invention offers the advantage of an improved statistic (using the above mentioned tests), which in turn allows identification of more polynucleotides encoding active biomolecules. An appropriate experimental design with respect to the number of replicates can, for example, be achieved with a statistical power analysis, which can be done along the same state of the art principles that are used for power analysis in determining frequency changes in naturally expressed RNA transcripts (Ching et al. 2014).

Importantly, the mentioned statistic tests can be not only employed if replicates are used. They may also be employed if only one sample is analyzed. The replicates are used to estimate the measuring error, biological and technical. For a single sample one would use a standardized measuring error, based on the standardized statistical distribution of errors, or obtained from previous similar experiments, or from using the range of differences for each polynucleotide between the measurements of the two time points as an error distribution function, or from using changes in frequency of neutral control polynucleotides (as specified above) to estimate an error distribution function. However, the power of detection of significant changes could potentially be

lower when no replicates are used. As mentioned elsewhere herein it is, thus, preferred to employ replicate samples in the context of the methods according to the present invention

It is preferred in the context of the methods of the invention that the cell cultivation/cell growth step is performed under optimal culturing conditions for the population of host cells to be employed. In other words, the culturing conditions may be non-selective, or at least as little selective as possible. Growth may only be limited by physiological parameters of the cells (and, for example, not by features of the culture medium). Growth may be at the maximum rate (in relation to host cells not comprising the polynucleotide to be identified). Notably, application of antibiotics or other substances selecting for presence of a marker gene, which may, for example, be used to select for presence of a vector comprising a polynucleotide of the polynucleotide library, is considered as non-selective or at least as little selective as possible. A person skilled in the art can readily select said optimal/non-selective culturing conditions based on his common general knowledge depending on the respective host cells used. In particular, standard culturing conditions known in the art for the respective host cell may be employed (optionally with adding a respective antibiotic to select for presence of a vector of the random polynucleotides; e.g. a vector comprising the same). For a number of host cells such culturing conditions are, for example reviewed in the eBook "Recombinant protein expression in microbial systems" (Rosano and Ceccarelli eds., Frontiers in Microbiology, 2014, DOI 10.3389/978-2-88919-294-6). In context of the host cells being *E. coli* cells, for example, an LB medium (in 1L $H_2O$: 10g Bacto-tryptone, 5g yeast extract, 10g NaCl, adjusted to pH7.5 with NaOH) or SOB medium (in 1L $H_2O$: 20g Bacto-tryptone, 5g yeast extract, 10mM NaCl, 2.5mM KCl, 10mM $MgCl_2$, 10 mM$MgSO_4$ adjusted to pH7.0 with NaOH) or SOC medium (same as SOB plus 20mM glucose) may be employed. Furthermore the *E. coli* may be cultivated at any temperature allowing for cell division. E.g., the *E.coli* cells may be cultivated at 16°C-42°C, but preferably at 37°C. If required for selecting for the presence of the random or partially random polynucleotides the respective medium (e.g. LB, SOB or SOC) may further comprise a respective antibiotic.

The cell cultivation/cell growth step may also be performed under suboptimal conditions for said population of host cells. However, it is preferred that at least non-toxic culturing conditions are to be employed. In one application, however, during

cultivating, conditions and/or agents which somehow inhibit the growth of the host cell population and/or which are partially toxic for said host cell population may be applied (less preferred). For example, this applies to selective agents which suppress cells which do not comprise a polynucleotide to be identified. In some embodiments the methods of the present invention may also be used to compare the spectra of polynucleotide or peptide changes between two host cell populations, of which one is pharmacologically or genetically perturbed, using the same general approach as it is done, for example, in shRNA library screening procedures (Sims et al. 2011) and as described elsewhere herein. For example one may split the host cell population containing the random polynucleotides into two groups, one that is grown under optimal culturing conditions, the other under the same conditions, but with a pharmacological substance of interest added. The polynucleotide or peptide changes may then be recorded for both cell populations after equivalent time points during cultivation. Differences in the profiles/fingerprints of polynucleotide frequencies (or corresponding frequencies of peptides encoded thereby) point then to polynucleotides or peptides that interact with processes that are targeted by the pharmacological substance. Instead of using a pharmacological substance, one could also genetically disrupt a specific gene in the second host cell population and would thus obtain polynucleotides or peptides that act within the same networks as the disrupted gene.

The biological activity of the identified biomolecules (e.g. peptides or RNAs) may, in principle, be limited to the host cells used for screening. This means that the biological activity of the biomolecules may, for example, be a biological activity in the host cells. However, it is envisaged in the context of the invention that also biomolecules are identified, the biological activity of which is not limited to particular host cells (e.g. the particular host cells used for screening). In one aspect, it is even preferred that the biological activity can be generally applied, e.g. to many different cell types derived from different species. In other words, the biomolecule may be biologically active in general, e.g. in many different cell types. This particularly applies in case the biological activity impacts cellular pathways and/or pertains to biomolecules that may be conserved throughout different host cell types and species. This can, for example, be assessed by performing the method(s) of the present invention in different host cell populations. In particular, the method(s) may also be performed with polynucleotide

libraries in which only polynucleotides identified in a different host cell population and one or more respective unchanged control polynucleotides are comprised.

The methods according to the present invention comprises determining the frequencies of polynucleotides in said library comprised in said population of host cells at a first time point during cultivation and at a subsequent second time point during cultivation. In other words this may also be expressed as follows: determining the frequencies of polynucleotides in said library comprised of said population of host cells at a first time point during cultivation and at a subsequent second time point during cultivation. The biomolecules encoded by polynucleotides of the library are expressed in said population between said first and said second time point. In other words, the method according to the present invention comprises determining the frequencies of polynucleotides in said library comprised in said population of host cells at a first time point during cultivation and at a second time point during cultivation, wherein said second time point is after said first time point, and wherein biomolecules encoded by the library are expressed in said population between said first and said second time point.

For example, and non-limiting, the population of host cell may pass between (about) 4 and (about) 35 more preferably between (about) 16 and (about) 25 cycles of cell divisions between the first and the second time point (see more detailed examples further below). In the appended non-limiting examples (about) 16 to (about) 25 cycles of cell divisions have passed between said first and said second time point during cultivation. A cycle of cell division means that the number of cells has doubled.

The first time point during cultivation can in principle be selected as any time point between the start of cultivation and the end of cultivation. In particular, the first time point can be a time point during culturing at which the biomolecules encoded by the library of polynucleotides are expressed or are not (yet) expressed.

The expression of the polynucleotides of the library of polynucleotides, i.e. the transcription into messenger RNA (mRNA) and optionally the translation of the transcribed mRNA into a peptide/polypeptide/protein, can, in principle, be driven by a constitutive promoter (i.e. a promoter that is active during the complete cultivation).

Such constitutive promoters for different host cells are well known in the art. In such case (but also in general in accordance with the invention), the first time point is preferably close to the start of culturing; i.e., for example, less than 1, less than 2, less than 3, less than 4, less than 5, less than 6, less than 7, or less than 8 cell division cycles (i.e. doubling of cell number) after start of culturing or transformation/transfection with the polynucleotide library. Similarly, in such case the first time point may be less than 0.5 h, less than 1h, less than 2h, less than 4h, less than 8h, less than 12h, or less than 24h after start of culturing or transformation/transfection with the polynucleotide library. Selecting the first time point at an early stage during culturing in the case of constitutive expression of the polynucleotides of the library has the advantage that potential advantages or disadvantages in proliferation or survival conferred by the polynucleotides encoding for biomolecules with biological activity do not yet dramatically alter the frequencies of different polynucleotides in the library. In particular, this also prevents that host cell clones comprising polynucleotides that inhibit cell proliferation or (in particular) survival get depleted completely from the population of host cells. Another advantage is that the changes in frequency between the first and the second time point can be higher as if the first time point is selected at a later time point during culturing.

As mentioned elsewhere herein, in the methods according to the present invention the polynucleotides of the library are preferably expressed under the control of an inducible promoter. This allows inducing the expression of the polynucleotides, i.e. the expression of the biomolecules encoded by the polynucleotides, at a defined time point during culturing. Thus, in one embodiment of the method according to the present invention the host cell population does not express the library of polynucleotides before said first time point. Before induction of expression the polynucleotide is not expressed or only mildly expressed. Mild expression can in some cases occur, e.g. due to a residual minimal activity of the inducible promoter in the absence of the trigger inducing the promoter activation. If the expression of the polynucleotide library is controlled by a inducible promoter system the first time point during culturing may, for example, be selected closely before the induction of the expression of the polynucleotides of the polynucleotide library, at the time point of inducing the expression of the polynucleotides of the polynucleotide library, or preferably shortly after inducing the expression of the polynucleotides of the polynucleotide library. An advantage of selecting the first time point only after the induction of the promotor is

that culturing conditions are identical between this time point and later time points. A further advantage can be that peptides with unspecific disruption of the cell physiology through forming aggregations are reduced or removed beforehand. On the other hand, selecting the first time point before inducing the promoter/expression of the polynucleotides of the library may have the advantage that potential advantages or disadvantages in proliferation or survival conferred by the polynucleotides encoding for biomolecules with biological activity do not yet dramatically alter the frequencies of different polynucleotides in the library. In particular, this may also prevent that host cell clones comprising polynucleotides that inhibit cell proliferation or (in particular) survival get depleted completely from the population of host cells.

The first time point during cultivation in the methods of the present invention can, however, in principle be any time point during cultivation before the second time point during cultivation. In particular, also a time point at which the polynucleotides of the library are expressed can be the first time point during cultivation.

In one aspect of the methods according to the present invention, the expression of the library of polynucleotides is induced closely before, closely after or preferably at said first time point. In particular, the expression of the library of polynucleotides can, for example, be induced by addition and/or depletion of one or more substances or by physical means. The substances or by physical means can in particular be substances or by physical means that activate an inducible promoter used for the expression of the polynucleotides of the library.

The second time point during cultivation in the context of the methods according to the present invention can in principle be any time point during culturing the host cell population at which the polynucleotides of the library are expressed (in any case, however, the second time point is after the first time point). In the context of the present invention culturing the host cells can mean maintaining cells *in vitro*. Preferably, the host cells are, however maintained under conditions allowing the proliferation of the population of host cells, i.e. the cells are maintained under conditions allowing for cell division of the host cells.

In principle, the host cell population may undergo any number of cell divisions between said first and said second time point during cultivation. In a preferred embodiment of the methods according to the present invention, the host cell population undergoes between (about) 1 and (about) 50, preferably between (about) 4 and (about) 35 and most preferably between (about) 16 and (about) 25 cycles of cell division between said first and said second time point. In particular, the host cell population may undergo (about) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49 or 50 cycles of cell division between said first and said second time point. In another aspect the host cell population may undergo less than 2, less than 3, less than 4, less than 5, less than 6, less than 7, less than 8, less than 9, less than 10, less than 11, less than 12, less than 13, less than 14, less than 15, less than 16, less than 17, less than 18, less than 19, less than 20, less than 21, less than 22, less than 23, less than 24, less than 25, less than 26, less than 27, less than 28, less than 29, less than 30, less than 31, less than 32, less than 33, less than 34, less than 35, less than 36, less than 37, less than 38, less than 39, less than 40, less than 41, less than 42, less than 43, less than 44, less than 45, less than 46, less than 47, less than 48, less than 49, or less than 50 cycles of cell division between said first and said second time point. It is particularly preferred that less than 25 cycles of cell division are between said first and said second time point. Using less than 25 cycles has the advantage that the chance of occurrence of secondary mutations in the polynucleotides of the library during cultivation are avoided. Either alone or in combination with this aspect, the host cell population may undergo at least 1, at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 21, at least 22, at least 23, at least 24, at least 25, at least 26, at least 27, at least 28, at least 29, at least 30, at least 31, at least 32, at least 33, at least 34, at least 35, at least 36, at least 37, at least 38, at least 39, at least 40, at least 41, at least 42, at least 43, at least 44, at least 45, at least 46, at least 47, at least 48, at least 49, or at least 50 cycles of cell division between said first and said second time point. The wording “the population of host cells undergoes [...] cycles of cell division between said first and said second time point” as used in the context of the present invention can alternatively be expressed as follows: the population of host cells has passed [...] cycles of cell division/cell doublings between said first and said second time point.

In principle, said second time point during cultivation can be 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49 or 50 days after said first time point during culturing. In particular, the time difference can be selected based on the host cell employed. For host cells with a short doubling time of e.g. 10 min to 3 hours or 10 min to 5h or 10 min (e.g. bacteria such as *E.coli*, or yeast cells) a shorter time difference between the first and second time point during cultivation is preferred. By contrast, if host cells that have longer doubling time are employed a longer time difference between the first time point and the second time point during cultivation is preferred. Most preferably the time difference is selected dependent on the doubling time of the respective host cells used in order to achieve between 1 and 50, preferably between 4 and 35 and most preferably between 16 and 25 cycles of cell division between said first and said second time point. The time difference can, for example, be calculated by multiplying the number of cell division cycles with the doubling time of the respective host cells. Doubling times of the host cells that can be employed in the context of the present invention are either well known in the art or can be tested for new cell types before setting up the experiment by methods that are well known in the art. Thus, for any host cells the time difference between the first and the second time point can be calculated and is disclosed herewith.

In the context of the present invention, the host cell population can, in principle, be in any kind of growth phase at the first time point during cultivation and/or said second time point during cultivation. In particular, the host cell population is in a logarithmic growth phase at said first and/or second time point. However, the host cell population can also be in a stationary growth phase at said second time point. Accordingly, in one embodiment the first time point during culturing is selected during logarithmic growth phase of the host cells and the second time point is selected during stationary growth phase of the host cells. This also includes scenarios in which the cells undergo several cycles of logarithmic growth phase and stationary growth phase; e.g. by diluting the culture in culturing media.

In the context of the methods according to the present invention the biomolecules encoded by the polynucleotides of the library of polynucleotides are expressed between the first and the second time point during cultivation. This means that the

biomolecules encoded by the library of polynucleotides are expressed at least at the second time point during cultivation. Thus, in other words, in the context of the present invention the biomolecules encoded by the polynucleotides of the library of polynucleotides are expressed at the second time point during cultivation. The wording "the biomolecules encoded by the library are expressed in said population between said first and said second time point" includes in principle any scenario in which the biomolecules are expressed for a defined time interval between said first and said second time point during cultivation. In principle, this time interval can only be a part of the time between the first and the second time point. Preferably, the time interval is however identical with the time difference between the first and the second time point during culturing. In one aspect, the biomolecules encoded by the library are expressed in said population between said first and said second time point, wherein said biomolecules are expressed at said first time point and said second time point. In another aspect, the biomolecules encoded by the library are expressed in said population between said first and said second time point, wherein said biomolecules are not expressed at said first time point.

The methods of the present invention involves determining the frequencies of polynucleotides in said library comprised in said population of host cells at a first time point during cultivation and at a subsequent second time point during cultivation. In principle, the frequency may also be determined at more than two time points during cultivation (e.g. at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, or at least 10 time points; preferably at least 4 time points). In any case the time difference (also relating to the definition by cell division cycles) between each of the time points is selected *mutatis mutandis* as specified for the first and the second time point. If more than two time points are determined, the change in frequency may be determined between any pair of two time points as described exemplary for the first and the second time point herein elsewhere.

The present invention also provides the polynucleotides as identified in accordance with the methods of the invention and the biomolecules (e.g. RNAs or polypeptides) encoded by these polynucleotides. Likewise, the present invention provides the biomolecules (e.g. RNAs or polypeptides) with biological activity and the polynucleotides encoding for these biomolecules. In particular, the present invention

provides novel or so far not known polynucleotides encoding for biomolecules (e.g. peptides or RNAs) with biological activity identified by the methods of the present invention. In particular, the present invention provides the polynucleotides and biomolecules (e.g. peptides) referred to herein elsewhere and, more particular, in Tables 3 and 4. Particular examples of the polynucleotides provided by the present invention are in particular also depicted in SEQ ID NOs: 69 to 128. Likewise, particular examples of the polypeptides (encoded by these polynucleotides) are depicted in SEQ ID NOs: 9 to 68. If not explicitly described in different manner herein elsewhere, the randomized part of the biomolecules being peptides (the randomized core section) starts at (about) amino acid position 5 and ends at (about) amino acid position 54 of the herein disclosed peptides, in particular of the peptide sequences as depicted in Tables 3 and 4 and/or in the SEQ ID NOs: 9 to 68. Similarly, if not explicitly described in different manner herein elsewhere, the randomized part (the randomized core section) of the polynucleotides encoding the biomolecules with biological activity starts at (about) nucleic acid position 13 and ends at (about) nucleic acid position 162 of the herein disclosed polynucleotides that encode for biomolecules (e.g. RNA or peptides) having biological activity, in particular of the polynucleotides that encode for biomolecules (e.g. RNA or peptides) having biological activity as depicted in Tables 3 and 4 and/or SEQ ID NOs: 69-128.

As mentioned elsewhere herein a biomolecule having a biological activity according to the present invention may also be an RNA. Thus, of course, the present invention also provides any of the RNAs encoded by the polynucleotides disclosed herein, and in particular encoded by the polynucleotide as depicted in Tables 3 and 4 and/or in SEQ ID NOs: 69 to 128. An "RNA encoded by a polynucleotide" of the present invention may alternatively also be referred to as an RNA "complementary to" or "transcribed from" a polynucleotide herein. As well known in the art, the sequences of the provided RNAs are identical to the DNA sequences shown in Tables 3 and 4 and/or in SEQ ID NOs: 69 to 128 with the only exception that the nucleotide thymine (T) is replaced by uracil (U). If not explicitly described in different manner herein elsewhere, the start and end of the randomized core section of the provided RNA molecules having biological activity provided herein are the same as indicated for the polynucleotide above.

The present invention is in particular also directed to a biomolecule having biological activity that comprises a biomolecule encoded by any one the polynucleotide sequences depicted in Tables 3 and 4 and/or in SEQ ID NOs: 69 to 128 (or variants of said biomolecule). Similarly, the present invention also relates to any one of the biomolecules (e.g. an RNA or a peptide) depicted in Tables 3 and 4 and/or in SEQ ID NOs: 9 to 68 (or variants thereof), or a biomolecule as encoded by any one the polynucleotide sequences depicted in Tables 3 and 4 and/or in SEQ ID NOs: 69 to 128 (or variants of said biomolecule). Moreover, the present invention also relates to a polynucleotide comprising any of the polynucleotides as depicted in Tables 3 and 4 and/or in SEQ ID NOs: 69 to 128 (or variants of said polynucleotide).

In particular, the present invention provides the peptides as depicted in Tables 3 and 4 (or variants thereof; see below) and/or in SEQ ID NOs: 69 to 128 (or variants thereof; see below). Similarly, the present invention provides the RNAs encoded by the polynucleotide sequences as depicted in Tables 3 and 4 (or variants thereof; see below) and/or in SEQ ID NOs: 69 to 128 (or variants thereof; see below).

The present invention also relates to polynucleotides, RNAs and peptides comprising, essentially consisting of or consisting of the randomized parts of the polynucleotides, RNAs and peptides, respectively (see, for example, the respective boundaries defined above) (or variants thereof). Accordingly, the present invention also relates to polynucleotides, RNAs and peptides comprising, essentially consisting of or consisting of the randomized parts of the polynucleotides (i.e. nucleotides 13 to 162 of the polynucleotides shown in Tables 3 and 4 and/or SEQ ID NOs: 69 to 128), RNAs (i.e. ribonucleotides 13 to 162 of the RNA encoded by the polynucleotides shown in Tables 3 and 4 and/or SEQ ID NOs: 69 to 128) and peptides (i.e. amino acids 5 to 54 as shown in Tables 3 and 4 and/or SEQ ID NOs: 9 to 68), respectively (or variants thereof). Preferred are those polynucleotides (or variants thereof) and biomolecules (or variants thereof) that have biological activity.

The present invention also relates to RNAs (or variants thereof) that comprise, essentially consist of or consist of an RNA sequence as encoded by any one of SEQ ID NOs: 69 to 128, having an RNA sequence encoded by a 5'-UTR (directly) fused to its 5'-end and/or an RNA sequence encoded by a 3'-UTR (directly) fused to its 3'-end.

Along this line, also any of the RNAs disclosed herein may further comprise or have a 5'-UTR (directly) fused to its 5'-end and/or an RNA sequence encoded by a 3'-UTR (directly) fused to its 3'-end. 5'-UTRs (5'-untranslated regions) and 3'-UTRs (3'-untranslated regions) are well known in the art. In the context of the present invention the RNA sequence encoded by a 5'-UTR is preferably the sequence as encoded by the RNA sequence shown in SEQ ID NO: 136. The RNA sequence encoded by a 3'-UTR is preferably the RNA sequence as encoded by the sequence as shown in SEQ ID NO: 137. A 3'-UTR may in the context of the present invention include a stop codon or may alternatively not include a stop codon. Specifically, the term preferably includes a stop codon when it is directly fused to a polynucleotide that does not comprise a stop codon at its 3'-end (as e.g. the polynucleotides depicted in SEQ ID NOs: 69 to 128).

An RNA as disclosed herein, may also be a m-RNA, preferably a bacterial m-RNA and most preferably an *E. coli* m-RNA. Accordingly, an RNA as disclosed herein may, for instance, comprise a 5'-cap or a polyA-tail.

Likewise, the present invention in one aspects also provides for any one of the polynucleotides (or variants thereof) as shown in Tables 3 and 4 and/or SEQ ID NOs: 69 to 128, wherein said polynucleotide further comprises a 5'-UTR (directly) fused to its 5'-end and/or a 3'UTR directly fused to its 3' end. Said 3'-UTR may preferably also comprise a stop codon at its 5'-end. A preferred 5'-UTR is the one shown in SEQ ID NO: 136 (or variants thereof). A preferred 3' UTR is the one shown in SEQ ID NO:137 (or variants thereof). Moreover, the present invention also provides for any one of the polynucleotides (or variants thereof) as shown in Tables 3 and 4 and/or SEQ ID NOs: 69 to 128, wherein said polynucleotide further comprises a stop codon directly fused to its 3'-end. The present invention also provides the biomolecules (or variants thereof) encoded by the polynucleotides defined in this paragraph.

Specifically, the present invention provides the members of a first set of biomolecules (e.g. novel RNAs and novel peptides) that promote cell proliferation, e.g. the biomolecules that comprise or consist of any of the biomolecule sequence as encoded by the DNA sequences shown in Table 3 and/or in SEQ ID NOs: 69 to 87 or in SEQ ID NOs: 69, 71 to 73, and 75 to 87. Further the present invention provides the biomolecules that comprise or consist of any of the biomolecule sequences as

depicted in Table 3 and/or in SEQ ID NOs: 9 to 27, or in SEQ ID NOs: 9, 11 to 13 and 15 to 27. Provided are also the polynucleotides encoding these biomolecules (e.g. polynucleotides comprising or consisting of any of the polynucleotide sequences as depicted in Table 3 and/or in SEQ ID NOs: 69 to 87 or in SEQ ID NOs: 69, 71 to 73, and 75 to 87). Said first set of biomolecules and/or polynucleotides encoding said biomolecules also includes biomolecules comprising or consisting of only the randomized part (see above) that promote cell proliferation. The polynucleotides encoding for these biomolecules comprising or consisting of only the randomized part (see above) are also provided herein. The biomolecules (e.g. RNAs and/or peptides) of said first set of biomolecules (or the polynucleotides encoding therefore) can particularly promote cell proliferation of *E. coli* cells but may also be able to promote cell proliferation of other cells.

Moreover, the present invention also provides the members of a second set of biomolecules (e.g. novel RNAs and/or peptides) that are inhibiting cell proliferation, e.g. the biomolecules that are encoded by the DNA sequences shown in Table 4 and/or SEQ ID NOs: 88 to 128, or the biomolecules as depicted in Table 4 and/or the biomolecules that comprise or consist of any of the biomolecule sequences as depicted in SEQ ID NOs: 28 to 68 or encoded by a polynucleotide as depicted in any one of SEQ ID NOs: 88 to 128. Moreover, also polynucleotides encoding these biomolecules are provided (e.g. as depicted in Table 4 and/or SEQ ID NOs: 88 to 128). Said second set of biomolecules and/or polynucleotides encoding said biomolecules also includes biomolecules comprising or consisting of only the randomized part (see above). Also the polynucleotides encoding for these biomolecules comprising or consisting of only the randomized part (see above) are provided herein. The second set of biomolecules (or the polynucleotides encoding therefore) can particularly inhibit cell proliferation of *E. coli* cells but may also be able to inhibit cell proliferation of other cells.

In a specific embodiment, the present invention provides an RNA (having biological activity), wherein said RNA comprises or consists of SEQ ID NO: 130 (or variants thereof), preferably SEQ ID NO: 131 (or variants thereof), or most preferably of SEQ ID NO: 132 (or variants thereof). Likewise, the present invention relates to an RNA (having biological activity) that comprises or consists of RNA encoded by the

polynucleotide sequence depicted in SEQ ID NO: 70 (or variants thereof). In a preferred embodiment RNA (having biological activity) that comprises or consists of an RNA sequence encoded by the randomized part of the RNA depicted in SEQ ID NO: 130 (or variants thereof), i.e. ribonucleotides 13 to 162 of SEQ ID NO: 70 (or variants thereof) is provided. The RNAs as defined in this paragraph may further comprise or have an RNA sequence encoded by a 5'-UTR (directly) fused to its 5'-end and/or an RNA sequence encoded by a 3'-UTR (directly) fused to its 3'-end. In the context of the present invention the RNA sequence encoded by a 5'-UTR is preferably the sequence as encoded by the sequence shown in SEQ ID NO: 136 (or variants thereof). The RNA sequence encoded by a 3'-UTR is preferably the sequence as encoded by the sequence shown in SEQ ID NO: 137 (or variants thereof). As suggested by the appended examples, the RNA as defined in this paragraph has biological activity. More specifically, it promotes cell growth in *E. coli* cells. The RNA may, however, also promote cell growth of other (host) cells. Accordingly, the present invention also provides for the use of any of the RNAs mentioned in this paragraph for promoting cell growth/proliferation, preferably *E. coli* cell growth/proliferation.

In a specific embodiment, the present invention provides an RNA (having biological activity), wherein said RNA comprises or consists of SEQ ID NO: 133 (or variants thereof), preferably SEQ ID NO: 134 (or variants thereof), or most preferably of SEQ ID NO: 135 (or variants thereof). Likewise, the present invention relates to an RNA (having biological activity) that comprises or consists of RNA encoded by the polynucleotide sequence depicted in SEQ ID NO: 74 (or variants thereof). In a preferred embodiment RNA (having biological activity) that comprises or consists of an RNA sequence encoded by the randomized part of the RNA depicted in SEQ ID NO: 133 (or variants thereof), i.e. ribonucleotides 13 to 162 of SEQ ID NO: 74 (or variants thereof) is provided. The RNAs as defined in this paragraph may further comprise or have an RNA sequence encoded by a 5'-UTR (directly) fused to its 5'-end and/or an RNA sequence encoded by a 3'-UTR (directly) fused to its 3'-end. In the context of the present invention the RNA sequence encoded by a 5'-UTR is preferably the sequence as encoded by the sequence shown in SEQ ID NO: 136 (or variants thereof). The RNA sequence encoded by a 3'-UTR is preferably the sequence as encoded by the sequence shown in SEQ ID NO: 137 (or variants thereof). As suggested by the appended examples, the RNA as defined in this paragraph has biological activity. More

specifically, it promotes cell growth in *E. coli* cells. The RNA may, however, also promote cell growth of other (host) cells. Accordingly, the present invention also provides for the use of any of the RNAs mentioned in this paragraph for promoting cell growth/proliferation, preferably *E. coli* cell growth/proliferation.

The present invention further relates to a peptide (having biological activity), wherein said peptide comprises or consists of SEQ ID NO: 24 (or variants thereof), i.e. the amino acid sequence encoded by SEQ ID NO: 84 (or variants thereof). In a preferred embodiment a peptide (having biological activity) that comprises or consist of the randomized part of the peptide sequence depicted in SEQ ID NO: 24 (or variants thereof), i.e. amino acids 5 to 54 of SEQ ID NO: 84 (or variants thereof) is provided. Likewise, the present invention relates to a peptide having biological activity that comprises or consists of a peptide (or variants thereof) encoded by the polynucleotide sequence depicted in SEQ ID NO: 24. In a preferred embodiment, a peptide having biological activity that comprises or consists of a peptide encoded by the randomized part of the nucleic acid depicted in SEQ ID NO: 24, i.e. nucleotides 13 to 162 of SEQ ID NO: 24 (or variants thereof) is provided. As suggested by the appended examples, the peptide as defined in this paragraph has biological activity. More specifically, it promotes cell growth in *E. coli* cells. The peptide may, however, also promote cell growth of other (host) cells. Accordingly, the present invention also provides for the use of any of the peptides mentioned in this paragraph for promoting cell proliferation, preferably *E. coli* cell proliferation.

The present invention also relates to a biomolecule (e.g. an RNA or peptide) comprising, or consisting of a biomolecule sequence (or variants thereof) as encoded by a polynucleotide (or the randomized part thereof) as depicted in SEQ ID NO: 98, 107 or 118, wherein said biomolecule has growth inhibiting activity (e.g. in E.coli). The present invention also relates to a biomolecule (or variants thereof) comprising, or consisting of the amino acid sequence (or variants thereof) as depicted in SEQ ID NO: 38, 47 or 58 (or the randomized parts thereof). The biomolecule as described in this paragraph has been employed e.g. in Example 4 of the present invention and has been shown to have cell growth inhibiting activity, e.g. in E.coli.

Also variants of polynucleotides as identified in accordance with the methods of the invention (e.g. the polynucleotides as depicted in SEQ ID NOs: 69 to 128 or the randomized part thereof between nucleotides 13 to 162) are provided and likewise provided are variants of the encoded biomolecules (e.g. see Tables 3 and 4 and/or SEQ ID NOs: 9 to 68). In particular, variants of the biomolecules (e.g. peptides) referred to herein elsewhere and, more particular, in Tables 3 and 4 referred to herein below and/or SEQ ID NOs: 9 to 68 are provided. More particular, variants of the randomized part of the biomolecules (e.g. RNAs or peptides) (the randomized core section; starting at (about) amino acid position 5 or ribonucleotide position 13 and ending at (about) amino acid position 54 or ribonucleotide position 162) as referred to herein elsewhere and, even more particular, in Tables 3 and 4 and/or SEQ ID NOs: 9 to 68 herein below are provided.

"Variant" in accordance with the invention particularly means that the respective biomolecule (e.g. RNA or peptide) has the biological activity in accordance with the invention (e.g. growth enhancing or decreasing activity and the like; see herein elsewhere) and that the variant polynucleotide encodes for such a biologically active biomolecule, respectively. For example, a variant polynucleotide or biomolecule in accordance with the invention is envisaged to share an identity (in particular sequence identity) of at least 60%, preferably at least 70%, preferably at least 80%, preferably at least 90%, preferably at least 95%, preferably at least 98% and even more preferably at least 99% identity (for example based on the number of nucleotides or amino acids comprised in the sequence, respectively) with a reference polynucleotide or biomolecule (e.g. the polynucleotide as described in Table 3 and 4 herein below and/or in SEQ ID NOs: 69 to 128 or the biomolecule as described in Table 3 and 4 herein below and/or in SEQ ID NOs: 9 to 68).

Those having skill in the art will know how to determine percent identity between/among sequences using, for example, algorithms such as those based on CLUSTALW computer program (Thompson (1994) Nucl. Acids Res. 2:4673-4680), CLUSTAL Omega (Sievers (2014) Curr. Protoc. Bioinformatics 48:3.13.1-3.13.16) or FASTDB (Brutlag (1990) Comp App Biosci 6: 237-245). Also available to those having skill in this art are the BLAST, which stands for Basic Local Alignment Search Tool, and BLAST 2.0 algorithms (Altschul, (1997) Nucl. Acids Res. 25:3389-3402; Altschul

(1990) J. Mol. Biol. 215:403-410). The BLASTN program for nucleic acid sequences uses as defaults a word length (W) of 11, an expectation (E) of 10, M=5, N=4, and a comparison of both strands. The BLOSUM62 scoring matrix (Henikoff (1992) Proc. Natl. Acad. Sci. U.S.A. 89:10915-10919) uses alignments (B) of 50, expectation (E) of 10, M=5, N=4, and a comparison of both strands.

In order to determine whether a nucleotide residue or a amino acid residue in a given nucleotide sequence or peptide sequence, respectively, corresponds to a certain position compared to another nucleotide sequence (e.g. one of the sequences shown in Tables 3 and/or 4) or peptide sequence (e.g. one of the sequences shown in Tables 3 and/or 4), respectively, the skilled person can use means and methods well known in the art, e.g., alignments, either manually or by using computer programs such as those mentioned herein. For example, BLAST 2.0 can be used to search for local sequence alignments. BLAST or BLAST 2.0, as discussed above, produces alignments of nucleotide sequences to determine sequence similarity. Because of the local nature of the alignments, BLAST or BLAST 2.0 is especially useful in determining exact matches or in identifying similar or identical sequences.

Another example of a variant polynucleotide in accordance with the invention is a polynucleotide that comprises or consists of a nucleic acid molecule hybridizing under stringent conditions to the complementary strand of a nucleic acid molecule (e.g. as depicted in Table 3 or 4, infra, and/or in SEQ ID NOs: 69 to 128) encoding a biomolecule of the invention (e.g. as depicted in Table 3 or 4, infra, and/or in SEQ ID NOs: 9 to 68). Also a biomolecule (e.g. protein or RNA) which is encoded by such as a polynucleotide is provided.

In the context of the present invention, "hybridizing" means that hybridization can occur between one nucleic acid molecule and another (complementary) nucleic acid molecule. Hybridization of two nucleic acid molecules usually occurs under conventional hybridization conditions. In the context of the invention, stringent hybridization conditions are preferred. Hybridization conditions are, for instance, described in Sambrook and Russell (2001), Molecular Cloning: A Laboratory Manual, CSH Press, Cold Spring Harbor, NY, USA.

As to a biomolecule being a peptide or protein it is, for example, envisaged that a variant is or comprises the amino acid sequence of the reference peptide or protein

having (about) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20 or even more amino acids inserted, deleted or substituted by a different amino acid (preferred is a conservative substitution). Any of the above-mentioned particular biomolecules may be a reference peptide or protein. A variant of a polynucleotide includes, in particular, also any polynucleotide that encodes for the biomolecule (e.g. peptide) encoded by the identified polynucleotide or a biomolecule variant (e.g. peptide variant) as defined herein. In particular, polynucleotides that are codon optimized (in order to ensure proper expression of a corresponding peptide) and/or that are different due to the degeneracy of the genetic code are included in the term "variant". As mentioned, any of the variants described herein is envisaged to have the biological activity in accordance with the invention (e.g. growth enhancing or decreasing activity and the like; see herein elsewhere).

A "variant" in accordance with the invention also encompasses a fragment of the polynucleotide (e.g. as depicted in Table 3 or 4 and/or in SEQ ID NOs: 69 to 128) to be identified or of the encoded biomolecule (e.g. as depicted in Table 3 or 4 and/or in SEQ ID NOs: 9 to 68). As to a polynucleotide, a fragment may be a nucleic acid sequence stretch of at least 30, at least 50, at least 70, at least 80, at least 90, at least 100, at least 110, at least 120, at least 130, at least 140 or at least 150 nucleotides. The same applies to a biomolecule being an RNA. As to a biomolecule, a fragment may be an amino acid stretch of at least 10, at least 20, at least 30, at least 40, at least 45, at least 46, at least 47, at least 48 or at least 49 amino acid residues. Most preferably, the fragment of the polynucleotide encodes for an amino acid sequence which exhibits biological activity in accordance with the invention and the fragment of the biomolecule exhibits biological activity in accordance with the invention, respectively.

A skilled person is well aware of methods to determine whether a biomolecule (e.g. an RNA or peptide), a variant thereof, or a fragment thereof has biological activity in the meaning of the present invention. For instance, to test for biological activity a skilled person may compare the growth of cells expressing said biomolecule ), a variant thereof, or a fragment thereof (i.e. cells comprising a corresponding polynucleotide encoding for said biomolecule, variant thereof, or a fragment thereof) with cells comprising a control polynucleotide (without biological activity or with a known

biological activity). Thereby, a skilled person can assess whether the growth of cells expressing said biomolecule a variant thereof, or fragment thereof has biological activity, e.g. promoting or inhibiting cell growth. If the biological activity of variants or fragments is assessed also comparative experiments using the respective "wild type" biomolecule may be performed.

To assess whether the biological activity of the biomolecule of the present invention results from a protein or an RNA (encoded by a respective polynucleotide), a method as described above as "a step to assess whether the biomolecule having biological activity is the RNA or the polypeptide encoded by a respective identified polynucleotide" in the context of the methods of the present inventions may be conducted. Examples for such a method are provided in the appended examples (in particular in example 4).

The meaning of the term "biomolecule" is known in the art and "biomolecule" is used herein accordingly. In particular, the term "biomolecule" as used herein refers to any molecule that can be encoded/expressed by a polynucleotide sequence, in particular by a polynucleotide as employed in the context of the present invention. A particular biomolecule to be employed in accordance with the invention is RNA and, preferably, a protein or a peptide. However, also biomolecules like DNA or PNA (peptide nucleic acid) may, in principle, be employed. A peptide nucleic acid (PNA) is a polyamide type of DNA analog. The biomolecule may be naturally occurring, synthetic or semisynthetic or it may be a derivative, such as a PNA (Nielsen (1991), Science 254, 1497-1500) or a phosphorothioate.

In one aspect, the methods of the invention further comprise (a step of) determining the structure of the identified biomolecule, in particular the amino acid sequence of the peptide encoded by the identified polynucleotide or the ribonucleic acid sequence of the RNA encoded by the identified polynucleotide. Means and methods for identifying the structure in accordance with this aspect are known in the art and are, for example, crystallography, NMR, electron microscopy, DNA or RNA sequencing (e.g. method according to Maxam und Gilbert, method according to Sanger, pyrosequencing), polypeptide sequencing (e.g. Edman degradation) and the like.

In the context of the present invention "a population of host cells" refers to a plurality of (identical) host cells of a defined cell type. A host cell population may preferably at least comprise a number of host cells that is at least as high as or higher (e.g. at least 2-times more, at least 3-times more, at least 4-times more, at least 5-times more, at least 10-times more or at least 20-times more) than the number of polynucleotides of the library of polynucleotides. In particular, the number of host cells in the population of host cells is preferably high enough to ensure that each of the polynucleotides of the polynucleotide library is at least present in one host cell, but preferably in more host cells to allow repeated sampling of subsets.

The term "a population of host cells capable of expressing a library of polynucleotides", as used herein, includes but is not limited to a scenario in which all cells of that population of host cells are capable of expressing one or more (most preferably one) of the polynucleotides of the library. This term, in particular, also includes that not all host cells of said population are capable of expressing the library of polynucleotides but that said population of host cells as a whole is capable of expressing a library of polynucleotides; i.e. that the population comprises for each polynucleotide of the library at least one cell capable of expressing the same. For example, the population of host cells may in particular also comprise host cells that have not been successfully transformed with one of the polynucleotides of the library. In other words, this term also includes and/or may mean that a population of host cells comprises (a first population) of host cells capable of expressing a library of polynucleotides (e.g. host cells transformed with a vector comprising a polynucleotide of the library and not transformed cells with a vector comprising a polynucleotide of the library). Preferably, each of the host cells capable of expressing a library of polynucleotides as comprised in the population of host cells comprises one polynucleotide of the library and or is capable of expressing the same. As mentioned also elsewhere, the term "a population of host cells capable of expressing a library of polynucleotides" may also include and/or mean that a population of host cells comprises cells capable of expressing a library of polynucleotides and comprises host cells comprising a control polynucleotide.

Where it is referred herein to a population of host cells capable of expressing a library of polynucleotides and comprising one or more control polynucleotide(s), this means that the population of host cells comprises a first subpopulation of host cells capable of

expressing a library of polynucleotides (preferably only one of the polynculeotides per cell) and a second subpopulation of cells that comprises cells expressing said one or more of said control polynucleotides (preferably only one per cell). Most preferably, said first and said second subpopulation are not identical.

The expression "the biomolecules encoded by the polynucleotides of said library (of polynucleotides) are expressed" may mean that the biomolecules are physically present and/or detectable in the population of host cells. However, this expression particularly also includes that some (e.g. at least 10%, 20%, 30%, 40%, 50%, 60%, 70%, or 80%) of the biomolecules are not physically present and or detectable. This may in particular be the result of highly instable biomolecules that may be encoded by some of the random or partially random polynucleotides of the library of polynucleotides. The term should thus be understood as providing conditions that in principle allow for expression of the biomolecules encoded by the polynucleotides of the library (e.g. inducing the transcription from an inducible promoter).

The "host cell" or "host cells" employed in the context of the present invention can in principle be any prokaryotic or eukaryotic cell(s) that can be engineered to express exogenous polynucleotides. The term "prokaryotic" is meant to include all bacteria and archaea, which can be transformed or transfected with polynucleotides. The term "eukaryotic" is meant to include in particular yeast, algae, higher plant, insect and mammalian cells. In the context of the present invention in particular prokaryotic and eukaryotic host cells are employed that are well known to be easily genetically modifiable and/or suitable for recombinant protein expression. For example, any host cell or cell type that is described in Yin et al. (Yin et.al., 2007, Journal of Biotechnology 127,335-347) can be employed as host cell in the context of the present invention. Prokaryotic host cells employed in the context of the present invention can, for example, be selected from the group consisting of *Escherichia coli, Bacillus subtilis, Caulobacter crescentus, Mycoplasma genitalium, Alvibrio fischeri, Ralstonia eutropha* (formerly known as *Alcaligenes eutrophus*), *Synecocystis* and *Pseudomonas*-based systems (e.g. the systems developed by DOW Chemical company) or any cultivable pathogenic strain for which novel ways of growth suppression (antibiotics) are to be developed. In the context of the present invention eukaryotic host cells can, for example, be any yeast cells (e.g. *Saccharomyces cerevisiae, Schizosaccharomyces*

*pombe, Pichia pastoris, Arxula adeninivorans, Hansenula polymorpha, Kluyveromyces lactis* or *Yarrowia lipolytica*), any filamentous fungus (e.g. a member of the *Aspergillus* genus such as *Aspergillus niger*, *Aspergillus oryzae* or *Aspergillus nidulans, Cunninghamella elegans, Neurospora crassa or Ustilago maydis*), any protist or algae (e.g. *Chlamydomonas reinhardtii, Dictyostelium discoideum, Tetrahymena thermophila, Emiliania huxleyi, Thalassiosira pseudonana)* any insect cells that can be cultured (e.g. any cell lines derived from the Drosophila origin such as S2 and Kc cells (Cherbas, L. and Gong, L. 2014. Cell lines. Methods 68: 74-81.) and any mammalian cells that can be cultured. Mammalian cells may be, for example, be any cells derived from rodents (rats, mice, guinea pigs, or hamsters) such as CHO, BHK, NSO, SP2/0, YB2/0 cells, cells derived from other mammals (e.g. COS cells, mouse L-cells), cells derived from human tissues (e.g. human embryonic kidney (HEK) 293 cells; HELA cells, myeloma cell lines such as J558L or Sp2/0 cells etc.), or cancer cell lines derived from the Cancer Cell Line Encyclopedia (Barretina et al. 2012). In one embodiment, also primary eukaryotic cells or induced pluripotent stem cells and their differentiated derivatives can be employed.

Preferably, in the context of the present invention cells having a high rate of cell division are employed. By doing so the culturing time between the first and the second time point at which the frequencies of the polynucleotides in the library are determined can be shortened. In a preferred embodiment of the present invention the host cell population is a population of *E. coli* cells (e.g. *E. coli* DH10B).

The host cells in accordance with the invention may be cells which are growing in solution or are adherent to a surface. Respective means and methods are known in the art and are exemplified herein.

In one embodiment a host cell may be a primary cell (preferably a primary mammalian cell) that can only be cultured for a limited time period (e.g. *in vitro*), or in other words have a time-limited viability and/or proliferation capability in (e.g. *in vitro*) culturing. A limited time period or time-limited may mean less than 1 week, preferably less than 2 week, more preferably less than 3 weeks, even more preferably less than 4 weeks, even more preferably less than 6 weeks or most preferably less than 12 weeks culturing. In the context of this embodiment the biomolecules identified to enhance cell survival and growth by the methods of the present invention can be used to enhance the time the cells can be cultured *in* vitro, preferably even to immortalize these primary

cells. This has the advantage of better and longer availability of these cells (e.g. for research purposes) without the continuous need of isolating fresh cells. Accordingly, the present invention also relates to a method for identifying polynucleotides and/or biomolecules encoded thereby for prolonging (e.g. *in vitro*) cell survival/proliferation of primary cells (preferably primary mammalian cells) and/or immortalizing primary cells (preferably primary mammalian cells) for (e.g. *in vitro*) culturing, wherein said method comprises the steps of any of the methods as described herein elsewhere. Similarly, the present invention relates to the use of the identified polynucleotides (e.g. as provided herein) encoding biomolecules with biological activities and the corresponding biomolecules (e.g. RNAs and/or peptides) for prolonging (e.g. *in vitro*) cell survival and/or increasing the proliferation potential of primary cells (preferably primary mammalian cells) and/or immortalizing primary cells (preferably primary mammalian cells). The present invention also relates to the so identified polynucleotides and biomolecules.

In principle, it is preferred in accordance with the invention that each, or essentially each, of the host cells of the host cell population comprises, is capable of expressing, or expresses one of the polynucleotides of the library. It is even more preferred that each, or essentially each, of the host cells of the host cell population comprises, is capable of expressing, or expresses exactly one of the polynucleotides of said library. In one aspect it is also envisaged that each, or essentially each, of the host cells of the first subpopulation of host cells comprised in the population of host cells comprises, is capable of expressing, or expresses one of the polynucleotides of the library. "Essentially each", for example, mean that at least 40%, preferably at least 50%, preferably at least 60%, preferably at least 70%, preferably at least 80%, preferably at least 85%, preferably at least 90%, preferably at least 95%, preferably at least 98%, preferably at least 99% and most preferably about 100% of the host cells comprises, is capable of expressing, or expresses one, preferably exactly one, of the polynucleotides of the library. In the appended examples, for example about 60% of the host cells comprises, is capable of expressing, or expresses one of the polynucleotides of the library. Referring to "one polynucleotide" in this paragraph means in particular that only one polynucleotide sequence is comprised in the respective host cells. It is, however also particularly envisaged that this one

polynucleotide sequence per cell is present in more than one copy. But is, in principle, also not excluded that more than one polynucleotide may be expressed in a single cell.

In one embodiment the present invention also provides for a population of host cells (as specified above and elsewhere herein) being transformed with a library of polynucleotides (as specified elsewhere herein) comprising a random or almost random nucleic acid sequence (e.g. in a vector such as an expression vector). In particular, the present invention relates to a population of *E. coli* cells (e.g. in form of a glycerol stock that can be stored at -80°C; e.g. comprising 20% glycerol) comprising said polynucleotides.

The general meanings of the terms "polynucleotide(s)", "polynucleotide sequence(s)", "nucleotide sequence(s)" and "nucleic acid sequence(s)" and the like are well known in the art and are used accordingly in context of the present invention.
For example, when used throughout this invention, these terms refer to all forms of naturally occurring or recombinantly generated types of polynucleotides and/or polynucleotide sequences/molecules as well as to chemically synthesized nucleotide sequences and/or nucleic acid sequences/molecules. In particular, these terms refer to deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). However, DNA is the preferred form of the polynucleotides to be identified in accordance with the methods of the invention. The polynucleotides may be made by synthetic chemical methodology known to one of ordinary skill in the art, or by the use of recombinant technology, or by a combination thereof. The DNA and RNA may optionally comprise unnatural nucleotides and may, in principle, be single or double stranded. In general, "polynucleotide" also refers to sense and anti-sense DNA and RNA, that is, a nucleotide sequence which is complementary to a specific sequence of nucleotides in DNA and/or RNA.
Furthermore, these terms and, in particular, the term "polynucleotide(s) may refer to DNA or RNA or hybrids thereof or any modification thereof that is known in the state of the art (see, e.g., US 5525711, US 4711955, US 5792608 or EP 302175 for examples of modifications). In principle, a polynucleotide of the invention may be single- or double-stranded, linear or circular, natural or synthetic, and/or, in principle, without any size limitation. For instance, the "polynucleotide(s)" may be genomic DNA, cDNA, mRNA, antisense RNA, ribozymal or a DNA encoding such RNAs or chimeroplasts

(Cole-Strauss Science 1996 273(5280) 1386-9). They may be in the form of a vector/plasmid or of viral DNA or RNA. "Polynucleotide(s)" may also refer to (an) oligonucleotide(s), wherein any of the state of the art modifications such as phosphothioates or peptide nucleic acids (PNA) are included. However, as mentioned above, DNA is the particularly preferred form of the polynucleotide to be identified in accordance with the invention.

In the context of the invention, (a) the polynucleotide(s) to be employed (for example DNA fragments encoding the biomolecule) may be cloned into or are comprised in a vector, in particular in an expression vector. The subsequent recombinant expression of the encoded polynucleotide(s) can be achieved by a routine procedure in molecular genetics and gene technology. Various commercial solutions are available for this purpose, including expression in prokaryotic and eukaryotic host cells. In general, expression vectors containing promoter sequences which facilitate the efficient transcription of the inserted polynucleotide are used in connection with the host. The expression vector typically contains an origin of replication, a promoter, and a terminator, as well as specific genes which are capable of providing phenotypic selection of the transformed cells.

A vector may be employed for the purpose of cloning. The vector may be a cloning vector and, in particular, an expression vector. For example, the vector may be a phage, plasmid (preferred), viral or retroviral vector. Retroviral vectors may be replication competent or replication defective. In the latter case, viral propagation generally will occur only in complementing host/cells. The herein provided nucleic acid molecule may be joined to a particular vector containing selectable markers for propagation in a host. Generally, a plasmid vector is introduced in a precipitate, such as a calcium phosphate precipitate or rubidium chloride precipitate, or in a complex with a charged lipid or in carbon-based clusters, such as fullerens. Should the vector be a virus, it may be packaged in vitro using an appropriate packaging cell line prior to application to host cells. The vector which comprises the polynucleotide in accordance with the invention may be (a vector that is capable to be) integrated into the genome of the host cell. In particular, the vector may be propagated or capable of being propagated in prokaryotic and/or eukaryotic host cells and/or may give rise to the expression of the polynucleotides of the library in prokaryotic and/or eukaryotic host cells. Suitable vectors can be chosen and ready obtained by the skilled person.

Particularly suitable expression vectors and (host) cell systems are described in the review by Yin et al. (2007), both for prokaryotes and eukaryotes. Some eukaryotic cell systems, such as yeast, insect cells or CHO SES cells allow using serum free medium for growth (Yin et al. 2007). This may further ensure that unknown peptide components from the serum cannot interfere with the screen. As shown in the appended examples, the pFLAG-CTC™ is a suitable expression vector that may be employed in *E. coli* cells.

For example, an expression system in eukaryotic cell cultures may be build on the same principles as it is done for shRNA library screens (Sims et al. 2011). For these, the initial library of polynucleotides may be cloned/inserted in vectors/plasmids, which may be packaged into lentiviral vectors by transfection of the plasmid DNA into eukaryotic packaging cells. Infectious virus particles may then be obtained from the supernatant, titered and then may be used at defined infection ratios in cells that express the cloned insert. Similar, as in the competitive shRNA screen (Sims et al. 2011), one may strive to keep cells in log growth, i.e. not more than 70% confluence before re-plating.

To identify polynucleotides or peptides with an activity spectrum in a broader range of organisms, one may use expression vectors that are designed to allow expression in different, i.e. multiple, species, in particular prokaryotic and eukaryotic ones. This enables to test the same library in different organismic backgrounds. Without being bound by theory, clones that show comparable fitness effects in different/multiple cells are likely to express polynucleotides or peptides that target conserved parts of metabolic or regulatory pathways.

Preferably, the polynucleotide to be employed in the context of the invention is operatively linked to expression control sequences (e.g. within the herein disclosed vector). These control sequences allow expression in prokaryotic or eukaryotic cells or isolated fractions thereof. Expression of said polynucleotide comprises transcription of the nucleic acid molecule, preferably into a translatable mRNA. Regulatory elements ensuring expression in eukaryotic cells, preferably mammalian cells, are well known to those skilled in the art. They usually comprise regulatory sequences ensuring initiation of transcription and optionally poly-A signals ensuring termination of transcription and stabilization of the transcript. Additional regulatory elements may include

transcriptional as well as translational enhancers. Possible regulatory elements permitting expression in prokaryotic host cells comprise, e.g., the *lac*, *trp* or *tac* promoter in *E. coli*, and examples for regulatory elements permitting expression in eukaryotic host cells are the *AOX1* or *GAL1* promoter in yeast or the CMV-, SV40-, RSV-promoter (Rous sarcoma virus), CMV-enhancer, SV40-enhancer or a globin intron in mammalian and other animal cells. Beside elements which are responsible for the initiation of transcription such regulatory elements may also comprise transcription termination signals, such as the SV40-poly-A site or the tk-poly-A site, downstream of the polynucleotide. In this context, suitable expression vectors are known in the art such as Okayama-Berg cDNA expression vector pcDV1 (Pharmacia), pCDM8, pRc/CMV, pcDNA1, pcDNA3 (Invitrogen), pSPORT1 (GIBCO BRL). The vector may be a gene transfer vector. Expression vectors derived from viruses such as retroviruses, adenoviruses, vaccinia virus, adeno-associated virus, herpes viruses, or bovine papilloma virus, may be used for delivery of the polynucleotides or vector of the invention into a targeted cell population. Methods which are well known to those skilled in the art can be used to construct a vector, in particular a vector comprising the control sequences, in accordance with this invention; see, for example, the techniques described in Sambrook, Molecular Cloning A Laboratory Manual, Cold Spring Harbor Laboratory (1989) N.Y. and Ausubel, Current Protocols in Molecular Biology, Green Publishing Associates and Wiley Interscience, N.Y. (1994). Alternatively, the polynucleotides and vectors to be employed in accordance with the invention can be reconstituted into liposomes for delivery to target cells.

In one aspect, the polynucleotides to be employed in accordance with the invention may comprise or may be positioned next to a transcription start site and a transcription termination signal; and/or a translation start site and translation termination site. As such, they may be comprised in the (expression) vector in accordance with the invention.

In a specific aspect, the polynucleotides to be identified may be operatively linked to a promoter sequence, wherein said promoter regulates the expression of said polynucleotides. It is preferred that the promoter is an inducible promoter. The (inducible) promoter may be a promoter that is capable of being activated. Activation may, for example, be achieved by addition and/or depletion of one or more substances or by physical means. Inducible promoters for prokaryotic and eukaryotic host cells

and means and method to induce those, e.g. be adding substances or by physical means are described in Yin et al. (2007). Respective examples for inducible promoters and promoters that can be activated, which can be employed in the context of the present invention are well known in the art and are, for example, described in Yin et. al. (2007). If bacterial cells, in particular *E. coli* cells are used as host cells, for example, a Ptac promoter (a hybrid of the trp and lac promoters from *E. coli*) may be employed. This is also illustrated in the appended examples.

In particular, it is envisaged in the context of the invention that the inducible promoter is activated at the first time point (or slightly before or after that time point, e.g. about 1, 2, 3, 4 or 5 cell division cycles before or after the time point, wherein the lower numbers of cell division cycles are preferred).

In a specific aspect, the expression vectors for polynucleotides/biomolecules may be designed such that they include the co-synthesis of defined biomolecule (e.g. peptide) fragments as tags. This can ease the subsequent functional analysis of identified biomolecules (e.g. bioactive peptides), for example by providing an epitope that is recognized by antibodies and that can be used for co-immunoprecipitation of molecular complexes to which the new peptide binds. A preferred tag in the context of the present invention is the FLAG tag. Other types of tags, such as a GFP domain, could be used to study intracellular localisation (Davis, 2004).
Furthermore, the (random) biomolecules (e.g. peptides) may be targeted to specific compartments of the cell, e.g. by co-synthesis of a targeting domain. In this context, the polynucleotides to be employed in accordance with the invention may further comprise a nucleic acid sequence that encodes a peptide and/or RNA targeting the expressed polypeptide and/or RNA to one or more cellular compartments. This includes, for example, extracellular secretion signals, transmembrane domains, nuclear localisation signals or targeting signals for mitochondria or plastids. Also, the addition of domains that target the peptide to specific DNA or RNA sequences (i.e. specific DNA or RNA binding domains) or other macromolecular complexes is possible. The use of such targeting vectors would allow to focus on the identification of new biomolecules (e.g. bioactive peptides) to specific cellular processes.

The methods of the present invention may comprise a step of generating the host cell population capable of expressing the library of polynucleotides prior to culturing. Said generating may comprises transforming said host cells with expression vectors for said library of polynucleotides, for example with the (particular) expression vectors as described herein elsewhere. The expression vectors may be integrated into the genome of the respective transformants. Alternatively and more preferably, the vector is kept as a plasmid without integration into the host cell genome. In particular, said generating may comprise cloning the polynucleotides of the libraries into expression vectors. The generating may also comprise generating the library of polynucleotides by chemical polynucleotide synthesis (e.g. DNA synthesis). In particular, the chemical synthesis may be a nucleoside phosphoramidite solid state synthesis. Such a nucleoside phosphoramidite solid state synthesis is offered by numerous commercial suppliers.

The transformed (prokaryotic or eukaryotic) hosts or the population of host cells can be grown in fermenters. They may be cultured according to techniques known in the art, in particular techniques which achieve optimal cell growth. The expressed biomolecules (e.g. peptides or RNA) may then be isolated from the growth medium, from cellular lysates, or from cellular membrane fractions and the like. The isolation and purification of the (microbially or otherwise) expressed biomolecules (e.g. peptides) may be by any conventional means such as, for example, preparative chromatographic separations and immunological separations such as those involving the use of monoclonal or polyclonal antibodies (Ausubel, Current Protocols in Molecular Biology, Green Publishing Associates and Wiley Interscience, N.Y. (1994)).

The method for identifying according to the invention may comprise detecting the expression of said biomolecule (e.g. peptide and/or RNA) by Western blots or Northern blots, chromatographic enrichment or affinity purification via a co-transcribed affinity tag in the case of RNA or a co-translated protein tag for peptides.

The polynucleotides of the library to be employed in accordance with the invention may further comprise a defined nucleic acid sequence. It is preferred that all or mostly all (e.g. at least 80%, preferably at least 90%, preferably at least 95%, preferably at least 99%) of the polynucleotides may comprise a defined nucleic acid sequence. The

defined nucleic acid sequence may be such that can be used as primer annealing site for amplification of said polynucleotides by PCR.

Means and methods for generating randomized polynucleotides (e.g. DNA or RNA fragments) are well known in the art. For example, means and methods using chemical synthesis of oligo-nucleotides has been described previously. Such a synthesis may proceed in a step-wise manner, where one nucleotide is added after each other to a growing chain. To obtain a random nucleotide sequence, one can add more than one (up to four) different nucleotides to each synthesis step. However, because of differences in the reactivity of the different nucleotides, it may be difficult to obtain completely random sequences in this way. Different synthesis schemes can, however, compensate for this, at least partly. Further, one may change the relative ratio of nucleotides at each synthesis step to bias its coding potential towards certain amino acids, or to reduce the incidence of premature stop codons. One of the most common approaches to construct random biomolecules (e.g. peptides) is to use (NNK)n codon degeneracy, where N indicates an equimolar mixture of all four nucleotides (A, G, C and T), and K indicates a 1:1 mixture of G and T (Omidfar & Daneshpour, 2015).

The evolutionary principle of natural selection is based on differences in reproductive fitness in a common environment (Boero 2015). Individuals that have an advantage will show an enhanced reproductive rate, while individuals that have a disadvantage will show a decreased reproductive rate. Even very small differences in reproductive rate can become effective when the growth continues, e.g. for multiple generations (for example for the number of cell cycles as described herein elsewhere), in the same environment. For example, assuming a 5% fitness difference between two populations, one would have a 1.2-fold difference in the number of individuals after 4 generations of uninhibited growth, 1.5-fold at 10% and 5.1-fold at 50% (Table 1). In the present invention, the differential fitness is provided by random polynucleotides/polypeptides (delivered by vectors) to the cells that have otherwise the same genetic background. When they grow in the same environment, their relative advantage can be assessed from the ratio of cells (e.g. measured by the polynucleotides of the library comprised in those cells) after a certain number of cell divisions.

During a second growth phase of the experiments, those clones having a growth advantage or disadvantage conferred by the expressed biomolecule (e.g. peptide and/or RNA) will get enriched or depleted, respectively, within the culture. In accordance with the invention, a growth advantage and or disadvantage can only be conferred if the corresponding biomolecule (e.g. peptide and/or RNA) has a biological activity.

In accordance with the invention, the determining of the frequencies of polynucleotides in the library may comprise a polynucleotide (e.g. DNA) sequencing. Said determining may further comprise isolating polynucleotides from a sample of the population of host cells to be employed. Said determining may further comprise amplification of said polynucleotides, for example by PCR.
The frequencies of polynucleotides of the library at the first and the second time point may be determined in parallel, for example by multiplex nucleotide (e.g. DNA) sequencing.

In one aspect, the polynucleotides of the library to be employed in accordance with the invention may further comprise a second/further nucleic acid sequence that encodes a second/further biomolecule (e.g. a second RNA or peptide segment). Said second/further biomolecule (e.g. a second RNA or peptide segment may allow for capturing and/or detecting a biomolecule (e.g. an RNA or a peptide) resulting from expression of the polynucleotides of the library. The second/further peptide segment may in particular be an affinity tag like a His tag, FLAG epitope tag, GFP-tag, GST-tag, or any other epitope tag known in the art.

In (one), however non-limiting, aspect(s), the invention may be performed as follows (by applying one or more of the items below):

1. Generation of a clone library containing randomized nucleotide sequences in suitable expression vectors that allow these sequences to be expressed as RNAs and with the option that the RNAs can be translated into polypeptides by providing an appropriate start codon and ribosome initiation site.
2. Transformation of the library into host cells in which the RNA expression from the expression vector can be activated such that the cloned insert becomes expressed in the cell; the transformation process is set up in a way to ensure that any cell

receives not more than one vector with one sequence variant (this is implicit in the term "cloning").

3. Optional amplification of the library under conditions where the expression vector does not express the insert sequence.

4. Growth of samples of the library under conditions where the expression of the insert from the expression vector is activated; the growth conditions should not be restrictive, i.e. should provide all cells good conditions of growth and the same probability of replicating, such that their relative replication rate is only influenced by the expression vector they carry.

5. Setting up replicates under the same growth conditions; the number of replicates depends on a statistical power analysis that estimates the probability to detect a significant change in frequency in dependence of the number of cell divisions surveyed, number of different clones in the library, the dispersion (variance) between replicates and depth of sequencing (point 6).

6. Sequencing of the inserts of the expression vectors by a parallel sequencing procedure to determine the relative frequency of the clones in the original library at the start of the experiment, as well as at the end of the experiment (optionally also at intermediate stages).

7. Identification of clones that have changed their relative frequency among the cells during the course of the experiment; three types of clones can be expected - type 1: clones that have raised in frequency are expected to express a polynucleotide or peptide that is beneficial for the growth of the cell; type 2: clones that became reduced in frequency are expected to express a polynucleotide or peptide that is detrimental for the growth of the cell; type 3: clones that have not significantly changed their frequency are expected to express a polynucleotide or peptide that is neutral for the growth of the cell

8. Determination of significance values for the increase or decrease of peptide frequencies; these significance assessments are based on running parallel experimental setups (step 5) to estimate the variance in the experiment and to set a cutoff for significant outliers.

9. Confirmation of the results by repeating the whole experiment under the same conditions; clones that are found in at least two independent experiments to show the same direction and magnitude of frequency change are considered as confirmed.

10. Further verification of the bioactivity (i.e. relative growth advantage or disadvantage) of the type 1 or type 2 polynucleotides or peptides through isolation of the respective clone from the library (or *de novo* synthesis based on the determined sequence) and repetition of the experiment under the same conditions, or with a reduced set of competitor clones, optionally with a standardized set of such clones.

11. Test for a given clone whether the polynucleotide or the peptide conveys the bioactivity by introducing an artificial early stop codon into the predicted peptide sequence of the insert and run the verification test of step 10; the activity of a peptide would be abolished by the stop codon, while the activity of a RNA would be expected to be only be mildly affected by one changed nucleotide.

Without being bound by theory, the polynucleotides or biomolecules (e.g. peptides) that are identified in the context of the means and methods of the invention may interact with a component in the cell that influences growth conditions in a positive or negative way. This makes them analogous to genetic mutations. As such, they are potential tools to study the function of cellular pathways. They are also candidate molecules to actively interfere with cellular pathways for pharmaceutical or diagnostic or other technical applications. For example, positively acting biomolecules (e.g. RNA or peptides) may be used to improve microbial production systems, to improve plant and animal strains and varieties used for agricultural or other bio-production, to increase fermentation yields, to increase viability and/or proliferation potential of cultured cells (in particular primary cells) or even to immortalize cells (e.g. primary cells); negatively acting biomolecules (e.g. RNA or peptides) may become useful for targeting disease causing organisms or cancer cells; either class of peptides or polynucleotides may become useful for generating new pharmaceutical drugs. Moreover, negatively-acting biomolecules (e.g. RNA or peptides) may be used as novel anti-bacterial drugs (e.g. antibiotics).

The present invention is also directed to methods that employ any of the methods for identifying a polynucleotide or a biomolecule (e.g. peptide or polynucleotide) encoded thereby and subsequent analysis of said identified polynucleotide or biomolecule. A given polynucleotide or peptide may be further studied, for example by identifying the cellular process with which it interferes. The whole established repertoire of genetic and biochemical techniques is available for this. In a first step, one may, for example, test transcriptomic and proteomic changes found in cells carrying the polynucleotide or

peptide versus control cells that do not comprise the respective polynucleotide. This would provide a first insight into the cellular networks that are affected. Alternatively, or in addition co-immunoprecipitation (optionally after crosslinking) could be used to identify proteins that interact with a peptide (in this case a co-translated tag may be used as a target for the antibody). For polynucleotide interactions, one could use capture by hybridization with a tag sequence. Co-purified proteins may be detected by proteomic approaches (mass spectrometry of defined fragments), co-purified nucleic acids by sequencing. Subcellular localization of peptides may be studied by fusing them with fluorescent domains, such as GFP. The identification of interaction partners in the cell may then guide further experiments to reveal the actual function of the new polynucleotide or peptide.

Apart of studying the function of individual polynucleotides or peptides, it would also be possible to prepare large sets of polynucleotides or peptides that were found to have bioactivity through the present invention. These may further be used for screening influences on specific biological processes of interest, much in the same way as sets of chemicals are used to screen for relevant functions (chemical library drug screening - see Archer 2005 for a review).

Accordingly, the present invention is also directed to a pool of polynucleotides that encode for a biomolecule having biological activity that were identified with one of the methods according to the present invention. In particular, also a library of host cells (e.g. *E. coli*) being transformed or comprising such library is also provided.

The polynucleotides/biomolecules to be identified in the accordance with the invention may be further optimized in established procedures of, for example, phage display, aptamer and SELEX approaches. The skilled person is well aware of such procedures/approaches.

The present invention also relates to the following items:

1. A method for identifying a polynucleotide encoding a biomolecule with biological activity, said method comprising:

   a) cultivating a population of host cells capable of expressing a library of polynucleotides, wherein each of said polynucleotides comprises a random nucleic acid sequence; and

b) determining the frequencies of individual polynucleotides in said library comprised in said population of host cells at a first time point and at a subsequent second time point during cultivation, wherein the biomolecules encoded by the library are expressed in said population between said first and said second time point,

wherein a change in the frequency of a polynucleotide in said library of polynucleotides between said first and said second time point determined according to step b) identifies a polynucleotide with biological activity.

2. The method according to item 1, wherein said change is an increase by at least 1.1-fold, preferably by at least 3-fold and most preferably by at least 10-fold; or said change is a decrease by at least 1.1-fold, preferably by at least 3-fold and most preferably by at least 10-fold.
3. The method according to item 1 or 2, wherein said change is a statistically significant increase or a statistically significant decrease.
4. The method according to item 3, wherein said statistical significance of said change is determined with a statistic test with a false discovery rate (FDR) of at least 50%, preferably 10% and most preferably 5%.
5. The method according to item 4, wherein said statistic test is a Wald test or any other statistic test based on assessing the probability of an observation within a random distribution.
6. The method according to item 1, wherein said method is performed with at least 2, preferably at least 5 more preferably at least 10 and most preferably 25 separate host cell populations in parallel.
7. The method according to any one of the preceding items, wherein said cultivating is performed under optimal culturing conditions for said population of host cells.
8. The method according to any one of items 1 to 6, wherein said cultivating is performed under suboptimal but non-toxic culturing conditions for said population of host cells.
9. The method according to any of the preceding items, wherein said biomolecule is an RNA or a peptide and said biomolecules are RNAs and/or peptides.
10. The method according to item 9, said method further comprising determining the amino acid sequence of the peptide encoded by the identified polynucleotide or

the ribonucleic acid sequence of the RNA encoded by the identified polynucleotide.

11. The method according to item 9 or 10, said method comprising detecting the expression of said peptide by Western blot or a chromatographic technique and/or RNA by Northern blot or a chromatographic technique.
12. The method according to any one of the previous items, wherein said biological activity is cell growth promoting or inhibiting activity and/or cell survival promoting or inhibiting activity.
13. The method according to any one of the preceding items, wherein each or essentially each of the host cells of said host cell population comprises and/or expresses one or not more than one of the polynucleotides of said library.
14. The method according to any one of the preceding items, wherein the host cell population does not express the library of polynucleotides before said first time point.
15. The method according to any one of the preceding items, wherein the expression of the library of polynucleotides is induced at said first time point.
16. The method of item 15, wherein the expression of the library of polynucleotides is induced by addition and/or depletion of one or more substances or by physical means.
17. The method according to any one of the preceding items, wherein the host cell population is in a logarithmic growth phase at said first and/or second time point.
18. The method according to any one of the preceding items, wherein the host cell population is in a stationary growth phase at said second time point.
19. The method according to any one of the preceding items, wherein said host cell population undergoes between 1 and 50, preferably between 4 and 35 and most preferably between 16 and 25 cycles of cell division between said first and said second time point.
20. The method according to any one of the preceding items, wherein said polynucleotides of said library are operatively linked to a promoter sequence, wherein said promoter regulates the expression of said polynucleotides.
21. The method according to item 20, wherein said promoter is an inducible promoter.
22. The method according to item 21, wherein said inducible promoter is capable of being activated.

23. The method according to item 21 or 22, wherein said promoter is capable of being activated by addition and/or depletion of one or more substances or by physical means.
24. The method according to any one of items 21 to 23, wherein said inducible promoter is activated at said first time point.
25. The method according to any one of the preceding items, wherein said host cell population is a population of eukaryotic or prokaryotic host cells.
26. The method according to any one of the preceding items, wherein said host cells are growing in solution or are adherent to a surface.
27. The method according to any one of the preceding items, wherein said host cell population is a population of *Escherichia coli (E.coli)* cells, *Bacillus subtilis* cells, *Ralstonia eutropha* cells or cells of a member of the *Pseudomonas genus.*
28. The method according to any one of the preceding items, wherein said host cell population is a population of *E. coli* cells.
29. The method according to any one of items 1 to 26, wherein said host cell population is a population of primary eukaryotic cells.
30. The method according to any one of items 1 to 26, wherein said host cell population is a population of CHO cells, BHK cells, NSO cells, SP2/0 cells, YB2/0 cells, COS cells, mouse L-cells), human embryonic kidney (HEK) 293 cells, HELA cells, J558L, Sp2/0 cells, Drosophila Kc, Drosophila S2 cells, induced pluripotent stem cells, differentiated cell lines derived from stem cells, or cancer cells.
31. The method according to any one of the preceding items, wherein said random nucleic acid sequence has a length of 18 to 300 nucleotides, preferably 36 to 250 nucleotides and most preferably 120 to 180 nucleotides.
32. The method according to any one of the preceding items, wherein the random sequence is expressed in frame.
33. The method according to any one of the preceding items, wherein the polynucleotides of said library comprise or are positioned next to a transcription start site and a transcription termination signal; and/or a translation start site and translation termination site.
34. The method according to any one of the preceding items, wherein each of the polynucleotides of said library is comprised in a vector.
35. The method according to item 34, wherein said vector is an expression vector.

36. The method according to item 34 or 35, wherein said vector is a plasmid.

37. The method according to any one of items 34 to 36, wherein said vector is integrated into the genome of a host cell.

38. The method according to any one of items 34 to 37, wherein said vector can be propagated in prokaryotic and/or eukaryotic host cells and/or can give rise to the expression of the polynucleotides of said library in prokaryotic and/or eukaryotic host cells.

39. The method according to any one of the preceding items, wherein the polynucleotides of said library further comprise a second nucleic acid sequence that encodes a second RNA and/or second peptide segment.

40. The method according to item 39, wherein said second RNA segment and/or said second polypeptide segment allows for capturing and/or detecting an RNA and/or a peptide resulting from expression of the polynucleotides of said library.

41. The method according to item 39 or 40, wherein said second peptide segment is an affinity tag.

42. The method according to any one of the preceding items, wherein the polynucleotides of said library further comprise a nucleic acid sequence that encodes a peptide and/or RNA targeting the expressed polypeptide and/or RNA to one or more cellular compartments.

43. The method according to any one of the preceding items, wherein said method further comprises generating said host cell population capable of expressing said library of polynucleotides prior to culturing.

44. The method according to item 43, wherein said generating comprises transforming said host cells with expression vectors for said library of polynucleotides.

45. The method according to item 44, wherein said expression vectors are integrated into the genome of the respective transformants.

46. The method according to of item 43 to 45, wherein said generating comprises cloning the polynucleotides of said libraries into expression vectors.

47. The method according to any one of items 43 to 46, wherein said generating comprises generating said library of polynucleotides by chemical DNA synthesis.

48. The method according to any one of the preceding items, wherein said determining of the frequencies of polynucleotides in said library comprises DNA sequencing.

49. The method according to item 48, wherein said determining of the frequencies of polynucleotides in said library further comprises isolating polynucleotides from a sample of said population of host cells.

50. The method according to item 48 or 49, wherein said determining of the frequencies of polynucleotides in said library further comprises amplification of said polynucleotides by PCR.

51. The method according to any one of the preceding items, wherein all polynucleotides of said library further comprise a defined nucleic acid sequence.

52. The method according to item 51, wherein said defined nucleic acid sequence can be used as primer annealing site for amplification of said polynucleotides by PCR

53. The method according to any one of the preceding items, wherein the frequencies of polynucleotides of said library at said first and said second time point are determined in parallel by multiplex DNA sequencing.

54. The method according to any one of the preceding items, wherein said method is performed with at least 2, preferably at least 5, more preferably at least 10 and most preferably 25 separate host cell populations in parallel.

55. The method according to item 54, wherein only a polynucleotide being enriched or depleted in 2, preferably 5, more preferably at least 10 or most preferably 25 of said host cell populations is identified as a polynucleotide with biological activity.

56. The method according to any one of the preceding items, wherein said random nucleic sequence of the polynucleotides of said library has an unequal representation of the four nucleotides A, C, G and T at each position, whereby the representation can include one, two, three or all four different nucleotides at either position.

57. The method according to any one of the preceding items, wherein said cultivating is performed for two subsets of said population of host cells, whereby one subset is cultivated under optimal culturing conditions and the second subset is treated with a pharmacological substance or genetically modified.

58. A method for identifying a polynucleotide encoding a biomolecule with biological activity, said method comprising:

a) cultivating a population of host cells capable of expressing a library of polynucleotides in at least 2, preferably at least 5, more preferably at least 10 or most preferably at least 25 replicates, wherein each of said polynucleotides comprises a different random nucleic acid sequence; and

b) determining for each of said replicates the frequencies of polynucleotides in said library comprised in said population of host cells at a first time point and at a subsequent second time point during cultivation, wherein the biomolecules encoded by the library are expressed in said population between said first and said second time point,

wherein a significant change in the frequency of a polynucleotide in said library of polynucleotides between said first and said second time point determined according to step b) identifies a polynucleotide encoding a biomolecule with biological activity,

and wherein the significance of change is assessed by a statistic test based on a Student t-test or a Wald test with a false discovery rate (FDR) of at least 50%, preferably 10% and most preferably 5%.

59. A method for identifying a polynucleotide encoding a biomolecule with biological activity (e.g. related to the effect of a certain treatment of host cells, such as addition of a chemical substance, and/or to a genetic manipulation), said method comprising:

a) cultivating two (or more) population of host cells capable of expressing a (similar or identical) library of polynucleotides, wherein each of said polynucleotides comprises a different random nucleic acid sequence, and wherein the two population of host cells are identical but are treated differently (e.g. by addition of a chemical substance); and/or the two population of host cells are of the same cell type but have a pre-defined genetic difference; and

b) determining the frequencies of individual polynucleotides in said library comprised in said two populations of host cells at a first time point and at a subsequent second time point during cultivation, wherein the biomolecules encoded by the library are expressed in said two populations between said first and said second time point,

c) determine the changes in the frequency of a polynucleotide in said library of polynucleotides between said first and said second time point for both populations of host cells,

wherein polynucleotides which with a difference said change in the frequency between said two populations of host cells are identified as a polynucleotide encoding a biomolecule with biological activity (e.g. related to the effect of a certain treatment of host cells, such as addition of a chemical substance, and/or to a genetic manipulation).

60. The method according to item 59, wherein said difference is an increase by at least 1.1-fold, preferably by at least 3-fold and most preferably by at least 10-fold; or

said change is a decrease by at least 1.1-fold, preferably by at least 3-fold and most preferably by at least 10-fold.

61. The method according to item 59 or 60, wherein said difference is a statistically significant increase or a statistically significant decrease.

62. The method according to item 61, wherein said statistical significance of said difference is determined with a statistic test with a false discovery rate (FDR) of at least 50%, preferably 10% and most preferably 5%.

63. The method according to item 62, wherein said statistic test is a Wald test or any other statistic test based on assessing the probability of an observation within a random distribution.

64. Polynucleotide (e.g. randomized part as defined herein elsewhere) as identified by a method according to any one of items 1 to 63.

65. Biomolecule (e.g. peptide or RNA) encoded by the polynucleotide of item 64 (it is particularly envisaged that the biomolecule exhibits biological activity in accordance with the invention).

66. Biomolecule, in particular protein, peptide or RNA, selected from the group consisting of:

(a) a biomolecule which comprises or consists of a biomolecule being at least about 60%, 70%, 85%, 90%, 95%, 98%, 99% or 100% identical to a biomolecule referred to herein elsewhere and, in particular, in Tables 3 and 4 herein below and/or in SEQ ID NOs: 9 to 68 (or as encoded by SEQ ID NOs: 69 to 128), or to the randomized core section of such a

biomolecule (it is particularly envisaged that the biomolecule exhibits biological activity in accordance with the invention);

(b) a biomolecule which comprises or consists of a biomolecule encoded by a nucleic acid molecule hybridizing under stringent conditions to the complementary strand of a nucleic acid molecule encoding a biomolecule referred to herein elsewhere and, in particular, in Tables 3 and 4 herein below and/or in SEQ ID NOs: 69 to 128, or encoding the randomized core section of such a biomolecule (it is particularly envisaged that the biomolecule exhibits biological activity in accordance with the invention);

(c) a biomolecule which comprises or consists of a fragment of the biomolecule of any one of (a) or (b) (said fragment comprising, for example, at least 10, 20, 30, 40, 45, 46, 47, 48 or 49 amino acid residues) (it is particularly envisaged that the biomolecule exhibits biological activity in accordance with the invention).

67. Polynucleotide encoding the biomolecule of item 66, polynucleotide as depicted in Tables 3 and 4 herein below and/or in SEQ ID NOs: 69 to 128, or variant of a polynucleotide as depicted in Tables 3 and 4 and/or in SEQ ID NOs: 69 to 128 herein below.

68. A polynucleotide comprising or consisting of a polynucleotide selected from the group consisting of:

(i) a polynucleotide as depicted in any one of SEQ ID NOs: 69 to 128, or a polynucleotide as defined by nucleotide positions 13 to 162 of a polynucleotide as depicted in any one of SEQ ID NOs: 69 to 128; and

(ii) a polynucleotide being at least about 60%, 70%, 85%, 90%, 95%, 98%, 99% or 100% identical to a polynucleotide as defined in (i),

wherein said polynucleotide encodes for a biomolecule with biological activity.

69. The polynucleotide of item 68, wherein said polynucleotide comprises or consists of a polynucleotide as defined in (i).

70. The polynucleotide of item 68 or 69, wherein said polynucleotide as defined in (i) is selected from a group consisting of:

a polynucleotide as depicted in any one of SEQ ID NOs: 69 to 87, preferably in any one of SEQ ID NOs: 70, 74 and 84;

a polynucleotide as defined by nucleotide positions 13 to 162 of a polynucleotide as depicted in any one of SEQ ID NOs: 69 to 87, preferably in any one of SEQ ID NOs: 70, 74 and 84;

and wherein said biological activity is a cell growth promoting acitivity preferably a cell growth promoting activity in *E. coli*.

71. The polynucleotide of item 68 or 69, wherein said polynucleotide as defined in (i) is selected from a group consisting of:

a polynucleotide as depicted in any one of SEQ ID NOs: 88 to 128, preferably in any one of SEQ ID NOs: 98, 107 or 118;

a polynucleotide as defined by nucleotide positions 13 to 162 of a polynucleotide as depicted in any one of SEQ ID NOs: 88 to 128, preferably in any one of SEQ ID NOs: 98, 107 or 118;

and wherein said biological activity is a cell growth inhibiting activity preferably a cell growth inhibiting activity in *E. coli*.

72. The polynucleotide of item 68 or 69, wherein said polynucleotide as defined in (i) is selected from a group consisting of:

a polynucleotide as depicted in any one of SEQ ID NOs: 70 and 74;

a polynucleotide as defined by nucleotides 13 to 162 of a polynucleotide as depicted in any one of SEQ ID NOs: 70 and 74;

wherein said biomolecule is an RNA, and, optionally, wherein said biological activity is a cell growth promoting activity (e.g., a cell growth promoting activity in *E. coli)*.

73. The polynucleotide of item 68 or 69, wherein said polynucleotide as defined in (i) is selected from a group consisting of:

a polynucleotide as depicted in SEQ ID NO: 84;

a polynucleotide as defined by nucleotides 13 to 162 of a polynucleotide as depicted in SEQ ID NO: 84,

wherein said biomolecule is a polypeptide, and, optionally, wherein said biological activity is a cell growth promoting activity (e.g. a cell growth promoting activity in *E. coli)*.

74. A vector, preferably an expression vector (e.g. a pFLAG-CTC™ expression vector) comprising the polynucleotide as defined in any one of items 68 to 73.

75. A biomolecule (e.g. an RNA or a peptide) encoded by the polynucleotide as defined in any one of items 68 to 73 or the vector of claim 74.

76. A cell, preferably an *E. coli* cell comprising a polynucleotide of any one of items 68 to 73, a vector of item 74 or a biomolecule of item 75.

77. A biomolecule, wherein said biomolecule is selected from the group consisting of:

a polypeptide comprising or consisting of

(i) an amino acid sequence as defined by amino acid positions 5 to 54 of SEQ ID NOs: 9 to 68 or preferably as defined by any one of SEQ ID NOs: 9 to 68;

(ii) an amino acid sequence having at least about 60%, 70%, 85%, 90%, 95%, 98%, 99% or 100% identity to the amino acid sequence as defined in (i); and

(iii) an amino acid sequence as defined in (i) having 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, or 20 amino acids inserted, deleted or substituted by a different amino acid; or

an RNA comprising or consisting of:

(iv) the RNA sequence encoded by positions 13 to 162 of SEQ ID NOs: 69 to 128, preferably by any one of SEQ ID NOs: 69 to 128; or most preferably by any one of SEQ ID NOs: 69 to 128 having SEQ ID NO: 136 directly fused to its 5'-end and/or having SEQ ID NO: 137 directly fused to its 3'-end; or

(v) the RNA sequence having at least about 60%, 70%, 85%, 90%, 95%, 98%, 99% or 100% identity to the RNA sequence as defined in (iv);

wherein said biomolecule has biological activity.

78. The biomolecule of item 77, wherein the amino acid sequence as defined in (i) is an amino acid sequence as defined by amino acid positions 5 to 54 of SEQ ID NOs: 9 to 27 or as defined by any one of SEQ ID NOs: 9 to 27; and wherein the RNA sequence as defined in (iv) is encoded by positions 13 to 162 of SEQ ID NOs: 69 to 87, any one of SEQ ID NOs: 69 to 87; or by any one of SEQ ID NOs: 69 to 87 having SEQ ID NO: 136 directly fused to its 5'-end and/or having SEQ ID NO: 137 directly fused to its 3'-end.

79. The biomolecule of item 77, wherein said biomolecule is a polypeptide, and wherein the amino acid as defined in (i) is the amino acid sequence as defined by amino acid positions 5 to 54 of SEQ ID NO: 24 or preferably as defined by SEQ ID NO: 24.

80. The biomolecule of item 77, wherein said biomolecule is an RNA, and wherein the RNA sequence as defined in (iv) is encoded by positions 13 to 162 of SEQ ID NOs: 70 or 74, preferably by any one of SEQ ID NOs: 70 or 74; or most preferably by any one of SEQ ID NOs: 70 or 74 having SEQ ID NO: 136 directly fused to its 5'-end and/or having SEQ ID NO: 137 directly fused to its 3'-end.

81. The biomolecule of any one of items 78 to 80, wherein said biological activity is a cell growth promoting activity, preferably a cell growth promoting activity in *E. coli*.

82. Use of a biomolecule as defined in any one of items 78 to 81, a polynucleotide encoding said biomolecule, an expression vector comprising said polynucleotide, or a cell comprising said biomolecule, said polynucleotide or said expression vector for increasing cell proliferation of a bacterial production strain (e.g. E.coli) in a fermentation process to produce a substance of interest (e.g. a chemical compound, a protein, an antibody, an amino acid etc.) and/or enhancing the yield of said substance of interest.

83. The biomolecule of item 77, wherein the amino acid sequence as defined in (i) is an amino acid sequence as defined by amino acid positions 5 to 54 of SEQ ID NOs: 28 to 68 or preferably as defined by any one of SEQ ID NOs: 28 to 68; and wherein the RNA sequence as defined in (iv) is an RNA sequence encoded by positions 13 to 162 of SEQ ID NOs: 88 to 128, preferably by any one of SEQ ID NOs: 88 to 128; or most preferably by any one of SEQ ID NOs: 88 to 128 having SEQ ID NO: 136 directly fused to its 5'-end and/or having SEQ ID NO: 137 directly fused to its 3'-end.

84. The biomolecule of item 77, wherein the amino acid sequence as defined in (i) is an amino acid sequence as defined by amino acid positions 5 to 54 of SEQ ID NO: 38, 47, or 58, or preferably as defined by SEQ ID NOs: 38, 47, or 58; and wherein the RNA sequence encoded by positions 13 to 162 of SEQ ID NOs: 98, 107 or 118, preferably by SEQ ID NOs: 98, 107 or 118; or most preferably by SEQ ID NOs: 98, 107 or 118 having SEQ ID NO: 136 directly fused to its 5'-end and/or having SEQ ID NO: 137 directly fused to its 3'-end.

85. The biomolecule of item 83 or 84, wherein said biological activity is a cell growth inhibiting activity, preferably a cell growth inhibiting activity in *E. coli*.

86. Use of a biomolecule as defined in any one of items 83 to 85, a polynucleotide encoding said biomolecule, an expression vector comprising said

polynucleotide, or a cell comprising said biomolecule, said polynucleotide or said expression vector as an anti-bacterial agent (e.g. antibiotics), preferably as an anti-*E.coli* agent.

87. The biomolecule of any one of items 77 to 81 and 83 to 85, wherein said polypeptide consists of (i), (ii) or (iii), preferably (i).
88. The biomolecule of any one of items 77 to 81, 83 to 85 and 87, wherein said RNA consists of (iv) or (v), preferably (iv).
89. The biomolecule of any one of items 77 to 81, 83 to 85 and 87 to 88, wherein said polypeptide comprises or consists of (i).
90. The biomolecule of any one of items 77 to 81, 83 to 85 and 87 to 89, wherein said RNA comprises or consists of (iv).
91. The biomolecule of any one of items 77 to 81, 83 to 845 and 87 to 90, wherein said biomolecule is a polypeptide.
92. The biomolecule of any one of items 77 to 81, 83 to 85 and 87 to 90, wherein said biomolecule is an RNA.
93. A polynucleotide encoding the biomolecule as defined in any one of items 77 to 81, 83 to 85 and 87 to 92.
94. An expression vector comprising the polynucleotide of item 93.
95. A host cell, preferably an *E.coli* cell, comprising the biomolecule of any one of items 77 to 81, 83 to 85 and 87 to 92, the polynucleotide of item 93 or the expression vector of item 94.

In principle, items 2 to 57 can also be applied *mutatis mutandis* to items 59 or 58.

**Literature (cited herein by reference to the author(s) (and the year of publication))**

Archer JR (2005). History, Evolution, and Trends in Compound Management for High Throughput Screening. ASSAY and Drug Development Technologies 2: 675-681.

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jané-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P Jr, de Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M,

Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483: 603-607. doi: 10.1038/nature11003.

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B 57 (1): 289–300.

Boero F (2015). From Darwin's Origin of Species toward a theory of natural history. F1000Prime Rep 7:49. doi: 10.12703/P7-49. eCollection 2015.

Bruno JG (2015). Predicting the Uncertain Future of Aptamer-Based Diagnostics and Therapeutics. Molecules. 20(4): 6866-6887. doi:10.3390/molecules20046866.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nature Methods 12: 59-60.

Darmostuk M, Rimpelova S, Gbelcova H, Ruml T (2015). Current approaches in SELEX: An update to aptamer selection technology. Biotechnol Adv. 33(6 Pt 2):1141-1161. doi: 10.1016/j.biotechadv.2015.02.008. Epub 2015 Feb 20.

Ching T, Huang S, Garmire LX (2014). Power analysis and sample size estimation for RNA-Seq differential expression. RNA 20: 1684-1696. doi:10.1261/rna.046011.114. Epub 2014 Sep 22.Chen Z, Liu J, Ng HK, Nadarajah S, Kaufman HL, Yang JY, Deng Y (2011). Statistical methods on detecting differentially expressed genes for RNA-seq data. BMC Syst Biol. 5; Suppl 3:S1.

Davis TN (2004). Protein localization in proteomics. *Current opinion in chemical biology* 8: 49-53.

Edgar RC (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26: 2460-2461. doi: 10.1093/bioinformatics/btq461

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data P. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079. doi: 10.1093/bioinformatics/btp352

Love MI, Huber W, Anders S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**, 550.

Neme T, Tautz D. 2016. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to *de novo* gene emergence. *eLife* 5:e09977.

Omidfar K & Daneshpour M (2015) Advances in phage display technology for drug discovery, Expert Opinion on Drug Discovery, 10:6, 651-669

Oshlack A, Robinson MD, Young MD (2010). From RNA-seq reads to differential expression results. Genome Biol. 11(12):220.

Schlötterer C (2015). Genes from scratch--the evolutionary fate of de novo genes. Trends Genet 31(4):215-9. doi: 10.1016/j.tig.2015.02.007. Epub 2015 Mar 12

Sedlazeck FJ, Rescheneder P, von Haeseler A. 2013. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. Bioinformatics 29: 2790-2791. doi: 10.1093/bioinformatics/btt468

Sims D1, Mendes-Pereira AM, Frankum J, Burgess D, Cerone MA, Lombardelli C, Mitsopoulos C, Hakas J, Murugaesu N, Isacke CM, Fenwick K, Assiotis I, Kozarewa I, Zvelebil M, Ashworth A, Lord CJ. High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing. Genome Biol. 2011 Oct 21;12(10):R104.

Tompa P, Schad E, Tantos A, Kalmar L (2015). Intrinsically disordered proteins: emerging interaction specialists. Curr Opin Struct Biol. 2015 Sep 21;35:49-59. doi: 10.1016/j.sbi.2015.08.009.

Wald A (1943). Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. Transactions of the American Mathematical Society, 54, 426-482.

Woolfson DN, Bartlett GJ, Burton AJ, Heal JW, Niitsu A, Thomson AR, Wood CW (2015). De novo protein design: how do we expand into the universe of possible protein structures? Curr Opin Struct Biol. 33:16-26. doi: 10.1016/j.sbi.2015.05.009.

Yin J1, Li G, Ren X, Herrler G (2007). Select what you need: a comparative evaluation of the advantages and limitations of frequently used expression systems for foreign genes. J Biotechnol 127: 335-347.

Xiao N, Cao DS, Zhu MF, Xu QS. 2015. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics 31: 1857-1859

The present invention is further described by reference to the following non-limiting figures and examples.

The Figures show:

**Figure 1** shows a scheme of the plasmid expression vector employed in the examples of the present invention. The promoter-regulatory region of the strong Ptac promoter (a hybrid of the trp and lac promoters from *E. coli*) drives transcription of ORF-FLAG fusion constructs. Control of transcription is regulated by the presence of the lacO sequences and inclusion of the lac repressor gene (lacI) on the plasmid. RBS is ribosome binding site, ATG is the start codon, MCS is the multiple cloning site, FLAG is the epitope peptide, STOP is the stop codon.

**Figure 2** shows an exemplary scheme of the general setup of a growth experiment performed with a method according to the present invention and using *E. coli* as a host cell. The growth medium is LB supplemented with ampicillin (AMP) (to select for plasmid-bearing *E. coli* cells) in the pre-culture and further supplemented with IPTG to induce the expression from the vector in the further cycles. The cultures are grown to stationary phase in each cycle. The scheme is shown up to cycle #3, the described experiments used an additional cycle.

**Figure 3** Induction of peptide expression drives changes in peptide frequency over time. Plots of fold-change (compared to the first cycle) versus mean counts across pairwise comparisons. Negative fold changes are indicative of depletion compared to the first cycle, and positive fold changes are indicative of enrichment compared to the first cycle. Left, center and right panels indicate comparisons with the $2^{nd}$, $3^{rd}$ and $4^{th}$ cycle (24 hours per cycle). Top panels show the experiments induced with IPTG; bottom panels the experiments without IPTG induction. Grey dots indicate peptides with significant fold changes (5% FDR), positive and negative, respectively. Black crosses indicate peptides with non-significant fold changes. The number in the lower-right corner of each plot indicates the total number of peptides with significant changes. Both experiments were derived from the same stock culture, and performed simultaneously.

**Figure 4** Examples of peptides with significant changes in frequency over time (in this case each time point represents a 24 h cycle). A and B show peptides increasing in frequency (as normalized counts). C and D show peptides decreasing in frequency. Boxplots show the median and the interquartile ranges (outliers as dots) across the 10 replicates of each cycle.

**Figure 5** Assessment of read depth on detection power. Progression of significant fold changes with sampling depth in experiment E7 (from 10% of the reads to 100% of the reads). Circles left to the dotted line indicate peptides with significant decreases in frequency, circles right to the dotted line indicate peptides with significant increases in frequency, black crosses indicate peptides with non-significant fold changes. Significance was set at 5% FDR. The X-axis represents the log2-fold changes, the Y-axis the abundance of the respective peptides in the sequence sample.

**Figure 6**: Average amino acid composition comparisons. E: Experimental peptides, (+): peptides enriched in all experiments, (–): peptides depleted in all experiments, R: computationally generated random peptides, B: biological sequences from *E. coli*. Significantly different amino acid composition between experimental and random sequences are shown as R* and between experimental and biological sequences are shown as B* (Wilcoxon-rank test, 5% FDR, corrected across all pairwise comparisons and all amino acids). All comparisons between enriched and depleted peptides to the experimental set and between each other are non-significant.

In this specification, a number of documents including patent applications are cited. The disclosure of these documents, while not considered relevant for the patentability of this invention, is herewith incorporated by reference in its entirety. More specifically, all referenced documents are incorporated by reference to the same extent as if each individual document was specifically and individually indicated to be incorporated by reference.

**Figure 7:** Expression of peptides. Western blot with antiFLAG antibody for the three individual clones and the whole library. Left side: after induction of the promoter with IPTG, right side control without induction.

**Figure 8.** Growth competition experiment with three selected clones providing a growth advantage (clones PEPNR00000000004 , PEPNR00000000032, PEPNR00000000600, see Table 3). A) Competition against empty vector separately with each of the clones. B) Competition in different combinations of vector and clones. C) Competition between vector and stop codon version of the respective clones. v = vector, 4 = clone 4, 32 = clone 32, 600 = clone 600. Note that in each experiment we mixed the cells of the corresponding clones in equal starting amounts, but the subsequent PCR favored different fragments to different extents. Hence, only the trajectories are comparable, not the absolute values. The X-axis denotes the four experimental cycles, the Y-axis the relative fraction of the respective sample, normalized to 1.

**Figure 9.** Competition experiments between clones with and without stop codons at the start of the respective random sequences. Experiments were conducted in the same way as the competition experiments described in Figure 8, but the resulting PCR fragments were sequenced in the end. The figure shows the trace files in the relevant regions where the two clone variants differ (i.e. at the engineered stop codon). The nucleotides that differ in the sop codon clones are indicated in small letters at the bottom of each panel. For clone 600 it is evident that the respective double peaks decline during the experiment in favor of the non-stop codon version, while this is not evident for clones 32 and 4.

**Figure 10.** Growth competition experiment with three selected clones providing a growth disadvantage. PEPNR00000000159, PEPNR00000000292, and PEPNR00000000419 from Table 4. Agarose gels showing PCR amplified fragments from different stages of the competition experiment between clone and empty vector - each experiment in three replicates. The top panel shows the input ratios (they were aimed to be equal but show some variation in individual replicates). The middle panel shows that the insert containing clones become reduced compared to the vector and the bottom panel from the end of the experiment shows a mere absence of the insert containing clones, showing that their growth was strongly suppressed compared to the vector growth.

**Figure 11.** Assessment of read depth on detection power based on rarefaction analysis. Top: Counts of enriched and depleted clones (green, red) at increasing read sampling depth. Bottom: Fraction of the total enriched and depleted clones found at 100% sampling detected at various sampling levels. At 10%, most of the depleted clones can be found (60%) and at 60% almost all have been detected (corresponds to 600,000 clones per experiment). From the clones detected as depleted in subsamples that are not present in the complete sample it was possible to estimate the percentage of detection error. The highest estimate is 5%, and this decreases with deeper sequencing.

The invention will now be described by reference to the following examples which are merely illustrative and are not to be construed as a limitation of the scope of the present invention.

**Example 1: Construction of a library of polynucleotides comprising different random nucleic acid sequences and cloning the same in an expression vector**

To construct a pool of expression vectors for the expression of a library of polynucleotides that comprise different random nucleic acid sequences, a library of polynucleotides with different random nucleic acid sequence was generated and subsequently cloned in the pFLAG-CTC™ expression vector.

The library of polynucleotides with different random polynucleotide sequences was generated as follows:

First, a pool of oligonucleotides, wherein each of said oligonucleotides comprises a different random nucleotide sequences of 150 nucleotides (N150) in length and 5' and 3' overhangs of a defined nucleic acid sequence (comprising restriction sites) has been generated by chemical oligonucleotide synthesis by state of the art nucleoside phophoramidite solid state synthesis through a commercial supplier (Metabion, Germany). To achieve a pool of oligonucleotides with randomized/random nucleotide sequences an equimolar mix of A, C, G and T was provided during successive chemical synthesis of each of the 150 random nucleotide positions. For synthesis of the nucleic acid sequences at the 5' and 3' end with defined nucleic acid sequences only the respective nucleotide of this position (e.g. A for position 1) was provided.

The synthesized oligonucleotides of the library have the following general/generic nucleotide sequence:

5´-ACGTCCAAGCTTAGC(N150)GCATTGGTCGACGTA-3´ (SEQ ID NO: 1)

wherein (N150) represents a nucleic acid sequence of 150 nucleotides and wherein each of said nucleotides is selected from the group consisting of A, C, T and G. Because of the large number of combinatorial possibilities, it is not expected that every possible combination of the four nucleotides at every position is actually to be found in the resulting individual molecules.

The synthesized pool of oligonucleotides was subsequently purified on a 8% acrylamide gel and amplified in a PCR reaction conditions using the following primers:

Oligo forward: 5'-ACGTCCAAGCTTAGC-3' (SEQ ID NO: 2)
Oligo reverse: 5'-TACGTCGACCAATGC-3' (SEQ ID NO: 3)

and AmpliTaq polymerase (Promega) under the buffer conditions recommended by the supplier (Promega) and a first PCR cycle containing only the reverse primer of 95°C 2min, 58°C 1min, 72°C 30sec; followed by adding the forward primer and 10 cycles of 95°C 2min, 58°C 1min, 72°C 3min.

The product resulting from this PCR was a corresponding pool/library of double-stranded polynucleotides as also the complementary strands to the oligonucleotides of the library was synthesized by the PCR.

Next, this library of double-stranded polynucleotides resulting from the PCR (subsequently referred to as the library of polynucleotides) was inserted into the pFLAG-CTC™ expression vector by cloning.

The pFLAG-CTC™ expression vector (see Figure 1) was commercially supplied by Sigma-Aldrich (catalog no. E8408). It is a 5.3 kb *E. coli* expression vector, which is typically used for cytoplasmic expression of a properly inserted open reading frame (ORF) as a fusion protein with a C-terminal FLAG® epitope tag. The FLAG epitope tag is a small, hydrophilic 8 amino acid tag (DYKDDDDK (SEQ ID NO: 4)) that provides for sensitive detection and high quality purification using ANTI-FLAG products (e.g. anti-FLAG antibodies). The pFLAG-CTC™ vector further comprises a tac promoter (Ptac),

which regulates the transcription of the nucleic acid sequence encoding for the respective ORF-FLAG fusion protein. The tac promoter is a hybrid of the trp and lac promoters from *E. coli*, which is regulated by the presence of the lacO sequences and inclusion of the lac repressor gene (lacI) on the vector. Importantly, the tac promoter allows for induction of the expression of a respective ORF-FLAG fusion protein by addition of *e.g.* IPTG to the culture media. In the absence of IPTG no or essentially no transcription takes place.

For the cloning, the pFLAG-CTC™ expression vector was subjected to a restriction digest with the restriction enzymes *Hind*III and *Sal*I. This removed the following sequence fragment from the multiple cloning site of the vector:

```
5'-AGCTTCTCGAGAATTCCCGGGTACCAGATCTG     -3' (SEQ ID NO: 8)
3'-    AGAGCTCTTAAGGGCCCATGGTCTAGACAGCT-5' (SEQ ID NO: 129)
```

Similarly, also the library of polynucleotides resulting from the above-mentioned PCR was subjected to a restriction digest with *HindIII* and *SalI*.

Subsequently, the digested library of polynucleotides and the digested pFLAG-CTC™ expression vector (lacking the above-mentioned sequence) were ligated using T4 DNA ligase at 16°C over night, following the experimental recommendations of the supplier (Promega). The product of the ligation reaction was a library of expression vectors, which allows expressing the generated polynucleotide library. Each expression vector comprising an insert, i.e. a polynucleotide of the library is in principle capable of giving rise to expression of a peptide with the following predicted generic peptide sequence (see lower sequence):

*MetLysLeu*Ser (aa50) AlaLeu*ValAspTyrLysAspAspAspAspLysSTOP* (SEQ ID NO: 5)

wherein (aa50) represents an amino acid sequence of 50 amino acids having a random amino acid composition at each of the 50 positions. In principle, each of these 50 positions can be represented by any amino acid. The 50 amino acids are encoded by the random nucleic acid sequence of the polynucleotides of the library.

Positions in italics represent amino acid sequences that are encoded by nucleic acid sequences provided by the vector and/or restriction sites, including the C-terminal FLAG sequence (the very C-terminal 8 amino acids).

**Example 2: Screening for polynucleotides having biological activity**

The library of expression vectors constructed in example 1 was subsequently employed to screen for novel polynucleotides that encode for peptides (or in principle also RNAs) with biological activity.

To this end, *E. coli* cells were transformed with the constructed library of expression vectors and subsequently cultured in LB-medium comprising the antibiotic selecting for transformants (here ampicillin) but otherwise non-selective (optimal) growth conditions. In this first growth phase the cells were pre-cultured in the absence of IPTG in order to amplify the transformed cells without expression of the peptides (and/or RNAs) encoded by the respective expression vectors, i.e. peptides comprising random amino acid sequences (and/or RNAs comprising random nucleic acid sequences). In a second growth phase, involving four culturing cycles (of different length depending on the experiment), the cells were cultured in the presence of IPTG, i.e. conditions under which the peptides (and/or RNAs) encoded by the expression vectors are expressed. During this second growth phase, samples of the culture were collected at different time points during culturing, as described further below. Finally, the frequency of different clones at these different time points was assessed by isolating the vectors/plasmids from each of the collected samples and assessing the frequency of individual vectors/plasmids in that sample by DNA sequencing of the respective insert sequences. The insert sequences expressed by those plasmids that get enriched or depleted during culturing are polynucleotides with biological activity as they promote or inhibit the proliferation of the respective host cell clone. In particular, the growth advantage in such cases is conferred by the RNA and/or peptide expressed from that random polynucleotide sequence.

Specifically the screening has been performed as described in the following:

**Cultivation experiments in *E. coli* cells transformed with the library of polynucleotides with randomized nucleic acid sequences**

*Library amplification*

The expression vector library prepared in Example 1 was transformed into *E. coli* DH10B cells through electro-transformation using the recommended procedure of the supplier of the competent cells (New England Biolabs). The transformed cells were then proliferated in LB medium plus ampicillin (AMP), but in the absence of IPTG until stationary phase (approx. $2x10^9$ cells per mL). In the absence of IPTG the cells could be amplified without expression of the library of polynucleotides comprised in the expression vectors. This has the advantage that potential growth advantages or disadvantages conferred by expression of individual polynucleotides of said library did not yet get apparent during the host cell library amplification step. After this initial amplification of the host cells comprising the expression vector library, the transformed cells were frozen at -80°C in 20% glycerol and subsequently used as a *library stock* for the growth competition experiments described below. Starting the different experiments as explained below from the *library stock* had the advantage that all experiments were started with the same host cell population comprising the same library of polynucleotides.

*Cultivation experiments*

Next, the *E. coli* cells transformed with the library of polynucleotides were subjected to cultivation experiments. Specifically, the cultivation experiments aimed to identify clones that would consistently show a frequency change across multiple culturing cycles, whereby all culturing cycles were run under the same culturing conditions. The following general setup was used (also illustrated in Figure 2) for all cultivation experiments (except for the indicated modifications):

- generate a pre-culture by inoculating 25mL LB medium (Invitrogen; cat. 12780-052) supplemented with 50µg/ml ampicillin (AMP) (Roth; cat. K029.1 - to restrict growth to *E. coli* cells bearing a vector of the expression vector library) with 500µL from the *library stock* (generated as described in above paragraph - volume corresponds to about $1x10^9$ cells) and grow over night at 37°C at 250rpm shaking conditions in Erlenmeyer flask to generate a *pre-culture*.

- inoculate up to 10 replicates, wherein each replicate is inoculated with 500µL of the *pre-culture* in 5mL fresh LB medium supplemented with AMP (50µg/ml) and IPTG (Sigma; cat. I1284; 1 mM final concentration) in 14mL tubes with snap lid (Falcon, 17 x 100mm, cat. 352057); grow over night at 37°C at 250rpm shaking conditions; this is culturing cycle #1. The presence of IPTG in the LB medium induces the expression of the RNAs and/or peptides encoded by the expression vector library.
- repeat the last step with the exception that one culture is inoculated per replicate using 500µL of the culture of the respective replicate resulting from culturing cycle#1. The overnight growth phase of these cultures represents culturing cycle 2#.
- repeat the last step, but use 500µL of the culture from the previous cycle until the end of the experiment. In the present example in total 4 culturing cycles have been performed for each of the experiments.

For the subsequent analysis, after each cycle 2 mL of the culture were used for plasmid isolation. Furthermore, 1 mL was stored in 20% glycerol at -80°C.

The above mentioned general experimental setup implies that when the cells from culturing cycle #1 have already grown in the presence of IPTG overnight, i.e. under conditions, in which the expression of the RNAs and/or peptides encoded by the library of polynucleotides has been induced, before the samples for the first time point of the subsequent analysis (which was used as a respective reference to assess the change in frequency of a individual polynucleotide) are removed from the replicates. Thus, the frequency of clones within this sample may already have slightly changed at this first time point in comparison to the starting frequency in the library or the pre-culture, because the expression of certain RNAs and/or peptides encoded by the polynucleotides of the library may confer a growth advantage or disadvantage to the respective clones compared to other clones. Hence, under these settings using the samples taken from culturing cycle 1# is used as first reference for the comparison of frequencies relies on growth competition trends that establish across the further culturing cycles (cycles 2# to cycles 4# in this particular example) Alternatively, it would in principle also be possible to take an additional sample of the *pre-culture* grown overnight and to use this sample as a respective first comparison reference. With this modification it would also be possible to monitor changes in frequency of clones within the first culturing cycle. However, in this case one would compare

cultures that grew under different culturing conditions (namely without and with IPTG). Further, the first growth cycle under IPTG is expected to remove peptides that would show an aggregation tendency with an unspecific blocking of cell growth. Such peptides are not desired since they would lack specificity for interacting with a specific cellular process.

Notably, in the above-mentioned setup inoculation at the beginning of each cycle was performed with a high number of cells (approx. $10^9$ cells, resulting in a final concentration of about $10^8$ cells per mL in the culture at the beginning of the next cycle) in order to assure that there is no dilution effect with respect to the number of clones in the library (approx. $5 \times 10^6$ - this number is estimated from the expected limit for the transformation efficiency of the cells; it was approximately confirmed by the sequencing results across all experiments). The inoculation with the high number of cells implies also that there are only about 4-5 generations until the new stationary phase is reached. In other words, the cells undergo between 4-5 cell divisions during each culturing cycle before the stationary growth phase is reached in the respective cycle.

In the present example in total seven experiments were performed according to this scheme. All experiments comprised four cycles. Specifically, three experiments in which each cycle lasted 4 hours (Exp1, Exp2, Exp3) and four experiments in which each cycle lasted 24 hours (i.e. a prolonged stationary phase) (Exp4, Exp5, Exp6, Exp7) were performed. Six of these 7 experiments were done with 10 replicates each (Exp1, Exp2, Exp3, Exp4, Exp5 and Exp6) and one with five replicates (Exp7). Further, a control experiment with 24h cycles and 10 replicates was performed. This control experiment was performed without IPTG induction; i.e. under similar conditions as the other seven experiments with the exception that during the 4 cycles medium without IPTG was employed.

**Example 3: Analysis of the samples collected from the cultivation experiments by DNA sequencing**

Next the samples collected during the cultivation experiments were analyzed. Specifically, this analysis aimed to determine the frequencies of the different clones resulting from the transformation of *E. coli* cells with the expression vector library at

each of the cycles in each replicate. By doing so, respective changes in frequency could be determined.

In order to determine the frequency of clones the frequency of the different expression vectors was determined. As the expression vectors differ in their inserts, in particular the nucleic acid sequences and respective frequencies of these inserts, which comprise random nucleic acid sequences, the frequencies were determined by DNA sequencing.

To this end, first, the plasmids comprised in the 2 mL samples collected after each cultivation cycle (also referred to as time point below) were isolated using the QIAGEN Plasmid Mini Kit (cat. 12125). The plasmid extraction was performed for each replicate and time point as described in the manufacturer's protocol.

After plasmid extraction, a DNA library for subsequent sequencing was prepared using the Illumina amplicon sequencing kit, with primers targeted to the periphery of the insert on the plasmid, and providing the primers for subsequent sequencing, using the following primers:

Forward primer (3`-end located 85nt before ATG start codon):
5'-ATGATACGGCGACCACCGAGATCTACACNNNNNNNNTATGGTAATTGTCATCATAACGGTTCTGGCAAATATTC-3' (SEQ ID NO: 6)
Reverse primer: (3´-end located 24nt after the stop codon)
5'-CAAGCAGAAGACGGCATACGAGATNNNNNNNNAGTCAGTCAGCCCTGTATCAGGCTGAAAATCTTCT-3' (SEQ ID NO: 7)

wherein the hexanucleotide with Ns indicates a region where sequencing barcodes were placed to be able to distinguish the different replicates if sequenced in parallel. After PCR, the samples were pooled together. Sequencing was carried out with an Illumina MiSeq sequencer, following the standard amplicon sequencing protocol as provided with the MiSeq Reagent Kit v3 for 600 cycles (cat. MS-102-3003) to produce 300 bp paired-end reads from each sequenced fragment.

*Data processing and analysis*

To derive the nucleic sequences and the frequencies of the polynucleotides the data received from the sequencing reaction were subsequently processed. Specifically,

Fastq paired-end reads were collapsed into a single fasta sequence using usearch (Edgar 2010). Whenever conflicting bases were detected between pairs, those with the best quality score were retained. Since the example experiment specifically aimed to analyse functional peptides encoded by the polynucleotide library (rather than the polynucleotides), open reading frames (ORFs) were obtained from each sequence, as well as the translated protein sequences using Unix shell commands. Only those ORFs starting and ending with the expected sequences (see above), and encoding exactly 65 amino acid residues (50 from the randomized sequences and 15 from the vector, including the tag) were retained for downstream analyses. Hence, although nucleic acid sequences were obtained by sequencing, we use the term peptides or protein sequence in the following, referring to the predicted translation product of the respective DNA sequence. A non-redundant database was constructed with usearch (Edgar 2010) for all experiments using protein sequences at 100% identity, i.e., similar non-identical sequences are treated as independent entries. This implies that this database includes translated sequences with possible sequencing errors. It was possible to estimate the error rates per sequencing run using the first 85nt of plasmid sequences in the reads. The reads were cropped to this length using Unix shell scripts, mapped to the reference plasmid sequence with NextGenMap (Sedlazeck et al. 2013), and determined the percentage of mismatches using samtools fillmd (Li et al. 2009) to assess substitutions as a proxy for errors. We found error rates in the range between 0.12-0.56%. Given these low rates, we did not try to curate the database further, although this implies that some peptides in the database are not real, but it is expected that these would not contribute to the analysis, since they should occur as singletons.
The sequences of each replicate in each experiment were matched to the database using diamond (Buchfink et al. 2015). This provided a quantitative representation of each sequence in each cycle and each replicate, as well as across experiments. These counts were used to statistically compare the changes in number and frequency of each peptide sequence over time.

*Statistical procedure*

The number of times a peptide was observed was recorded and the counts for each time point were determined. Very low frequency peptides cannot be statistically analyzed. For this reason, occurrence of at least five times or more in any one replicate of an experiment to consider a statistical analysis of a given peptide was

required. The further statistical analysis was based on the procedures described by Love et al. (2014) designed for differential gene expression, but applicable to any type of count data, in particular those derived from high-throughput sequencing experiments. The analysis was done using the R package DESeq2 (Love et al. 2014) accessible at https://bioconductor.org/packages/release/bioc/html/DESeq2.html

Size factors were applied to each replicate to account for differences in depth of sampling (arising from the experiment or sequencing procedures) to allow a comparison across sequencing data of different depth for different replicates. The size factors were estimated using the median of ratios between the individual peptides in each replicate and a representative pseudo-sample obtained from the geometric mean across all pair-wise peptide comparisons of all replicates in an experiment. Then, a generalized linear model was fitted to each peptide in an experiment (i.e. across replicates) using the negative binomial distribution, assuming that the mean of the observations is representative of the frequency of the peptide. From the fitting it was possible to obtain the overall frequency of the peptide, an estimate of the log2-fold change of frequency between the tested time points, and the standard error of the log2-fold change. A Wald test (Wald 1943) was performed to test whether the log2-fold change was significantly different from zero: the log2-fold change was divided by the standard error, and the resulting z-statistic was compared to a normal distribution to obtain a p-value. P-values were corrected for multiple testing after testing all clusters using the Benjamini-Hochberg (1995) procedure for a false discovery rate of 5%.

Power analysis was done according to Ching et al. (2014). Based on the observed dispersion in our data (median ranging between 0.04-0.09 for 4h cycles and 0.13-0.47 in 24-hour cycles). Under these conditions and with 5-10 replicates, the depth of sequencing should be at least 5 million sequenced fragments to reach statistical significance with saturation at approximately 20 million fragments. Our experiments were sequenced at a minimum depth of about 15 million fragments (paired-end sequenced) in Exp4, and reached up to 41 million in Exp7.

A control experiment was performed to test whether any major frequency changes could be detected when the expression was not induced by IPTG. Figure 3 shows the effect of induction with IPTG compared to replicate cultures of the same experiment

without induction. The induced experiment showed major significant shifts in peptide frequencies over time, both negative and positive, while the non-induced experiment showed only minor non-significant variation. This proves that expression of the peptides is strictly required to cause frequency changes.

For all experiments peptides with significantly different frequencies between the first time point and the last time point were recorded. By including comparisons at the other two time points, they were further categorized into increasing or decreasing in frequency across time points when the tendency was consistent. Two clear examples each for peptides that increase and decrease in frequency during culturing over time are shown in Figure 4.

It is particularly advantageous that the general procedure allows a parallel confirmation across all significant peptides by repeating the whole experiment with the same number of replicates within the experiment. Clusters (and their corresponding peptides) are considered as having been confirmed when they show the same trends in a second experiment.

*Rarefaction analyses*

Experiment E7, which was sequenced intensively, was used to estimate the effects of sampling on the discovery of enrichment or depletion. All replicates were normalized at 50,000 peptides each, thus giving the whole experiment a total number of 1 million sequences. Random subsamples were obtained at 10% intervals. Subsampled experiments were analyzed in the same way as full experiments (described above).

*Sequence properties*

To estimate whether the enriched or depleted peptides in the experiments behave like random sequences or biological sequences, random 150nt sequences were simulated, and the vector information was added to obtain a translation like the one performed in the experiments. Further, all protein sequences from *E. coli* deposited in the GenBank were used as 65aa fragments to act as biological controls. Compositional properties were calculated using the package protr (Xiao et al. 2015) and Wilcoxon rank tests were run to compare properties of experimental peptides to random and biological sequences.

*Results of experiments*

Across all experiments we found a large number of full length peptides in the library. However, most of these occurred only at low levels, the majority were detected only once across all experiments. Since these rare clones preclude statistical analysis, we focused only on those for which at least five counts were observed in at least one of the parallel 5-10 replicates. This reduced the number to 1,082 peptides that could be evaluated in at least one experiment, and all further statistical analyses are based on this number of peptides.

Among the 1,082 peptides statistically analysed, a large number showed an increase or decrease in frequency in the experiments under induction conditions. Most of them showed a consistent increase or decrease across the four experimental cycles.

The overall results are summarized in Table 2. In the three experiments with the 4h cycles (Exp1-Exp3), around 70% of the analyzed peptides showed a significant change in frequency, with about 3 to 4-fold more peptides going down in frequency than going up. The 24h cycle experiments are more heterogeneous with respect to these overall percentages, but maintain the same overall trend.

The observation of a higher proportion of depleted peptides depends on sequencing depth, combined with the statistical requirement of a certain depth of coverage for each peptide to detect a significant change. Identifying statistically significant depletion is easier when a peptide has already a high starting frequency - and such peptides will also be found at lower sequencing depth. Significant increases in frequency, on the other hand, can start at very low initial frequencies and these will become more visible at higher sequencing depth. To test this, an additional experiment (Exp7) was conducted with four 24h cycles, including five replicates (instead of 10) and a deeper sequencing effort. At this higher sequencing depth we find an overall fraction of significant peptides similar to most of the other experiments (79%), but indeed also a higher relative number of enriched peptides (Table 2).

Using the data from experiment E7 we could also estimate the power to detect depleted versus enriched peptides/clones. We performed analyses for subsets at 10%, 50%, 80% and 100% of the total reads in the experiment. Figure 5 shows the fold-change plots for four depths of sampling. We find that from 50% onwards more and more initially low frequency peptides/clones become significantly enriched, as is

expected from the higher statistical power. Figure 11 shows the corresponding rarefaction analysis where the detection of depleted clones is more or less complete at 60% coverage, while the detection of enriched clones keeps rising.

The peptides with significant changes in frequencies do not share any similarities with each other, in line with the notion that they constitute independent draws out of a random sample. The comparison of the amino acid composition of all analyzed peptides with random and with biological controls showed for almost every case where the biological control deviates much from random expectation, that the peptides in our study are closer to the random control (Figure 6). Similarly, there is no significant difference between up and down regulated peptides with respect to amino acid composition (Figure 6). Still, minor differences were found for some comparisons in the relative frequency of certain amino acids such as less R, D, C, G, S, V and more N, E, Q, I and T in the real *E. coli* protein sequences.

The list of confirmed peptides that showed significant increase or decrease in frequency is provided in Tables 3 and 4 (and/or in SEQ ID NOs: 9 to 68). These tables also show the respective corresponding polynucleotide sequences encoding the peptide sequences (see SEQ ID NOs: 69 to 128). The respective RNA sequences encoded by the polynucleotides (see SEQ ID NOs: 69 to 128 are identical to the polynucleotide sequences shown with the only exception that T has to be replaced by U.

**Example 4: Validation of individual clones**

In a next step to illustrate the high reliability of the screening method of the present invention the identified bioactivity confirmed by some of the identified polynucleotides was assessed. Specifically, three expression vectors of the screened polynucleotide library comprising random polynucleotides that have been identified to be enriched in frequency during cultivation (i.e. confer a growth advantage to the cells) and three expression vectors of the screened polynucleotide library comprising random polynucleotides that have been identified to be decreased in frequency during cultivation (i.e. confer a growth disadvantage to the cells) were isolated and retransformed into the original host *E. coli* cells. Using these freshly transformed *E. coli* cells either in isolation or in less complex mixtures (e.g. together with cells comprising an empty vector control) the bioactivity of the selected candidates was again assessed in culturing assays.

Moreover, to assess whether the growth-promoting bioactivity of the three random polynucleotides retested in the validation experiments depends on the encoded peptide sequence or rather on the RNA sequence encoded by the random polynucleotide sequence also experiments employing variants of the identified polynucleotides in which a premature stop codon has been introduced before the start of the random polynucleotide sequence part were conducted. In these experiments it was expected that if the bioactivity was conferred by the encoded peptide sequence the bioactivity should be absent in the presence of a premature stop codon.

For the validation experiment the following three clones identified as enriched (i.e. as growth-promoting) in the screening of Examples 1 to 3 above were analyzed: PEPNR00000000004 (also referred to as clone 4 herein elsewhere), PEPNR00000000032 (also referred to as clone 32 herein elsewhere), PEPNR00000000600 (also referred to as clone 600 herein elsewhere).

The sequences of the peptides encoded by these clones and of the polynucleotides encoding these peptides are the following (amino acid sequence: nucleotide sequence; randomized part thereof is underlined) and are also shown in Table 3:

PEPNR00000000004

MKLSPVSWIHGATAQSGGLSLRLAVRSGIDGCAWFIRAECGGARALSDGPGVSYALVDYKDDDDK(SEQ ID NO:10):ATGAAGCTTAGCCCCGTCTCCTGGATTCACGGTGCTACCGCTCAGTCTGGAGGATTATCCCTCAGGCTTGCAGTCCGCTCAGGAATAGATGGGTGTGCATGGTTCATCAGGGCTGAATGCGGAGGGGCTCGTGCGCTTTCAGACGGGCCTGGGGTAAGCTATGCATTGGTCGACTACAAGGACGATGACGACAAG(SEQ ID NO:70)

PEPNR00000000032

MKLSYWNSSMASGDIRALVFDSGGGLIFLRHQLAGWWACLFPLLASREARFDTDALVDYKDDDDK(SEQ ID NO:14):ATGAAGCTTAGCTACTGGAATAGCTCTATGGCGTCGGGGGATATCCGTGCTCTTGTGTTTGATTCAGGCGGAGGCTTAATATTCCTTCGGCATCAGCTGGCGGGGTGGTGGGCCTGTTTGTTTCCGCTACTGGCATCGCGGGAGGCACGGTTTGATACCGACGCATTGGTCGACTACAAGGACGATGACGACAAG(SEQ ID NO:74)

PEPNR00000000600

MKLSRGIHLGRTSTCVNASYALCHTYRSARRGKSRKRGRSSPPIGTSLVHWVLDALVDYKDDDDK(SEQ ID NO:24):ATGAAGCTTAGCCGCGGTATTCACCTAGGTCGGACGAGTACATGCGTCAACGCTTCGTACGCACTCTGCCACACGTACCGTTCAGCCCGCCGTGGCAAGTCCAGGAAGAGGGGGAGGAGTTCACCACCGATCGGGACCTCTTTAGTACACTGGGTTTTGGACGCATTGGTCGACTACAAGGACGATGACGACAAG(SEQ ID NO:84)

and the following three clones identified as decreased in the screening of Examples 1 to 3 above were analyzed: PEPNR00000000159 (also referred to as clone 159 herein elsewhere), PEPNR00000000292 (also referred to as clone 292 herein elsewhere), PEPNR00000000419 (also referred to as clone 419 herein elsewhere).

The sequences of the peptides encoded by these clones and of the polynucleotides encoding these peptides are the following (amino acid sequence: nucleotide sequence; randomized part thereof is underlined):

PEPNR00000000159
MKLSVGKPDTWLHRARGAVWVRIAGNVSLGMGLRGLSGRLPCVCGPLRTFGAFEALVDYKDDDDK(SEQ ID NO:38):ATGAAGCTTAGCGTTGGGAAACCGGATACATGGCTCCATAGAGCCAGAGGAGCAGTTTGG GTTAGGATTGCCGGGAACGTGTCATTGGGTATGGGACTGCGTGGTTTGTCTGGTCGGCTCCCATGTGTAT GTGGGCCTCTCAGGACCTTTGGGGCCTTCGAGGCATTGGTCGACTACAAGGACGATGACGACAAG(SEQ ID NO:98)
PEPNR00000000292
MKLSAATWVASLRVAFGGDLILRLIRYQAAGRSGALDQFYEANSILGVHRRTRDALVDYKDDDDK(SEQ ID NO:47):ATGAAGCTTAGCGCGGCTACCTGGGTCGCGAGTCTCCGAGTTGCCTTCGGTGGGGACCTT ATTCTGCGGTTAATCAGATATCAGGCGGCAGGGCGAAGCGGAGCGCTCGACCAGTTTTATGAAGCGAACT CCATACTAGGTGTCCACAGGCGTACGCGAGATGCATTGGTCGACTACAAGGACGATGACGACAAG(SEQ ID NO:107)
PEPNR00000000419
MKLSCPFPDTHGAICCRVWSGFALIVLRLLDAIGSCGRHGGVGHALAEHVFWVCALVDYKDDDDK(SEQ ID NO:58):ATGAAGCTTAGCTGTCCATTTCCGGATACCCATGGCGCGATCTGCTGCCGTGTTTGGTCC GGGTTCGCGTTGATTGTTCTGCGTTTGCTCGACGCCATAGGGTCCTGCGGCAGGCATGGTGGTGTGGGCC ACGCTCTAGCCGAGCACGTCTTCTGGGTGTGTGCATTGGTCGACTACAAGGACGATGACGACAAG(SEQ ID NO:118)

In the present example the following methods have been used

*Single clone recovery*

To isolate the expression vectors comprising the three above-mentioned selected polynucleotide sequences from the random polynucleotide library employed for the screening as described in Examples 1 to 3, above. The isolation was achieved with specific primer pairs and a PCR amplification of the whole expression plasmids.

Moreover, in parallel also expression plasmids comprising the above-mentioned mutant variants of the selected polynucleotides that comprise premature stop codons were generated. The mutation(s) resulting in a premature stop codon were introduced

by using different primers for amplifying the expression plasmids from the polynucleotide library that introduced the respective mutation(s) to introduce a premature stop codon.

To obtain single clones from the library, PCR primers based on the determined sequences of the clones facing outward of each other were used. Stop codons at the desired positions were engineered by modifying one of the primers at its 5`-end to code for a stop codon. Amplification then yields the full vector that needs only to be religated. However, to ensure that the vector had not suffered a mutation, the inserts of the recovered clones were re-cloned into the original vector and transformed into the original *E. coli* host cells.

*Confirmation that encoded peptides are expressed in the newly generated clones*

Western blotting was employed to check for the expression of the respective peptides in the newly generated clones. The analysis was exemplary performed for the clones comprising a polynucleotide with cell growth promoting activity. To this end 1.4 mL overnight culture were spun down and resuspended in 100µL Laemmli buffer with 5% β-mercaptoethanol, then samples were incubated at 99°C for 5 minutes and debris was centrifuged down. 30µL were loaded onto a 4-20% tris-glycine gel (Bio-Rad) and run for 1 hour 40 minutes at 70 volts. The proteins were then transferred to PVDF membrane for 15 minutes at 13 volts using a Bio-Rad semi-dry electroblot unit. The membrane was washed 2 x 10 minutes with gentle shaking in PBS with 0.1% tween 20 (PBST) and then blocked in 5% powdered milk (1% fat) dissolved in PBST with shaking at room temperature for one hour. The monoclonal mouse anti-FLAG M2 antibody (F1804 Sigma) was added, diluted 1 in 2000 in 2.5% milk PBST. The membrane was incubated overnight with shaking in a cold room (approx. 6 °C). The membrane was washed 3 x 10 minutes in PBST with shaking. Goat-anti mouse HRP (A16072 Thermo-Fisher) diluted 1 in 2500 in 2.5% milk PBST was added and incubated with shaking at room temperature for one hour. The membrane was washed 3 x 10 minutes in PBST with shaking. ECL (Clarity Western ECL from Bio-Rad) was pipetted onto the blot (approximately 3 mL per blot) and incubated for 5 minutes, then blotted with thick filter paper and protected from light. The membrane was the imaged using a digital imager (Alpha Innotech) with increasing exposures until bands were well visible.

Single clone competition experiments

The competition experiments with individual clones or combinations thereof were done under the same conditions as described in Example 2. The competition assays comprised the following steps: (1) Create a starting culture from each clone in 25ml LB plus Ampicillin (Amp, 50µg/ml) by growing overnight at 37°C, 500 rpm; (2) mix equal volumes of the clones to be tested in a total volume of 500µl; (3) add this mixture to 4.5 ml of LB+Amp+IPTG(1mM) to make it a total 5 ml (1/10 dilution); (4) incubate 3h or 24h at 37°C, 500 rpm; (5) take 500µl of the respective culture and repeat step 3; (6) generate a total of four cycles.

But instead of proceeding with a sequencing step as with the original experiments with the whole library (see above), the expression vector inserts comprising the random polynucleotide part were amplified by PCR and the products were ran on an agarose gel and an Agilent Biochip (DNA 7500) when quantification was required. The PCR primers used were chosen such that a 349bp fragment would be generated from the vector without insert and a 449bp fragment when the vector contained an insert. To distinguish the sizes of the clones with inserts, we digested them first with diagnostic restriction enzymes. The Agilent software for quantification of the bands was then used to obtain concentration differences of the fragments between time points.

**Results**

All the experiments presented in Examples 2 and analyzed in Example 3 above were conducted in the context of a large mixture of clones. To see whether the bioactivity patterns identified in these screening experiments could be confirmed, six clones were assessed individually or in less complex mixtures. The clones mentioned above were selected and the respective expression plasmids were isolated from the polynucleotide library by PCR as described above. To re-generate the respective *E. coli* clones, these expression plasmids were transformed into *E. coli* cells, respectively.. Subsequent Western blotting of a whole cell lysate of three of the clones showed that all three of the regenerated clones express a peptide in dependence of IPTG, albeit at different steady state levels, which could reflect different overall stability of the expressed peptides (Figure 7).

To test whether the clones have an advantage with respect to clones harboring only an empty plasmid, a standard 4-cycle culturing experiment as described above was ran, the inserts of the expression vector were amplified and the DNA was quantified on an

agarose gel or an Agilent gel chip. In line, with the results of the screening experiments shown in Example 2 and 3, it was found that the three clones providing a growth advantage show an increase in frequency over time (Figure 8A) and the three clones providing a growth disadvantage get lost over time (Figure 10) when cultured together with a clone comprising an empty vector. For the three clones providing a growth advantage, it was tested whether they would also show this in competition with each other in different combinations. This was indeed the case; all were better than the empty vector control (Figure 8B).

Given that the bioactivity could be conveyed either by the transcribed RNA or the translated peptide, versions of the three clones providing a growth advantage harboring a premature stop codon directly at the start of the random part of the sequence (i.e. only the first four amino acids that are common among all clones would be translated) were produced (see above) and analyzed. These mutated clones were also tested in pairwise competition assays with the empty vector as described above for the respective originally identified sequences. Only one of the clones (PEPNR00000000600 - Table 3 herein) showed a clear difference between the mutated and the non-mutated version (Figure 8C), which would suggest that only this clone exerts its effect via the encoded peptide, while the two other clones might act through the RNA encoded by the respective polynucleotide alone. To study this in more detail, an experiment with a direct competition of each original clone with its stop codon counterpart was conducted. Confirming the previous result, the same qualitative results were obtained in these experiments (see Figure 9).

For the three clones providing a growth disadvantage, the results were already clear from the analysis of the corresponding agarose gel, i.e. no relative quantification was required. Figure 10 shows the respective gel representing the input mixture, the change after the first cycle and the end result after the last cycle. It is clearly evident that the polynucleotides providing a growth disadvantage are depleted over time during the culturing experiments.

*Summary*

In summary, the validation experiments shown in the present Example illustrate the power of the screening method above to identify random polynucleotides encoding for

RNAs and/or peptides having biological activity. All of the re-tested clones showed the expected effect on cell growth, also when introduced into a novel *E. coli* cell. Accordingly, the present Example also illustrated the power to affect cell growth by applying the polynucleotides identified in the context of the present invention to encode for bioactive biomolecules such as RNA and/or peptides. Further, the achieved results suggest that both polynucleotides encoding for bioactive peptides and bioactive RNAs may be identified with a method according to the present invention.

**Discussion/Conclusions**

The experiments shown herein above demonstrate that an unexpectedly large fraction of random RNA or peptide sequences are bioactive, at least in the sense of influencing relative growth rates in *E. coli* cells. In line with this finding, the present invention provides a number of such bioactive or growth rate-influencing RNAs, peptides and polynucleotides encoding these RNAs and/or peptides (summarized in Tables 3 and 4; peptides depicted in SEQ ID NOs: 9 to 68; polynucleotides depicted in SEQ ID NOs: 69 to 128). The results imply that it could be either the RNA encoded by random DNA sequences itself, or the corresponding translated protein that conveys the bioactivity. Although two of the three growth promoting clones individually tested in Example 4 suggest that the RNA function could be more important than the protein function, this constitutes at present only a small sample and may not be indicative of the true ratio between RNA and peptide functions. However, this observation fits well with the notion that an active RNA may precede an active peptide during de novo gene evolution of genes (Ruiz-Orera, J., Messeguer, X., Subirana, J. A. & Alba, M. M. Long non-coding RNAs as a source of new peptides. *Elife* **3**, doi:10.7554/eLife.03523 (2014)).

Previous studies have shown that specific biochemical activities or specific resistance against stress conditions can be recovered from random peptides using very large sample sizes and directed selection experiments (Stepanov, V. G. & Fox, G. E. Stress-driven in vivo selection of a functional mini-gene from a randomized DNA library expressing combinatorial peptides in Escherichia coli. *Molecular Biology and Evolution* **24**, 1480-1491, doi:10.1093/molbev/msm067 (2007); Keefe, A. D. & Szostak, J. W. Functional proteins from a random-sequence library. *Nature* **410**, 715-718, doi:10.1038/35070613 (2001). However, this occurred only at low frequencies and required multiple rounds of selection. Our results show that a non-directional approach without selective pressure in which only proxies for biological fitness are considered,

rather than specific activities, recovers a much higher fraction of bioactive RNAs and peptides and thus allows a better understanding of the functional potential of the unexplored random sequence space for molecular innovation.

Almost any random RNA could fold into a higher order structure, or interact with other RNAs via base pairing, although the free energies and interaction would be expected to be weak. For peptides, one could expect that they interact via charged or hydrophobic interactions with other molecules. They would not need to fold into a stable structure to do this. Many proteins exist that are partly or fully made up of intrinsically disordered protein regions (Tompa, P., Schad, E., Tantos, A. & Kalmar, L. Intrinsically disordered proteins: emerging interaction specialists. *Current Opinion in Structural Biology* **35**, 49-59, doi:10.1016/j.sbi.2015.08.009 (2015)). One can assume that such disordered peptides or proteins can associate with molecular complexes and can influence their activity. This can be more or less specific, i.e. only a single complex, or multiple complexes are affected (Cumberworth, A., Lamour, G., Babu, M. M. & Gsponer, J. Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. *Biochemical Journal* **454**, 361-369, doi:10.1042/bj20130545 (2013),Tompa, P., Davey, N. E., Gibson, T. J. & Babu, M. M. A Million Peptide Motifs for the Molecular Biologist. *Molecular Cell* **55**, 161-169, doi:10.1016/j.molcel.2014.05.032 (2014)). But as long as the specific effect or activity is reproducible, it becomes functionally - and thus also evolutionarily - relevant.

Negative effects of expressed peptides may not always be be very specific, given that a strong promoter is used in the expression vector and that some peptides might simply aggregate and thus harm the cell. However, since the first frequency measurement in the experiments was taken at the end of cycle #1 where cells have already grown under induction conditions, it can be expect that very strongly deleterious peptides are already mostly lost and cannot be detected. Hence, it is assumed that even a lot of the negative effects on cell growth are specific, i.e. in the sense that they do not simply block the whole cell physiology. In general, it should be emphasized that irrespective of the mechanism, the polynucleotides and the encoded RNA and/or peptides with positive or negative effects on cells growth are commercially interesting and may be employed for the uses indicated herein elsewhere.

In the evolution experiments of the present invention a very high reproducibility both with respect to trends across cycles, but also between and within experimental setups

was observed. Further, it could be demonstrated for six clones that their effects are also measurable in isolation. Thus, the results suggest that a large fraction of all possible random sequences may have some biochemical activity of biological relevance, in particular in *E. coli*, but possibly also in any other cellular organisms. It should be noted that the findings presented and the newly identified biomolecules also have practical implications. Such molecules can be seen as novel probes for studying cell physiology and growth, in particular when the target cellular pathways are identified. Random expression libraries described herein can also be used for screening approaches similar to the ones that have been developed for short hairpin RNA libraries (shRNA; Sims, D. *et al.* High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing. *Genome Biology* **12**, doi:10.1186/gb-2011-12-10-r104 (2011)), to identify specific RNAs or peptides that influence particular pathways or physiological states. This could also lead to novel procedures to identify pharmaceutically relevant molecules, e.g. with anti-bacterial activity.

The present invention refers to the following tables:

**Table 1: Cellular growth dynamics with different levels of growth advantage**

A single dividing cell produces about $10^9$ descendants after 30 cell divisions (column "normal") within a given timeframe. A cell with a 5% growth advantage would have given raise to 4.3 times as many cells at this stage (columns "adv. 5%" and "fold 5%") and a cell with 10% growth advantage 17.5 times as many (etc.). All numbers are rounded.

| cell divisions | normal [N] | adv. 5% [N] | fold 5% | adv. 10% [N] | fold 10% | adv. 50% [N] | fold 50% |
|---|---|---|---|---|---|---|---|
| **0** | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **4** | 16 | 19 | 1.2 | 23 | 1.5 | 81 | 5.1 |
| **16** | $6.5x10^4$ | $1.4x10^5$ | 2.2 | $3.0x10^5$ | 4.6 | $4.3x10^7$ | 657 |
| **25** | $3.4x10^7$ | $1.1x10^8$ | 3.4 | $3.6x10^8$ | 10.9 | $8.5x10^{11}$ | 25251 |
| **30** | $1.1x10^9$ | $4.6x10^9$ | 4.3 | $1.9x10^{10}$ | 17.5 | $2.1x10^{14}$ | 191751 |

**Table 2: Summary across the seven different experiments in the appended examples.**

| | 4h cycles | | | 24h cycles | | | | |
|---|---|---|---|---|---|---|---|---|
| experiment | E1 | E2 | E3 | E4 | E5 | E6 | E7[3] | **total[4]** |
| reads [1] | 457,212 | 324,090 | 678,823 | 293,638 | 300,807 | 812,315 | 4,142,836 | **7,009,721** |
| analyzed [2] | 623 | 529 | 607 | 616 | 499 | 618 | 1,061 | **1,082** |
| significant at FDR 0.05 | 83% | 84% | 85% | 65% | 47% | 80% | 79% | 77% |
| enriched | 115 | 90 | 115 | 130 | 71 | 181 | 320 | **402** |
| depleted | 399 | 353 | 399 | 272 | 162 | 313 | 518 | **566** |

[1] total number of full length peptide reads in each experiment
[2] number of peptides with at least 5 reads in at least one replicate in any experiment
[3] experiment with only five replicates, resulting in ten-fold higher read coverage per replicate
[4] for the reads the sum of all reads; for the peptides the sum of different peptides across experiments

**Table 3: List of clones/peptides that increased in frequency in at least two experiments - designated as (up_conf), or in only one experiment designated as (up). The corresponding nucleotide sequences are listed after the colon for each peptide. The randomized part of the amino acid sequences and polynucleotide sequences are underlined.**

PEPNR00000000003 (up_conf)
MKLSSGGTRGLSLVRGLLAVAGGSHAVLPVSGSVPSPRSQSPSFRASIMGQGLAALVDYKDDDDK**(SEQ ID NO:9)**: ATGAAGCTTAGCTCTGGTGGAACCCGTGGTTTAAGTTTGGTGCGCGGGCTGCTCGCTGTCGC CGGTGGTTCTCATGCAGTGTTACCTGTGTCTGGTAGCGTGCCTTCTCCTCGTTCACAGTCTCCGTCTTTC CGCGCGTCGATTATGGGCCAGGGTCTTGCGGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:69)**

PEPNR00000000004 (up_conf)
MKLSPVSWIHGATAQSGGLSLRLAVRSGIDGCAWFIRAECGGARALSDGPGVSYALVDYKDDDDK**(SEQ ID NO:10)**: ATGAAGCTTAGCCCCGTCTCCTGGATTCACGGTGCTACCGCTCAGTCTGGAGGATTATCCC TCAGGCTTGCAGTCCGCTCAGGAATAGATGGGTGTGCATGGTTCATCAGGGCTGAATGCGGAGGGGCTCG TGCGCTTTCAGACGGGCCTGGGGTAAGCTATGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:70)**

PEPNR00000000012 (up_conf)
MKLSSDQDRYVKSELWSGTSFSGGRGPVSRWRHRRLKCLAQHLAYVRDACVDDAALVDYKDDDDK**(SEQ ID NO:11)**: ATGAAGCTTAGCTCCGATCAGGATAGATATGTCAAGTCGGAGCTTTGGAGTGGTACTTCCT TCTCTGGGGGGAGGGGCCCTGTTAGCCGGTGGCGGCACCGGAGACTAAAGTGCCTTGCCCAGCATCTTGC TTACGTTCGGGACGCTTGTGTGGATGACGCAGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:71)**

PEPNR00000000024 (up_conf)
MKLSGSRRVYPGRTSSGCKARRGSCLVCCYSFSALPGGISELFRARTLSSGGSAALVDYKDDDDK**(SEQ ID NO:12)**: ATGAAGCTTAGCGGCAGCCGGCGGGTATATCCAGGTCGCACGAGCTCGGGTTGTAAAGCTC GTCGCGGCTCATGCTTAGTATGTTGTTACTCCTTCAGTGCGTTGCCAGGCGGGATATCTGAGCTGTTCAG AGCTCGCACCCTCAGTTCTGGTGGATCCGCAGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:72)**

PEPNR00000000029(up_conf)
MKLSLCKSYAYFAYPTTSSEHSMSALGVSGGSAVTSSGLNECATTLACSSGFSCALVDYKDDDDK**(SEQ ID NO:13)**: ATGAAGCTTAGCCTCTGTAAATCCTATGCGTATTTCGCGTATCCGACCACAAGTAGCGAGC ACTCGATGTCGGCCCTTGGTGTGTCTGGGGGCTCTGCAGTAACGTCTTCGGGCCTGAACGAGTGCGCCAC TACATTGGCGTGTTCGTCGGGTTTTTCATGTGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:73)**

PEPNR00000000032 (up_conf)
MKLSYWNSSMASGDIRALVFDSGGGLIFLRHQLAGWWACLFPLLASREARFDTDALVDYKDDDDK**(SEQ ID NO:14)**: ATGAAGCTTAGCTACTGGAATAGCTCTATGGCGTCGGGGGATATCCGTGCTCTTGTGTTTG ATTCAGGCGGAGGCTTAATATTCCTTCGGCATCAGCTGGCGGGGTGGTGGGCCTGTTTGTTTCCGCTACT GGCATCGCGGGAGGCACGGTTTGATACCGACGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:74)**

PEPNR00000000037 (up_conf)
MKLSWKWLAASPSCQSARCQADGVSGGNYRGSLAPRGPRVGCFTRILVASMTAVALVDYKDDDDK**(SEQ ID NO:15)**: ATGAAGCTTAGCTGGAAATGGCTAGCAGCATCCCCTTCGTGTCAGTCTGCTCGTTGTCAAG CAGATGGTGTTTCTGGCGGGAATTACAGGGGCAGTTTGGCGCCTCGAGGACCGCGAGTTGGCTGTTTTAC GCGCATCCTCGTCGCTTCTATGACCGCGGTTGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:75)**

PEPNR00000000055 (up_conf)
MKLSVSSNDRKVAGVWASPRRSGEVPSTDGCGLRRAWARSFRLVSTSHFVDAKCALVDYKDDDDK**(SEQ ID NO:16)**: ATGAAGCTTAGCGTATCTTCTAATGATAGAAAAGTTGCAGGCGTATGGGCCTCGCCTCGAC GAAGTGGCGAGGTTCCGTCAACCGACGGATGTGGTCTGCGGAGAGCCTGGGCACGGTCATTTCGCTTGGT TTCTACTTCGCATTTTGTAGATGCTAAATGTGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:76)**

PEPNR00000000086 (up_conf)
MKLSLQVSRRSTVPSHMKCTCLLGAPRAWPGLPSIGRALVPVVEVCNWYLLFPPALVDYKDDDDK**(SEQ ID NO:17)**: ATGAAGCTTAGCTTGCAGGTATCTCGGCGTAGCACTGTTCCTTCGCATATGAAGTGCACTT GTCTCCTCGGGGCGCCGCGTGCGTGGCCAGGCCTTCCGTCTATCGGGCGGGCTTTAGTGCCCGTGGTAGA GGTATGTAATTGGTATCTCCTATTTCCTCCGGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:77)**

PEPNR00000000102 (up_conf)
MKLSWAPRPGKQAECTCNFKLVRIPRRSFREEIAKPSWFCFSQLVAWRGGLLFSALVDYKDDDDK**(SEQ ID NO:18)**: ATGAAGCTTAGCTGGGCACCGCGACCCGGTAAGCAAGCTGAATGCACGTGCAACTTTAAAT TGGTGCGTATACCAAGGCGTTCGTTCCGAGAGGAAATAGCGAAACCCAGTTGGTTTTGCTTCTCACAGCT GGTGGCCTGGCGCGGCGGCCTGCTATTTTCGGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:78)**

PEPNR00000000104 (up_conf)
MKLSCGNSGTMYAALVVAGRGRGGAMPPGVGGSRRRVSFCQVGTRKTLQSRPFHALVDYKDDDDK**(SEQ ID NO:19)**: ATGAAGCTTAGCTGTGGTAACTCTGGAACCATGTACGCCGCTCTTGTGGTAGCGGGTCGCG GGCGTGGGGGTGCCATGCCGCCGGGGGTTGGTGGGTCCCGACGAAGGGTATCTTTCTGTCAGGTCGGGAC AAGAAAAACACTCCAGTCGCGTCCCTTTCACGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:79)**

PEPNR00000000110 (up_conf)
MKLSGVGSDLRTTLAENLGKPLRLQREPLIVVVIGRVGIPSSKYPHAACAWSRCALVDYKDDDDK**(SEQ ID NO:20)**: ATGAAGCTTAGCGGAGTGGGGTCCGATTTGCGGACTACCCTCGCTGAGAATCTAGGCAAAC CCCTGAGGCTTCAACGCGAGCCTCTTATCGTGGTTGTAATAGGAAGGGTGGGGATCCCTTCGTCCAAATA CCCTCATGCCGCCTGCGCCTGGAGTAGATGCGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:80)**

PEPNR00000000171 (up_conf)
MKLSVGVAERISSCGAQIFVRGDYGASAVYSGSRQVVFTCVRRGHAPGLLGGHCALVDYKDDDDK**(SEQ ID NO:21)**: ATGAAGCTTAGCGTAGGAGTAGCTGAGAGGATATCTAGTTGCGGCGCTCAAATCTTTGTGC GGGGCGACTACGGAGCTTCGGCGGTCTACTCTGGCAGTCGCCAGGTCGTATTCACCTGCGTTCGTAGAGG TCATGCTCCCGGATTGCTTGGGGGGCATTGCGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:81)**

PEPNR00000000179 (up_conf)
MKLSDPCLGFVAHRYLLSLERGQSPERICSCRKVIIAGLLQFQPRAKSGCSLGMALVDYKDDDDK**(SEQ ID NO:22)**: ATGAAGCTTAGCGACCCCTGTCTAGGTTTCGTAGCGCACAGGTATCTGCTTTCACTTGAAC GCGGTCAATCGCCGGAGCGCATTTGCAGCTGCCGCAAGGTTATCATTGCAGGCCTGTTGCAGTTTCAGCC GAGGGCTAAGTCGGGATGTAGCCTCGGCATGGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:82)**

PEPNR00000000579 (up)
MKLSDRIVQMLRNVSLRAWSRTGAPWETQCPSGPINYNLFCYNGLLGSQKSSSCALVDYKDDDDK**(SEQ ID NO:23)**: ATGAAGCTTAGCGACCGGATCGTGCAAATGCTTCGGAATGTGAGCTTGAGGGCATGGTCGC GGACCGGCGCGCCCTGGGAGACCCAGTGTCCCAGTGGCCCGATTAATTACAATCTTTTTTGCTATAACGG GCTCTTGGGCAGTCAGAAGTCATCTAGTTGTGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:83)**

PEPNR00000000600 (up)
MKLSRGIHLGRTSTCVNASYALCHTYRSARRGKSRKRGRSSPPIGTSLVHWVLDALVDYKDDDDK**(SEQ ID NO:24)**: ATGAAGCTTAGCCGCGGTATTCACCTAGGTCGGACGAGTACATGCGTCAACGCTTCGTACG CACTCTGCCACACGTACCGTTCAGCCCGCCGTGGCAAGTCCAGGAAGAGGGGGAGGAGTTCACCACCGAT CGGGACCTCTTTAGTACACTGGGTTTTGGACGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:84)**

PEPNR00000000612 (up)
MKLSNYNVVVYLAASTQIHRDSCLVGCVSWGGGRLTWFGTGGMHLLCRVWHYAFALVDYKDDDDK**(SEQ ID NO:25)**: ATGAAGCTTAGCAACTACAACGTGGTGGTGTATCTTGCAGCAAGTACGCAGATACACCGGG ACTCCTGCCTAGTAGGATGCGTGTCTTGGGGGGGGGGCCGTCTCACATGGTTCGGGACAGGCGGGATGCA TCTGCTGTGTCGAGTTTGGCATTACGCATTCGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:85)**

PEPNR00000000734 (up)
MKLSCGVRSKSDRSLPGSKRACVKVSAEPGKFAWSHVRAWRASTPGKGLSRLWPALVDYKDDDDK**(SEQ ID NO:26)**: ATGAAGCTTAGCTGCGGCGTGAGGTCAAAATCTGACCGTAGTTTGCCTGGTTCGAAGCGCG CCTGTGTAAAAGTGTCAGCTGAGCCTGGCAAGTTTGCGTGGTCGCATGTGCGCGCGTGGCGTGCTAGCAC GCCGGGTAAGGGATTAAGCCGTCTGTGGCCTGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:86)**

PEPNR00000000756 (up)
MKLSGRAWGRRSRRRVRPGSCAWRSRRGRRVRFDRSATQGIGCMVLELLSGLVVALVDYKDDDDK**(SEQ ID NO:27)**: ATGAAGCTTAGCGGTCGCGCGTGGGGCCGACGATCCCGCCGTCGCGTGCGCCCAGGGTCGT GCGCGTGGCGTTCGCGGCGAGGGAGACGGGTACGTTTTGATCGTTCGGCGACGCAAGGGATTGGGTGTAT GGTCCTCGAACTCCTGTCAGGTTTGGTAGTAGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:87)**

**Table 4: List of peptides that decreased in frequency in at least two experiments - designated as (down_conf), or in only one experiment designated as (down). The corresponding nucleotide sequences are listed after the colon for each peptide. The randomized part of the amino acid sequences and polynucleotide sequences are underlined.**

REPNR00000000060 (down_conf)
MKLSQHSAGPLGRGLGSFACLTRETFRRPRIAAFPGFTEANSGVYPRGLNTTPFALVDYKDDDDK**(SEQ ID NO:28)**: ATGAAGCTTAGCCAACATAGTGCTGGACCACTGGGCCGCGGGCTCGGTAGCTTTGCGTGTT TAACTCGCGAGACATTTCGCAGGCCACGTATTGCAGCGTTTCCTGGGTTCACGGAGGCGAACAGCGGCGT CTACCCCCGCGGCCTTAATACAACGCCCTTTGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:88)**

REPNR00000000061 (down_conf)
MKLSGMVIIASVSYGTKGMCPGTEGSPAVPSRIVLCGCFVIATALGSSYMRRHAALVDYKDDDDK**(SEQ ID NO:29)**: ATGAAGCTTAGCGGAATGGTTATTATTGCTTCAGTAAGTTATGGGACTAAGGGTATGTGTC CGGGCACCGAGGGCTCCCCGGCGGTGCCATCGCGCATAGTGTTGTGCGGCTGCTTTGTGATCGCAACAGC TCTGGGTAGTAGTTACATGAGAAGGCACGCCGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:89)**

REPNR00000000065 (down_conf)
MKLSVRCRNGSRLVPPWCARVCLELGLMVVTGFFRPCRMGPSVGILGAAYDGRLALVDYKDDDDK**(SEQ ID NO:30)**: ATGAAGCTTAGCGTACGCTGTAGGAACGGTTCTAGACTTGTCCCGCCATGGTGTGCTCGCG TGTGCTTGGAACTTGGATTGATGGTCGTGACAGGATTCTTCCGCCCGTGCAGAATGGGGCCGTCGGTAGG GATTTTGGGCGCAGCCTATGACGGACGACTGGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:90)**

REPNR00000000070 (down_conf)
MKLSIREGWTVRREGHRSQVTLDLHAPHRVTYRASAPAYSRLTVHGQPGAQRFRALVDYKDDDDK**(SEQ ID NO:31)**: ATGAAGCTTAGCATTCGCGAGGGATGGACAGTTAGGAGAGAGGGTCATCGTTCGCAAGTGA CCCTCGATTTGCATGCGCCGCATCGAGTTACATACAGGGCCTCCGCACCAGCGTATTCACGATTGACTGT GCATGGCCAGCCGGGCGCTCAGAGGTTTCGAGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:91)**

REPNR00000000101 (down_conf)
MKLSFHTTAKCLTAPVWRSACGIRLYSAGGGWVLGVPSVHGCSCVTYLAALRGCALVDYKDDDDK**(SEQ ID NO:32)**: ATGAAGCTTAGCTTCCACACGACTGCAAAATGCCTAACAGCACCCGTCTGGCGGTCGGCCT GTGGAATACGGCTTTACTCAGCGGGAGGAGGGTGGGTTCTGGGCGTCCCTTCGGTTCATGGGTGTTCGTG CGTGACCTATTTGGCCGCGCTTCGGGGTTGTGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:92)**

REPNR00000000111 (down_conf)
MKLSRRGAMCWCDGQRRGLDSQFCVRQERNHGGKGARARSTISRTGRPGRRAMCALVDYKDDDDK**(SEQ ID NO:33)**: ATGAAGCTTAGCAGGCGTGGAGCCATGTGTTGGTGCGACGGGCAGCGTAGAGGTCTTGATT CGCAGTTTTGTGTGCGGCAAGAGCGGAATCACGGGGGCAAGGGAGCGCGAGCGCGGAGCACTATATCTCG TACAGGGCGGCCTGGGCGAAGGGCCATGTGCGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:93)**

REPNR00000000124 (down_conf)
MKLSSRPHDVRCMGLFDEIPIFWPFLAALRLALCSGSALMCRGPDVSACYHWVYALVDYKDDDDK**(SEQ ID NO:34)**: ATGAAGCTTAGCAGTCGCCCTCACGACGTGCGGTGCATGGGTCTTTTCGACGAAATTCCAA TATTCTGGCCCTTCCTCGCGGCCTTACGACTGGCTTTGTGCTCAGGCTCTGCTCTGATGTGTAGGGGGCC AGACGTCAGTGCGTGTTATCATTGGGTCTATGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:94)**

REPNR00000000141 (down_conf)
MKLSRSFPLPRFSSHMVGHAVWREPRHFQRRTCGWSAACELSVATLIRRCRESLALVDYKDDDDK**(SEQ ID NO:35)**: ATGAAGCTTAGCCGCAGCTTTCCTCTGCCTCGCTTCTCCTCACACATGGTAGGCCACGCAG TTTGGAGAGAGCCTCGTCATTTCCAGCGTAGGACATGCGGATGGTCGGCGGCCTGTGAGTTGTCAGTCGC GACGTTAATTCGGAGGTGTCGGGAGTCATTAGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:95)**

REPNR00000000149 (down_conf)
MKLSVMLDFAQGGVRCVKHKAGLCGWTVNSCSSIFPSSSRPNDCATVRSHTPRYALVDYKDDDDK**(SEQ ID NO:36)**: ATGAAGCTTAGCGTCATGCTTGACTTCGCGCAAGGTGGTGTGCGCTGTGTAAAGCACAAGG CAGGATTGTGTGGCTGGACGGTAAATTCTTGTAGTAGTATATTCCCCTCTTCTTCCCGGCCAAACGATTG CGCTACGGTGAGGTCACATACGCCGCGATACGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:96)**

REPNR00000000158 (down_conf)
MKLSSGSAYQERFIDPGFQPSHGRFSYNCSPSRALGPERLNGSIALASYIGCLKALVDYKDDDDK**(SEQ ID NO:37)**: ATGAAGCTTAGCAGTGGAAGTGCGTATCAGGAGCGGTTCATTGATCCGGGGTTCCAGCCGT CGCACGGACGGTTCAGCTATAATTGTAGCCCGTCGCGGGCTTTGGGTCCGGAGCGCCTTAACGGATCGAT AGCGCTTGCGTCCTACATAGGGTGTTTGAAAGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:97)**

REPNR00000000159 (down_conf)
MKLSVGKPDTWLHRARGAVWVRIAGNVSLGMGLRGLSGRLPCVCGPLRTFGAFEALVDYKDDDDK**(SEQ ID NO:38)**: ATGAAGCTTAGCGTTGGGAAACCGGATACATGGCTCCATAGAGCCAGAGGAGCAGTTTGGG TTAGGATTGCCGGGAACGTGTCATTGGGTATGGGACTGCGTGGTTTGTCTGGTCGGCTCCCATGTGTATG TGGGCCTCTCAGGACCTTTGGGGCCTTCGAGGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:98)**

REPNR00000000162 (down_conf)
MKLSVASQFLFPEVGLARNGSRSARRSEFRVNVNLCGCVEQCAKPRLYAALLRVALVDYKDDDDK**(SEQ ID NO:39)**: ATGAAGCTTAGCGTGGCCTCGCAATTCCTCTTCCCAGAAGTTGGACTGGCCAGGAACGGGT CCCGGAGCGCTCGCCGTTCAGAGTTTCGGGTCAATGTTAATTTGTGTGGTTGCGTGGAGCAATGTGCGAA GCCTAGGCTCTATGCAGCTCTTTTGCGTGTTGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:99)**

REPNR00000000164 (down_conf)
MKLSRADLKLFSEHRAFVTGRRRGGMGACCLRARVEMRGLSTRQCTPVGLVYVSALVDYKDDDDK**(SEQ ID NO:40)**: ATGAAGCTTAGCCGAGCGGACTTGAAATTGTTTTCGGAGCATCGCGCCTTCGTCACGGGAC GGAGACGAGGGGGTATGGGGGCATGTTGCTTGCGAGCTAGAGTGGAGATGAGAGGCCTGAGCACACGGCA GTGTACTCCGGTTGGCTTAGTTTATGTCTCTGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:100)**

REPNR00000000193 (down_conf)
MKLSRPPFRARASCELRFSAAGRAVLDLGSQVIGLAINVISFLLCWRHEGTAYFALVDYKDDDDK**(SEQ ID NO:41)**: ATGAAGCTTAGCCGGCCACCTTTTCGTGCGCGAGCGTCATGCGAGTTGAGATTCTCGGCTG CAGGCCGTGCCGTTTTAGACCTCGGGTCACAGGTGATCGGATTGGCTATTAACGTAATAAGTTTTCTCCT TTGCTGGCGACACGAAGGGACGGCATATTTCGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:101)**

REPNR00000000206 (down_conf)
MKLSPLSRIDRFDWQWTCPGRCSCHNRLRSTPGNGVRVIYGVAGKSGRPLRISPALVDYKDDDDK**(SEQ ID NO:42)**: ATGAAGCTTAGCCCGCTTTCGCGCATCGATCGATTCGATTGGCAGTGGACTTGTCCGGGGGA GGTGTAGTTGCCACAACCGCCTAAGATCTACACCTGGTAATGGTGTGCGTGTTATTTATGGTGTCGCCGG TAAGTCGGGCCGACCGTTACGCATCTCGCCCGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:102)**

REPNR00000000209 (down_conf)
MKLSKCNPLCCTVHITVPRGLWMSGSQRVGSNRGEVLFPESRHYARKSAALTGRALVDYKDDDDK**(SEQ ID NO:43)**: ATGAAGCTTAGCAAATGTAATCCGCTATGCTGTACCGTGCATATAACGGTACCTAGGGGTT TGTGGATGAGCGGGTCACAACGGGTGGGCTCTAATCGAGGCGAGGTGCTCTTCCCGGAAAGCAGGCATTA CGCACGGAAATCTGCCGCGTTGACAGGCCGAGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:103)**

REPNR00000000271 (down_conf)
MKLSSRDVRLPQNINLSRRCLTRASPAIRALRWAMGGGAAWLMRPVRGSDPSWLALVDYKDDDDK**(SEQ ID NO:44)**: ATGAAGCTTAGCTCACGCGACGTTAGGTTGCCACAAAATATAAACTTGTCACGCAGATGTT TGACTCGGGCGTCTCCGGCCATACGCGCTTTACGGTGGGCGATGGGTGGAGGGGCGGCGTGGTTGATGCG TCCCGTCCGCGGTAGCGATCCATCCTGGCTTGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:104)**

REPNR00000000289 (down_conf)
MKLSYSYFQAGLRFRTGRQFVASVGNGARGREEPLSRRRRFCASRIRAGWAHDAALVDYKDDDDK**(SEQ ID NO:45)**: ATGAAGCTTAGCTACAGTTACTTTCAGGCAGGCCTACGATTCAGGACAGGGAGACAGTTTG TGGCGTCGGTAGGAAACGGGGCGAGAGGGCGGGAGGAGCCGCTGTCTCGTCGCAGAAGGTTTTGCGCAAG CCGAATTAGAGCAGGTTGGGCTCATGATGCAGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:105)**

REPNR00000000291 (down_conf)
MKLSSWAAKPQNELQARGYEDVCKAYLMLAHGGNELRCCTAALRALLRAALRIVALVDYKDDDDK**(SEQ ID NO:46)**: ATGAAGCTTAGCTCTTGGGCAGCGAAACCACAGAACGAGCTTCAAGCAAGGGGTTACGAAG ATGTGTGCAAGGCCTACTTAATGTTGGCCCATGGGGGCAACGAACTGCGCTGCTGCACAGCGGCACTGCG GGCGCTTTTGCGCGCGGCGCTTAGGATAGTCGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:106)**

REPNR00000000292 (down_conf)
MKLSAATWVASLRVAFGGDLILRLIRYQAAGRSGALDQFYEANSILGVHRRTRDALVDYKDDDDK**(SEQ ID NO:47)**: ATGAAGCTTAGCGCGGCTACCTGGGTCGCGAGTCTCCGAGTTGCCTTCGGTGGGGACCTTA TTCTGCGGTTAATCAGATATCAGGCGGCAGGGCGAAGCGGAGCGCTCGACCAGTTTTATGAAGCGAACTC CATACTAGGTGTCCACAGGCGTACGCGAGATGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:107)**

REPNR00000000295 (down_conf)
MKLSILALLAYLTRKTVCRAIVRGNQSWLDRIRGGPWGSRSIRGGVPNGHSARAALVDYKDDDDK**(SEQ ID NO:48)**: ATGAAGCTTAGCATTTTGGCGTTGCTTGCGTACCTTACACGCAAGACTGTGTGTCGGGCGA TCGTGCGCGGTAATCAGTCATGGCTTGACAGAATACGGGGCGGTCCTTGGGGAAGCCGCTCTATTCGGGG TGGGGTTCCAAACGGGCATTCGGCGAGGGCTGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:108)**

REPNR00000000304 (down_conf)
MKLSPHRLRRLYRLGAGCVGLVGIKGWLWTVGYLLELRVVGVRRGAIVGEALNTALVDYKDDDDK**(SEQ ID NO:49)**: ATGAAGCTTAGCCCACATCGGCTTCGTCGCCTTTACCGCTTGGGAGCTGGTTGCGTCGGTT TAGTCGGGATTAAAGGTTGGTTGTGGACGGTGGGGTACTTGTTAGAGTTGCGCGTGGTTGGGGTGCGTCG AGGGGCGATTGTAGGGGAAGCCTTGAATACAGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:109)**

REPNR00000000319 (down_conf)
MKLSATLWAKGGLILLSASLALRLYMVAFVLRITVIVLWEGMYAHRGTGPVAGRALVDYKDDDDK**(SEQ ID NO:50)**: ATGAAGCTTAGCGCCACGCTGTGGGCAAAAGGAGGGCTAATCCTGTTAAGCGCTAGTCTGG CACTCCGGTTGTACATGGTGGCATTTGTGTTACGGATTACAGTGATCGTTCTGTGGGAAGGGATGTATGC GCATCGGGGGACCGGACCGGTGGCGGGGAGGGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:110)**

REPNR00000000347 (down_conf)
MKLSCLLYLTRICLFRSEVLMRGLLLRRVPLGSAYAMVPVRAEMGLWLVVQFERALVDYKDDDDK**(SEQ ID NO:51)**: ATGAAGCTTAGCTGCCTATTGTATTTAACGCGTATTTGTCTGTTTCGAAGTGAAGTTCTGA TGCGAGGATTGCTCCTCCGTCGCGTTCCGTTGGGGTCGGCTTATGCCATGGTGCCAGTGCGGGCAGAAAT GGGCCTTTGGCTGGTTGTCCAGTTTGAGCGTGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:111)**

REPNR00000000348 (down_conf)
MKLSGARYRTIFHPVCQFGVLFCLDWAACAVGPNGGFGLARGSQGGETLMGEVQALVDYKDDDDK**(SEQ ID NO:52)**: ATGAAGCTTAGCGGGGCACGATATCGGACGATATTTCACCCAGTGTGCCAGTTCGGCGTTT TATTTTGTCTTGACTGGGCCGCGTGCGCGGTGGGTCCCAACGGGGGTTTTGGTTTGGCGCGTGGGTCTCA AGGGGGAGAGACGTTGATGGGGGAGGTTCAGGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:112)**

REPNR00000000350 (down_conf)
MKLSCCPVPPTSGQWPVGAASLQIVPRVCKSLIAPLSGARQLTVGGQYPAIEWRALVDYKDDDDK**(SEQ ID NO:53)**: ATGAAGCTTAGCTGCTGTCCAGTCCCGCCAACATCGGGCCAATGGCCGGTGGGTGCGGCGT CTCTTCAAATAGTCCCCAGGGTTTGTAAGAGTTTGATTGCCCCTCTCAGCGGAGCTCGTCAGCTCACTGT GGGAGGTCAATACCCAGCGATAGAATGGAGGGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:113)**

REPNR00000000352 (down_conf)
MKLSSKPLIRTHTNPRTMVILGGADLRAFALRLLFVWGERPTAESGCGKRTVTCALVDYKDDDDK**(SEQ ID NO:54)**: ATGAAGCTTAGCAGTAAACCATTAATCCGAACACATACGAACCCTAGGACAATGGTAATAC TTGGCGGTGCTGATCTCCGTGCATTCGCGCTACGGCTGCTGTTTGTGTGGGGGGAACGGCCAACCGCCGA GTCCGGGTGTGGGAAAAGGACCGTAACTTGTGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:114)**

REPNR00000000355 (down_conf)
MKLSFEFYRDKVIFARMCLLVPVRGPKRQAGYCEPQVPNDYAHLLDSVGLPAWVALVDYKDDDDK**(SEQ ID NO:55)**: ATGAAGCTTAGCTTTGAGTTCTATCGCGATAAGGTAATTTTCGCACGAATGTGTTTACTAG TGCCGGTTCGTGGGCCTAAGCGCCAGGCTGGGTATTGTGAGCCCCAGGTGCCCAATGACTACGCTCACCT ACTTGACAGTGTGGGGTTGCCAGCGTGGGTAGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:115)**

REPNR00000000358 (down_conf)
MKLSVGCYICSRHKAFVWVMLLDRRKETASRGDVGAFILHACRYPTNLHALNRRALVDYKDDDDK**(SEQ ID NO:56)**: ATGAAGCTTAGCGTGGGATGTTATATTTGTTCTCGTCACAAGGCTTTCGTGTGGGTAATGC TACTGGACCGGCGCAAGGAAACGGCATCGCGCGGCGATGTCGGGGCGTTCATCTTGCACGCTTGTCGCTA TCCTACAAATTTGCATGCGCTGAATCGGCGGGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:116)**

REPNR00000000402 (down_conf)
MKLSTRKITVKGGILGGAWHSWGATGEVRLTGLQDYQLEPRSQAQNPMELGGMEALVDYKDDDDK**(SEQ ID NO:57)**: ATGAAGCTTAGCACTCGTAAAATAACTGTGAAGGGCGGGATCTTGGGGGGAGCTTGGCATT CGTGGGGTGCGACAGGTGAGGTACGCCTGACTGGACTGCAGGACTACCAATTAGAACCGCGGTCACAGGC TCAAAATCCTATGGAGCTCGGCGGAATGGAGGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:117)**

REPNR00000000419 (down_conf)
MKLSCPFPDTHGAICCRVWSGFALIVLRLLDAIGSCGRHGGVGHALAEHVFWVCALVDYKDDDDK**(SEQ ID NO:58)**: ATGAAGCTTAGCTGTCCATTTCCGGATACCCATGGCGCGATCTGCTGCCGTGTTTGGTCCG GGTTCGCGTTGATTGTTCTGCGTTTGCTCGACGCCATAGGGTCCTGCGGCAGGCATGGTGGTGTGGGCCA CGCTCTAGCCGAGCACGTCTTCTGGGTGTGTGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:118)**

REPNR00000000424 (down_conf)
MKLSRSGLSMVYKHRLNVSGSVSGRTACRLRVGRKDSVVGVAVVRWRAGWMLWAALVDYKDDDDK**(SEQ ID NO:59)**: ATGAAGCTTAGCAGGAGTGGTCTTTCTATGGTGTACAAGCATCGGTTAAACGTGTCTGGTT CGGTGAGCGGTCGTACGGCCTGCCGGTTGAGGGTCGGTCGGAAGGATAGCGTGGTGGGGGTAGCGGTTGT TAGGTGGCGTGCTGGATGGATGTTGTGGGCTGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:119)**

REPNR00000000468 (down_conf)
MKLSVATMPKTPLRPRLRRGSEVASSSGDQSTGLRGTCRWEEGLVGTRRALSEVALVDYKDDDDK**(SEQ ID NO:60)**: ATGAAGCTTAGCGTGGCAACTATGCCTAAAACGCCCCTCCGCCCCCGCTTGCGTCGGGGTT CAGAGGTGGCAAGTTCGTCGGGAGACCAATCTACTGGGCTAAGGGGGACTTGTAGGTGGGAGGAGGGTTT AGTTGGTACTCGGCGCGCTCTATCGGAAGTTGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:120)**

REPNR00000000469 (down_conf)
MKLSSEMASVKDLRLGVNRGMVCAAGRAFEYAGGVHARKEISRGRMLWSTSADFALVDYKDDDDK**(SEQ ID NO:61)**: ATGAAGCTTAGCAGTGAAATGGCGTCAGTGAAAGATCTGCGCCTCGGTGTTAATCGGGGAA TGGTTTGCGCTGCGGGGCGCGCATTCGAGTACGCGGGAGGGGTGCATGCGCGGAAGGAGATTAGTCGCGG TCGCATGCTGTGGTCAACGAGCGCGGATTTTGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:121)**

REPNR00000000493 (down_conf)
MKLSSRQYSVVSTFAIKRGGPTQSRKKSGAMLRGAVWLAHFGRDRNAGHALCVLALVDYKDDDDK**(SEQ ID NO:62)**: ATGAAGCTTAGCTCTCGGCAGTATAGCGTTGTATCCACTTTTGCTATTAAGAGGGGCGGGC CAACGCAATCGCGCAAAAAAAGCGGTGCCATGTTACGAGGCGCAGTGTGGTTGGCTCATTTCGGCCGTGA TCGGAACGCTGGCCATGCCTTGTGTGTCTTGGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:122)**

REPNR00000000515 (down_conf)
MKLSIGVNSRSCKRFWCLIVVVSRGPGHLGHTCKRGVARYRRAVWLRCWWTAADALVDYKDDDDK**(SEQ ID NO:63)**: ATGAAGCTTAGCATAGGGGTGAATTCTCGATCATGTAAGCGCTTCTGGTGTTTGATAGTTG TAGTTTCGCGCGGTCCGGGGCACTTAGGTCATACATGCAAGAGGGGGGTGGCTCGGTACCGTCGCGCGGT ATGGCTACGGTGTTGGTGGACGGCTGCTGACGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:123)**

REPNR00000000540 (down)
MKLSYDAPTEWRCVHINRWTALPYVPRATFCLNVCCWTEGVRRLLCEWWSVNHRALVDYKDDDDK**(SEQ ID NO:64)**: ATGAAGCTTAGCTATGACGCGCCGACAGAGTGGCGGTGCGTCCATATCAACAGGTGGACGG CTTTACCGTATGTCCCTCGGGCGACGTTCTGCCTGAATGTGTGTTGTTGGACGGAGGGGGTTAGGCGACT TCTTTGCGAATGGTGGAGTGTTAATCATCGTGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:124)**

REPNR00000000555 (down)
MKLSVYILTVQFCTGWGVPMATTYLYAGGLRRGHHQGRSESSYRSFRKRRANTLALVDYKDDDDK**(SEQ ID NO:65)**: ATGAAGCTTAGCGTGTATATTCTTACGGTCCAGTTCTGCACTGGCTGGGGGGTGCCGATGG CCACGACATACTTGTATGCTGGGGGGCTGCGGCGGGGTCATCACCAAGGCCGCTCTGAGTCTTCTTATCG GAGTTTTCGTAAGCGTCGGGCTAACACGCTGGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:125)**

REPNR00000000583 (down)
MKLSFDDVFSTLLTCTVQRKRGMVLILKLCGVLGVPGHSGCSGQPSRTRPRFSAALVDYKDDDDK**(SEQ ID NO:66)**: ATGAAGCTTAGCTTTGATGACGTGTTTTCAACCTTGTTGACTTGTACGGTCCAGCGAAAAA GAGGTATGGTTTTAATCCTAAAGCTCTGCGGCGTTCTGGGAGTGCCAGGGCATTCCGGTTGTTCCGGTCA GCCAAGTCGTACGCGACCGCGATTTTCCGCGGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:126)**

REPNR00000000628 (down)
MKLSSVCILVLVLRHRLDALWLRLRSEGAISIFSWHEESYRVGGDLCTERKPSRALVDYKDDDDK**(SEQ ID NO:67)**: ATGAAGCTTAGCTCAGTTTGCATCCTTGTCCTGGTTCTGAGGCACCGGTTAGACGCGTTGT GGCTAAGATTACGCTCGGAAGGGGCGATTAGCATCTTCAGTTGGCATGAGGAGAGCTATCGGGTCGGTGG CGATCTGTGCACAGAGCGCAAGCCCTCTCGGGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:127)**

REPNR00000000629 (down)
MKLSKVVYRRAAQSRARSGGLTGGRVEENDVLTGARVRLRALLCCAGVSVCVTSALVDYKDDDDK**(SEQ ID NO:68)**: ATGAAGCTTAGCAAAGTAGTTTATCGTCGCGCAGCTCAGTCCCGTGCTCGGTCCGGCGGCT TGACCGGGGGTCGCGTGGAGGAAAATGATGTCCTTACGGGTGCGAGGGTGAGATTACGGGCTTTACTTTG TTGCGCGGGAGTCAGTGTCTGTGTAACCTCGGCATTGGTCGACTACAAGGACGATGACGACAAG **(SEQ ID NO:128)**

# CLAIMS

1. A method for identifying a polynucleotide encoding a biomolecule with biological activity, said method comprising:
   a) cultivating a population of host cells capable of expressing a library of polynucleotides, wherein each of said polynucleotides comprises a random nucleic acid sequence; and
   b) determining the frequencies of individual polynucleotides in said library comprised in said population of host cells at a first time point and at a subsequent second time point during cultivation, wherein the biomolecules encoded by the polynucleotides of said library are expressed in said population between said first and said second time point,

   wherein a change in the frequency of a polynucleotide in said library of polynucleotides between said first and said second time point determined according to step b) identifies a polynucleotide encoding a biomolecule with biological activity.

2. The method according to claim 1, wherein said change is an increase by at least 1.1-fold, preferably by at least 3-fold and most preferably by at least 10-fold; or said change is a decrease by at least 1.1-fold, preferably by at least 3-fold and most preferably by at least 10-fold.

3. The method according to claim 1 or 2, wherein said change is a statistically significant increase or decrease, said statistical significance of said increase or decrease being determined with a statistic test with a false discovery rate (FDR) of at least 50%, preferably 10% and most preferably 5%.

4. The method according to claim 3, wherein said statistic test is a Wald test or any other statistic test based on assessing the probability of an observation within a random distribution.

5. The method according to any of claims 1 to 4, wherein said cultivating is performed under optimal culturing conditions for said population of host cells.

6. The method according to any of claims 1 to 5, wherein said biomolecule is an RNA or a peptide and said biomolecules are RNAs and/or peptides.

7. The method according to claim 6, said method further comprising determining the amino acid sequence of the peptide encoded by the identified polynucleotide or the ribonucleic acid sequence of the RNA encoded by the identified polynucleotide.

8. The method according to any one of claims 1 to 7, wherein said biological activity is a cell growth promoting or inhibiting activity and/or a cell survival promoting or inhibiting activity.

9. The method according to any one of claims 1 to 8, wherein each or essentially each of the host cells of said host cell population comprises and/or is capable of expressing not more than one, one or more than one of the polynucleotides of said library.

10. The method according to any one of claims 1 to 8, wherein said population of host cells comprises one or more control polynucleotide(s) and the method further comprises determining the frequency of one or more of said control polynucleotide(s) said population of host cells at a first time point and at a subsequent second time point during cultivation.

11. The method according to any one of claims 1 to 10, wherein said host cell population undergoes between 1 and 50, preferably between 4 and 35 and most preferably between 16 and 25 cycles of cell division between said first and said second time point.

12. The method according to any one of claims 1 to 11, wherein said polynucleotides of said library are operatively linked to a promoter sequence, wherein said promoter regulates the expression of said polynucleotides, and wherein said promoter is an inducible promoter.

13. The method according to claim 12, wherein said inducible promoter is activated at said first time point.

14. The method according to any one of claims 1 to 13, wherein said host cell population is a population of eukaryotic or prokaryotic host cells.

15. The method according to any one of claims 1 to 14, wherein said host cell population is a population of *E. coli* cells.

16. The method according to any one of claims 1 to 15, wherein said random nucleic acid sequence has a length of 18 to 300 nucleotides, preferably 36 to 250 nucleotides and most preferably 120 to 180 nucleotides.

17. The method according to any one of claims 1 to 16, wherein each of the polynucleotides of said library is comprised in a vector such as an expression vector.

18. The method according to any one of claims 1 to 17, wherein said determining of the frequencies of polynucleotides in said library comprises DNA sequencing.

19. A biomolecule, wherein said biomolecule is selected from the group consisting of:

a polypeptide comprising or consisting of

(i) an amino acid sequence as defined by amino acid positions 5 to 54 of any one of SEQ ID NOs: 24, 58, 9 to 23, 25 to 57 and 59 to 68, or preferably as defined by any one of SEQ ID NOs: 24, 58, 9 to 23, 25 to 57 and 59 to 68;

(ii) an amino acid sequence having at least about 60%, 70%, 85%, 90%, 95%, 98%, 99% or 100% identity to the amino acid sequence as defined in (i); and

(iii) an amino acid sequence as defined in (i) having 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, or 20 amino acids inserted, deleted or substituted by a different amino acid; or

an RNA comprising or consisting of:

(iv) the RNA sequence encoded by positions 13 to 162 of any one of SEQ ID NOs: 84, 118, 69 to 83, 85 to 117 and 119 to 128, preferably by any one of SEQ ID NOs: 84, 118, 69 to 83, 85 to 117 and 119 to 128; or most preferably by any one of SEQ ID NOs: 84, 118, 69 to 83, 85 to 117 and 119 to 128 having SEQ ID NO: 136 directly fused to its 5'-end and/or having SEQ ID NO: 137 directly fused to its 3'-end; or

(v) the RNA sequence having at least about 60%, 70%, 85%, 90%, 95%, 98%, 99% or 100% identity to the RNA sequence as defined in (iv);

wherein said biomolecule has biological activity.

20. The biomolecule of claim 19, wherein the amino acid sequence as defined in (i) is an amino acid sequence as defined by amino acid positions 5 to 54 of any one of SEQ ID NOs: 24, 9 to 23 and 25 to 27, or as defined by any one of SEQ ID NOs: 24, 9 to 23 and 25 to 27; and wherein the RNA sequence as defined in (iv) is encoded by positions 13 to 162 of SEQ ID NOs: 84, 69 to 83 and 85 to 87, preferably by any one of SEQ ID NOs: 84, 69 to 83 and 85 to 87; or most preferably by any one of SEQ ID NOs: 84, 69 to 83 and 85 to 87 having SEQ ID NO: 136 directly fused to its 5'-end and/or having SEQ ID NO: 137 directly fused to its 3'-end.

21. The biomolecule of claim 19, wherein said biomolecule is a polypeptide, and wherein the amino acid as defined in (i) is the amino acid sequence as defined by amino acid positions 5 to 54 of SEQ ID NO: 24, or preferably as defined by SEQ ID NO: 24.

22. The biomolecule of claim 19, wherein said biomolecule is an RNA, and wherein the RNA sequence as defined in (iv) is encoded by positions 13 to 162 of any

one of SEQ ID NOs: 70 and 74, preferably by any one of SEQ ID NOs: 70 and 74; or most preferably by any one of SEQ ID NOs: 70 and 74 having SEQ ID NO: 136 directly fused to its 5'-end and/or having SEQ ID NO: 137 directly fused to its 3'-end.

23. The biomolecule of any one of claims 20 to 22, wherein said biological activity is a cell growth promoting activity, preferably a cell growth promoting activity in *E. coli*.

24. The biomolecule of claim 19, wherein the amino acid sequence as defined in (i) is an amino acid sequence as defined by amino acid positions 5 to 54 of any one of SEQ ID NOs: 58, 28 to 57 and 59 to 68 or preferably as defined by any one of SEQ ID NOs: 58, 28 to 57 and 59 to 68; and wherein the RNA sequence as defined in (iv) is an RNA sequence encoded by positions 13 to 162 of any one of SEQ ID NOs: 118, 88 to 117 and 119 to 128, preferably by any one of SEQ ID NOs: 118, 88 to 117 and 119 to 128; or most preferably by any one of SEQ ID NOs: 118, 88 to 117 and 119 to 128 having SEQ ID NO: 136 directly fused to its 5'-end and/or having SEQ ID NO: 137 directly fused to its 3'-end.

25. The biomolecule of claim 19, wherein the amino acid sequence as defined in (i) is an amino acid sequence as defined by amino acid positions 5 to 54 of any one of SEQ ID NOs: 58, 38 and 47, or preferably as defined by any one of SEQ ID NOs: 58, 38 and 47; and wherein the RNA sequence as defined in (iv) is encoded by positions 13 to 162 of any one of SEQ ID NOs: 118, 98 or 107, preferably by any one of SEQ ID NOs: 118, 98 and 107; or most preferably by any one of SEQ ID NOs: 118, 98 and 107 having SEQ ID NO: 136 directly fused to its 5'-end and/or having SEQ ID NO: 137 directly fused to its 3'-end.

26. The biomolecule of claim 24 or 25, wherein said biological activity is a cell growth inhibiting activity, preferably a cell growth inhibiting activity in *E. coli*.

27. The biomolecule of any one of claims 19 to 26, wherein said polypeptide consists of (i), (ii) or (iii), preferably (i).

28. The biomolecule of any one of claims 19 to 27, wherein said RNA consists of (iv) or (v), preferably (iv).

29. The biomolecule of any one of claims 19 to 26, wherein said polypeptide comprises or consists of (i).

30. The biomolecule of any one of claims 19 to 26 and 29, wherein said RNA comprises or consists of (iv).

31. The biomolecule of any one of claims 19, 20, 23 to 30, wherein said biomolecule is a polypeptide.

32. The biomolecule of any one of claims 19, 20, 23 to 30, wherein said biomolecule is an RNA.

33. A polynucleotide encoding the biomolecule as defined in any one of claims 19 to 32.

34. An expression vector comprising the polynucleotide of claim 33.

35. A host cell comprising the biomolecule of any one of claims 19 to 32, the polynucleotide of claim 33 or the expression vector of claim 34.

36. The host cell of claim 35, wherein said host cell is an *E. coli* cell.

Figure 1.

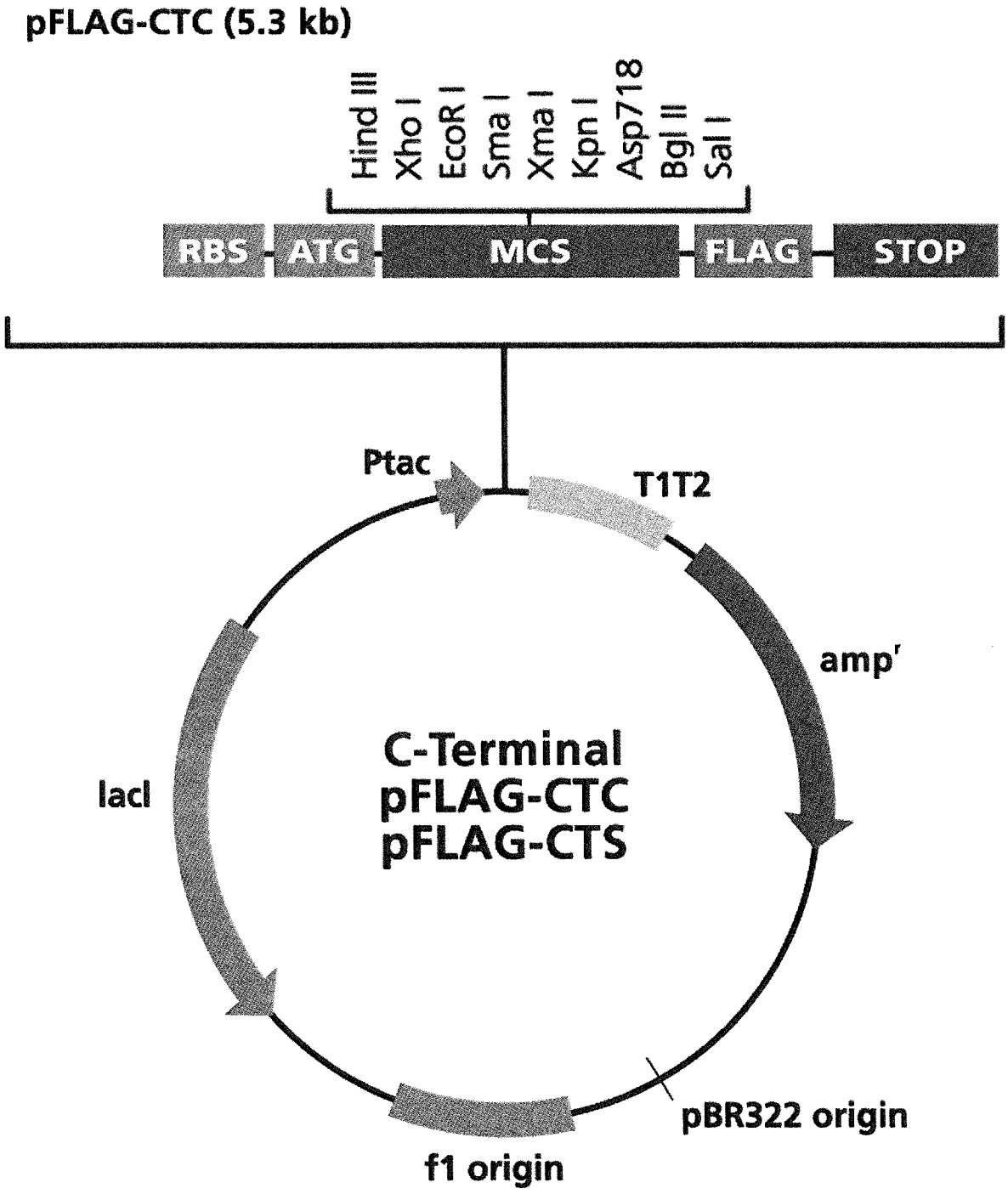

pFLAG-CTC (5.3 kb)
Hind III
Xho I
EcoR I
Sma I
Xma I
Kpn I
Asp718
Bgl II
Sal I
RBS
ATG
MCS
FLAG
STOP
Ptac
T1T2
amp'
C-Terminal
pFLAG-CTC
pFLAG-CTS
lacI
pBR322 origin
f1 origin

**Figure 2.**

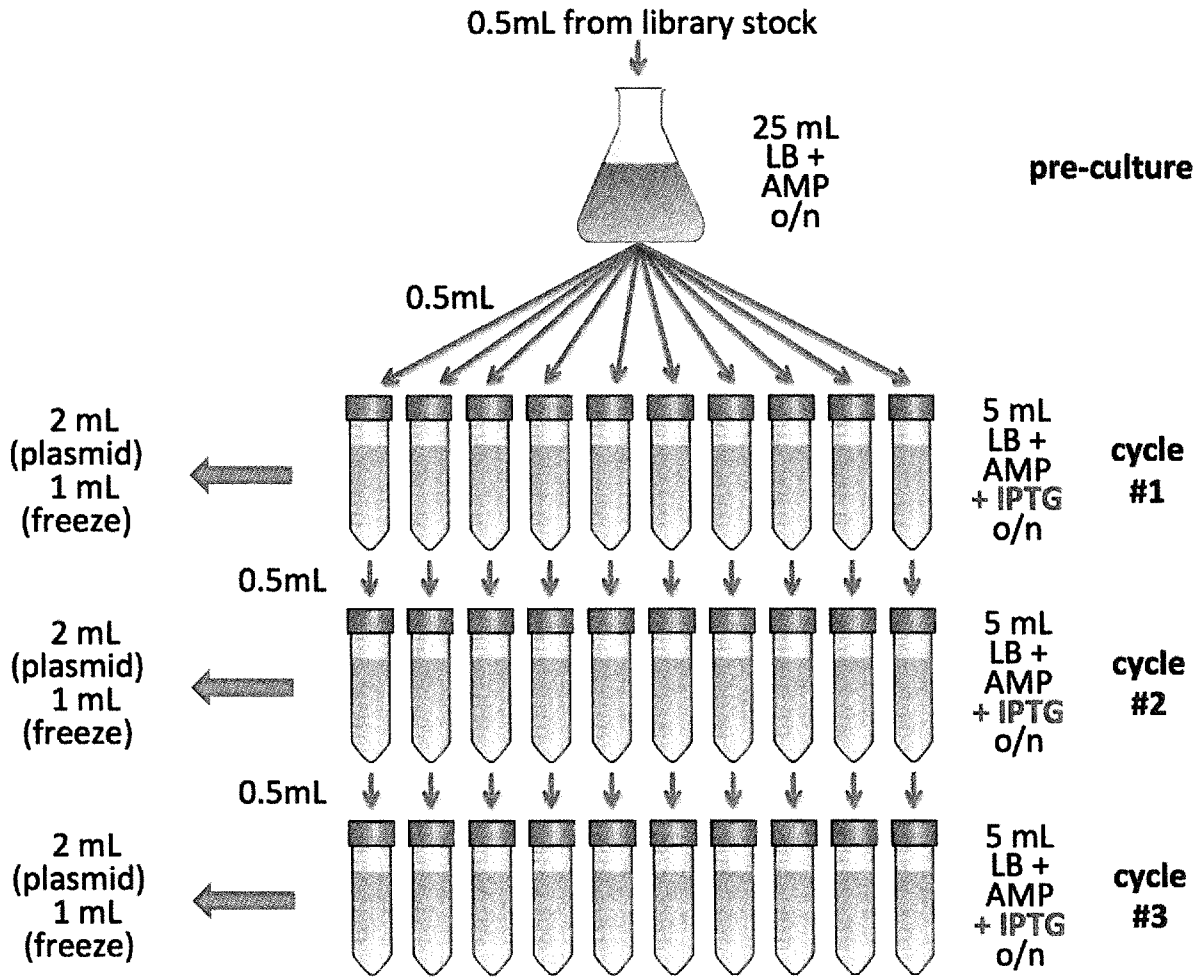

0.5mL from library stock
25 mL LB + AMP o/n
pre-culture
0.5mL
2 mL (plasmid) 1 mL (freeze)
5 mL LB + AMP + IPTG o/n
cycle #1
0.5mL
2 mL (plasmid) 1 mL (freeze)
5 mL LB + AMP + IPTG o/n
cycle #2
0.5mL
2 mL (plasmid) 1 mL (freeze)
5 mL LB + AMP + IPTG o/n
cycle #3

**Figure 3.**

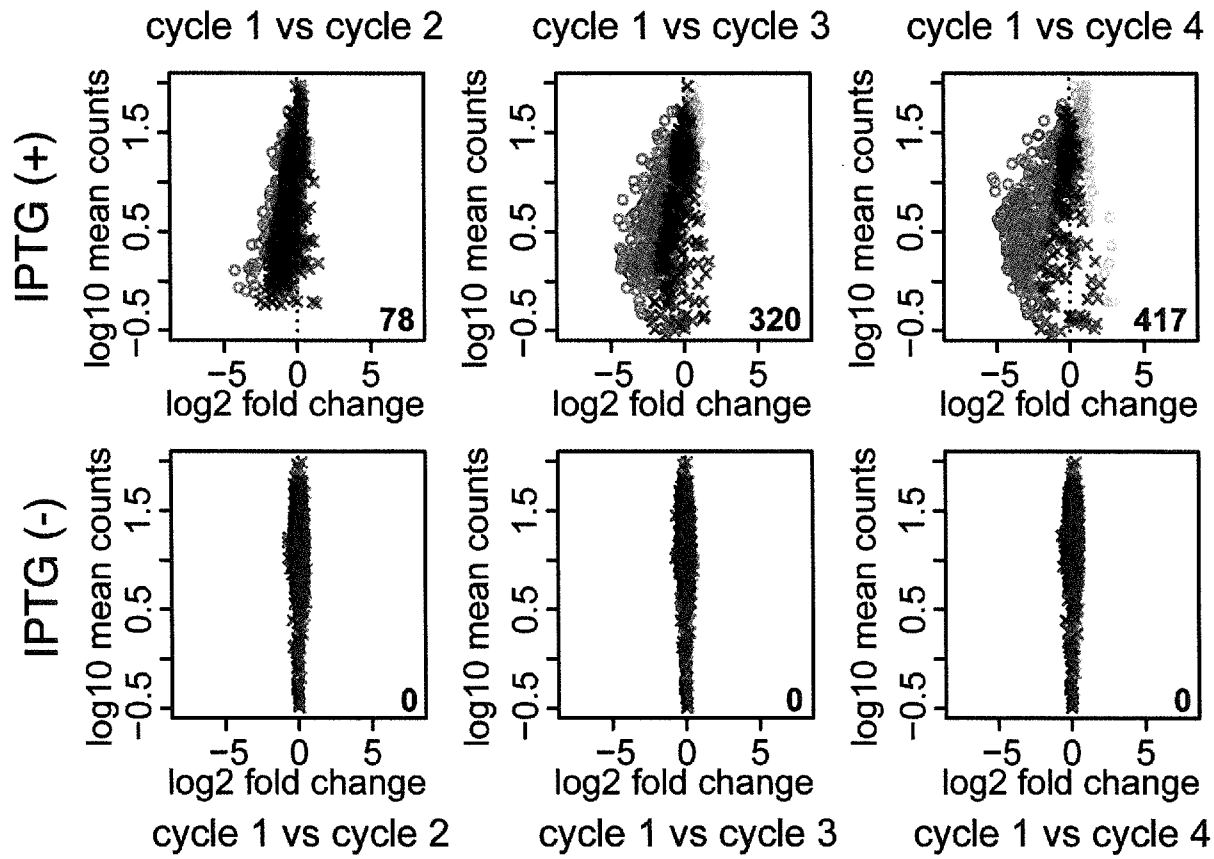

cycle 1 vs cycle 2
cycle 1 vs cycle 3
cycle 1 vs cycle 4
IPTG (+)
IPTG (-)
log10 mean counts
log2 fold change
78
320
417
0
0
0
−0.5
0.5
1.5
−5
0
5

**Figure 4.**

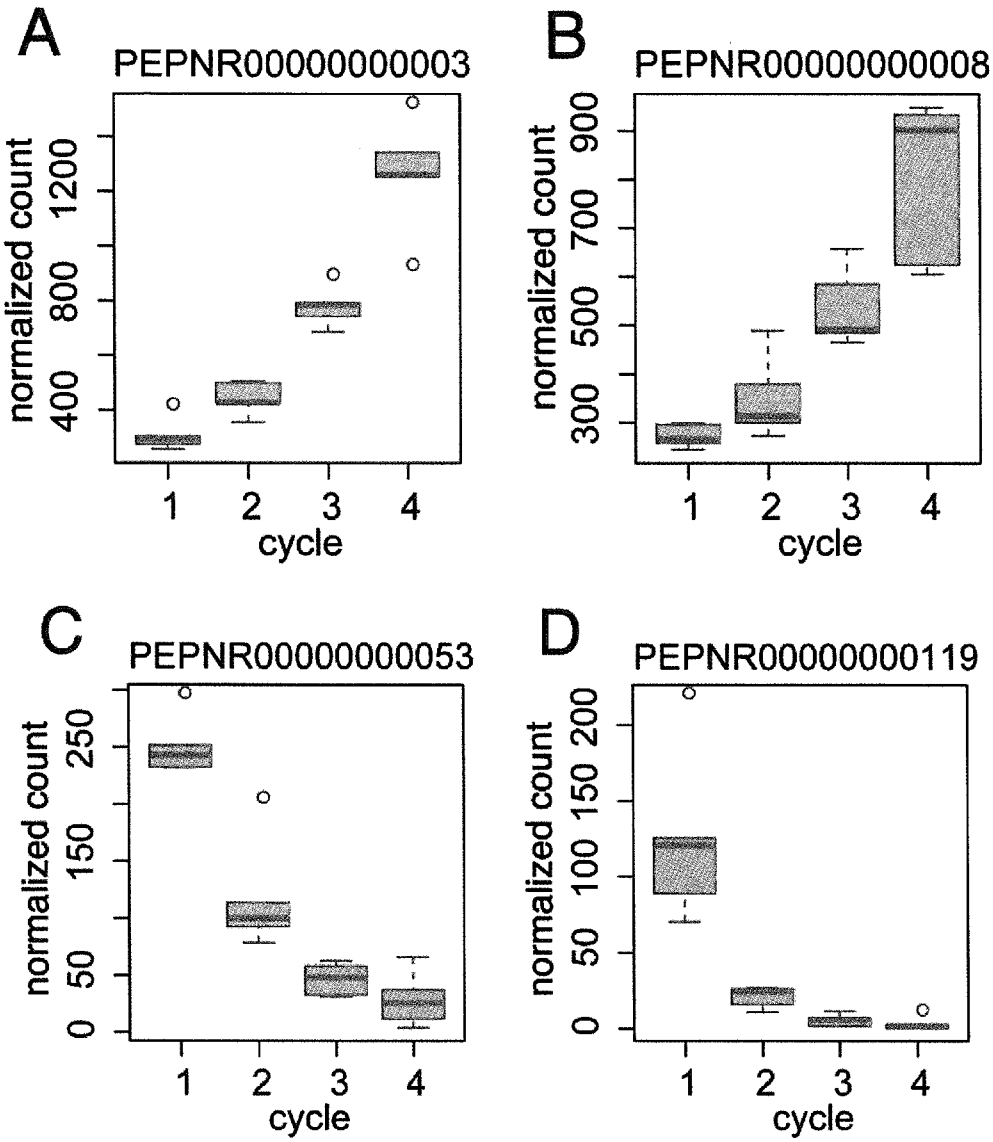



**Figure 5.**

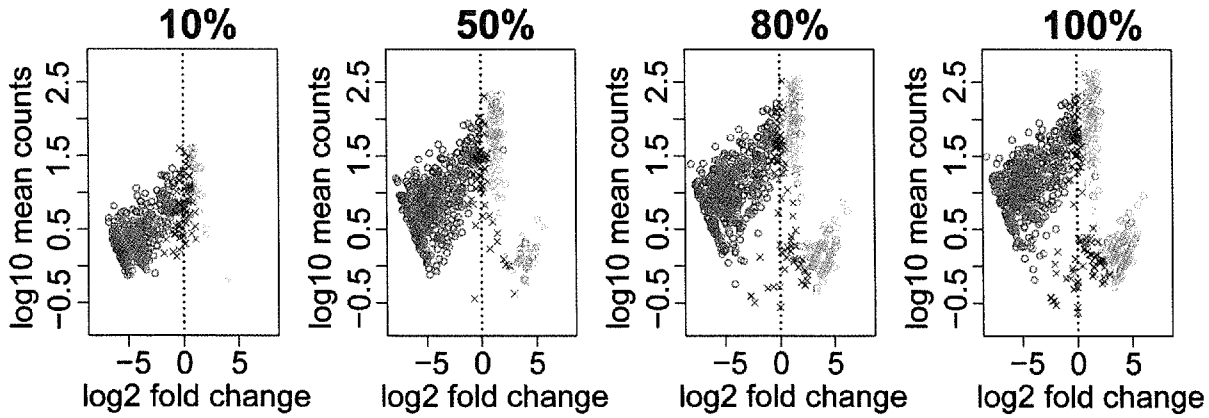


**Figure 6.**

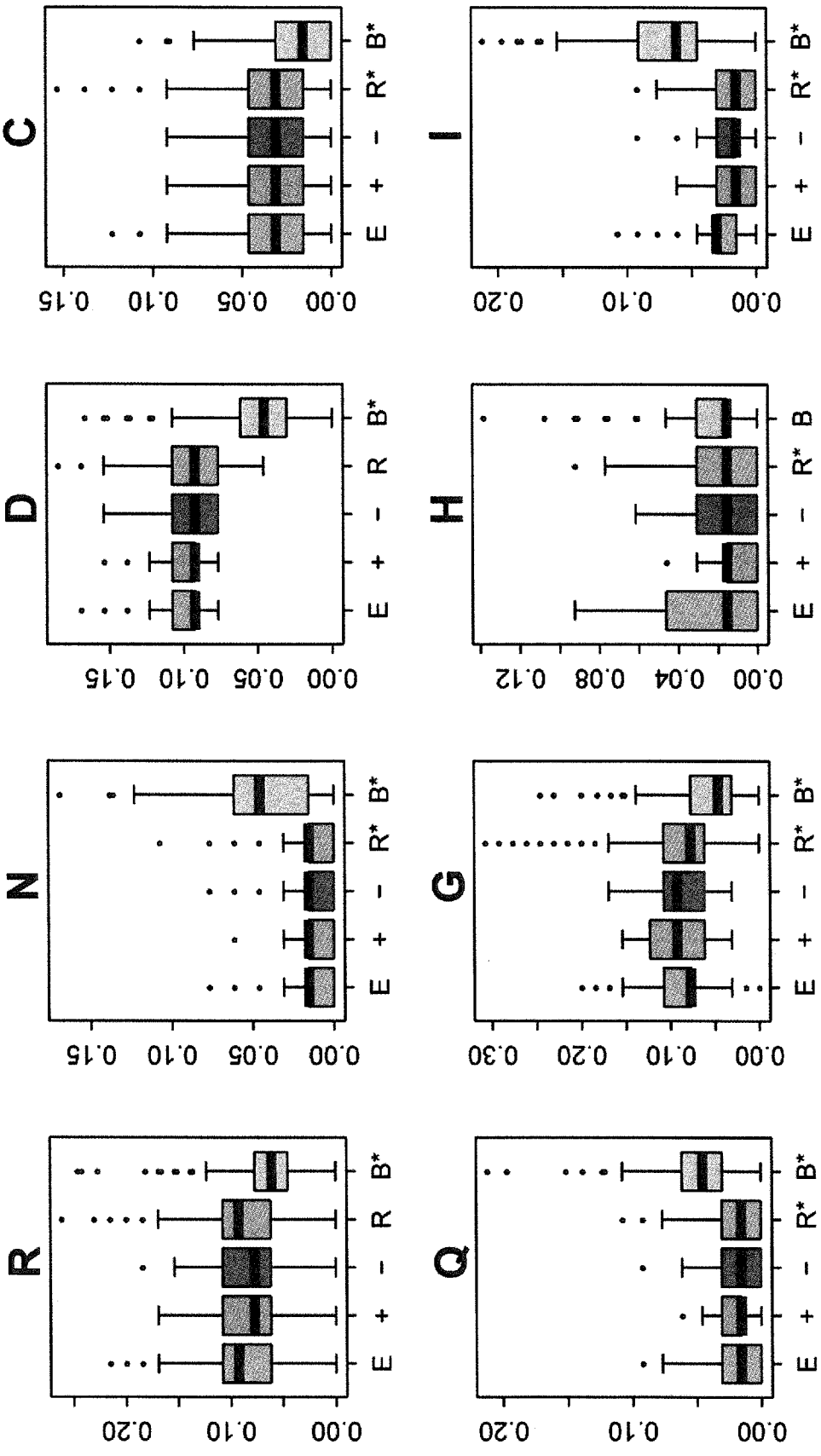


**Figure 6. continued**

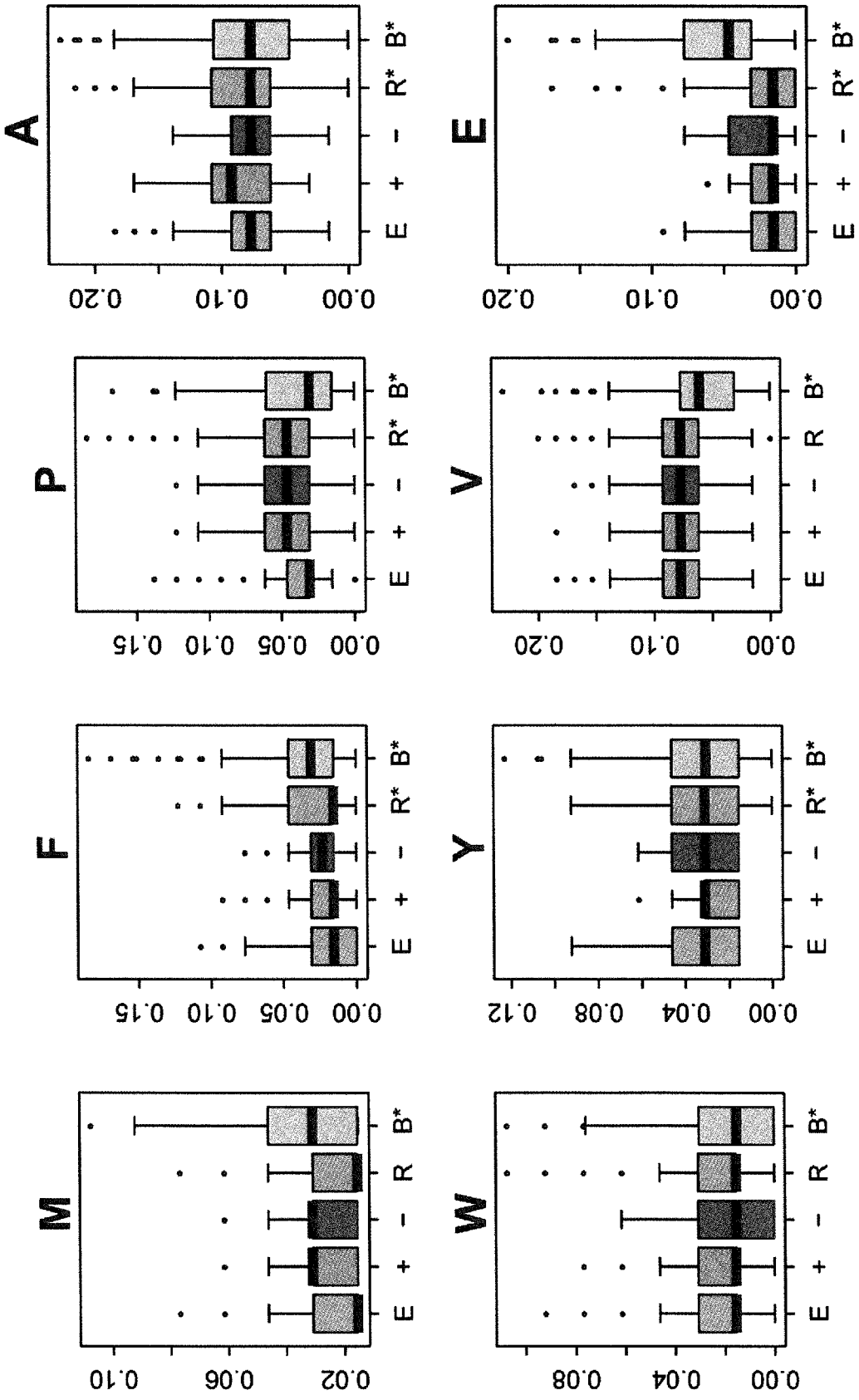


Figure 6. continued

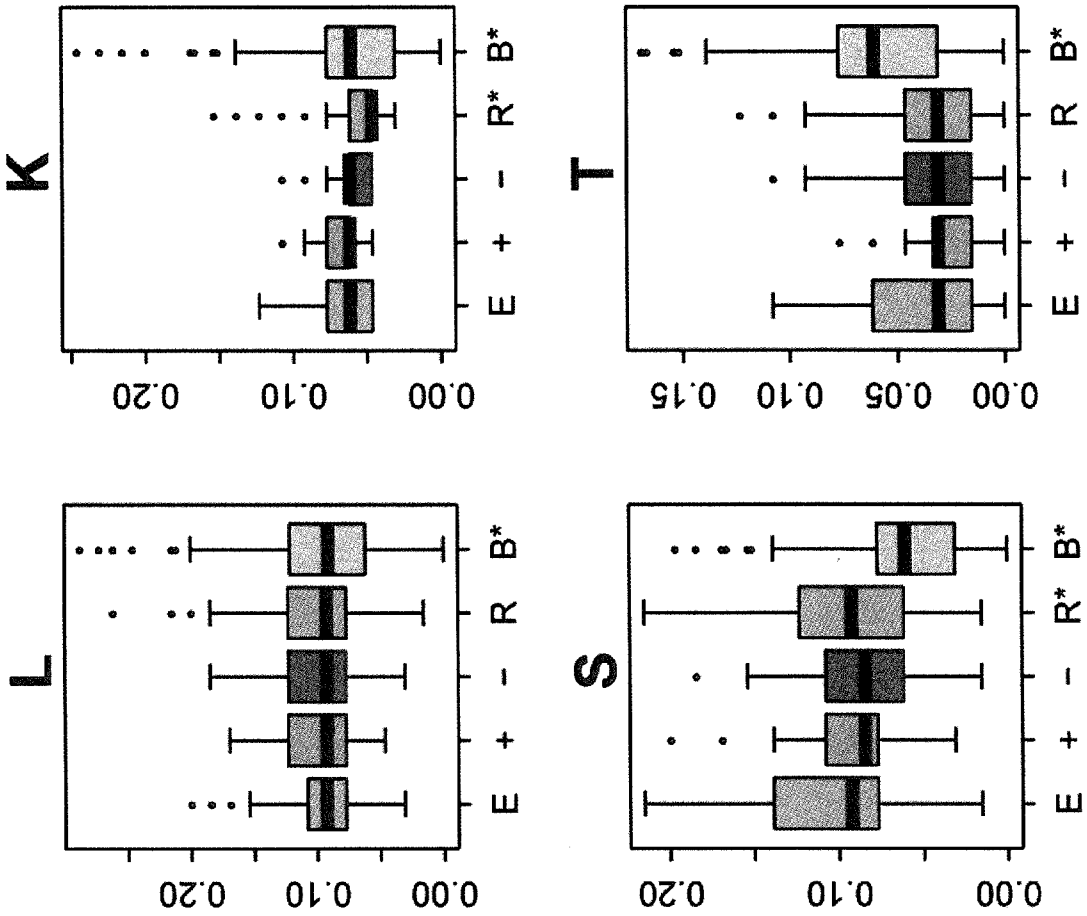



Figure 7.

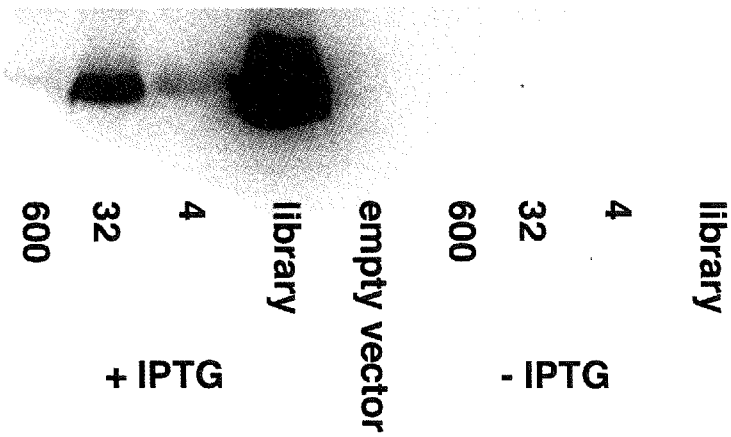


Figure 8.

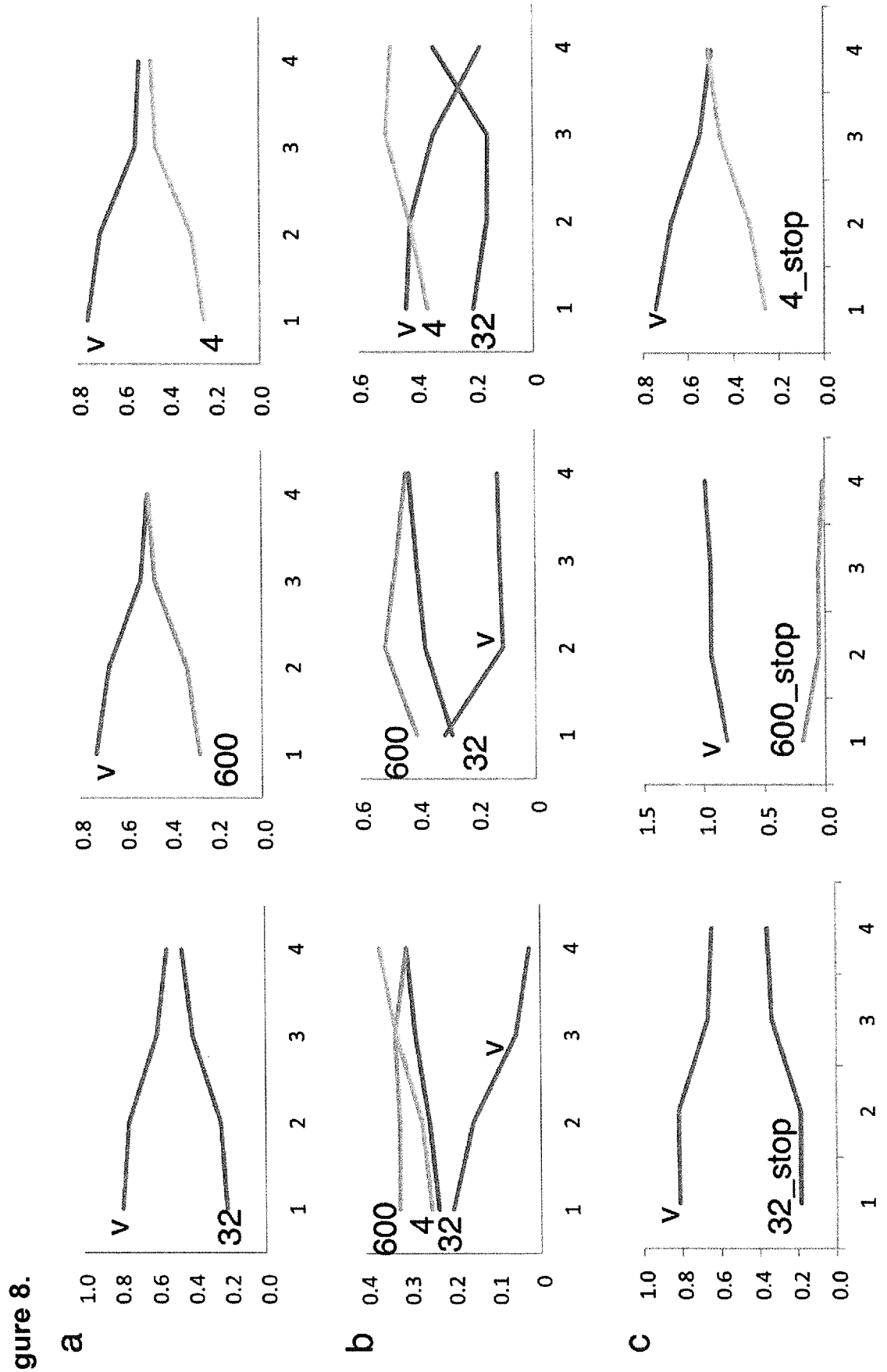
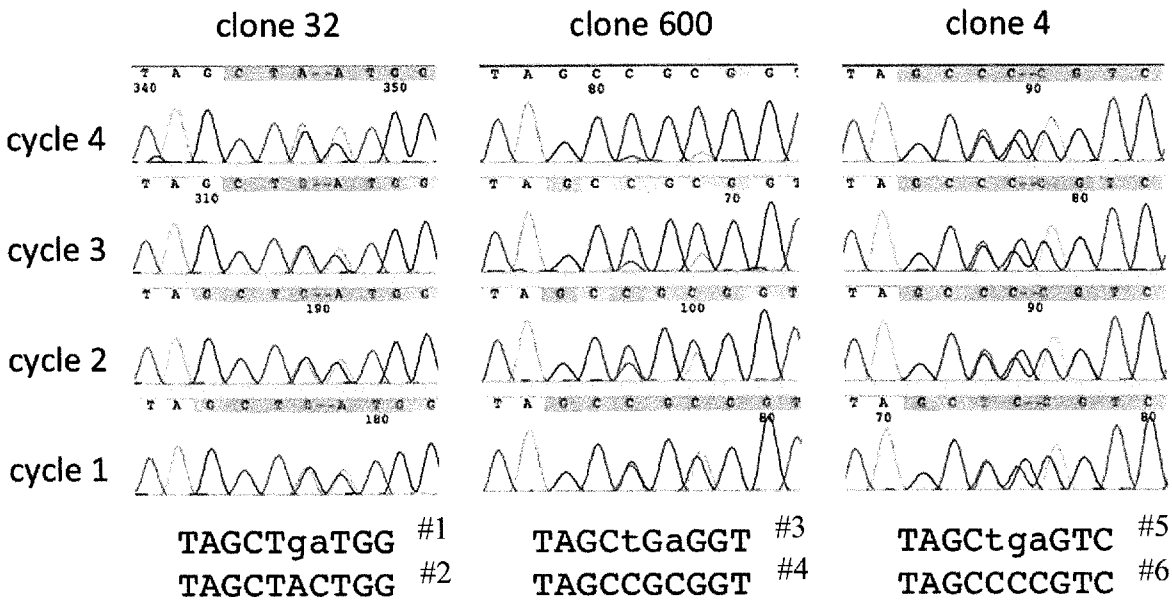


**Figure 9.**




#1= SEQ ID NO: 138

#2= SEQ ID NO: 139

#3= SEQ ID NO: 140

#4= SEQ ID NO: 141

#5= SEQ ID NO: 142

#6= SEQ ID NO: 143

**Figure 10.**

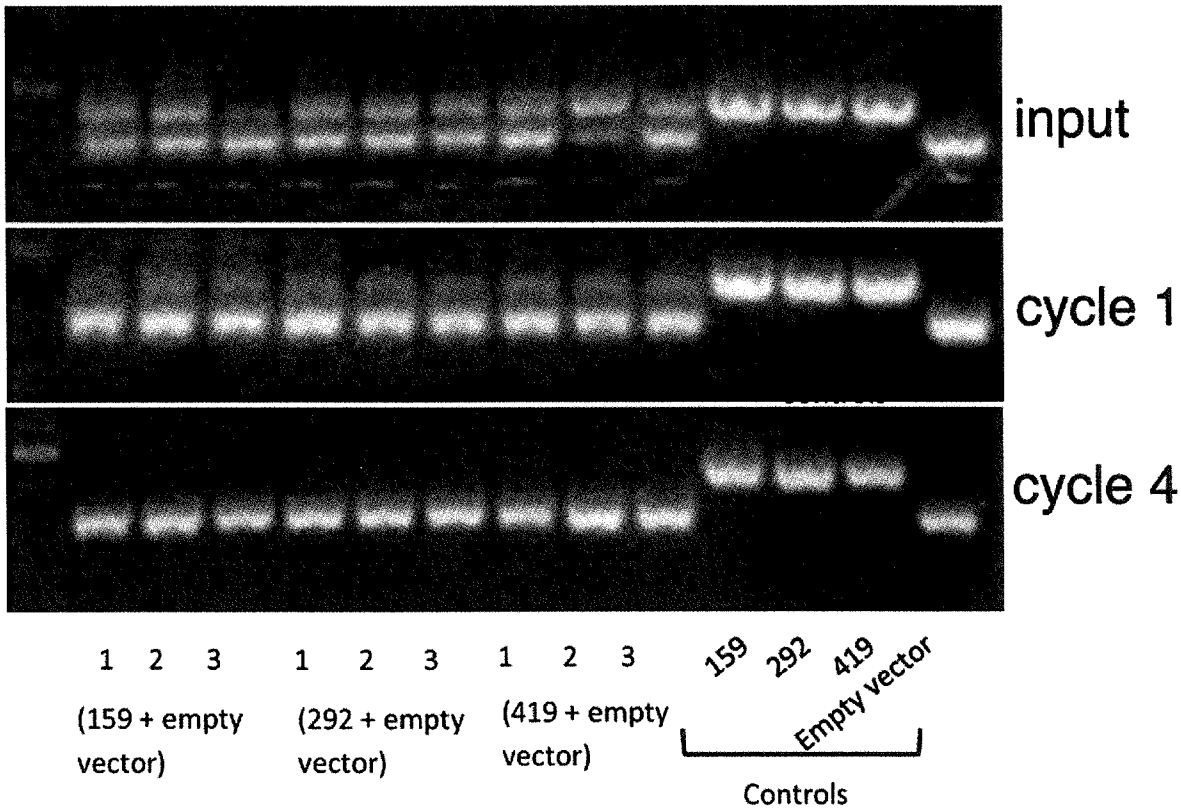

input
cycle 1
cycle 4
1 2 3
(159 + empty vector)
1 2 3
(292 + empty vector)
1 2 3
(419 + empty vector)
159
292
419
Empty vector
Controls

**Figure 11.**

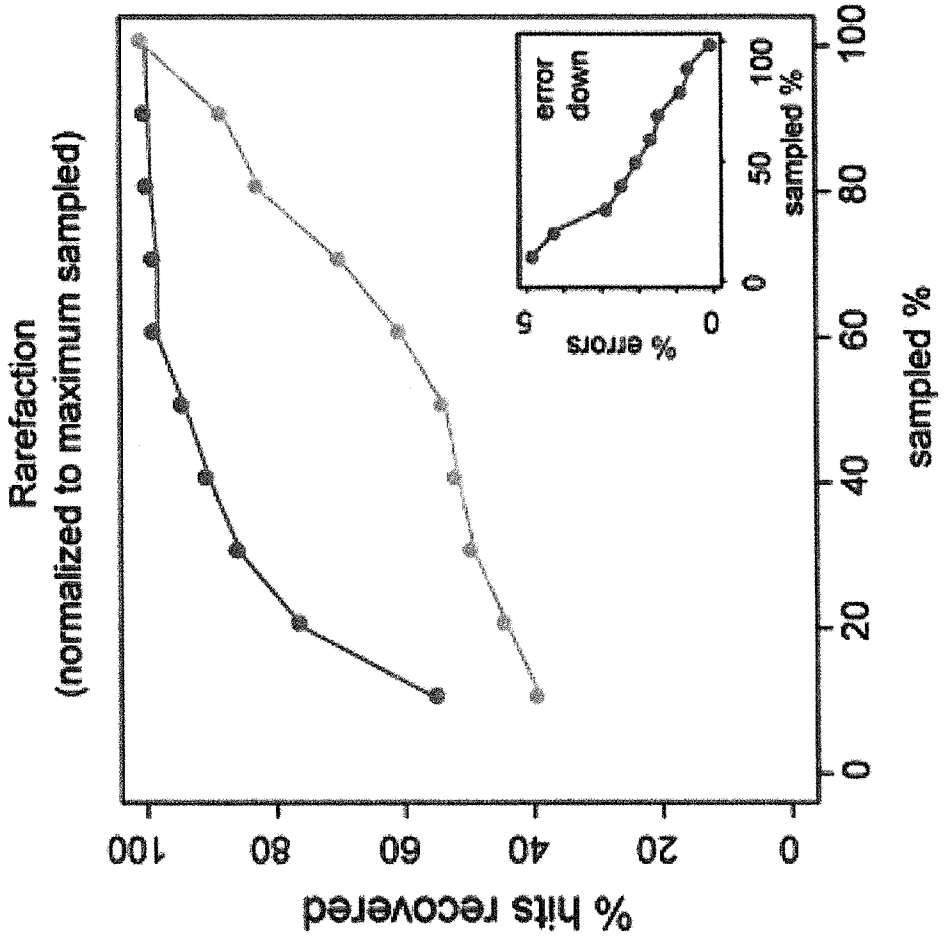
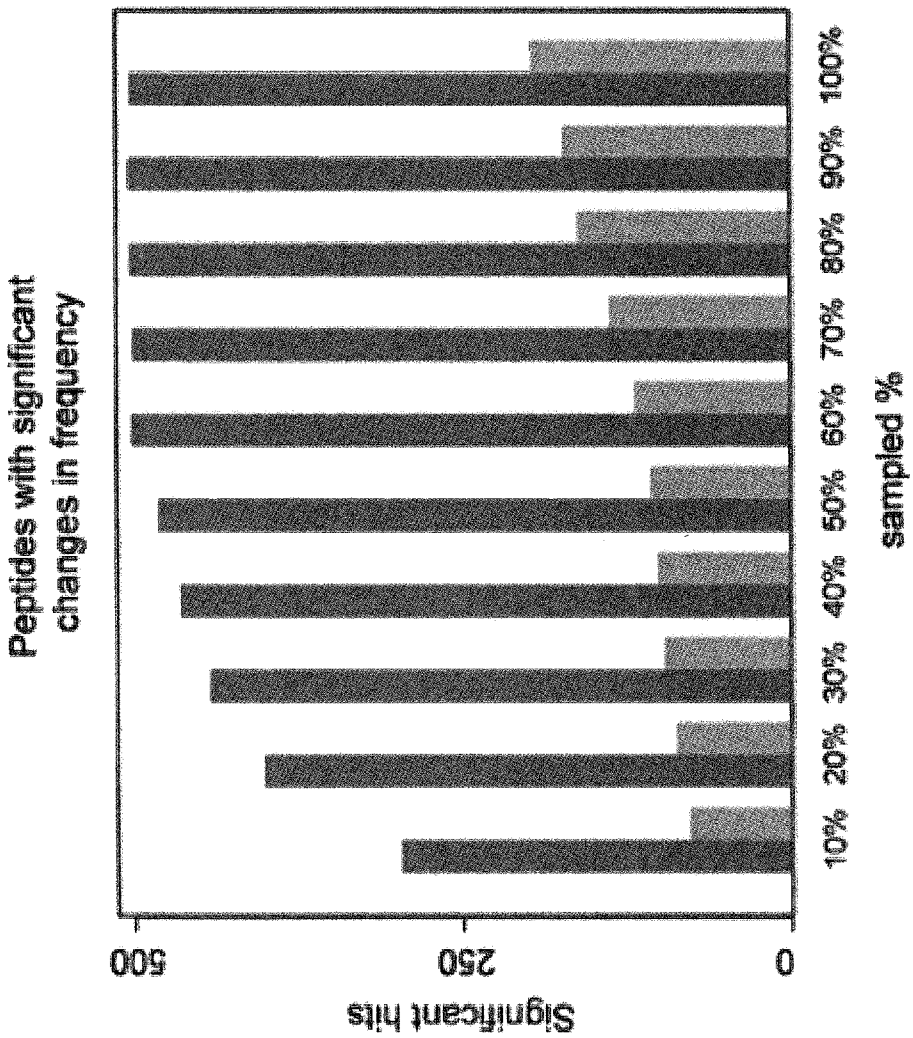

Rarefaction
(normalized to maximum sampled)
% hits recovered
0
20
40
60
80
100
sampled %
0
20
40
60
80
100
error down
% errors
0
5
sampled %
0
50
100

**Figure 11. continued**



# INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2017/056454

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12N15/10 C07K14/00 C12Q1/68 C12Q1/18 C12N15/63
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C12N C07K C12Q C12C

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, BIOSIS, EMBASE, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| Y | WO 2011/116138 A2 (UNIV WAYNE STATE [US]; CUNNINGHAM PHILIP R [US]; COLANGELO WES [US]) 22 September 2011 (2011-09-22) claims 1,23-28; sequences 7-461 | 1-18 |
| Y | WO 2016/011080 A2 (UNIV CALIFORNIA [US]) 21 January 2016 (2016-01-21) paragraphs [0126] - [0128]; claims 1,29-31,39; figure 1 | 1-18 |
| Y | US 2006/099571 A1 (ALTMAN ELLIOT [US]) 11 May 2006 (2006-05-11) paragraph [0007]; claims 67-72; examples II, III; tables 10,11,15,16 | 1-18 |

-/--

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| 12 July 2017 | 01/09/2017 |

| Name and mailing address of the ISA/ | Authorized officer |
|---|---|
| European Patent Office, P.B. 5818 Patentlaan 2<br>NL - 2280 HV Rijswijk<br>Tel. (+31-70) 340-2040,<br>Fax: (+31-70) 340-3016 | Schmitt, Anja |

3

Form PCT/ISA/210 (second sheet) (April 2005)

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| Y | US 2012/077681 A1 (RYAN T GILL [US] ET AL)<br>29 March 2012 (2012-03-29)<br>paragraphs [0072] - [0086]; claims 1-10;<br>example 1 | 1-18 |
| Y | US 2006/084098 A1 (GILL RYAN T [US] ET AL)<br>20 April 2006 (2006-04-20)<br>paragraphs [0029], [0030], [0031];<br>claims 1-18; sequences 1-5 | 1-18 |
| A | WO 2009/033024 A1 (GENENTECH INC)<br>12 March 2009 (2009-03-12)<br>page 2, line 31 - page 3, line 20; claims<br>1-7; figures 4,7; table 1<br>page 15, line 1 - page 17, line 4 | 1-18 |

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/EP2017/056454

**Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)**

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

**Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)**

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

1-18

**Remark on Protest**

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

**FURTHER INFORMATION CONTINUED FROM PCT/ISA/** 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1-18

Method for identifying a polynucleotide encoding a biomolecule with biological activity, said method comprising: a) cultivating a population of host cells capable of expressing a library of polynucleotides, wherein each of said polynucleotides comprises a random nucleic acid sequence; and b) determining the frequencies of individual polynucleotides in said library comprised in said population of host cells at a first time point and at a subsequent second time point during cultivation, wherein the biomolecules encoded by the polynucleotides of said library are expressed in said population between said first and said second time point, wherein a change in the frequency of a polynucleotide in said library of polynucleotides between said first and said second time point determined according to step b) identifies a polynucleotide encoding a biomolecule with biological activity.

---

2. claims: 21(completely); 19, 20, 23, 27-36(partially)

A biomolecule, wherein said biomolecule is selected from the group consisting of a polypeptide comprising or consisting of (i) an amino acid sequence as defined by amino acid positions 5 to 54 of SEQ ID NO:24; (ii) an amino acid sequence having at least about 60%, 70%, 85%, 90%, 95%, 98%, 99% or 100% identity to the amino acid sequence as defined in (i); and (iii) an amino acid sequence as defined in (i) having 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, or 20 amino acids inserted, deleted or substituted by a different amino acid; or an RNA comprising or consisting of: (iv) the RNA sequence encoded by positions 13 to 162 of SEQ ID NO:84; (v) the RNA sequence having at least about 60%, 70%, 85%, 90%, 95%, 98%, 99% or 100% identity to the RNA sequence as defined in (iv); wherein said biomolecule has biological activity. A polynucleotide encoding said biomolecule. An expression vector comprising said polynucleotide. A host cell comprising said biomolecule or said expression vector.

---

3-4. claims: 22(completely); 19, 20, 23, 27-36(partially)

A biomolecule, polynucleotide, expression vector, host cell as generally defined in invention 2, wherein the polypeptide is defined by SEQ ID NO:10 (14) and the RNA is defined by SEQ ID NO:70 (74).

---

5-19. claims: 19, 20, 23, 27-36(all partially)

A biomolecule, polynucleotide, expression vector, host cell as generally defined in invention 2, wherein the polypeptide is defined by SEQ ID NO:9 (11-13, 15-23, 25-27) and the RNA is defined by SEQ ID NO:69 (71-73, 75-83, 85-87).

---

20-22. claims: 25(completely); 19, 24, 26-36(partially)

A biomolecule, polynucleotide, expression vector, host cell as generally defined in invention 2, wherein the polypeptide is defined by SEQ ID NO:58 (38, 47) and the RNA is defined by SEQ ID NO:118 (98, 107).

---

23-61. claims: 19, 24, 26-36(all partially)

A biomolecule, polynucleotide, expression vector, host cell as generally defined in invention 2, wherein the polypeptide is defined by SEQ ID NO:28 (29-37, 39-46, 48-57, 59-68) and the RNA is defined by SEQ ID NO:88 (89-97, 99-106, 108-117, 119-128).

---

# INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2017/056454

| Patent document cited in search report | Publication date | Patent family member(s) | Publication date |
|---|---|---|---|
| WO 2011116138 A2 | 22-09-2011 | US 2013150260 A1 | 13-06-2013 |
| | | WO 2011116138 A2 | 22-09-2011 |
| WO 2016011080 A2 | 21-01-2016 | CA 2954791 A1 | 21-01-2016 |
| | | EP 3169776 A2 | 24-05-2017 |
| | | US 2017204407 A1 | 20-07-2017 |
| | | WO 2016011080 A2 | 21-01-2016 |
| US 2006099571 A1 | 11-05-2006 | AU 2003261305 A1 | 16-02-2004 |
| | | CA 2493306 A1 | 05-02-2004 |
| | | EP 1537137 A2 | 08-06-2005 |
| | | US 2003190740 A1 | 09-10-2003 |
| | | US 2006099571 A1 | 11-05-2006 |
| | | US 2009004696 A1 | 01-01-2009 |
| | | US 2012059145 A1 | 08-03-2012 |
| | | US 2014106341 A1 | 17-04-2014 |
| | | US 2017074858 A1 | 16-03-2017 |
| | | WO 2004011485 A2 | 05-02-2004 |
| US 2012077681 A1 | 29-03-2012 | US 2006084098 A1 | 20-04-2006 |
| | | US 2012077681 A1 | 29-03-2012 |
| US 2006084098 A1 | 20-04-2006 | US 2006084098 A1 | 20-04-2006 |
| | | US 2012077681 A1 | 29-03-2012 |
| WO 2009033024 A1 | 12-03-2009 | CA 2697519 A1 | 12-03-2009 |
| | | CN 101848925 A | 29-09-2010 |
| | | CN 104844690 A | 19-08-2015 |
| | | CY 1114618 T1 | 05-10-2016 |
| | | DK 2207790 T3 | 28-10-2013 |
| | | EP 2207790 A1 | 21-07-2010 |
| | | ES 2434829 T3 | 17-12-2013 |
| | | HK 1213916 A1 | 15-07-2016 |
| | | HR P20131108 T1 | 31-01-2014 |
| | | JP 5734653 B2 | 17-06-2015 |
| | | JP 2010538089 A | 09-12-2010 |
| | | PT 2207790 E | 04-11-2013 |
| | | SG 190602 A1 | 28-06-2013 |
| | | SI 2207790 T1 | 29-11-2013 |
| | | US 2009143248 A1 | 04-06-2009 |
| | | US 2011183413 A1 | 28-07-2011 |
| | | US 2017088813 A1 | 30-03-2017 |
| | | WO 2009033024 A1 | 12-03-2009 |