Targeted transcript profiling by sequencing

Pawel Zajac & Afshin Ahmadian

1Laboratory for Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Scheeles väg 1, SE-171 77 Stockholm, Sweden, 2Royal Institute of Technology (KTH), School of Biotechnology, Division of Gene Technology, Science for Life Laboratory (SciLifeLab) SE-171 21 Solna, Sweden.

In this work we present a targeted gene expression strategy employing trinucleotide threading (TnT) amplification and massive parallel sequencing. We have previously shown that TnT combined with array readout accurately monitors expression levels. However, with this detection strategy spurious products go undetected. Accordingly, we adapted the TnT protocol to massive parallel sequencing to acquire an unbiased view of the entire TnT-generated product population. In this manner we investigated the identity of undesired products, their extent at different oligonucleotide:RNA ratios and their effect on the expression levels. We demonstrate that TnT gene expression profiling with massive sequencing readout renders reliable expression data from as low as 3.5 ng of total RNA. Moreover, using 350 ng of total RNA results in only 0.7% to 1.1% undesired products. When lowering the amount of input material, the undesired product fraction increases but this does not influence the expression profiles.

Gene expression analysis provides an avenue to a wealth of information as the expression levels can shed light on various cellular processes and offer insights about the molecular underpinnings of diseases. As such, approaches for studying mRNA abundances represent a highly important and influential group of methods.

Over the years, several expression analysis methods, catering to different needs, have been developed. Some approaches are capable of analyzing a single gene (or at most a very limited number of genes), albeit at a high sample throughput. The 'gold standard' method of reverse transcription PCR (RT-PCR) falls within this category and is commonly recruited for validation purposes. At the other end of the spectrum are techniques allowing expression analysis of all transcripts. The most well-known of these global approaches relies on hybridization of a sample to DNA microarrays carrying a large number of probes corresponding to the transcripts of interest. However, microarrays suffer from cross-hybridization induced unspecificity and a rather limited dynamic range. Nevertheless, they have been widely applied to study different facets of the transcriptome over the last decade.

Currently, a group of methods collectively termed RNA-Seq, based on readout with massively parallel DNA sequencers, is being widely adopted at the expense of microarrays. In these approaches – originating from expressed sequence tag (EST) sequencing – RNA is converted into a DNA library, which is sequenced and the counts of all different transcripts converted to expression levels. The benefit of sequencing-based approaches is the highly reliable digital readout and a superior dynamic range. Additionally, the sequencing output illuminates transcript structure and reveals, for example, mutations and polymorphisms, thus further characterizing the transcriptional landscape. However, RNA-Seq is still too cost-prohibitive to be performed more routinely. Moreover, the majority of the RNA-Seq protocols do not present the possibility to only target a subset of particularly informative RNA species. Several sequencing instruments are commercially available with Illumina, Life Technologies (both Ion Torrent and SOLiD) and 454/Roche sequencers being the most widely employed. Moreover, there is fervent activity in developing novel sequencing approaches.

Spanning the divide between global and validation methods are intermediary techniques adapted to rapid analysis of moderate gene sets at a high throughput. BeadsArray for the Detection of Gene Expression (BADGE) utilizing the Luminex microsphere suspension arrays, and cDNA-mediated annealing, selection, extension and ligation (DASL) marketed by Illumina are two examples. Recently, the RNA-mediated oligonucleotide annealing, selection and ligation (RASL) method, a forerunner of the DASL technique where ligation occurs directly on RNA and the extension step is omitted, has been adapted for massive sequencing in a method dubbed RASL-seq. Additionally, capture of RNA species of interest with readout using sequencing, in a manner akin to enrichment of selected genomic regions or exomes, has been demonstrated.

Received 10 July 2012
Accepted 14 September 2012
Published 8 November 2012

Correspondence and requests for materials should be addressed to A.A. (afshin.ahmadian@scilifelab.se)
An alternative technique to the aforementioned intermediary methods is trinucleotide threading (TnT). TnT harnesses the specificity of a polymerase and a ligase, in conjunction with a restricted trinucleotide set, to faithfully amplify several genomic regions. Briefly, the expression version of the method involves two probes that are designed to target a region specific to the transcript of interest. These anneal to mRNA-derived single-stranded cDNA in a manner creating a small gap between them. The distinguishing feature of this gap is its composition as it only entails three out of the four possible nucleotides. The concerted action of a polymerase and a ligase bridges this gap and links the two segments creating a full DNA thread. Naturally, each transcript generates one type of thread, the amount of which is dependent upon the prevalence of that transcript. To increase sensitivity, each full DNA thread carries general amplification handles enabling a parallel amplification with a single universal primer pair. All threads are of similar lengths, addressing the length bias frequently observed in PCR. The original gene expression TnT investigation entailed a thread-specific primer extension coupled with hybridization of the extension product to a generic address tag arrays for readout.

Regardless of the choice of expression analysis method, it is important to consider the amount of oligonucleotides or primers (as required by the method) with the RNA input. A balanced oligonucleotide:RNA ratio allows generation of reliable expression profiles while maximizing the resources (for instance, enzymatic activity or sequencing capacity) allocated to actual transcripts, as opposed to undesired side-products such as primer-dimers.

In this study, we have used TnT in conjunction with 454 sequencing to investigate the effect of different ratios of oligonucleotides to input RNA on both the generated expression profiles and on the presence of unwanted products. The massively parallel sequencing readout enables a precise characterization of all obtained species – both desired and undesired – and hence clearly illuminates the outcome of the reaction. We found a direct relationship between higher oligonucleotide:RNA ratios and the occurrence of undesired products. Although the expression patterns were similar, the unwanted products used a significantly higher proportion of sequencing resources when an overabundance of oligonucleotides was employed.

Additionally, as the investigation featured a set of 32 selected genes, a combination of TnT and 454 represents a targeted RNA-Seq strategy whereby transcripts of particular interest can be enriched and analyzed in parallel against a background of the entire mRNA population. This is beneficial as only a small percentage of the about 10,000 protein-coding genes expressed at any given time displays considerable abundance differences when two samples are compared. Accordingly, targeting only these species can provide sufficient information while drastically lessening the dimensionality of the involved assays. The expression profiles obtained with TnT-454 correlate well with both TnT analyzed with an array-based strategy and with RT-PCR. However, the broad dynamic range of the 454 platform allows a reduction in input RNA requirements. Consequently, by choosing informative intermediary gene sets and barcoding transcripts from several different individuals in a combinatorial scheme, we envision TnT in combination with massively parallel sequencing platforms to enable a highly multiplexed analysis both with regard to transcript and sample number.

### Results

Trinucleotide threading is a method capable of multiplex amplification of genomic regions or transcripts in a single tube, while keeping the amount of spurious products to a minimum. Our previous renditions of the method have used arrays with user-selected, albeit fixed, content as readout. As such, potential undesired products went undetected. To investigate the spurious product formation, especially its extent at different oligonucleotide:RNA ratios and if it affects the obtained expression results, we adapted the TnT gene expression approach to a sequencing-based readout with the Roche/454 Genome Sequencer FLX Titanium instrument. The sequencing strategy, although more expensive than conventional arrays, offers an unbiased view of the TnT-generated product population detecting both desired and side products whereby shedding light on the events taking place during this reaction. Total RNA from two cell lines, EFO-21 and SK-MEL-30, was employed at three different oligonucleotide:RNA ratios each. The TnT oligonucleotide concentration was fixed (0.01 nM of each extension primer and 0.05 nM of each thread-joining primer), while progressively lowering the total RNA input (350 ng, 35 ng and 3.5 ng). Accordingly, two reaction triplets with gradually decreasing RNA amounts were set up. The reactions of each triplet were individually barcoded with 454/Roche Multiplex Identifiers (MIDs), pooled and loaded into a single lane of a 454/Roche Genome Sequencer FLX Titanium picotiter plate. The sequencing reads were partitioned based on their MID sequences and subsequently BLAST-searched against a custom database comprising all potential DNA threads. True thread reads and reads aligning to two database entries were isolated and counted. Short reads, reads producing short or inferior alignments and displaying no hits were removed. The criteria for true threads were rather stringent to avoid inclusion of undesired sequences in this category and, accordingly, to enable a reliable extraction of expression profiles. Naturally, with relaxation of the parameters a greater number of reads can be classified as threads.

The statistics of the sequencing run are given in Table 1. On average, the EFO-21 reactions generated about 39,100 reads that passed the filtering steps. For SK-MEL-30 the corresponding read number was 61,900. The distribution of reads among the pooled samples was not even. For the EFO-21 reaction triplet (occupying the same lane in the picotiter plate), the read distribution ranged from 25.6% to 45.2% per reaction. In the case of SK-MEL-30 triplet the interval was between 27.8% and 44.3%. A plausible explanation for this is that the barcoding scheme employing 454 Multiplex Identifiers is slightly biased. For example, the efficiencies of the individual barcoding reactions may be different. Nevertheless, the identifiers are useful if the output of one sequencing lane is greater than required by a single sample. True thread reads corresponded to 99.3%, 98.3% and 85.5% of the reads for the EFO-21 RNA dilution series (Table 1). Correspondingly, for SK-MEL-30 these percentages were 98.9%, 94.2% and 59.9%.

| Table 1 | Sequencing statistics and TnT product percentages |
|---------|---------------------------------------------------|
|         | **EFO-21**                                         | **SK-MEL-30**                                    |
|         | 350 ng | 35 ng | 3.5 ng | 350 ng | 35 ng | 3.5 ng |
| Reads   | 34268  | 29966  | 53020  | 82309  | 51874  | 51622  |
| True threads | 34019 (99.3%) | 29467 (98.3%) | 45357 (85.5%) | 81368 (98.9%) | 48852 (94.2%) | 30932 (59.9%) |
| Reads with double hits | 249 (0.7%) | 499 (1.7%) | 7663 (14.5%) | 941 (1.1%) | 3022 (5.8%) | 20690 (40.1%) |

The number of sequencing reads passing the filters for each of the six libraries is indicated. These reads have passed the 454 quality filters and have given above threshold values in a BLAST search against a database of all hypothetical threads. Furthermore, the number of true threads and of reads aligning to two hypothetical threads, as well as the corresponding percentages, are shown. Additional information about the data filtering and analysis is provided in the Methods section.
Next, we sought to investigate whether the different oligonucleotide:RNA ratios affected or skewed the obtained expression data. To this end, we calculated all possible pairwise Pearson correlation coefficients between the reactions employing total RNA from the same cell line (Table 2). The obtained true thread counts were used as input in this analysis. Overall, a high level of correlation was observed. For EFO-21 the correlations between the highest total RNA amount (350 ng) and the dilutions were 0.98 for the 35 ng and 0.93 for the 3.5 ng samples, respectively. The analogous coefficients for SK-MEL-30 were 0.97 and 0.90. Accordingly, reduction of input total RNA and, consequently, altered oligonucleotide:RNA ratios did not significantly change the generated expression profiles. Thereafter, expression data generated by the TnT-454 and TnT-array approaches was compared. Generally, a high level of correlation was observed for the total RNA of 350 and 35 ng, respectively (data not shown). The highly abundant transcripts were the same with both detection approaches. Equally, there was congruency with respect to lowly expressed genes. However, the array platform generated lower signals for the lowest input total RNA (3.5 ng). Taken together, we conclude that both readout platforms – massively parallel 454 sequencing and conventional arrays – produce comparable expression profiles when the concentration of target template is high while the sequencing approach renders reliable results with as low as 3.5 ng of total RNA.

To further validate the TnT-454 strategy, three genes exhibiting diverse expression behaviour between the two analyzed cell lines in the sequencing data were studied with the established RT-PCR method. One of the genes – DCT – was found to be expressed in SK-MEL-30, but was not detected in EFO-21. APLP1 displayed an inverse profile with expression in EFO-21, but not in SK-MEL-30. Finally, expression of LAMB1 was observed in both cell lines. The RT-PCR results displayed good accordance with the TnT-454 profiles (data not shown). As such, the TnT-454 method is a reliable means of measuring abundance levels of selected transcripts.

Having established that the sequencing-based readout of TnT reactions does produce dependable expression profiles, we next turned our attention to undesired products. These were defined as reads mapping to two potential DNA threads with alignment characteristics above a pre-defined threshold (see Methods). The fractions of such undesired entities were 0.7%, 1.7% and 14.5% for the EFO-21 reactions employing 350 ng, 35 ng and 3.5 ng total RNA, respectively (Table 1). For SK-MEL-30, the analogous percentages were 1.1%, 5.8% and 40.1% (Table 1). Importantly, although the fraction of undesired products increases as the input material entering the pipeline is reduced, it does not significantly affect the generated expression data, as evidenced by the >0.90 Pearson correlation coefficients between samples featuring different total RNA amounts (Table 2). Accordingly, the undesired products do not skew the obtained expression profiles.

Nevertheless, the side products warrant further attention as they have the potential to consume valuable reagents and use up sequencing resources. To gain insight into the creation of undesired entities, the most occurring side products were extracted from the sequencing data and analyzed. In this analysis, the reactions encompassing 3.5 ng of input material were considered. Five species correspond to the majority of the non-informative reads. For EFO-21 three species account for 95.0% of the undesired reads (BCL2L2-SCD5; 55.8%, SLC2A1-FADS1: 31.0% and S100A8-JUN 8.2%). In the case of SK-MEL-30, four species make up for 95.7% of the undesired products (BCL2L2-SCD5; 81.7%, SOX4-MYC; 6.7%, AQP3-DMKN 3.7% and S100A8-JUN 3.6%). Clearly, the product formed by the extension primer of BCL2L2 and thread-joining primer of SCD5 is the main culprit. This undesired entry entails side-by-side ligation of the above-mentioned TnT primers. However, there is not any apparent molecule able to prime this ligation event. The same pattern is observed for the extension primer of SLC2A1 and thread-joining primer of FADS1, as well as the extension primer of AQPI and thread-joining primer of DMKN. These three species are most probably formed in the trinucleotide threading reaction, making use of yet unidentified molecules for enabling of the ligation. A considerable amount of effort was used during TnT primer design to ascertain that no transcript, complete and/or partial thread could be able to join two mismatching TnT primers, but the presence of such priming molecules cannot be fully excluded. The fourth undesired moiety comprises the extension primer of S100A8 and the thread-joining primer of JUN with a single C base in-between. This structure is most likely also generated in the threading reaction, once again employing an unknown molecule responsible for bringing the two primers in close proximity. On the other hand, the SOX4-MYC moiety encompasses thread-joining primers of SOX4 and MYC and requires the reverse complement of one of the probes. Accordingly, it is created in the PCR amplification step. By employing an alternative clean-up strategy as presented in the Discussion section it could be eliminated.

**Discussion**

Gene expression analysis with trinucleotide threading (TnT) is capable of profiling abundance levels of intermediate transcript sets and, accordingly, complements comprehensive methods targeting the entire transcriptome and single-gene validation approaches. In a proof-of-concept investigation, TnT was shown to accurately monitor expression levels for an 18-gene set (15 targeted genes and 3 housekeeping genes), with the data correlating well to that of the extended primer of JUN with a single C base in-between. This structure is most likely also generated in the threading reaction, once again employing an unknown molecule responsible for bringing the two primers in close proximity. On the other hand, the SOX4-MYC moiety encompasses thread-joining primers of SOX4 and MYC and requires the reverse complement of one of the probes. Accordingly, it is created in the PCR amplification step. By employing an alternative clean-up strategy as presented in the Discussion section it could be eliminated.

| Table 2 | Pearson correlation coefficients |
|----------|--------------------------------|
| **EFO-21** | **SK-MEL-30** |
| | 454 | 454 | 350 ng | 35 ng | 3.5 ng | 350 ng | 35 ng | 3.5 ng |
| 350 ng | 1.00 | 0.98 | 0.93 | 1.00 | 0.97 | 0.90 |
| 35 ng | 1.00 | 0.96 | 1.00 | 1.00 | 0.92 |
| 3.5 ng | 1.00 |

The table shows the Pearson correlation coefficients for EFO-21 and SK-MEL-30 for different input total RNA amounts (350 ng, 35 ng, and 3.5 ng). The highest correlation coefficients are observed for the highest input total RNA amount, with slightly lower coefficients for the 35 ng samples and the lowest for the 3.5 ng samples. The correlation coefficients range from 0.90 to 1.00, indicating a high level of correlation between the expression profiles generated by different input total RNA amounts.
Recently, the RASL technique (RNA-mediated oligonucleotide annealing, selection and ligation) for targeted expression analysis, was combined with massively parallel sequencing. While the RASL-seq protocol recommends an originally relying on array detection, was adopted to readout with massive parallel 454 sequencing. As all generated species are sequenced, a comprehensive view of both the desired and undesired products can be obtained. A set comprising 32 genes was targeted and reliable expression data was generated starting from as low as 3.5 ng of total RNA. The combination of TnT and 454 sequencing produced reliable data as demonstrated by the good correlation with both RT-PCR and TnT read out – the formation of which is severely hampered as outlined above – are unable to participate in the primer extension reaction and are thus invisible on the array.

To study the spurious products of the TnT reaction, the TnT gene expression protocol was adapted to readout with massive parallel 454 sequencing. As all generated species are sequenced, a comprehensive view of both the desired and undesired products can be obtained. A set comprising 32 genes was targeted and reliable expression data was generated starting from as low as 3.5 ng of total RNA. The combination of TnT and 454 sequencing produced reliable data as demonstrated by the good correlation with both RT-PCR and TnT read out – the formation of which is severely hampered as outlined above – are unable to participate in the primer extension reaction and are thus invisible on the array.

Table 3 | Genes and trinucleotide threading (TnT) primers

| ID | Gene | Accession | Extension primer | Thread-joining primer |
|----|------|-----------|-----------------|----------------------|
| a01 | AQP3 | NP_004925.3 | gagggtgacatccacatagtccgaac | TTTAGAAAAGGGTGTTACACCTCCTTCCTTTTA |
| a02 | PCDH21 | NP_033010.1 | gagggtgacatccacatagtccgaac | TTTAGAAAAGGGTGTTACACCTCCTTCCTTTTA |
| a03 | NM_002890.2 | gagggtgacatccacatagtccgaac | TTTAGAAAAGGGTGTTACACCTCCTTCCTTTTA |
| a04 | C1orf57 | NP_024323.3 | gagggtgacatccacatagtccgaac | TTTAGAAAAGGGTGTTACACCTCCTTCCTTTTA |
| a05 | SPPN | NM_033311.2 | gagggtgacatccacatagtccgaac | TTTAGAAAAGGGTGTTACACCTCCTTCCTTTTA |
| a06 | DCT | NP_007152.2 | gagggtgacatccacatagtccgaac | TTTAGAAAAGGGTGTTACACCTCCTTCCTTTTA |
| a09 | DLL4 | NP_002964.3 | gagggtgacatccacatagtccgaac | TTTAGAAAAGGGTGTTACACCTCCTTCCTTTTA |
| a10 | LAMB1 | NP_002291.2 | gagggtgacatccacatagtccgaac | TTTAGAAAAGGGTGTTACACCTCCTTCCTTTTA |
| a13 | SLCA1 | NP_006516.2 | gagggtgacatccacatagtccgaac | TTTAGAAAAGGGTGTTACACCTCCTTCCTTTTA |
| a15 | AMPL1 | NM_001024681.1 | gagggtgacatccacatagtccgaac | TTTAGAAAAGGGTGTTACACCTCCTTCCTTTTA |
| bCLC2L2 | ECLC2L2 | NP_004052.3 | CAGTGGGAGAACATACGATAGAGGGGGGT | TCTGGATTCAGAGCCGATGTAAGC |
| dCTNISBP | C TNTNISBP1 | NP_002248.2 | CAGTGGGAGAACATACGATAGAGGGGGGT | TCTGGATTCAGAGCCGATGTAAGC |
| dCTNISBP | C TNTNISBP1 | NP_002248.2 | CAGTGGGAGAACATACGATAGAGGGGGGT | TCTGGATTCAGAGCCGATGTAAGC |
| m1 | LEPF1 | NP_016190.2 | CAGTGGGAGAACATACGATAGAGGGGGGT | TCTGGATTCAGAGCCGATGTAAGC |
| m2 | MYC | NP_002267.3 | CAGTGGGAGAACATACGATAGAGGGGGGT | TCTGGATTCAGAGCCGATGTAAGC |
| m4 | GLI1 | NP_002569.1 | CAGTGGGAGAACATACGATAGAGGGGGGT | TCTGGATTCAGAGCCGATGTAAGC |
| m6 | ECTA | NP_002513.3 | CAGTGGGAGAACATACGATAGAGGGGGGT | TCTGGATTCAGAGCCGATGTAAGC |
| v01 | FASC1D2 | NP_002394.1 | CAGTGGGAGAACATACGATAGAGGGGGGT | TCTGGATTCAGAGCCGATGTAAGC |
| v02 | GPD1 | NP_00126767.1 | CAGTGGGAGAACATACGATAGAGGGGGGT | TCTGGATTCAGAGCCGATGTAAGC |
| v04 | SDO5 | NP_00130347.1 | CAGTGGGAGAACATACGATAGAGGGGGGT | TCTGGATTCAGAGCCGATGTAAGC |
| v05 | SNC2 | NP_00136781.1 | CAGTGGGAGAACATACGATAGAGGGGGGT | TCTGGATTCAGAGCCGATGTAAGC |
| v06 | G0X4 | NP_0033107.2 | CAGTGGGAGAACATACGATAGAGGGGGGT | TCTGGATTCAGAGCCGATGTAAGC |
| v07 | MON1B | NP_014902.0 | CAGTGGGAGAACATACGATAGAGGGGGGT | TCTGGATTCAGAGCCGATGTAAGC |
| v09 | 1GF2B2 | NP_00007225.1 | CAGTGGGAGAACATACGATAGAGGGGGGT | TCTGGATTCAGAGCCGATGTAAGC |
| v14 | F2D8 | NP_0013046.1 | CAGTGGGAGAACATACGATAGAGGGGGGT | TCTGGATTCAGAGCCGATGTAAGC |
| v15 | C1R1F7 | NP_182632.2 | CAGTGGGAGAACATACGATAGAGGGGGGT | CTTCTGGATTCAGAGCCGATGTAAGC |
| v19 | PAF2B | NP_177414.1 | CAGTGGGAGAACATACGATAGAGGGGGGT | CTTCTGGATTCAGAGCCGATGTAAGC |
| v20 | NCD2 | NP_001364.1 | CAGTGGGAGAACATACGATAGAGGGGGGT | CTTCTGGATTCAGAGCCGATGTAAGC |
| uCSGA | CHG8 | NP_0011275.3 | CAGTGGGAGAACATACGATAGAGGGGGGT | CTTCTGGATTCAGAGCCGATGTAAGC |
| u1 | GL2 | NP_0013670.3 | CAGTGGGAGAACATACGATAGAGGGGGGT | CTTCTGGATTCAGAGCCGATGTAAGC |
| u2 | JUN | NP_002283.3 | CAGTGGGAGAACATACGATAGAGGGGGGT | CTTCTGGATTCAGAGCCGATGTAAGC |
| u3 | PTK2 | NP_002624.3 | CAGTGGGAGAACATACGATAGAGGGGGGT | CTTCTGGATTCAGAGCCGATGTAAGC |

The project ID, official symbol and accession number is provided for each of the 32 genes. All primer sequences are written from 5' to 3'. The universal amplification tag sequences are depicted in lowercase letters and the gene-specific regions in capital letters. Apart from a 5'-phosphate on the thread-joining primers the probes were unmodified.
from poor granularity as they are run in an 'all-or-none'-fashion. This implies that the number of reads generated even from a single sample to 39-end of each detection oligonucleotide carried a C6 amino

Methods

Gene and probes. 32 genes implicated in basal cell carcinoma were selected (Table 3). The design of the two TnT probes – the extension primer and the thread-joining primer – was performed using a custom script implemented in Java/BioJava (Table 3)\(^\text{15}\). As oligo dT was used to initiate cDNA synthesis, TnT regions in 3'-parts of mRNAs were favoured. 10–12 bp extension regions were used (14 bp in one instance). A NCBI BLAST-search against the human genomic plus transcript database (Build 36.3) was performed for each complete DNA thread (extension primer – extension region – thread-joining primer) to avoid the inclusion of DNA threads with several highly scoring hits to the genome and/or transcripts. Additionally, the NCBI nucleotide database was used to scan the DNA threads for SNPs to eliminate the risks of inefficient annealing.

The detection oligonucleotides for the direct hybridization of the TnT products to microarrays were designed using a custom Perl/BioPerl script (Table 4). The script converts the input of complete DNA threads to a list of oligonucleotides complementary to the extension regions and the necessary flanking bases. Each of these oligonucleotides carried a 5'-Amino C6 amino group modification and a 15' spacer. Probe sets for five of the genes (d02, d08, u01, u15 and u19) were ordered from Eurofins MWG Operon (Ebersberg, Germany). The remaining sets were synthesized by Thermo Fisher Scientific (Ulm, Germany). The thread-joining primers were ordered with 5'-phosphate groups. The necessary universal amplification primers were acquired from Eurofins MWG Operon (Table 3).

Total RNA, cDNA synthesis and purification. The employed total RNA originated from two human cell lines: EFO-21 (ovaries, serous cystadenocarcinoma) and SK-MEL-30 (skin, malignant melanoma).

Total RNA was diluted 100 times in a two-step series, each step diluting the sample 10-fold, with DEPC-treated water. The diluted samples were subsequently used as template for cDNA synthesis, with the individual reactions encompassing 2 µg, 200 ng and 20 ng total RNA, respectively. Firstly, the total RNA was combined with 0.7 nmol of biotinylated oligo dT (5'–Biotin–T20–V; Qiagen Operon, Huntsville, USA) and 1 unit of Ampligase DNA ligase (Epicentre Biotechnologies, Madison, WI, USA) in 1x Ampligase reaction buffer (20 mM Tris-HCl pH 8.3, 25 mM KCl, 10 mM MgCl\(_2\), 0.2 mM of each dNTP, 0.2 µM of each extension primer, 0.05 µM of each thread-joining primer, 0.2 mM of the ACG trinucleotide mix (0.2 mM each of dATP, dCTP and dGTP), 2 units of Ampligase DNA ligase (Epiconcept Biotechnologies, Madison, WI, USA) and 1 unit of AmpliTaq DNA polymerase Stoffel fragment (Applied Biosystems, Foster City, CA, USA). The cDNA was eliminated from the mixture by immobilization on 15 µl paramagnetic beads. This was performed in a Magnatrix 1200 biomagnetic workstation (NorDag AB, Hägersten, Sweden) employing a custom protocol\(^\text{15}\). Briefly, 30 µl Dynabeads M-270 Streptavidin (equalling approximately 10 x 10\(^\text{6}\) beads; Invitrogen) were introduced, the biotinylated cDNA immobilized at room temperature for 15 minutes and the beads collected and washed. Lastly, the bound cDNA was released by suspending the beads in pure water and raising the temperature to 80°C for 1 min. Trinucleotide threading. The 10 µl trinucleotide reaction included an amount of cleaned-up cDNA corresponding to either 350 ng, 35 ng and 3.5 ng of total RNA, respectively, captured on streptavidin-coated super-paramagnetic beads. This was performed in a Magnatrix 1200 biomagnetic workstation (Table 3)\(^\text{15}\). As oligo dT was used to initiate cDNA synthesis, TnT regions in 3'-parts of mRNAs were favoured. 10–12 bp extension regions were used (14 bp in one instance). A NCBI BLAST-search against the 'human genomic plus transcript' database (Build 36.3) was performed for each complete DNA thread (extension primer – extension region – thread-joining primer) to avoid the inclusion of DNA threads with several highly scoring hits to the genome and/or transcripts. Additionally, the NCBI nucleotide database was used to scan the DNA threads for SNPs to eliminate the risks of inefficient annealing.

The detection oligonucleotides for the direct hybridization of the TnT products to microarrays were designed using a custom Perl/BioPerl script (Table 4). The script converts the input of complete DNA threads to a list of oligonucleotides complementary to the extension regions and the necessary flanking bases. Each of these oligonucleotides carried a 5'-Amino C6 amino group modification and a 15' spacer. Probe sets for five of the genes (d02, d08, u01, u15 and u19) were ordered from Eurofins MWG Operon (Ebersberg, Germany). The remaining sets were synthesized by Thermo Fisher Scientific (Ulm, Germany). The thread-joining primers were ordered with 5'-phosphate groups. The necessary universal amplification primers were acquired from Eurofins MWG Operon (Table 3).

Total RNA, cDNA synthesis and purification. The employed total RNA originated from two human cell lines: EFO-21 (ovaries, serous cystadenocarcinoma) and SK-MEL-30 (skin, malignant melanoma).

Total RNA was diluted 100 times in a two-step series, each step diluting the sample 10-fold, with DEPC-treated water. The diluted samples were subsequently used as template for cDNA synthesis, with the individual reactions encompassing 2 µg, 200 ng and 20 ng total RNA, respectively. Firstly, the total RNA was combined with 0.7 nmol of biotinylated oligo dT (5'–Biotin–T20–V; Qiagen Operon, Huntsville, AL, USA) and 1 unit of Ampligase DNA ligase (Epicentre Biotechnologies, Madison, WI, USA). The remaining sets were synthesized by Thermo Fisher Scientific (Ulm, Germany). The thread-joining primers were ordered with 5'-phosphate groups. The necessary universal amplification primers were acquired from Eurofins MWG Operon (Table 3).

Table 4 | Detection oligonucleotides

| ID | Gene | Detection primer |
|----|------|-----------------|
| d01 | AQP3 | AAGTGGGGCTTTCCCATC |
| d02 | PCDH21 | CCAATCTGCTCCGAGC |
| d03 | FGGBP | GACGCTTCCTTCTCAGGAA |
| d04 | C19orf57 | GGATGGGCCCTCTGTCA |
| d06 | DMKN | CAGGCTCTGGTCGACT |
| d07 | DCT | CAGCCTCTCCTCTTAGG |
| d08 | S100A8 | CAGGCTCTGGCAGCAC |
| d09 | LAMB1 | GGCAGCTTCGCTTAAAT |
| d13 | SLCA2A | CAGGCTCTTCTCCATG |
| d16 | ARPI1 | GGAGGCTCTGCTGGACA |
| bCL2L2 | BCL2L2 | CCCTAATTCCTCCTGAGCA |
| dCTNNBIP | CTNNBIP1 | GTAGCTGTGTTCCCCACA |
| dFADS | FADS1 | CCGAGCTTTTGGTAC |
| m1 LEF1 | LEF1 | GACAGCTGTGTTCTCCAA |
| m2 MYC | MYC | TCAAGATGTTGTTCCAATC |
| m4 GU1 | GU1 | AGAGCTGCCCCGGTAC |
| m6 CTA | CTA | CTCAATTTGCTCGGGAAGA |
| u01 | TACSTD1 | AGATGCTTGCTCAGC |
| u02 | NPT | CAGGCTCTGCTTCTAC |
| u04 | SCDS | TAAGGCTCTCTGGCTCATA |
| u05 | BNC2 | GATAGTCTTCTTTTGATC |
| u06 | SOX4 | GAATCTCCCTTGGTAAAT |
| u07 | MON1B | CACTAGCTGGCCCTGGTG |
| u09 | IGF2BP2 | TATAGCTTCTGATGC |
| u14 | FZD8 | ATACGGTGCCGCTTAC |
| u15 | CL0117 | CAGTACCTGCTGGAAGA |
| u19 | PPAP2B | CAGGCTCTGCTTCTAC |
| u20 | PEC1 | AGAATGGTGCTGCA |
| uCHGA | CHGA | CAGGCTCTGCTTCA |
| uGL2 | GL2 | CATAGCTTCTGCTGGCTG |
| uJUN | JUN | ACCAAATCTGCTGGAAGAT |
| uPITCH | PITCH1 | AGGAAGTTCTGCTGTGAAG |

All sequences are shown 5' to 3'. The 5'-end of each detection oligonucleotide carried a C6 amino group and a 15' spacer.

Table 5 | RT PCR primers

| ID | Gene | Forward primer | Reverse primer |
|----|------|----------------|----------------|
| d07 | DCT | TAGGGTGCTCATGCTTTACC | CAACTCAAGAGAGAAGAGCCTAGG |
| d09 | LAMB1 | AGCGAGTTAGAGGAAATGGGG | CTTGCGACATTGCTCAGC |
| d16 | ARPI1 | TCAACCCTTTTTGAGAGC | GAGGCTGCGAGAGACTC |

All sequences are written from 5' to 3'.

www.nature.com/scientificreports

SCIENTIFIC REPORTS | 2 : 821 | DOI: 10.1038/srep00821

5
50 mM KCl (Invitrogen). The cycling parameters were: 95°C for 5 min, 30 cycles of 95°C for 30 s, 65°C for 30 s and 72°C for 30 s. The PCR reactions were purified with the MinElute PCR purification kit (Qiagen, Hilden, Germany) with elution in 25 μl EB buffer (10 mM Tris-Cl, pH 8.5; Qiagen). The products were validated on a 2100 Bioanalyzer instrument (Agilent Technologies, Santa Clara, CA, USA) with the DNA 1000 Series II kit (Agilent Technologies).

454 sequencing. The purified and quality assessed samples were enrolled into the single-end amplicon sequencing pipeline for the Genome Sequencer FLX Titanium instrument (Roche Applied Science / 454 Life Sciences, Branford, CT, USA). Standard Multiplex Identifiers (MIDs Roche Applied Science / 454 Life Sciences) were used to be able to pool several samples in each sequencing lane. An automated version of the protocol was utilized for library preparation.

454 data analysis. The 454 data output comprised a FASTA file of all reads that passed the built-in quality filters. The data analysis, resulting in a text file with the number of hits to ‘true’ DNA threads and to undesired products, was performed with custom Perl/BioPerl scripts. Firstly, the reads were sorted based on their MID sequences. In this step no mismatches in the MID sequences were allowed. Thereafter, a database containing DNA threads of all analyzed genes was set up. Each entry of the MID-sorted output files was BLAST searched against this database. Single hits against theoretical threads with an alignment length of over 47 bases were classified as ‘true’ threads and counted. Reads mapping to two thread entities (double-hits), indicating a possible amalgamation, were categorized according to the identities of the two matches, counted and the sequences of their reads printed into a separate file. The requirement for inclusion was that the sum of the two highest-ranking alignment lengths was over 40 bases. Reads producing short alignments or giving no hits were removed. Pearson correlation coefficients between different samples were calculated with the R environment using the obtained DNA thread counts.

The most prevalent double-hits were investigated further. The reads were aligned with ClustalW2 using default parameters. The obtained alignments were manually evaluated. Subsequently, the consensus sequence was BLAST-searched against the NCBI ‘human genomic plus transcript’ database (Build 36.3) employing default parameters.

Array experiments. The array fabrication has been previously described15,20. A parallel thread amplification reaction with the conditions described above, but using a Cy5-labeled forward universal amplification primer, was set up to incorporate the dye into the final product. 20 μl of each amplification reaction was combined with 20 μl hybridization buffer (5x SSC with 0.2% SDS). This mixture was heat denatured at 95°C for 30 seconds and introduced to the array. The slide was incubated for 75 minutes at 50°C and 85 rpm shaking. Subsequently, a three step wash procedure was implemented: 50°C 2x SSC with 0.1% SDS for 5 min, 0.2x SSC for 1 min at room temperature and 0.1x SSC for 1 min at room temperature. The slide was dried by centrifugation and scanned with an Agilent G2505B scanner (Agilent). The images were analyzed with GenePix Pro 5.1 (Molecular Devices, Sunnyvale, CA, USA).

RT-PCR. Three genes exhibiting different expression characteristics in the two analyzed cell lines were further investigated with RT-PCR: DCT (expressed in SK-LLN1 cell line), NF2 (expressed in U251 cell line) and LAMB1 (expressed in both cell lines). To enable a faithful comparison, the primers were designed to cover the TrnT regions, or at least partly overlap these regions (Table 5).

Reverse transcription was performed with SuperScript III First-Strand Synthesis System (Invitrogen) following the provided instructions. 50 pmol non-biotinylated oligo-T23 (Qiagen Operon) were used. 2 μg of total RNA acted as template. The cDNA was purified with the MinElute PCR Purification Kit (Qiagen) and analyzed with the 2100 Bioanalyzer instrument (Agilent Technologies) using the RNA Pico Series Kit (Agilent Technologies).

Real-time PCR was performed using the iQ SYBR Green Supermix (Bio-Rad Laboratories, Hercules, CA, USA) in an iCycler instrument (Bio-Rad Laboratories). 25 μl reactions were set up according to the manufacturer’s recommendations. 5 pmol each of the forward and reverse primers were used. The thermal cycling parameters were: 95°C for 3 min followed by 40 cycles of 95°C for 30 s, 55°C for 45 s and 72°C for 45 s. The threshold cycles (Ct) were determined with the iCycler software (version 3.0a; Bio-Rad Laboratories) and compared between the cell lines.

1. Ahmadian, A. & Svanh, H. A. Massively parallel sequencing platforms using lab on a chip technologies. *Lab on a chip* **11**, 2653–2655 (2011).
2. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nature reviews* **12**, 87–98 (2011).
3. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews* **10**, 57–63 (2009).
4. Pickrell, J. K. et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
5. Metzker, M. L. Sequencing technologies - the next generation. *Nature reviews* **11**, 31–46 (2010).
6. Niedringhaus, T. P., Milano, D., Kerby, M. B., Snyder, M. P. & Barron, A. E. Landscape of next-generation sequencing technologies. *Analytical chemistry* **83**, 4327–4341 (2011).
7. Ozsolak, F. Third-generation sequencing techniques and applications to drug discovery. *Expert opinion on drug discovery* **7**, 231–243 (2012).
8. Yang, L., Tran, D. K. & Wang, X. BADGE, Beads Array for the Detection of Gene Expression, a high-throughput diagnostic bioassay. *Genome research* **11**, 1888–1898 (2001).
9. Fan, J. B. et al. A versatile assay for high-throughput gene expression profiling on universal array matrices. *Genome research* **14**, 878–885 (2004).
10. Li, H. et al. Versatile pathway-centric approach based on high-throughput sequencing to anticancer drug discovery. *Pro Natl Acad Sci U S A* **109**, 4609–4614 (2012).
11. Li, H., Qiu, J. & Fu, X. D. RASL-seq for massively parallel and quantitative analysis of gene expression. *Curr Protoc Mol Biol Chapter 4*, Unit 4.13 11–19 (2012).
12. Mercer, T. R. et al. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nature biotechnology* **30**, 99–104 (2011).
13. Pettersson, E., Lindskog, M., Lundeberg, J. & Ahmadian, A. Tri-nucleotide threading for parallel amplification of minute amounts of genomic DNA. *Nucleic acids research* **34**, e49 (2006).
14. Pettersson, E. et al. Allelotyping by massively parallel pyrosequencing of SNP-carrying trinucleotide threads. *Human mutation* **29**, 323–329 (2008).
15. Zajac, P., Pettersson, E., Gry, M., Lundeberg, J. & Ahmadian, A. Expression profiling of signature gene sets with trinucleotide threading. *Genomics* **91**, 209–217 (2008).
16. Neiman, M., Lundin, S., Savolainen, P. & Ahmadian, A. Decoding a substantial set of samples in parallel by massive sequencing. *PloS one* **6**, e17785 (2011).
17. Patterson, N. & Gabriel, S. Combinatorics and next-generation sequencing. *Nature biotechnology* **27**, 826–827 (2009).
18. Holmberg, A. et al. The biotin-streptavidin interaction can be reversibly broken using water at elevated temperatures. *Electrophoresis* **26**, 501–510 (2005).
19. Lundin, S., Stranenhein, H., Pettersson, E., Klevebring, D. & Lundeberg, J. Increased throughput by parallelization of library preparation for massive sequencing. *PloS one* **5**, e10029 (2010).
20. Hultén, E., Kaller, M., Ahmadian, A. & Lundeberg, J. Competitive enzymatic reaction to control allele-specific extensions. *Nucleic acids research* **33**, e48 (2005).