Characterization of Transcriptional Activities

Adison Wong[1], Maurice HT Ling[2]

[1] Chemical Engineering and Food Technology Cluster, Singapore Institute of Technology, Republic of Singapore

[2] Colossus Technologies LLP, Republic of Singapore

Corresponding Authors:

Adison Wong	Maurice HT Ling
Chemical Engineering and Food Technology	Colossus Technologies LLP
Cluster, Singapore Institute of Technology	8 Burn Road, #15-13, TRIVEX
10 Dover Drive	Singapore 369977
Singapore 138683	

Abstract

Transcription is the first stage of gene expression, leading to the eventual determination the protein abundance and affecting metabolism. Hence, there is a need to measure and characterize transcriptional activities accurately. This article discusses the experimental techniques to characterizing transcriptional activities from a single-gene approach (quantitative PCR) to high-throughput methods (microarray technology and next generation sequencing). Computational approaches to predicting relative transcript abundance from sequence features and gene co-expression network will be described. As a proxy to protein / enzyme abundance, transcriptional activities are critical in developing simulatable biochemical models, which can then to test biological hypotheses before laboratory experimentation.

Keywords: Quantitative PCR, RNA-Seq, Next Generation Sequencing, Microarrays, Reporter Proteins, Sequence Features, Transcriptome-Phenotype Correlation, Gene Expression Prediction, Modeling and Simulating Metabolism, Perturbation Effects, *In silico* Testing.

Background

Information in biological systems are encoded in a string of biopolymers known as deoxyribonucleic acids or DNA. The Central Dogma, comprising of the processes DNA replication, transcription and translation, describes the sequential flow of information by which living cells enzymatically decode and process latent genetic information into messenger ribonucleic acids (mRNA) and further into active proteins for essential and secondary cellular metabolism. The ability to accurately measure and characterize transcriptional activities – the study of DNA promoter elements driving mRNA transcription – both *in vitro* and *in vivo* could potentially revolutionize the way in which biology and medicine are studied. Conventionally, transcriptional activities are characterized in low throughput by real time quantitative polymerase chain reaction (qPCR) and semi-qualitative Northern blot assay. Recent progress in DNA microarray and RNA sequencing have enabled large scale and high quality whole cell transcriptome analysis to be performed. When complemented with gene ontology and principled analysis software, both technologies could rapidly isolate important genetic determinants for further validation by qPCR. Besides the measurement of transcription at the mRNA level, other studies have sought to quantify gene expression using reporter proteins in synthetic genetic circuits both whole cell and cell free. The results of these studies are usually integrated into computational models to enable the prediction of gene expression at larger scale. In the following sections of this review, the various methods for characterizing transcriptional activities will be further discussed.

Experimental Techniques

In this section, three major experimental techniques; namely, quantitative PCR, microarrays, and RNA-seq (which is an example of next generation sequencing technology); for analyzing transcriptomic activities will be discussed. This is followed by a discussion on reporter proteins, which are instrumental for experimental validation of transcriptomic activity.

Reverse Transcription & qPCR. qPCR is a widely adopted molecular biology workflow that exploits the sensitivity and sequence specificity of PCR reactions for targeted gene expression analysis (Figure 1). Through this method, transcriptional activities could be characterized either in terms of the absolute transcripts abundance, or as relative quantitation between samples in different biological contexts. qPCR is also used to validate differential gene expression results obtained from high throughput DNA microarray and RNA sequencing technologies, despite the limit to

analyze a maximum of 384 samples per run. In the qPCR workflow, mRNA is first isolated from the biological sample using commercial RNA extraction kit. An RNA stabilizing agent, RNAlater®, is often added to the biological sample prior to cell disruption to protect RNA integrity as the extraction process may take at least an hour or more. TRIzol® reagent, a monophasic solution of phenol and guanidine isothiocyanate, is routinely used for the extraction of total RNA of mammalian, plant, yeast or bacteria origin. When chloroform is added to tissue samples or cells homogenized in TRIzol®, RNA is preferentially extracted into the top aqueous phase. Afterwards, isopropanol is added to the aqueous layer to precipitate out RNA. Purified RNA in sterile, diethylpyrocarbonate (DEPC)-treated water, comprising mainly of rRNAs and tRNAs and about 1% – 5% mRNA, is analyzed for RNA integrity and quantity using synthetic dyes that emit fluorescence when bound to nucleic acids or the NanoDrop spectrophotometer. As a rule of thumb, the 260:280 nm ratio of the total purified RNA should be between 1.9 – 2.1 and that of 260:230 nm should be 2.0 – 2.2. Appreciable lower values may indicate the presence of contaminants by either proteins or RNA extraction reagents. Clear and distinct rRNA bands observed on agarose gel electrophoresis also indicates the availability of good quality RNA for downstream qPCR. To generate cDNA templates for qPCR, the isolated mRNA (mixed in total RNA) is reversed transcribed into single stranded cDNA using random primers, dNTPs, reaction buffer and reverse transcriptase from commercial kits. Depending on the target gene expression level ranging from low to high expression, between 1 ng – 100 ng of total RNA may be required. Random primers are made of short oligomers with six to nine nucleotides and are able to bind to any region in the RNA pool of most organisms. Other types of primers that could be used for cDNA synthesis are oligo(dT)s which anneal non-specifically to the 3’ poly(A) tails of mRNA, and sequence specific primers which anneal to only targeted mRNA regions. However, unlike random primers, oligo(dT)s are seldom used in prokaryotic cDNA synthesis as the poly(A) tails of prokaryotic mRNA are generally shorter and less polyadenated (Sarkar, 1997). During reverse transcription, bound DNA primer is extended towards the 5’ direction of mRNA. Thus, even though nonspecific primers may generate truncated cDNAs from internal mRNA sequence, gene expression results would not be much affected so long as qPCR primers are designed targeting the 5’ end of the mRNA. Sequence specific primers for cDNA synthesis are commonly used for synthetic RNA standards or where analysis only involves a few gene targets.

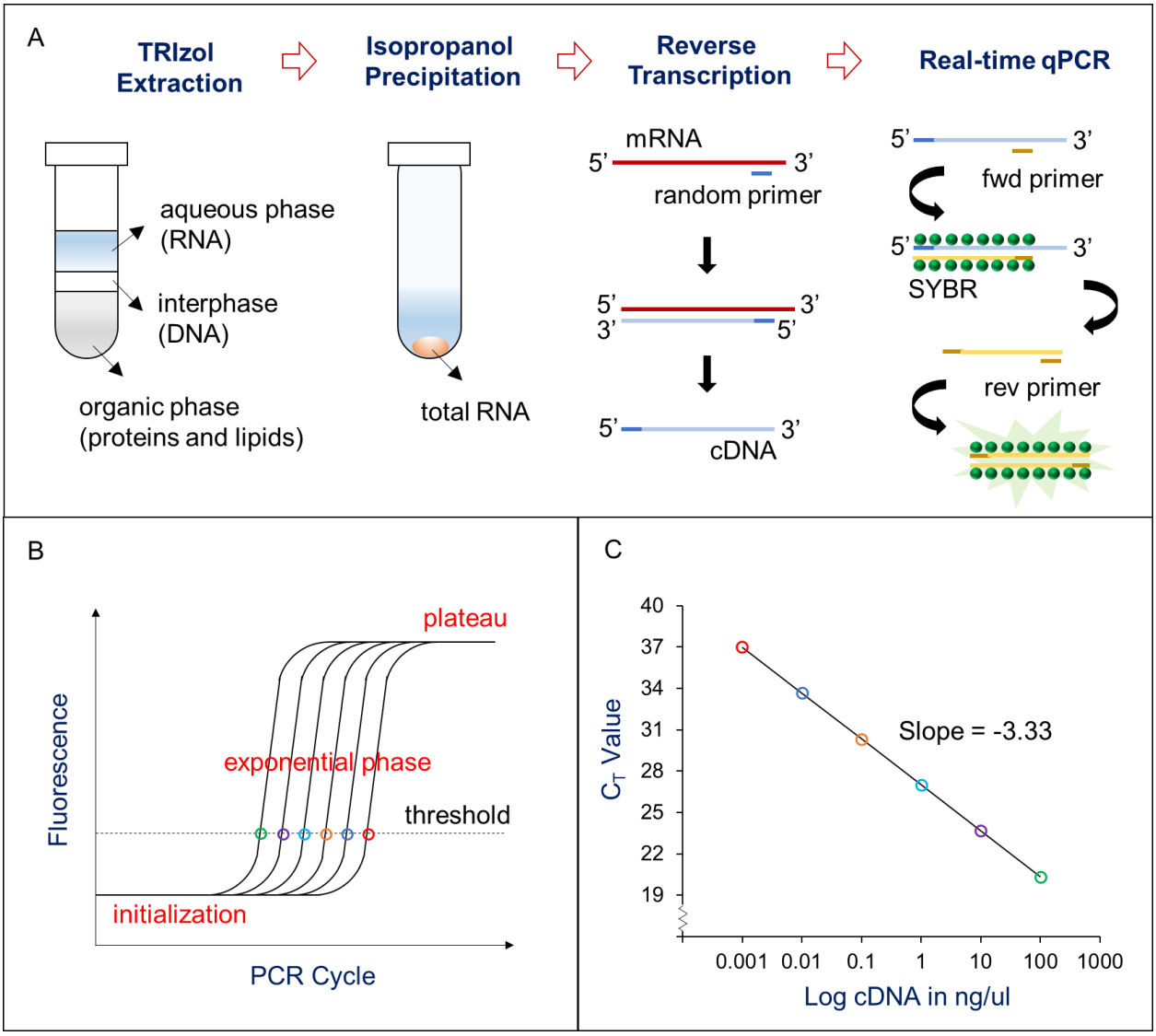

Fig 1. qPCR analysis of targeted transcripts. (A) Sample preparation for qPCR. Total RNA is chemically extracted from biological sample and then reverse transcribed to generate single stranded cDNA molecules. Thereafter, the cDNA are used as templates for qPCR in real time thermocyclers. SYBR green bind to double stranded cDNA during the extension process and fluorescence readings are recorded. (B) Sigmodal amplification of DNA during qPCR. A threshold is set above noise level in the exponential phase and used to determine the threshold cycle, Ct. (C) Relationship between cDNA concentration and Ct values under ideal condition, i.e. DNA exactly doubles with each PCR cycle. Hypothetical Ct values from (B) is reflected in (C) by colored circles.

To initiate the qPCR reaction, appropriate amounts of cDNA template, sequence specific primers and nuclease-free water are added to qPCR master mix, comprising of DNA-binding dye (SYBR), dNTPs, magnesium and Taq polymerase, according to the manufacturer's instruction, before importing into the qPCR thermocycler for gene expression measurement. qPCR could be

performed in standard PCR tubes, 96-well PCR plates, or 384-well PCR plates. qPCR primers, typically 20 bp – 30bp are designed to amplify approximately 60 bp – 150 bp of the sense strand of the target cDNA at the 5' end. Amplicons above 150 bp usually suffers from poor PCR efficiency, resulting in gross underestimation of the actual amount of transcripts. qPCR thermocycles are equipped with optical instruments to detect fluorescent signals from DNA-binding dye. As qPCR cycle proceeds, double stranded amplicons are generated, resulting in more PCR products for SYBR to bind and consequential increase in green fluorescence emission. Fluorescence intensity correlates directly with the amount of DNA amplicons, and in turn gives an indication of the amount of cDNA template (mRNA transcripts). As a standardized practice, gene expressions are calculated based on the threshold cycle (Ct), or the number of PCR cycles it takes for fluorescence to be at a critical value above background, which also indicates the start of PCR exponential phase. Ct values are inversely proportional to initial template amount. Variations in sampling procedures, RNA extraction, reverse transcription and qPCR reaction may significantly lead to bias and compound error (Sanders et al., 2014). To mitigate the extent of potential bias in qPCR analysis, gene expression results are normalized to internal reference genes. One common practice is setting genes of core genomic functions, such as constitutive housekeeping genes as the normalization reference. These genes are assumed to have constant expression (mRNA levels) at all times, despite several studies have shown otherwise (Bustin, 2000; Dundas and Ling, 2012; Thellin et al., 1999). To improve the accuracy of measurement, target gene expression could be normalized to the geometric average of two or more internal control genes (Vandesompele et al., 2002). It is also good practice to validate the internal control genes with statistical software such as geNorm, NormFinder and BestKeeper (Andersen et al., 2004; Pfaffl et al., 2004; Vandesompele et al., 2002). The relative expression ratio of two samples could be determined using the generalized Pfaffl method as illustrated in Example 1 (Hellemans et al., 2007; Pfaffl, 2001). To determine the absolute amount of mRNA transcript, a standard curve of threshold cycle number plotted against varying amount of cDNA standard, Ct versus log [cDNA], could be developed by performing qPCR on synthesized cDNA standards with six or more serial dilutions. Importantly, the synthesized standard and target cDNA should be similar in PCR efficiency, amplicon length and primer binding efficiency, ensuring that the standard curve is applicable for the gene of interest. Known amounts of synthetic RNA standards could also be spiked into the extracted RNA to

account for reverse transcription bias. Recommendations on experimental design, results reporting and selection of reference genes are described in the MIQE guidelines (Bustin et al., 2009, 2010).

Example 1: Fold-expression Calculation. Real time quantitative PCR was performed to analyze how the gene expression of an efflux pump would change when cells were exposed to antibiotic stress. Results of the qPCR run is shown in Table 1 below. In this experiment, three different housekeeping genes (Ref Gene A, B and C) were used as reference genes. "Treated" and "Control" represent the mRNA extracted from cells with and without antibiotics treatment, respectively. For each gene, threshold cycles are the arithmetic mean of technical replicates. Determine the relative fold expression change of the target gene and comment if the efflux pump (gene of interest; GOI) is up- or down-regulated following antibiotics exposure.

Table 1. qPCR analysis of gene encoding for efflux pump after exposure to antibiotics

	GOI	**Ref Gene A**	**Ref Gene B**	**Ref Gene C**
Efficiency	2	2	2	2
Treated	28.50	21.25	19.00	23.00
Control	34.00	20.50	18.75	22.00

The general formula for normalized relative quantities (NRQ) with multiple reference genes is given as, $NRQ = \frac{{E_{GOI}}^{\Delta Ct_{GOI}\,(control-treated)}}{\sqrt[f]{\prod_0^f {E_{ref0}}^{\Delta Ct_{ref0}\,(control-treated)}}}$

In this analysis, we assumed that the efficiency of amplification is 100% and that DNA amount exactly doubles with every qPCR cycle in the exponential phase. This gives an efficiency value of 2 for both the GOI and reference genes, $NRQ = \frac{{2_{GOI}}^{\Delta Ct_{GOI}\,(control-treated)}}{\sqrt[f]{\prod_0^f {2_{ref0}}^{\Delta Ct_{ref0}\,(control-treated)}}}$

The $2^{\Delta Ct}$ value of GOI is 45.25 ($2_{GOI}{}^{\Delta Ct_{GOI}\,(control-treated)} = 2^{(34.00-28.50)} = 45.25$). Similarly, the $2^{\Delta Ct}$ values of reference genes A, B and C are 0.59, 0.84 and 0.5, respectively. Geometric mean of $2^{\Delta Ct}$ values of the three reference genes is 0.63.

$$\sqrt[f]{\prod_{0}^{f} 2_{ref0}{}^{\Delta Ct_{ref0}\,(control-treated)}} = \sqrt[3]{0.59 \times 0.84 \times 0.5} = 0.63$$

Fold expression ratio of GOI / reference genes, $NRQ = \frac{45.25}{0.63} = 71.83$. Hence, the efflux pump gene increased by ~72-folds upon antibiotic exposure, implying that the gene was up-regulated in the experiment.

DNA Microarray. The DNA microarray is a structured 2D array of short, single-stranded DNA molecules that are systematically deposited or synthesized on solid surfaces as discrete spots (Figure 2). These DNA molecules, also known as oligo probes, are designed to be complementary to target gene sequence so that hybridization could readily occur between a fluorescent labeled target DNA and the immobilized probe when in contact. Each spot contains at least picomoles of oligo probes between 25 bp – 80 bp. In a typical microarray experiment, total RNA is first isolated from both test and control biological samples using commercially available RNA extraction kits. Thereafter, the extracted RNAs are assessed for microarray suitability using both capillary electrophoresis and NanoDrop. RNA quality is validated by the RNA integrity number (RIN) measured on the Agilent 2100 Bioanalyzer, and a value of 7 or higher is considered suitable for microarray experiment. Depending on the microarray technology, cell type and nature of the study, a range of between 200 ng – 20 μg of total RNA (typical concentration ~1 μg/μl) may be required (Kang and Chang, 2012; Ling et al., 2013; Trevino et al., 2007). In Affymetrix arrays, the extracted mRNAs are used to generate labeled cDNAs in a one pot reverse transcription reaction. Comparatively, in Agilent arrays, mRNAs are first retro-transcribed into T7 RNA promoter tailed cDNAs, and then converted into labeled cRNAs by the Eberwine reaction (Van Gelder et al., 1990). In either workflows, the cDNA and cRNA molecules are labeled using fluorescent labeled nucleotides. Common fluorescent dyes used for distinguishing gene expression profiles includes the red Cy5 and the green Cy3. Following the generation of single stranded cDNA (cRNA) molecules, equal amounts of both test and control cDNAs (cRNAs) are pooled together and allowed to hybridize on the microarray slides. At this point, competitive hybridization occurs in which the test and control cDNAs (cRNAs), each tagged to different fluorescent dye, compete for binding to discrete spots of the microarray. Successful hybridization events will result in an

increase of local fluorescent intensity that is proportional to the abundance of the target mRNA transcript. These results could be measured using a robotic fluorescent scanner which consist of a device similar to that of a fluorescence microscope coupled with excitation laser and digital camera for image capturing. Captured fluorescent images are overlapped and computationally processed to convert spot intensities into numerical outputs. Where Cy5 is used for the labeling of cDNAs generated from the control transcripts and Cy3 for that generated from test transcripts, effective green and red fluorescence indicates up and down -regulation of the target genes, respectively. Yellow spots, implying the lack of any significant changes in gene expression, are observed when transcripts from both the test and control biological sample are of comparable abundance. In general, genes are considered differentially expressed when test and control hybridizations exhibit 2- or more fold differences. Microarray results are usually reported based on the MIAME guidelines created by the Functional Genomics Data Society (Brazma et al., 2001).

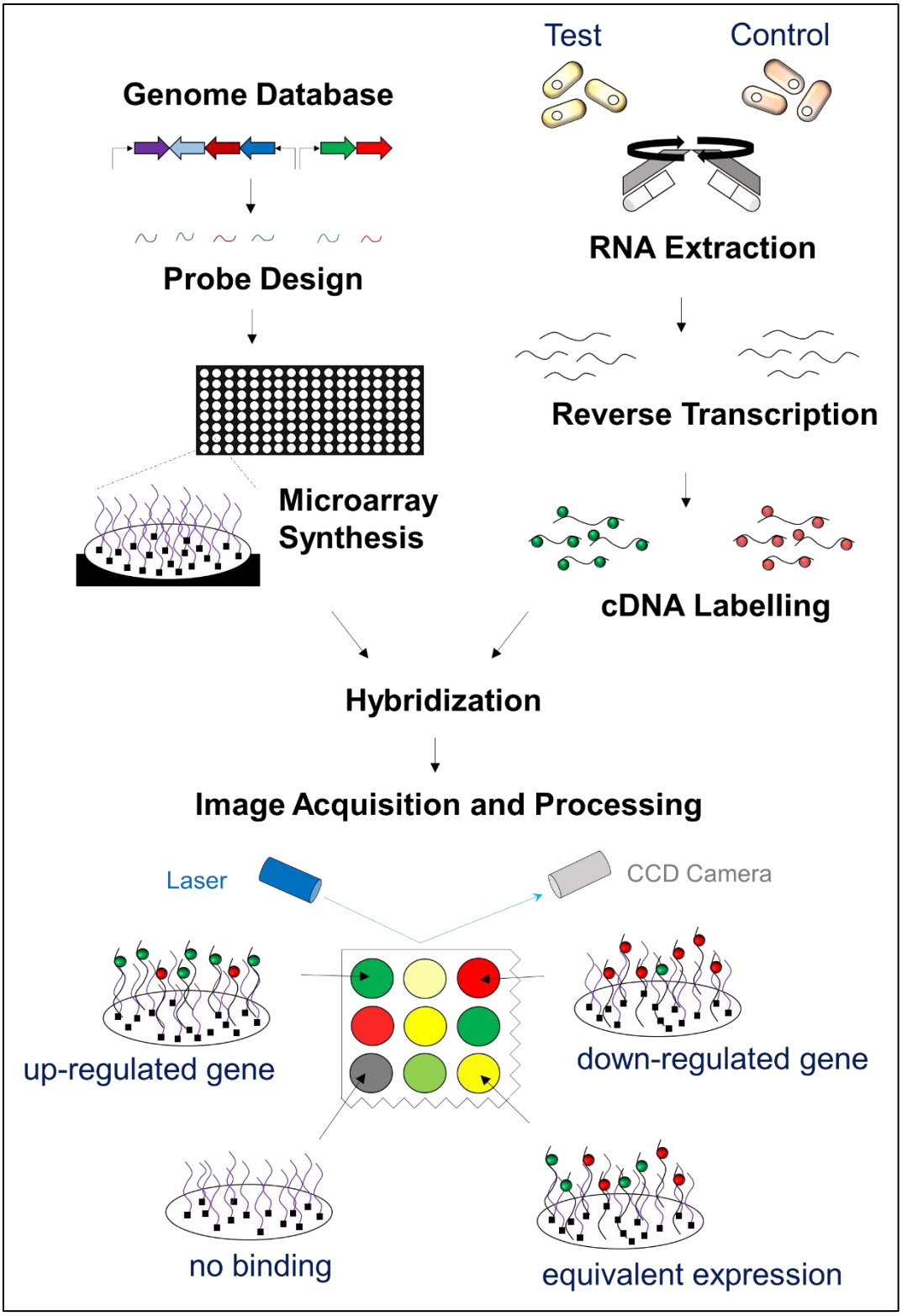

Fig 2. Schematic of a DNA microarray workflow.

Example 2: Statistical Analysis of Microarray Results. Two-sample microarrays, also known as two-channel microarrays are used determine differential gene expressions between two samples; such as, healthy tissues versus diseased tissues (as shown in Figure 2). It is common for a gene to be represented more than once within a microarray for the purpose of statistical calculations. In this example, we assume a hypothetical microarray of only five genes and each gene is represented by five probes. The summarized results are given in Table 2. The t-statistic of each gene and its p-value can be calculated using the following equation, where n is the number of probes per gene.

$$t - statistic = \frac{|\overline{Healthy}_{Gene(i)} - \overline{Diseased}_{Gene(i)}|}{s_{Healthy}/\sqrt{n}}$$

Table 2. Microarray Summarized Values and Statistics

Gene Name	Healthy Tissue Signal		Diseased Tissue Signal		t-statistic	p-value	Significance
	Mean Intensity	Standard Deviation	Mean Intensity	Standard Deviation			
Gene A	14.09	1.703	14.82	0.023	1.242	0.2821	Not Significant
Gene B	13.64	0.642	15.43	0.375	4.983	0.0076	Significant
Gene C	14.14	0.053	15.99	0.796	18.002	5.6E-05	Significant
Gene D	8.18	0.836	9.73	1.487	3.785	0.0194	Not Significant
Gene E	12.72	1.102	13.39	1.227	1.413	0.2303	Not Significant

The threshold (α) for determining statistical significance is taken as 0.05. From this point of view, Gene D is significantly upregulated in diseased tissue compared to healthy tissue (p-value = 0.0194). However, as there are multiple t-tests to be carried out (one test for each gene), the total threshold will increase – when 2 and 3 t-tests are carried out for an experiment, the threshold increases from 0.05 to 0.0975 ($1 - 0.95^2$) and 0.143 ($1 - 0.95^2$) respectively. Hence, there is a need to maintain the overall threshold at 0.05 and a simple way is by Bonferroni correction, which is to reduce the threshold to the quotient of 0.05 and the number of statistical tests required. In this case, the Bonferroni corrected threshold is 0.01 (0.05 / 5) as there are 5 t-tests required. As a result, Gene D is not significantly upregulated in diseased tissue compared to healthy tissue.

RNA-Seq. In 1977, a method to decode extracted DNA sequences at single nucleotide level was developed by Nobel Prize chemist Frederick Sanger and his colleagues (Sanger et al., 1977). The

Sanger method, which was based on the selective incorporation of chain-terminating fluorescent dideoxynucleotide, was used as the core technology to complete the Human Genome Project (Schmutz et al., 2004). Since then, a remarkable transformation is witnessed in the field of DNA sequencing with newly developed chemistries and chip-based analytical platforms, giving rise to second and third generation sequencing technologies (Chen et al., 2013; Heather and Chain, 2016; Schadt et al., 2010; Voelkerding et al., 2009). These developments led to higher throughput, sequencing depth and reliability, enabling applications beyond functional genomics, including the transcriptome analysis of prokaryotic and higher-order eukaryotic organisms, also known as RNA sequencing (RNA-Seq) (Croucher and Thomson, 2010; Mäder et al., 2011; McGettigan, 2013). RNA-Seq measures transcript abundance by repeated, direct sequencing of reverse-transcribed cDNAs. Sequence data are then mapped back to a reference genome and the number of mapped reads are normalized to quantify output expression as reads per kilobase of transcript per million mapped reads (RKPM) or transcripts per million (TPM) (Mortazavi et al., 2008; Wagner et al., 2012). Example 3 shows the different procedures to normalize raw transcripts output into RKPM or TPM. The normalization procedures are required because sequencing depth, i.e. the number of times a sample is read, and transcript length result in higher raw output readings on the detection platform. TPM provides a more straightforward breakdown of gene expression profile, although both normalization approaches are equally accepted by the scientific community (Wagner et al., 2012).

Over the years, RNA-Seq has gradually displaced microarrays for transcriptomic analysis. Without the need to hybridize to transcript complementary probes, RNA-Seq avoided common problems associated with microarray gene expression measurement. Firstly, microarray experimental design relied on the availability of known genome sequence for chip synthesis. Hence, when genome sequences are not available, the only way forward is by using chips originally intended for other closely related specie to capture cDNA. Due to sequence variability; however, this approach may miss out on capturing the full transcriptome profile of the target. In contrast, RNA-Seq could be performed for unknown samples in parallel with whole genome sequencing to access both functional genomics and transcriptome data. This greatly reduces experiment down time while also enables the identification of novel transcript isoforms (Trapnell et al., 2010). A second advantage is in the reliability of transcriptome data. RNA-Seq is not affected by the non-specific

hybridization of cDNA to probes. Instead, multiple reads are performed on the same sample to ensure sequence reliability. This procedure allows transcription to be accurately mapped down to single nucleotide resolution, and could be particularly useful when characterizing transcriptomes with frequent repeats. Additionally, because transcript abundance is measured by direct sequencing reads, RNA-Seq exhibits a larger dynamic range of detection, with better sensitivity at both low and high gene expression levels. On the other hand, the upper detection limit of microarray is bounded by the optical sensitivity of microarray scanner under saturating condition (blinding effect). Despite the apparent advantages, there are a few considerations on the use of RNA-Seq, namely data management and storage, speed, and cost. Data generated from RNA-Seq are typically in the giga- to terabytes range while those from microarray are in tens of megabytes. This implies that sequencing data is inherently more complicated to work with and would require a reliable suite of bioinformatics tools for data normalization and analysis (Trapnell et al., 2012). In terms of speed, a typical sequencing run takes between 2 – 14 days to complete while microarray experiment could be done within a day. Microarray consumables are also cheaper than those required for RNA-Seq by ~10-folds. In the longer run, however, these concerns would likely be reconciled as new sequencing platforms with higher throughput, sequence accuracy and longer read lengths are developed. Ongoing efforts to correct experimental bias and consolidate best practices in RNA-Seq would eventually pave the way for dominance of this technology in high throughput transcriptomics study (Conesa et al., 2016).

In the RNA-Seq workflow, total RNA is first extracted using commercial kits and analyzed for amount and quality on NanoDrop and Agilent 2100 Bioanalyzer. RNA requirements for sequencing are more stringent than that for microarray; at least a RIN > 8 for eukaryotic cells and > 9.5 for prokaryotic cells should be satisfied. Illumina also recommends to evaluate RNA suitability in terms of the percentage of RNA fragments > 200 nucleotides (DV_{200}) on Agilent 2100 Bioanalyzer. A few hundred nanograms total RNA with DV_{200} > 30% would be appropriate for downstream application. Before RNA-Seq library preparation, it is prudent to remove unwanted rRNA transcripts which competes with mRNA for sequencing capacity. For this purpose, terminator exonuclease (TEX), which processively degrades only rRNA substrates with 5'-monophosphate ends, could be added to the total RNA pool (Croucher et al., 2009; Sharma et al., 2010). Alternatively, subtractive-hybridization kits sold by Illumina (Ribo-Zero™), Qiagen

(GeneRead™) and Thermofisher Scientific (RiboMinus™) could be used to deplete prokaryotic and eukaryotic rRNAs. Oligonucleotide probes on magnetic beads support are used to capture rRNAs by hybridization, followed by subsequent ethanol extraction of mRNA from supernatant after beads pull down (Croucher et al., 2009; O'Neil et al., 2013; Petrova et al., 2017; Stewart et al., 2010; Westermann et al., 2016). Depending on the sequencing equipment and the purpose of the study, different methods have been developed to prepare RNA-Seq cDNA libraries, which in turn determines if the directional information of transcriptional activities could be effectively captured (Croucher and Thomson, 2010). A common workflow involves fragmenting rRNA-depleted samples by ultra-sound sonication or nebulization into 200 bp – 400 bp RNA fragments. This is followed by dephosphorylation of the fragmented transcripts on the 5' end and ligation with the first sequencing adaptor on the 3' end. Transcript 5' ends are then re-phosphorylated with T4 polynucleotide kinase and ligated with the second sequencing adaptor. Processed transcripts are reverse transcribed using 3' adaptor as primer to generate first strand cDNA (Westermann et al., 2016). Alternatively, rRNA-depleted samples could be 3' polyadenylated with poly(A) polymerase and then treated with tobacco acid pyrophosphatase. This effectively converts 5' triphosphate into monophosphate for 5' adaptor ligation. Processed transcripts are reverse transcribed using oligo(dT)-adapter primers to generate first strand cDNA (Westermann et al., 2016). Noteworthy, in both methods, the adaptors are attached stepwise in specific orientation to preserve directional information. Besides mechanical fragmentation, transpososome could be used to fragment and incorporate modified adaptors onto double stranded cDNA in a one pot reaction known as tagmentation (Adey et al., 2010; Kia et al., 2017). In this process, rRNA-depleted samples are reverse transcribed and amplified to generated double stranded cDNA, using uracil instead of thymine for second strand synthesis. The resultant cDNA molecules are transposed with sequencing adapters and treated with USER enzyme mix to deplete the second cDNA strand, retaining only tagged single stranded cDNA. This cDNA library is selectively amplified by Phusion polymerase (Phusion polymerase do not extent well on DNA templates with uracil residues) and purified (Gertz et al., 2012).

The next step in RNA-Seq involves hybridizing cDNA libraries to support material, typically insoluble beads or glass slide, which contains oligonucleotides complementary to the adaptors. With sufficient dilution, each bead is expected to capture only one cDNA molecule. In the case of

glass slide, the hybridized cDNAs would be sparsely distributed on a two-dimensional axis. Bead capture is adopted in Ion Torrent and SOLiD sequencing workflows, while glass slide in the form of a flow cell is adopted by Illumina. Captured transcripts are then clonally amplified by emulsion PCR into bead-bound libraries, or by isothermal bridge amplification into clusters of DNA clones. Finally, sequencing primers are added to initiate the actual sequencing process. Raw sequence reads are mapped to specific transcripts of the reference genome and normalized to provide accurate expression output data in RKPM or TPM. Currently, the Illumina sequencing-by-synthesis platform (SBS) is considered the most dominant next generation sequencing (NGS) technology (Figure 3) (Greenleaf and Sidow, 2014). Other NGS platforms include (1) Ion Torrent that also works on the principle of SBS but detects H^+ proton instead of fluorescence during nucleotide addition, (2) ABI SOLiD that relies on ligation chemistry (synthesis-by-ligation), and (3) single molecule sequencing technologies from PacBio and Oxford Nanopore. For more information on the various NGS platforms, the reader is referred to comprehensive reviews published elsewhere (Heather and Chain, 2016; Mardis, 2017; Schadt et al., 2010).

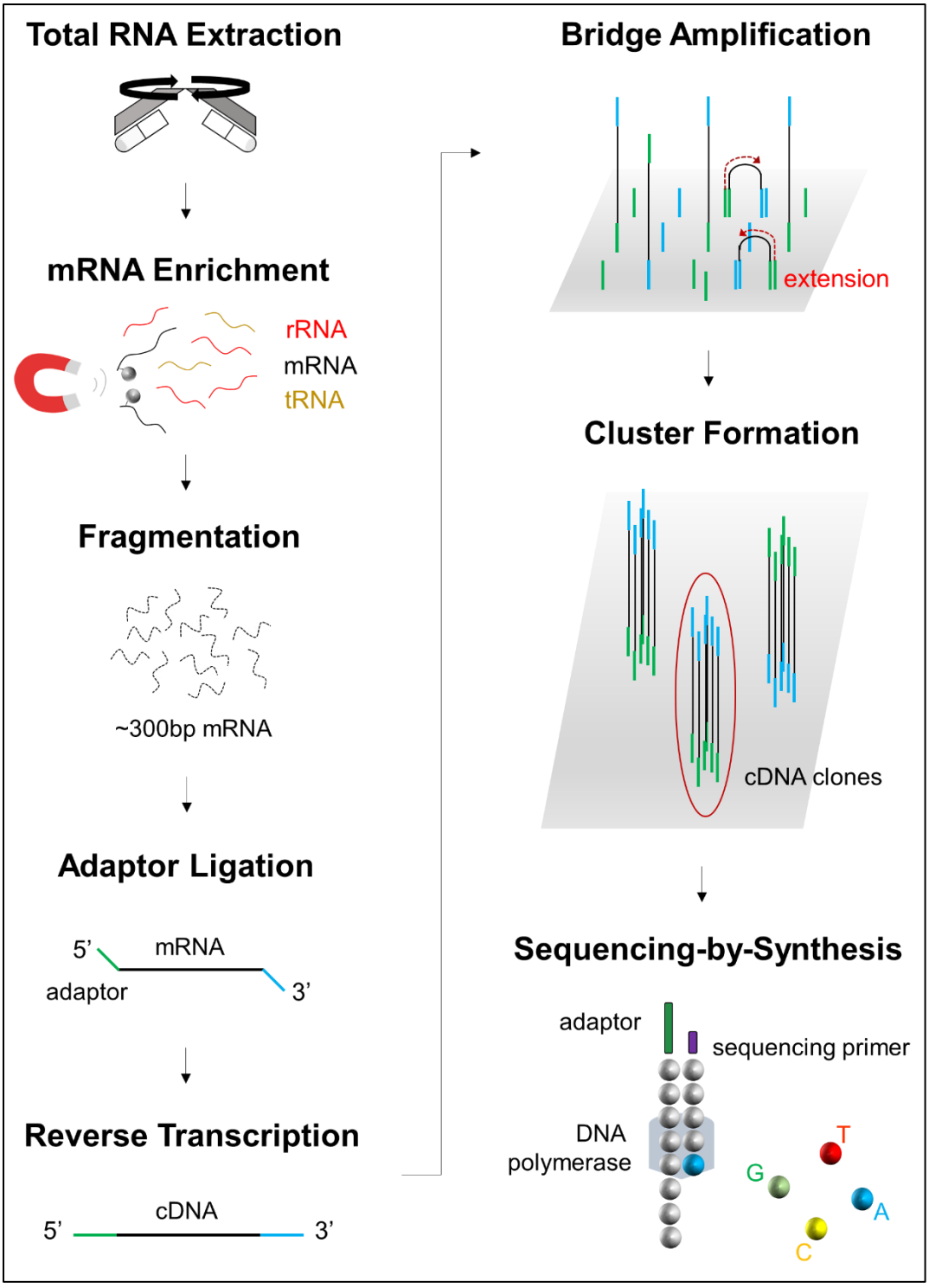

Fig 3. Schematic of Illumina RNA-Seq workflow.

Example 3: RKPM and TPM Calculation. An RNA-Seq experiment was performed to determine the transcript abundance in a cell culture with three technical replicates. From simplicity, it is assumed that the cell has expressed 5 different genes. Results of the sequencing run are presented in Table 3 as raw transcript counts. Determine the normalized TPM and RKPM values.

Table 3. Transcript abundance data from RNA-Seq

		Mapped Reads ($\times 10^6$)		
	Size (kb)	**Replicate I**	**Replicate II**	**Replicate III**
Gene A	0.5	1.2	1.5	1.6
Gene B	1	3.0	3.2	2.9
Gene C	2	4.0	4.2	4.2
Gene D	3	2.0	1.8	2.2
Gene E	4	0.5	0.5	0.6

TPM is obtained after two normalization steps, first with gene size and then by sequencing depth. Sample calculation for Replicate I, Gene A with size of 0.5 kb.

$$\frac{seq\ counts}{gene\ size} = \frac{1.2}{0.5} = 2.40$$

		Mapped Reads ($\times 10^6$)			
1st Normalization by Gene Size		**Size (kb)**	**Replicate I**	**Replicate II**	**Replicate III**
	Gene A	0.5	2.40	3.00	3.20
	Gene B	1	3.00	3.20	3.30
	Gene C	2	2.00	2.10	2.10
	Gene D	3	0.67	0.60	0.73
	Gene E	4	0.13	0.13	0.15
	Total Count		8.19	9.03	9.48

Total sequencing depth of Replicate I after 1st normalization is 8.19 million reads.

$$TPM = \frac{normalized\ seq\ counts}{total\ counts} = \frac{2.40}{8.19} = 0.29$$

2nd Normalization by Sequencing Depth		Mapped Reads in TPM		
	Size (kb)	**Replicate I**	**Replicate II**	**Replicate III**
Gene A	0.5	0.29	0.33	0.34
Gene B	1	0.37	0.35	0.35
Gene C	2	0.24	0.23	0.22
Gene D	3	0.08	0.07	0.08
Gene E	4	0.02	0.01	0.02
Total Count		1.00	1.00	1.00

RKPM is obtained after two normalization steps in the reverse order of TPM, first with sequencing depth and then by gene size. Total sequencing depth of Replicate I before normalization is 10.7 million reads ($Total\ seq\ counts = 1.2 + 3.0 + 4.0 + 2.0 + 0.5 = 10.7$).

Sample calculation for Replicate I, Gene A (1st normalization by sequencing depth).

$$\frac{seq\ counts}{total\ counts} = \frac{1.2}{10.7} = 0.11$$

1st Normalization by Sequencing Depth		Mapped Reads ($\times 10^6$)		
	Size (kb)	**Replicate I**	**Replicate II**	**Replicate III**
Gene A	0.5	0.11	0.13	0.13
Gene B	1	0.28	0.29	0.28
Gene C	2	0.37	0.38	0.35
Gene D	3	0.19	0.16	0.18
Gene E	4	0.05	0.04	0.05
Total Count		1.00	1.00	1.00

2nd normalization by gene size for the same gene, $RKPM = \frac{normalized\ seq\ counts}{gene\ size} = \frac{0.11}{0.5} = 0.22$

2nd Normalization by Gene Size			Mapped Reads in RKPM		
		Size (kb)	**Replicate I**	**Replicate II**	**Replicate III**
	Gene A	0.5	0.22	0.27	0.27
	Gene B	1	0.28	0.29	0.28
	Gene C	2	0.19	0.19	0.18
	Gene D	3	0.06	0.05	0.06
	Gene E	4	0.01	0.01	0.01
	Total Count		0.75	0.81	0.80

Note that RKPM can be converted to TPM by dividing by the total RKPM values. Sample calculation for Replicate I, Gene A with RKPM value of 0.22, $TPM = \frac{RKPM}{\sum RKPM} = \frac{0.22}{0.75} = 0.29$.

Reporter Proteins. Reporter proteins are fluorescent proteins, or enzymes that are able to convert specific substrates into colorimetric or luminescent products. To quantify gene expression, genes that encode for reporter proteins are assembled downstream of DNA promoters in synthetic genetic circuits, and then introduced into the biological host of choice by either electroporation or chemical transformation. This approach has been adopted to characterize promoter strengths in diverse genetic context, including bacterial, yeast, plant, mammalian cells, and cell-free extracts (Auslander et al., 2012; Canton et al., 2008; Chappell et al., 2013; Huang et al., 2010; Redden and Alper, 2015; Reeve et al., 2016; Schaumberg et al., 2016; Wong et al., 2015). The amount of reporter proteins generated during gene expression is assumed to be proportional to background-subtracted readings measured on optical instruments; such as, flow cytometry and fluorescent plate readers. Reporter proteins could also be purified and measured at different concentrations to obtain calibration plots of the optical readers. Direct quantification of gene expression by reporter proteins circumvent the need to isolate mRNA from cell cultures, a procedure that is often time consuming, labor intensive, costly and prone to human error. For this reason, the use of reporter proteins is preferred over qPCR for quantifying gene expression, especially for the large-scale

validation of DNA promoters. Importantly, the results of this approach reflect the combined effect of transcription, translation and post-translational activities, rather than a true measurement of mRNA transcription alone. Incorporating the results of reporter protein outputs into computational models could give estimate of the amount of transcript produced. This will require either additional experiments or probing through systems biology databases to determine key modelling parameters, including the rates of mRNA degradation, protein maturation, protein degradation, and cell growth (Canton et al., 2008; Milo et al., 2010). Reporter proteins could also be used for functional analysis. For example, promoter sequences could be mutated and then assayed for changes in gene expression. When complemented with DNA sequencing, this procedure could rapidly identify core promoter regulatory regions and develop synthetic promoter libraries of varying strengths (Alper et al., 2005; Hartner et al., 2008; Rud et al., 2006; Siegl et al., 2013).

Due to the ease of use, fluorescent proteins are among the most commonly used reporter proteins in gene expression studies. The superfolder green fluorescent protein (sfGFP) in particular, is a highly evolved variant of wild type GFP with unmatched performance in terms of folding kinetics, solubility and tolerance to high temperature and chemicals (Pédelacq et al., 2006). A major limitation of GFP and its variants however, is the need for molecular oxygen as cofactor during fluorophore formation, thus restricting its application to only aerobic biological systems. Flavin mononucleotide-based fluorescent proteins (FbFP) was later developed to enable fluorescent detection in both aerobic and anaerobic condition, but the low output fluorescence remains a technical hurdle that hindered widespread adoption (Drepper et al., 2007). For anaerobic biological systems, the use of non-fluorescent reporters such as light emitting luciferases and absorbance changing enzymes are still commonplace. Popular examples of enzyme-based reporter proteins include β-galactosidase, β-glucosidase and β-glucuronidase (Partow et al., 2010; Siegl et al., 2013). The recent development of Spinach2, an RNA aptamer that binds to 3,5-difluoro-4-hydroxybenzylidene imidazolinone (fluorophore mimic of GFP), provides an exciting platform to study gene expression at the mRNA level *in vivo* (Strack et al., 2013).

Bioinformatics Analysis / Applications

Bioinformatics can be applied to the study of transcriptional activities in two major ways – (1) the prediction of gene expression using sequence features (such as, response elements in promoters)

or from the expression of other genes, and (2) constructing mathematical models for further analyses.

Predicting Gene Expressions. In prokaryotes, it is common to find several genes with related functions grouped together and regulated by a common promoter and a set of enhancers, forming an operon. Hence, it is likely that the expression of genes within the same operon to be correlated (Sabatti et al., 2002). This is supported by a study (De Hoon et al., 2004) examining the gene expressions between genes within operons and across operons, and found that genes within operons show higher correlation than genes across operons. Using expressional correlation, De Hoon et al. (2014) can predict whether the genes are from the same operon, with 79.9% accuracy. This concept of expressional correlation from genes within the same operon has been used to improve microarray analysis by noise reduction (Xiao et al., 2006). However, despite having the same promoter, the correlation is not absolute as the relative position of the genes within the operon may affect expression. A study on *Streptomyces coelicolor* operons found that the expression of genes decreases along the relative position on the operon by normalizing expression of all genes in the operon to that of the first gene in the operon (Laing et al., 2006) but this expression decline is not statistically significant up to the 5^{th} gene on the operon. This is not observed in which was not observed in *Escherichia coli* where the expression of all genes in the operon is constant regardless of positioning in the operon. However, a study using synthetic operon found that gene expression increases linearly with the distance from the start of a gene to the end of the operon (Lim et al., 2011). Nevertheless, these studies suggest that expression of genes within the same operon can be predicted by knowing the expression of one of the genes within the operon, subjected to species-specific variability as in the case of *S. coelicolor*.

In both prokaryotes and eukaryotes, several genes may be activated by the same transcription factor. Similar to the reason why genes within the same operon is likely to be expressionally correlated, a set of genes activated by the same set of transcription factors are also likely to be expressionally correlated. Binding of transcription factors to the promoter may affect chromatin accessibility (Lamparter et al., 2017), leading to correlated expression patterns from genes affected by the same transcription factors (Mahdevar et al., 2013). A study (Zhang and Li, 2017) examining more than 1000 human transcription factors found that transcription factor usage can be used to

predict gene expression level (r^2 of up to 0.617). Genes with similar sequence features in the promoter sequence may also exhibit correlated expression profiles. For example, it has been shown that genes with dioxin-response element in their promoter demonstrate correlated expression (Kim et al., 2006). Similar cases have also been discovered in other response elements; such as, immune response (Care et al., 2015). It is plausible to conceive that several genes must be expressed in response to a given stimulus (Kim et al., 2003), and the regulatory signaling of such stimulus may involve the same set of transcription factors or response elements.

This is supported by a previous study examining the functional aspects of genes with correlated expressions and found that genes with correlated expression demonstrate functional correspondence (Reverter et al., 2005). Moreover, a study on transcriptomics (Ling et al., 2008) also found large number of genes that are correlated. Hence, it may be possible to predict entire transcriptome from known expression of a handful of genes by constructing gene co-expression network. This concept is demonstrated by a study (Ling and Poh, 2014), which uses the expression values of 59 genes to predict the expression of the entire *E. coli* transcriptome. The correlation between predicted and actual expression value is 0.467, which is similar to the microarray intra-array variation. This suggests that intra-array variation accounts for a substantial portion of the transcriptome prediction error and further strengthen the potential of this approach with more reliable experimental data. Ling and Poh (2014) demonstrated the application of their predictive model using a case study – hydrogenase 2 to hydrogen production during glucose (Maeda et al., 2007) or glycerol fermentation (Trchounian et al., 2013) and hydrogenase 2 maturation endopeptidase (hybD) is involved in the maturation of hydrogenase 2. The model predicted 87 genes significantly affected by the 56% knockdown of hybD, and was subjected to Gene Ontology Enrichment Analysis (Zheng and Wang, 2008). All 5 significant molecular functions enriched were of carbon/sugar transferase-typed activity, which corresponds to the expected activity as previously described (Maeda et al., 2007; Trchounian et al., 2013). Knowing that 56% knockdown of hybD significantly affects the expression of 87 genes, it is then possible to ask for the range of hybD expression variation that will not significantly affect any other genes; that is, the expression buffer of hybD. Using 10% stepwise changes of *hybD* from 100% knockout to 2x over-expression the number of affected genes is symmetrical and fits a quadratic model and solving the roots of

the quadratic model, the expressional buffer of hybD in *E. coli* MG1655 is estimated to be we estimate to be between 73.88% and 124.52% of its mean expression.

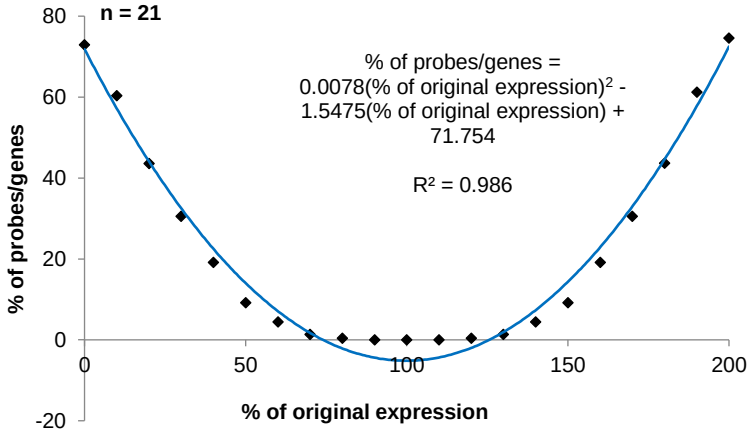

Fig 4. Percentage of reachable genes affected by varying levels of hydrogenase 2 maturation endopeptidase (hybD) expressions.

Two sequence features of the gene, GC content and codon usage, have been found to be useful in predicting gene expressions. In a study of more than four thousand human promoters across eight different cell lines (Landolin et al., 2010), it was found that the GC content of promoters can be predictive of gene expression as promoters high in GC content (more than 50% GC) demonstrating constitutive expression and promoters with low GC content (less than 50% GC) more inclined towards cell-specific expression. This resulted in significant correlation (r = 0.43) between promoter activity and reporter gene expression. Of the 789 genes expressed in all eight cell lines, 719 (91%) were genes with high GC promoters and only 70 (9%) were from genes with low GC promoters. Conversely, 378 of the 483 genes (78.3%) that were expressed in only one of the eight cell lines were from low GC promoters. This suggests that GC content of human promoters can be indicative of cell specificity.

However, the relative impact of GC content and codon usage on gene expression and transcript stability is a subject of debate. A study (Barahimipour et al., 2015) attempted to resolve this by designing 4 yellow fluorescent protein (YFP) genes encoding the same amino acid sequence for expression in *Chlamydomonas reinhardtii* using differing GC content and codon usages. Using this, Barahimipour et al. (2015) found that codon usage plays a more important role in expression

efficiency and transcript stability compared to GC content in *C. reinhardtii*. Expression efficiency is contributed by translational efficiency as the relative abundance between RNA and protein are highly correlated using RNA blot analysis, suggesting that expression efficiency is a result of transcriptional efficiency and the codon usage plays an important role in transcriptional efficiency. Moreover, this study also adds strength to the usefulness of transcriptomics analysis as a proxy for proteomics analysis as it depends on high correlation between transcript abundance and peptide abundance, given that it is experimentally easier to conduct transcriptomics studies (using experimental tools such as microarrays and RNA-seq) than proteomics studies (using experimental tools such as mass spectrometry).

There are a number of metrics to calculate codon usages bias. These include codon adaptation index (Sharp and Li, 1987), relative codon bias strength (Das et al., 2012), relative codon adaptation (Fox and Erill, 2010), and modified relative codon bias strength (Sahoo and Das, 2014). Fox and Erill (2010) used relative codon usage bias to predict the expression levels of *E. coli* genes of more than 1000 bp, achieving a correlation of 0.489 between predicted and actual expression. Similarly, a study (Das et al., 2017) found correlation between these metrics but modified relative codon bias strength was found to have the highest correlation with transcriptomics ($r = -0.31$) and proteomics ($r = -0.44$). Although the correlations are not high, these studies do suggest the potential of using codon usage bias to predict transcriptional activities.

Modeling and Simulation. The main reason for studying transcriptional activities of a cell; and by extension, the entire transcriptome at large; on the premise of correlation between transcript abundance and peptide / protein abundance; is that the knowledge of transcriptional activities allows for a door towards elucidating the genetic and biochemical basis of various conditions. For example, it is the fundamental step for analyzing differential gene expression between various conditions (such as, disease versus healthy, or different treatment conditions) or across various time points (Creecy and Conway, 2015; Du et al., 2017; Herrera-Marcos et al., 2017; Kato et al., 2005; Lin et al., 2016; Liu et al., 2015), or to construct models to predict the phenotype given a transcriptomic change (transcriptome-phenotype correlation).

A biochemical reaction is commonly written as (where A, B, and C are the substrates; P, Q, and R are the products), with the following rate laws,

$$A \xrightarrow{Enzyme} P \qquad rate\ law\ 1 = \frac{k_{cat}[Enzyme][A]}{K_m + [Enzyme][A]}$$

$$B + C \xrightarrow{Enzyme} Q + R \qquad rate\ law\ 2 = \frac{k_{cat}[Enzyme][B][C]}{K_m + [Enzyme][B][C]}$$

Where k_{cat} and K_m are the turnover number (per unit time) and Michaelis-Menten constant (concentration) of the enzyme respectively, and $k_{cat}[Enzyme]$ corresponds to the V_{max} of the reaction (per unit time). When there is more than one substrate or product, the mechanism can take more complicated forms; such as ternary-complex mechanisms (Yang et al., 2011) or ping-pong mechanisms (Nakamura et al., 1994). However, these more complicated mechanisms will be unwieldy to be extended to enzymes using more than 2 substrates. In addition, these mechanisms will require a larger set of enzyme kinetics which is not easily obtainable. Hence, we propose to use an approximation where we consider the enzyme can only work when all the required substrates are present at the same location and with a random probability. Hence, the approximated rate law in generalized form can be written as

$$\frac{k_{cat}[enzyme](\prod_{i=1}^{N}[substrate_i])}{K_m + (\prod_{i=1}^{N}[substrate_i])}$$

The concentration of enzymes, which can be directly represented by the transcriptional activity of the gene encoding the enzyme, play a crucial role in the rate laws. The concentrations of both substrate(s) and product(s) over time can be defined as ordinary differential equations (ODEs) as

$$\frac{d[A]}{dt} = -rate\ law\ 1 \qquad \frac{d[P]}{dt} = rate\ law\ 1$$

$$\frac{d[B]}{dt} = \frac{d[C]}{dt} = -rate\ law\ 2 \qquad \frac{d[P]}{dt} = \frac{d[Q]}{dt} = rate\ law\ 2$$

As enzymes are the result of gene expression, the concentration of an enzyme can be modelled using ODEs developed in previous studies (Jayaraman et al., 2016; Saeidi et al., 2011). Briefly, expressed protein concentration from an inducible promoter can be modelled as

$$For\ constitutive\ expression: \frac{d[RNA]}{dt} = v_0 - \gamma_{RNA}[RNA]$$

$$For\ inducible\ expression: \frac{d[RNA]}{dt} = v_0 + \frac{v_{max} + [inducer]^n}{K_m + [inducer]^n} - \gamma_{RNA}[RNA]$$

$$For\ repressible\ expression: \frac{d[RNA]}{dt} = v_0 + \frac{v_{max}}{K_m + [repressor]^n} - \gamma_{RNA}[RNA]$$

$$\frac{d[Enzyme]}{dt} = \frac{d[protein]}{dt} = \beta[RNA] - \gamma_{protein}[protein]$$

Where v_0 and v_{max} are the baseline and maximum expression (concentration per second) of the promoter, K_m is the Michaelis-Menten constant (concentration), n is the Hill coefficient, β is the relative ribosome binding site (RBS) strength (Wang et al., 2017), and γ_{RNA} and $\gamma_{protein}$ are the degradation rates (per unit time) of RNA and protein respectively.

Once created, models can be used as both a repository of knowledge and an *in silico* platform for hypothesis testing and analysis (Ling, 2016) for the purpose of advising the experimentalist on the parameters to optimize for the optimal yield (MacDonald et al., 2011). This approach had been used in several studies. For example, an enzyme pathway kinetic model of mevalonate pathway had been constructed to study the crucial steps for the production of amorphadiene from mevalonate in *E. coli* (Weaver et al., 2015). After model construction, global sensitivity analysis was carried out to determine which of the six enzymes are important and found that amorphadiene synthase expression and activity are most critical. Another study attempted to increase carotenoid production in maize also constructed an enzyme kinetic model (Comas et al., 2016). While amorphadiene from mevalonate is a linear pathway (Weaver et al., 2015), many carotenoids can be synthesized from phytoene in a branched pathway with many enzymes (Comas et al., 2016); hence, it is not obvious which enzyme expression to overexpress or knockdown to result in increasing specific carotenoids. By analyzing the model, four independent maize lines were engineered using insights from the model analysis and validated experimentally.

These studies suggest that modeling and simulation can provide important computational tools to advise experimentalists but the success of these tools depends on the quality of characterization data from previous experiments. The quality of these experimental data can be further strengthened using bioinformatics approaches to reduce noise, as in the case of Xiao et al. (2006). Hence, there is a reflective process between laboratory and experimental experimentation. It can then be expected that increased experimental data will improve current computational methods, and improved computational methods will better inform laboratory experiments. This trend is likely to stay in the foreseeable future.

References

Adey, A., Morrison, H.G., Asan, Xun, X., Kitzman, J.O., Turner, E.H., Stackhouse, B., MacKenzie, A.P., Caruccio, N.C., Zhang, X., et al. (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. *11*, R119.

Alper, H., Fischer, C., Nevoigt, E., and Stephanopoulos, G. (2005). Tuning genetic control through promoter engineering. Proc. Natl. Acad. Sci. U.S.A. *102*, 12678–12683.

Andersen, C.L., Jensen, J.L., and Orntoft, T.F. (2004). Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. Cancer Research *64*, 5245–5250.

Auslander, S., Auslander, D., Muller, M., Wieland, M., and Fussenegger, M. (2012). Programmable single-cell mammalian biocomputers. Nature *487*.

Barahimipour, R., Strenkert, D., Neupert, J., Schroda, M., Merchant, S.S., and Bock, R. (2015). Dissecting the contributions of GC content and codon usage to gene expression in the model alga Chlamydomonas reinhardtii. Plant J. *84*, 704–717.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., et al. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat. Genet. *29*, 365–371.

Bustin, S.A. (2000). Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. J. Mol. Endocrinol. *25*, 169–193.

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., et al. (2009). The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin. Chem. *55*, 611–622.

Bustin, S.A., Beaulieu, J.-F., Huggett, J., Jaggi, R., Kibenge, F.S.B., Olsvik, P.A., Penning, L.C., and Toegel, S. (2010). MIQE précis: Practical implementation of minimum standard guidelines for fluorescence-based quantitative real-time PCR experiments. BMC Mol. Biol. *11*, 74.

Canton, B., Labno, A., and Endy, D. (2008). Refinement and standardization of synthetic biological parts and devices. Nat Biotechnol *26*.

Care, M.A., Westhead, D.R., and Tooze, R.M. (2015). Gene expression meta-analysis reveals immune response convergence on the IFNγ-STAT1-IRF1 axis and adaptive immune resistance mechanisms in lymphoma. Genome Medicine *7*, 96.

Chappell, J., Jensen, K., and Freemont, P.S. (2013). Validation of an entirely in vitro approach for rapid prototyping of DNA regulatory elements for synthetic biology. Nucleic Acids Res. *41*, 3471–3481.

Chen, F., Dong, M., Ge, M., Zhu, L., Ren, L., Liu, G., and Mu, R. (2013). The history and advances of reversible terminators used in new generations of sequencing technology. Genomics Proteomics Bioinformatics *11*, 34–40.

Comas, J., Benfeitas, R., Vilaprinyo, E., Sorribas, A., Solsona, F., Farré, G., Berman, J., Zorrilla, U., Capell, T., Sandmann, G., et al. (2016). Identification of line-specific strategies for improving carotenoid production in synthetic maize through data-driven mathematical modeling. Plant J. *87*, 455–471.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., et al. (2016). A survey of best practices for RNA-seq data analysis. Genome Biol. *17*, 13.

Creecy, J.P., and Conway, T. (2015). Quantitative bacterial transcriptomics with RNA-seq. Curr. Opin. Microbiol. *23*, 133–140.

Croucher, N.J., and Thomson, N.R. (2010). Studying bacterial transcriptomes using RNA-seq. Curr. Opin. Microbiol. *13*, 619–624.

Croucher, N.J., Fookes, M.C., Perkins, T.T., Turner, D.J., Marguerat, S.B., Keane, T., Quail, M.A., He, M., Assefa, S., Bähler, J., et al. (2009). A simple method for directional transcriptome sequencing using Illumina technology. Nucleic Acids Res. *37*, e148.

Das, S., Roymondal, U., Chottopadhyay, B., and Sahoo, S. (2012). Gene expression profile of the cynobacterium synechocystis genome. Gene *497*, 344–352.

Das, S., Chottopadhyay, B., and Sahoo, S. (2017). Comparative Analysis of Predicted Gene Expression among Crenarchaeal Genomes. Genomics Inform *15*, 38–47.

De Hoon, M.J.L., Imoto, S., Kobayashi, K., Ogasawara, N., and Miyano, S. (2004). Predicting the operon structure of Bacillus subtilis using operon length, intergene distance, and gene expression information. Pac Symp Biocomput 276–287.

Drepper, T., Eggert, T., Circolone, F., Heck, A., Krauss, U., Guterl, J.-K., Wendorff, M., Losi, A., Gärtner, W., and Jaeger, K.-E. (2007). Reporter proteins for in vivo fluorescence without oxygen. Nat. Biotechnol. *25*, 443–445.

Du, Y., Zhao, B., Liu, Z., Ren, X., Zhao, W., Li, Z., You, L., and Zhao, Y. (2017). Molecular Subtyping of Pancreatic Cancer: Translating Genomics and Transcriptomics into the Clinic. J Cancer *8*, 513–522.

Dundas, J., and Ling, M.H. (2012). Reference genes for measuring mRNA expression. Theory in Biosciences *131*, 1–9.

Fox, J.M., and Erill, I. (2010). Relative codon adaptation: a generic codon bias index for prediction of gene expression. DNA Res. *17*, 185–196.

Gertz, J., Varley, K.E., Davis, N.S., Baas, B.J., Goryshin, I.Y., Vaidyanathan, R., Kuersten, S., and Myers, R.M. (2012). Transposase mediated construction of RNA-seq libraries. Genome Res. *22*, 134–141.

Greenleaf, W.J., and Sidow, A. (2014). The future of sequencing: convergence of intelligent design and market Darwinism. Genome Biol. *15*, 303.

Hartner, F.S., Ruth, C., Langenegger, D., Johnson, S.N., Hyka, P., Lin-Cereghino, G.P., Lin-Cereghino, J., Kovar, K., Cregg, J.M., and Glieder, A. (2008). Promoter library designed for fine-tuned gene expression in Pichia pastoris. Nucleic Acids Res. *36*, e76.

Heather, J.M., and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics *107*, 1–8.

Hellemans, J., Mortier, G., De Paepe, A., Speleman, F., and Vandesompele, J. (2007). qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. Genome Biol. *8*, R19.

Herrera-Marcos, L.V., Lou-Bonafonte, J.M., Arnal, C., Navarro, M.A., and Osada, J. (2017). Transcriptomics and the Mediterranean Diet: A Systematic Review. Nutrients *9*.

Huang, H.-H., Camsund, D., Lindblad, P., and Heidorn, T. (2010). Design and characterization of molecular tools for a Synthetic Biology approach towards developing cyanobacterial biotechnology. Nucleic Acids Res. *38*, 2577–2593.

Jayaraman, P., Devarajan, K., Chua, T.K., Zhang, H., Gunawan, E., and Poh, C.L. (2016). Blue light-mediated transcriptional activation and repression of gene expression in bacteria. Nucleic Acids Res. *44*, 6994–7005.

Kang, A., and Chang, M.W. (2012). Identification and reconstitution of genetic regulatory networks for improved microbial tolerance to isooctane. Mol Biosyst *8*, 1350–1358.
Kato, H., Saito, K., and Kimura, T. (2005). A perspective on DNA microarray technology in food and nutritional science. Curr Opin Clin Nutr Metab Care *8*, 516–522.
Kia, A., Gloeckner, C., Osothprarop, T., Gormley, N., Bomati, E., Stephenson, M., Goryshin, I., and He, M.M. (2017). Improved genome sequencing using an engineered transposase. BMC Biotechnol. *17*, 6.
Kim, B.-R., Hu, R., Keum, Y.-S., Hebbar, V., Shen, G., Nair, S.S., and Kong, A.-N.T. (2003). Effects of Glutathione on Antioxidant Response Element-Mediated Gene Expression and Apoptosis Elicited by Sulforaphane. Cancer Res *63*, 7520.
Kim, W.K., In, Y.-J., Kim, J.-H., Cho, H.-J., Kim, J.-H., Kang, S., Lee, C.Y., and Lee, S.C. (2006). Quantitative relationship of dioxin-responsive gene expression to dioxin response element in Hep3B and HepG2 human hepatocarcinoma cell lines. Toxicology Letters *165*, 174–181.
Laing, E., Mersinias, V., Smith, C.P., and Hubbard, S.J. (2006). Analysis of gene expression in operons of Streptomyces coelicolor. Genome Biology *7*, R46–R46.
Lamparter, D., Marbach, D., Rueedi, R., Bergmann, S., and Kutalik, Z. (2017). Genome-Wide Association between Transcription Factor Expression and Chromatin Accessibility Reveals Regulators of Chromatin Accessibility. PLOS Computational Biology *13*, e1005311.
Landolin, J.M., Johnson, D.S., Trinklein, N.D., Aldred, S.F., Medina, C., Shulha, H., Weng, Z., and Myers, R.M. (2010). Sequence features that drive human promoter function and tissue specificity. Genome Res. *20*, 890–898.
Lim, H.N., Lee, Y., and Hussein, R. (2011). Fundamental relationship between operon organization and gene expression. Proc. Natl. Acad. Sci. U.S.A. *108*, 10626–10631.
Lin, M., Lachman, H.M., and Zheng, D. (2016). Transcriptomics analysis of iPSC-derived neurons and modeling of neuropsychiatric disorders. Mol. Cell. Neurosci. *73*, 32–42.
Ling, M. (2016). Of (Biological) Models and Simulations. MOJ Proteomics & Bioinformatics *3*, 00093.
Ling, M.H., and Poh, C.L. (2014). A predictor for predicting Escherichia coli transcriptome and the effects of gene perturbations. BMC Bioinformatics *15*, 140.
Ling, H., Chen, B., Kang, A., Lee, J.-M., and Chang, M.W. (2013). Transcriptome response to alkane biofuels in Saccharomyces cerevisiae: identification of efflux pumps involved in alkane tolerance. Biotechnol Biofuels *6*, 95.
Ling, M.H., Lefevre, C., and Nicholas, K.R. (2008). Filtering microarray correlations by statistical literature analysis yields potential hypotheses for lactation research. The Python Papers *3*, 4.
Liu, Y., Ai, N., Liao, J., and Fan, X. (2015). Transcriptomics: a sword to cut the Gordian knot of traditional Chinese medicine. Biomark Med *9*, 1201–1213.
MacDonald, J.T., Barnes, C., Kitney, R.I., Freemont, P.S., and Stan, G.B. (2011). Computational design approaches and tools for synthetic biology. Integr Biol (Camb) *3*, 97–108.
Mäder, U., Nicolas, P., Richard, H., Bessières, P., and Aymerich, S. (2011). Comprehensive identification and quantification of microbial transcriptomes by genome-wide unbiased methods. Curr. Opin. Biotechnol. *22*, 32–41.
Maeda, T., Sanchez-Torres, V., and Wood, T.K. (2007). Enhanced hydrogen production from glucose by metabolically engineered Escherichia coli. Appl. Microbiol. Biotechnol. *77*, 879–890.
Mahdevar, G., Nowzari-Dalini, A., and Sadeghi, M. (2013). Inferring gene correlation networks from transcription factor binding sites. Genes Genet. Syst. *88*, 301–309.

Mardis, E.R. (2017). DNA sequencing technologies: 2006-2016. Nat Protoc *12*, 213–218.
McGettigan, P.A. (2013). Transcriptomics in the RNA-seq era. Curr Opin Chem Biol *17*, 4–11.
Milo, R., Jorgensen, P., Moran, U., Weber, G., and Springer, M. (2010). BioNumbers--the database of key numbers in molecular and cell biology. Nucleic Acids Res. *38*, D750-753.
Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods *5*, 621–628.
Nakamura, A., Haga, K., and Yamane, K. (1994). The transglycosylation reaction of cyclodextrin glucanotransferase is operated by a Ping-Pong mechanism. FEBS Lett. *337*, 66–70.
O'Neil, D., Glowatz, H., and Schlumpberger, M. (2013). Ribosomal RNA depletion for efficient use of RNA-seq capacity. Curr Protoc Mol Biol *Chapter 4*, Unit 4.19.
Partow, S., Siewers, V., Bjørn, S., Nielsen, J., and Maury, J. (2010). Characterization of different promoters for designing a new expression vector in Saccharomyces cerevisiae. Yeast *27*, 955–964.
Pédelacq, J.-D., Cabantous, S., Tran, T., Terwilliger, T.C., and Waldo, G.S. (2006). Engineering and characterization of a superfolder green fluorescent protein. Nat. Biotechnol. *24*, 79–88.
Petrova, O.E., Garcia-Alcalde, F., Zampaloni, C., and Sauer, K. (2017). Comparative evaluation of rRNA depletion procedures for the improved analysis of bacterial biofilm and mixed pathogen culture transcriptomes. Sci Rep *7*, 41114.
Pfaffl, M.W. (2001). A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res. *29*, e45.
Pfaffl, M.W., Tichopad, A., Prgomet, C., and Neuvians, T.P. (2004). Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper--Excel-based tool using pair-wise correlations. Biotechnol. Lett. *26*, 509–515.
Redden, H., and Alper, H.S. (2015). The development and characterization of synthetic minimal yeast promoters. Nat Commun *6*, 7810.
Reeve, B., Martinez-Klimova, E., de Jonghe, J., Leak, D.J., and Ellis, T. (2016). The Geobacillus Plasmid Set: A Modular Toolkit for Thermophile Engineering. ACS Synth Biol *5*, 1342–1347.
Reverter, A., Barris, W., Moreno-Sanchez, N., McWilliam, S., Wang, Y.H., Harper, G.S., Lehnert, S.A., and Dalrymple, B.P. (2005). Construction of gene interaction and regulatory networks in bovine skeletal muscle from expression data. Australian Journal of Experimental Agriculture *45*, 821–829.
Rud, I., Jensen, P.R., Naterstad, K., and Axelsson, L. (2006). A synthetic promoter library for constitutive gene expression in Lactobacillus plantarum. Microbiology (Reading, Engl.) *152*, 1011–1019.
Sabatti, C., Rohlin, L., Oh, M.-K., and Liao, J.C. (2002). Co-expression pattern from DNA microarray experiments as a tool for operon prediction. Nucleic Acids Res. *30*, 2886–2893.
Saeidi, N., Wong, C.K., Lo, T.-M., Nguyen, H.X., Ling, H., Leong, S.S.J., Poh, C.L., and Chang, M.W. (2011). Engineering microbes to sense and eradicate Pseudomonas aeruginosa, a human pathogen. Mol. Syst. Biol. *7*, 521.
Sahoo, S., and Das, S. (2014). Analyzing gene expression and codon usage bias in diverse genomes using a variety of models. Current Bioinformatics *9*, 102–112.
Sanders, R., Mason, D.J., Foy, C.A., and Huggett, J.F. (2014). Considerations for accurate gene expression measurement by reverse transcription quantitative PCR when analysing clinical samples. Anal Bioanal Chem *406*, 6471–6483.
Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. *74*, 5463–5467.

Sarkar, N. (1997). Polyadenylation of mRNA in prokaryotes. Annu. Rev. Biochem. *66*, 173–197.
Schadt, E.E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. Hum. Mol. Genet. *19*, R227-240.
Schaumberg, K.A., Antunes, M.S., Kassaw, T.K., Xu, W., Zalewski, C.S., Medford, J.I., and Prasad, A. (2016). Quantitative characterization of genetic parts and circuits for plant synthetic biology. Nat. Methods *13*, 94–100.
Schmutz, J., Wheeler, J., Grimwood, J., Dickson, M., Yang, J., Caoile, C., Bajorek, E., Black, S., Chan, Y.M., Denys, M., et al. (2004). Quality assessment of the human genome sequence. Nature *429*, 365–368.
Sharma, C.M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., et al. (2010). The primary transcriptome of the major human pathogen Helicobacter pylori. Nature *464*, 250–255.
Sharp, P.M., and Li, W.H. (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. *15*, 1281–1295.
Siegl, T., Tokovenko, B., Myronovskyi, M., and Luzhetskyy, A. (2013). Design, construction and characterisation of a synthetic promoter library for fine-tuned gene expression in actinomycetes. Metab. Eng. *19*, 98–106.
Stewart, F.J., Ottesen, E.A., and DeLong, E.F. (2010). Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. ISME J *4*, 896–907.
Strack, R.L., Disney, M.D., and Jaffrey, S.R. (2013). A superfolding Spinach2 reveals the dynamic nature of trinucleotide repeat-containing RNA. Nat. Methods *10*, 1219–1224.
Thellin, O., Zorzi, W., Lakaye, B., De Borman, B., Coumans, B., Hennen, G., Grisar, T., Igout, A., and Heinen, E. (1999). Housekeeping genes as internal standards: use and limits. J. Biotechnol. *75*, 291–295.
Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.
Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc *7*, 562–578.
Trchounian, K., Soboh, B., Sawers, R.G., and Trchounian, A. (2013). Contribution of hydrogenase 2 to stationary phase H2 production by Escherichia coli during fermentation of glycerol. Cell Biochem. Biophys. *66*, 103–108.
Trevino, V., Falciani, F., and Barrera-Saldaña, H.A. (2007). DNA microarrays: a powerful genomic tool for biomedical and clinical research. Mol. Med. *13*, 527–541.
Van Gelder, R.N., von Zastrow, M.E., Yool, A., Dement, W.C., Barchas, J.D., and Eberwine, J.H. (1990). Amplified RNA synthesized from limited quantities of heterogeneous cDNA. Proc. Natl. Acad. Sci. U.S.A. *87*, 1663–1667.
Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., and Speleman, F. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biology *3*, RESEARCH0034.
Voelkerding, K.V., Dames, S.A., and Durtschi, J.D. (2009). Next-generation sequencing: from basic research to diagnostics. Clin. Chem. *55*, 641–658.

Wagner, G.P., Kin, K., and Lynch, V.J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. *131*, 281–285.

Wang, H., Ling, M.H., Chua, T.K., and Poh, C.L. (2017). Two cellular resource-based models linking growth and parts characteristics aids the study and optimisation of synthetic gene circuits. Engineering Biology *1*, 30–39.

Weaver, L.J., Sousa, M.M.L., Wang, G., Baidoo, E., Petzold, C.J., and Keasling, J.D. (2015). A kinetic-based approach to understanding heterologous mevalonate pathway function in E. coli. Biotechnology and Bioengineering *112*, 111–119.

Westermann, A.J., Förstner, K.U., Amman, F., Barquist, L., Chao, Y., Schulte, L.N., Müller, L., Reinhardt, R., Stadler, P.F., and Vogel, J. (2016). Dual RNA-seq unveils noncoding RNA functions in host-pathogen interactions. Nature *529*, 496–501.

Wong, A., Wang, H., Poh, C.L., and Kitney, R.I. (2015). Layering genetic circuits to build a single cell, bacterial half adder. BMC Biology *13*, 40.

Xiao, G., Martinez-Vaz, B., Pan, W., and Khodursky, A.B. (2006). Operon information improves gene expression estimation for cDNA microarrays. BMC Genomics *7*, 87.

Yang, Y., Yamashita, T., Nakamaru-Ogiso, E., Hashimoto, T., Murai, M., Igarashi, J., Miyoshi, H., Mori, N., Matsuno-Yagi, A., Yagi, T., et al. (2011). Reaction mechanism of single subunit NADH-ubiquinone oxidoreductase (Ndi1) from Saccharomyces cerevisiae: evidence for a ternary complex mechanism. J. Biol. Chem. *286*, 9287–9297.

Zhang, L.-Q., and Li, Q.-Z. (2017). Estimating the effects of transcription factors binding and histone modifications on gene expression levels in human cells. Oncotarget *8*, 40090–40103.

Zheng, Q., and Wang, X.-J. (2008). GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. Nucleic Acids Res. *36*, W358-363.