OPEN

## ORIGINAL ARTICLE

# Insights into psychosis risk from leukocyte microRNA expression

CD Jeffries[1], DO Perkins[2], SD Chandler[3], T Stark[4], E Yeo[5], J Addington[6], CE Bearden[7,8], KS Cadenhead[9], TD Cannon[10], BA Cornblatt[11], DH Mathalon[12], TH McGlashan[13], LJ Seidman[14], EF Walker[15,16], SW Woods[13], SJ Glatt[17] and M Tsuang[3]

Dysregulation of immune system functions has been implicated in schizophrenia, suggesting that immune cells may be involved in the development of the disorder. With the goal of a biomarker assay for psychosis risk, we performed small RNA sequencing on RNA isolated from circulating immune cells. We compared baseline microRNA (miRNA) expression for persons who were unaffected ($n = 27$) or who, over a subsequent 2-year period, were at clinical high risk but did not progress to psychosis ($n = 37$), or were at high risk and did progress to psychosis ($n = 30$). A greedy algorithm process led to selection of five miRNAs that when summed with $\pm 1$ weights distinguished progressed from nonprogressed subjects with an area under the receiver operating characteristic curve of 0.86. Of the five, miR-941 is human-specific with incompletely understood functions, but the other four are prominent in multiple immune system pathways. Three of those four are downregulated in progressed vs. nonprogressed subjects (with weight -1 in a classifier function that increases with risk); all three have also been independently reported as downregulated in monocytes from schizophrenia patients vs. unaffected subjects. Importantly, these findings passed stringent randomization tests that minimized the risk of conclusions arising by chance. Regarding miRNA–miRNA correlations over the three groups, progressed subjects were found to have much weaker miRNA orchestration than nonprogressed or unaffected subjects. If independently verified, the leukocytic miRNA biomarker assay might improve accuracy of psychosis high-risk assessments and eventually help rationalize preventative intervention decisions.

*Translational Psychiatry* (2016) **6**, e981; doi:10.1038/tp.2016.148; published online 13 December 2016

## INTRODUCTION

Schizophrenia affects about 1% of the general population, typically emerges in late adolescence and early adulthood, and is usually chronic, relapsing and disabling.[1,2] However early identification and treatment of psychosis is associated with better clinical outcome,[3] and interventions in persons experiencing high-risk symptoms show promise in preventing the development of psychosis.[4] Clinical diagnostic criteria for the psychosis prodrome identify persons with a 13–22% 2-year psychosis risk.[5–11] While much higher than the general population risk, this relatively low conversion rate hampers the development and implementation of preventative interventions. Thus, for persons at high risk a biomarker assay that improved risk prediction would be of great value. In addition, employed biomarkers may illuminate mechanisms involved in the emergence of schizophrenia and potentially point towards new therapeutic targets.

The tissue used for biomarker discovery should represent the disease pathology. Regarding schizophrenia, biomarker studies have often considered circulating immune cells as easily accessible proxies for the brain that may reflect an environmental or genetic vulnerability that is shared with the brain.[12,13] Moreover, peripheral immune cells are now known to regulate brain functions involved in schizophrenia, including cognition,[14] behavioral responses to stress,[15] neural plasticity[16] and neurogenesis.[17–19] In addition, converging reports from genetic, epidemiological, clinical and post-mortem studies implicate dysregulation of both the innate and adaptive immune systems in schizophrenia.[20] In particular meta-analyses have found schizophrenia to be associated with elevations in blood levels of specific analytes,[21,22] as well as shifts in adaptive immune cell populations.[23] We and others have found that alterations in plasma analyte levels predicted psychosis in persons meeting clinical criteria for psychosis risk.[24,25] Thus, peripheral immune cells may be of value for schizophrenia biomarker discovery because their dysregulation may be directly linked to emerging psychosis.

Here we report the results of small RNA sequencing on RNA isolated from circulating immune cells, emphasizing the comparison of expression in persons at clinical high risk for psychosis who progressed to psychosis (schizophrenia or a related disorder) vs. those who remained psychosis-free. We focused on microRNAs (miRNAs), ~ 22 nucleotide single-stranded RNA molecules that are

[1]Renaissance Computing Institute, University of North Carolina, Chapel Hill, NC, USA; [2]Department of Psychiatry, University of North Carolina, Chapel Hill, NC, USA; [3]Center for Behavioral Genomics, Department of Psychiatry, and the Institute of Genomic Medicine, University of California, San Diego, San Diego, CA, USA; [4]Division of Biological Sciences, University of California, San Diego, San Diego, CA, USA; [5]Department of Cellular and Molecular Medicine, University of California, San Diego, San Diego, CA, USA; [6]Hotchkiss Brain Institute, Department of Psychiatry, University of Calgary, Calgary, AB, Canada; [7]Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, Los Angeles, CA, USA; [8]Department of Psychology, University of California, Los Angeles, Los Angeles, CA, USA; [9]Department of Psychiatry, University of California, San Diego, San Diego, CA, USA; [10]Departments of Psychology and Psychiatry, Yale University, New Haven, CT, USA; [11]Department of Psychiatry, Zucker Hillside Hospital, Glen Oaks, NY, USA; [12]Department of Psychiatry, University of California, San Francisco, San Francisco, CA, USA; [13]Department of Psychiatry, Yale University, New Haven, CT, USA; [14]Department of Psychiatry, Harvard Medical School at Beth Israel Deaconess Medical Center and Massachusetts General Hospital, Boston, MA, USA; [15]Department of Psychology, Emory University, Atlanta, GA, USA; [16]Department of Psychiatry, Emory University, Atlanta, GA, USA and [17]Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, Syracuse, NY, USA. Correspondence: Dr CD Jeffries, Renaissance Computing Institute, University of North Carolina, 3127, UNC at Chapel Hill, Chapel Hill 27517, NC, USA.
E-mail: clark_jeffries@med.unc.edu
Received 24 August 2015; revised 17 June 2016; accepted 30 June 2016

2

now generally appreciated as regulators of mRNA processing in translation and doubtless involved in many developmental and pathological processes in animals.[26–28] In particular, alterations in miRNA abundance may indicate a shift in immune system state. The present work offers a preliminary connection of immune cell miRNA levels and the likelihood of transition from clinical high risk to psychosis, providing further evidence of association of immune dysregulation.

## MATERIALS AND METHODS

As described previously, the North American Prodrome Longitudinal Study, Phase 2 (NAPLS 2)[29] is an eight-site observational study of the predictors and mechanisms of conversion to psychosis in persons at elevated risk indicated by the Criteria of Psychosis-Risk States.[30] The full NAPLS 2 cohort includes 764 high-risk and 280 demographically similar unaffected subjects between the ages of 12–35. The study was approved by the Institutional Review Board at each site, and each subject provided written informed consent or assent, with a parent or guardian consenting for subjects $< 18$ years old.

In the present analysis, we included all high-risk subjects with RNA samples who had either progressed to psychosis within 2 years ($n = 30$) or who remained nonprogressed at 2-year follow-up ($n = 37$), as of February 2012. Also included were some unaffected comparison subjects ($n = 27$) who did not meet high-risk criteria and had no personal or family history of a psychotic disorder.

### Assessments

Clinical assessments were done every 6 months and subjects followed for up to 2 years. Participants were screened using the Structured Interview for Psychosis-Risk Syndromes and rated with the Scale of Psychosis-Risk Symptoms as defined by the Criteria of Psychosis-Risk States: attenuated psychotic symptoms, brief intermittent psychotic symptoms, substantial functional decline combined with a first-degree relative with a psychotic disorder, or schizotypal personality disorder in individuals $< 18$ years old.[30] The Structured Clinical Interview for DSM-IV[31] was used to determine psychiatric diagnoses.

Data on prescription medications were based on self-reports and/or parental reports. Socioeconomic status was estimated by maximum years of education of mother or father.

### Assays of leukocytes for miRNAs

Immediately after phlebotomy, leukocytes were isolated on a filter and RNA preserved with RNAlater (Qiagen, Venlo, The Netherlands). Samples were stored at $-20\,^{\circ}C$ until processing. RNA was extracted using a modification of the LeukoLOCK procedure (Life Technologies, Foster City, CA, USA).[32] Small RNA libraries were prepared with Illumina TruSeq kits (Illumina, San Diego, CA, USA) following manufacturer's protocol. Barcoded libraries were combined in equimolar amounts (10 nmol $l^{-1}$ each), then diluted to 4 pmol $l^{-1}$ for each flowcell lane and sequenced by Illumina HiSeq Sequencing Systems (Illumina).

The Illumina processing pipeline v1.5 (Illumina) was used for base-calling using the SCARF text format. Each of 2588 mature miRNA sequences from miRBase v21 was sought as an exact sequence match within each read. From an initial set of 100 samples, we excluded 6 (3 unaffected, 1 nonprogressed, 2 progressed) with low-abundance reads, leaving 27 unaffected and 67 high-risk subjects. The analysis was also repeated using trimmed subsequences of canonical miRNA sequences to account for isoform diversity.

### Normalisation

For all analyses, we included the 136 miRNAs that were robustly expressed, defined as $\geq 10\,000$ total reads in the 94 subjects. However, different subjects sometimes had many miRNAs with high read counts or many with low read counts. For one pair of miRNAs, this skewness would already imply high correlations. Thus, to avoid spuriously high correlations from skewness, we divided read counts for each sample by the average of read counts for the top 30 miRNAs of that sample, forming quotients (Supplementary Materials and Methods) and reducing the ratios maximum:minimum among the top 30 miRNAs. For each miRNA, we then used the average and s.d. over all unaffected subjects' quotients for that miRNA to process all quotients into z-scores; final values for each miRNA were in a $4\times$ range over all 94 samples.

One nonprogressed sample was confidentially submitted in technical duplicates for sequencing. After normalization, the 136 robustly expressed miRNAs had a Pearson correlation of 0.61, achieving the 98.4th percentile of all 4371 possible correlations over 94 samples. If the 20 least-correlated miRNAs were dropped, the correlation rose to 0.80. Visually, the quality of this correlation as a test of the normalization process can be appreciated from a graph in Supplementary Materials and Methods. Numerically, the correlation in 136-dimensional space of two vectors having entries generated from a normal distributions would exceed 0.61 with probability 9.3E − 15.

### Trimmed miRNA sequences

Each canonical mature miRNA sequence shown in miRBase is only a representative of multiple RNA species arising from the same precursor,[33] and other isoforms might be important in cell functions.[34] To investigate the potential impact of multiple isoforms, we trimmed two bases from the 5′ end and four bases from the 3′ end of each canonical sequence, completely retabulated matches, and reanalyzed read count data. This had the effect of multiplying the grand total of all miRNA matches by ~ 1.79. However, rerunning normalization and analyses with trimmed miRNA sequences led to similar choices of informative miRNAs for the full set of samples for trimmed vs. untrimmed data (Supplementary Materials and Methods). That is, the first, second, and fourth chosen miRNAs for the full data were, respectively, the first, second, fifth chosen for trimmed data, and so on.

### Construction of a classifier that best differentiated nonprogressed from progressed subjects

We detected $\geq 1$ reads for 1569 of the 2588 canonical miRNAs in miRBase v21 but limited our analyses to the 137 miRNAs with at least 10 000 reads over 94 subjects. One miRNA (miR-486-5p) was discarded as superabundant (62% of all reads), leaving 136 miRNAs.

We employed a greedy algorithm (Supplementary Materials and Methods) to develop our classifier. It selected first the one miRNA that best distinguished nonprogressed from progressed, based on the Student t-test P-value. The greedy algorithm then sought to add a weighted, second miRNA that best improved the overall Student t-test P-value, if possible. The greedy algorithm continued to add miRNAs to the sum until no improvement in the metric was possible or a limit was reached. We predicted (correctly) that classifier functions with low Student t-test P-values would also achieve high area under the curve of the receiver operating characteristic (AUC of ROC), thereby diversifying performance objectives.

Using the Student P-value as the selection metric was not logically the same as optimization of various geometric fits as in conventional linear regressions. That is, we sought to distinguish the groups as sets, not as abstract points in space separated by a hyperplane in some way. In our tests enforcing the same limits on the number of selected markers, the AUCs obtained were about the same or superior to those from standard geometric methods (data not shown).

Notably, the nonzero weights of miRNA values we used were all $\pm 1$. In practice, the greedy algorithm typically terminated after $\leq 10$ iterations (selection of $\leq 10$ of 136 miRNAs, each weighted $\pm 1$ in a classifier function). Considering its construction, we call this greedy algorithm 'Coarse Approximation Linear Function' (CALF); a similar algorithm has been developed for distinguishing pairs of markers.[35] CALF is now freely available at https://cran.r-project.org/web/packages/CALF/index.html. Although using real numbers as weights as in conventional regression would presumably yield better metric (and AUC) values, using $\pm 1$ avoided instability. That is, setting aside a few samples and re-computing optimal real weights in conventional methods generally would not yield exactly the same real numbers, but stable (identical) weights would be more likely when limited to $\pm 1$ (Supplementary Materials and Methods). In mathematical experiments with real data, we have found that in dimensions $\geq 5$, using few $\pm 1$ weights can approximate target functions almost as well as using the same number of real-valued weights. This surprising fact is due, in higher dimensions, to the exponentially increasing crowding of directions defined by all such coarse vectors (as rays from origin) among the directions defined by all real vectors as they penetrate the cube with vertex components $\pm 1$ (Supplementary Materials and Methods). However, using any classifier algorithm that automatically selected a small subset of markers and otherwise passed rigorous randomization tests
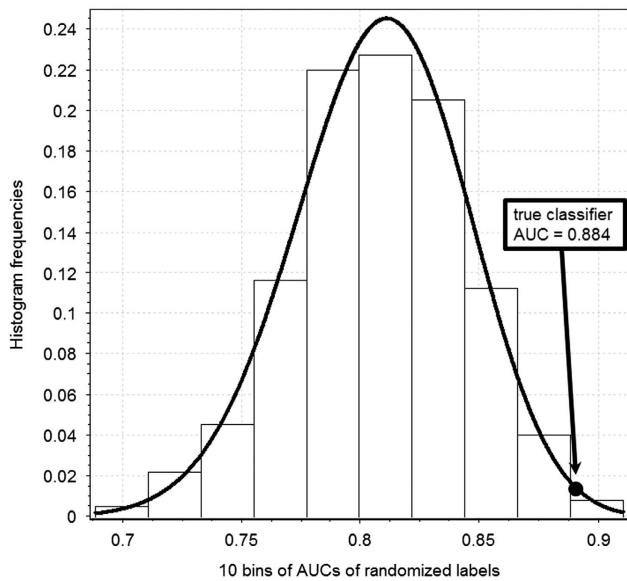
**Figure 1.** Histogram of one AUC from true data vs. 1000 AUCs of classifiers built by the same greedy algorithm applied to pseudo data (NP and P labels randomly permuted). Fitted with a beta distribution, the AUC from real data indicates a *P*-value of 0.012. Since 17 random AUCs of 1000 by chance exceed the true AUC, an algebraic method[36] gives alternative *P*-value = 0.018. Thus, the performance of the Greedy Algorithm limited to selection of at most six markers and applied to the full data set is unlikely to be chance. AUC, area under the curve of receiver operating characteristic.

(as in Figure 1 above) would have been acceptable from the standpoint of prudent classifier construction processes that support reproducibility.

## Assay validation

Real time quantitative PCR validation of sequencing results was done with a total of 37 clinical high-risk subjects who did not progress to psychosis within 2 years and 30 others who did progress. Following conventional reverse transcription, cDNA synthesis, and preamplification steps, samples were assayed with high-throughput real-time qPCR (HT-PCR) using Fluidigm technology (San Francisco, CA, USA). A total of 21 miRNAs were assayed, including the 5 miRNAs selected in the classifier function (equation (1)). Data for 32 nonprogressed subjects and 24 progressed subjects could be compared with read counts. The 56 Spearman correlation of PCR values vs. RNA-seq read values averaged 0.64 (s.d. 0.11); minimum was 0.30. It follows that the 56 Spearman correlations with such values would arise by chance with probability $< 5.2E-30$.

Furthermore, a simple (but likely suboptimal) way to modify the classifier (equation (1)) to use PCR data is to find for each of the subjects the average over the 21 miRNAs of Cq values. The averages can be subtracted from the raw Cq values to reduce gross sample-to-sample biases. A signal remains when selected miRNAs are relatively high within one group and low in the other. The Student *P*-value for such a classifier from equation (1) was 0.0012 and the AUC was 0.72.

In summary, PCR values and RNA-seq values were consistent. However, switching to PCR technology in an extension of the present work would properly assay anew a pool of the five selected miRNAs in equation (1), as well as other miRNAs that could be selected by CALF if the five were disallowed, all leading to a revised classifier function similar but not necessarily identical to equation (1).

## Randomization tests to assess significance of classification

Prudent case/control classifier construction can include six steps. First, a classifier algorithm is applied to true data, yielding a performance number (for example, AUC). Next, case and control memberships of all samples are randomly permuted, the model-building algorithm is re-applied to the pseudo data exactly as to the true data, and the performance number recorded—all repeated multiple times (for example, 1000 times). If such randomization tests indicate clear ability of the algorithm to distinguish

**Table 1.** Demographic and clinical characteristics of study subjects

| | Unaffected comparison (UC) n = 29 | Clinical high risk, not psychotic (CHR-NP) n = 37 | Clinical high risk, psychotic (CHR-P) n = 30 |
|---|---|---|---|
| Age, average (s.d.) | 19.3 (4.4) | 18.1 (3.8) | 18.7 (3.7) |
| *Ancestry* | | | |
| % Caucasian | 61%, | 65% | 52% |
| % African | 32% | 13.5% | 19% |
| % Asian | 7% | 13.5% | 19% |
| % Mixed | 0% | 8% | 10% |
| Sex, % male | 68% | 62% | 74% |
| SES, average (s.d.) | 7.5 (1.7) | 6.5 (1.7) | 6.2 (1.6) |
| *Peripheral blood mononuclear cells* | | | |
| Neutrophils % | 56 (11) | 55 (13) | 55 (10) |
| Lymphocytes % | 34 (9) | 35 (11) | 33 (9) |
| Monocytes % | 8 (2) | 7 (3) | 8 (3) |
| Eosinophils % | 2 (3) | 2 (2) | 2 (2) |
| Basophils % | 1 (1) | 1 (1) | 1 (1) |
| *SOPS scores, average (s.d.)* | | | |
| Total[a] | 4.8 (5.3) | 36.8 (12.4) | 45.0 (13.0) |
| Positive[a] | 1.3 (1.8) | 12.6 (4.4) | 13.9 (3.7) |
| Negative[a] | 1.3 (1.8) | 11.5 (5.9) | 14.0 (5.9) |
| Disorganized[a] | .8 (1.1) | 4.9 (2.6) | 6.2 (3.4) |
| General[a,b] | 1.4 (1.7) | 7.8 (4.5) | 10.9 (4.7) |
| *Prescription medication* | | | |
| Antipsychotic[c] | 0% | 27% | 13% |
| Antidepressant[d] | 3% | 24% | 23% |
| Stimulant | 0% | 7% | 6% |
| Mood stabilizer | 0% | 0% | 3% |
| Benzodiazepine[e] | 0% | 3% | 13% |
| NSAID | 0% | 0% | 0% |
| Antibiotic | 0% | 0% | 0% |
| *Substance use* | | | |
| Tobacco use[f] | 7% | 30% | 39% |
| Alcohol use | 41% | 38% | 35% |
| Marijuana use[g] | 7% | 24% | 32% |

Abbreviations: NSAID, non-steroidal anti-inflammatory drug; SES, socio-economic status. [a]CHR-P vs UC *t*-test *P*-value < 0.0001, CHR-NP vs UC *t*-test *P*-value < 0.0001. [b]CHR-P vs CHR-NP *t*-test *P*-value = 0.02. [c]CHR-P vs UC FET *P*-value = 0.047, CHR-NP vs UC FET *P*-value = 0.001. [d]CHR-P vs UC FET *P*-value = 0.011, CHR-NP vs UC FET *P*-value = 0.002. [e]CHR-P vs UC FET *P*-value = 0.047. [f]CHR-P vs UC FET *P*-value = 0.001, CHR-NP vs UC FET *P*-value = 0.02. [g]CHR-P vs UC FET *P*-value = 0.020, CHR-NP vs UC FET *P*-value = 0.056.

case from control, further development of the classifier is indicated. Then the same algorithm is applied multiple times (for example, 1000 times) to random 80% subsets of true cases and controls. The resulting classifiers are integrated to produce a final classifier. The integrated classifier is tested with true data. If the integrated classifier performs well, then it is applied to external data (beyond the means and scope of the present pilot study). This general process and its many modern enhancements are widely used in drug design, development of cancer biomarkers, genomic research and other fields.[37–41] It can be attempted with any classifier algorithm. We emphasise publication of histograms of randomization tests (see Figure 1) or the logical equivalent to reduce readers' scepticism. For example, a successful training set histogram would reduce likelihood of someone repeatedly adjusting an algorithm *post hoc* to optimize performance with both training and external data.

It should be emphasized that the randomization tests herein are applied to AUCs of true and pseudo classifiers; this is logically distinct from applying the true classifier to true and randomized data, as is often done.
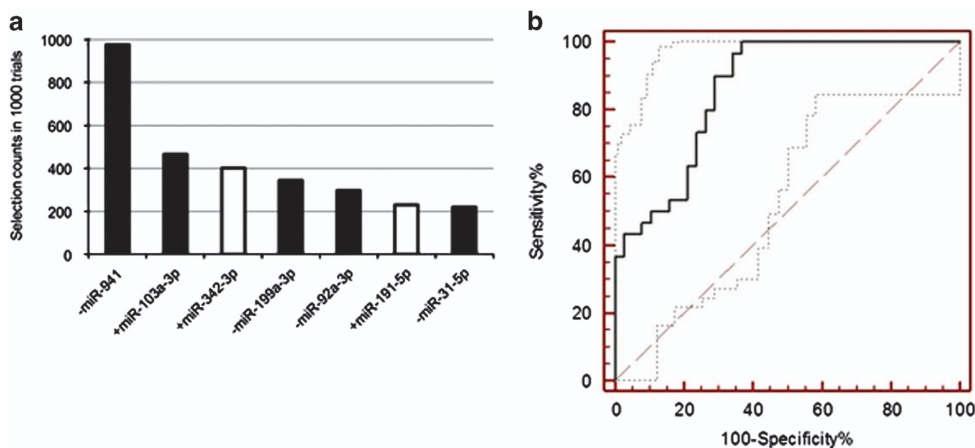
**Figure 2.** (a) The Greedy Algorithm was applied to 1000 selections of random 80% subsets of nonprogressed subjects and random 80% subsets of progressed subjects. Each time up to six markers could be chosen. The seven most frequently chosen markers are shown with their selection rates. The solid bars indicate the five markers that were also selected in the first six markers chosen by the Greedy Algorithm for the full data set (Figure 1), yielding as a sum of z-scores the classifier function in sum (equation (1)). This function applied to the full data yields AUC = 0.86. (b) ROC of the five-miRNA classifier function (equation (1)). Dotted lines are 95% confidence levels, and the dashed line is hypothetical performance of a random classifier. AUC, area under the curve; miRNA, microRNA; ROC, receiving operating characteristics.

Nonetheless, best practices do not insure that the only or best classifier has been developed. Furthermore, selected markers might only be surrogates for some deeper, causal markers unknown or inaccessible to the experimenter. It is always possible that systematic bias could have entered the analysis as misdiagnoses or some fundamental chemical bias in RNA-seq processing. Ultimately, the true utility of classification by miRNAs will rest with testing samples from additional subjects.

## RESULTS

### Participant characteristics

Table 1 provides a description of subjects. All at high-risk met attenuated psychosis criteria. The 30 progressed included: 14 with schizophrenia; 11 with psychosis, not otherwise specified; 2 with major depression with psychotic features; and 1 each with schizoaffective, delusional and psychotic bipolar disorder.

### Using individual miRNAs

The smallest Student t-test P-value for a miRNA for nonprogressed vs progressed data was $\alpha = 0.0053$ (miR-941) (see Supplementary Materials and Methods). Thus, Bonferonni control for multiple comparisons over 136 miRNAs would declare no individual miRNA as a statistically significant biomarker. However, as explained by Fredrickson et al.,[42] it can indeed happen that the association between a set of markers and phenotypes reaches a high level of reliability while individual markers in the same set fail to do so. This principle is heavily employed in the present work.

### Using sets of miRNAs for psychosis-risk prediction

The performance of the miRNA classifier developed from all nonprogressed and progressed subjects using the sum of the first six miRNAs chosen by CALF was AUC = 0.88. This value was superior to 983 of 1000 AUCs from exactly the same algorithm applied to randomized data (Figure 1). Next we applied CALF to 1000 random selections of 80% subsets of nonprogressed subjects and 80% subsets of progressed subjects. The 7 miRNAs that were selected in at least 225 of the 1000 trials are shown in Figure 2, and 5 of these were also among the six in the initial classifier developed from all subjects. Our integrated classifier was thus the sum of the five miRNAs chosen by both approaches, where the '+' and '-'

means the value (z-score of normalized data) is added and the '–' means the value is subtracted from the final score:

$$-miR\text{-}941 + miR\text{-}103a\text{-}3p - miR\text{-}199a\text{-}3p - miR\text{-}92a\text{-}3p - miR\text{-}31\text{-}5p \quad (1)$$

The classifier function (equation (1)) generally results in higher values for progressed (mean = 1.24, s.d. = 0.27) than nonprogressed subjects (mean = -0.78, s.d. = 0.22), with AUC = 0.86 (Figure 2). In addition (equation (1)) achieved AUC = 0.75 on application to unaffected vs progressed subjects. Regarding ranks of total read counts among the 136 miRNAs, those in equation (1) ranked 39, 18, 119, 1, 115, respectively, implying the 5 miRNAs in equation (1) have diverse frequencies among the top 136.

Random permutations of true data will have chance patterns that algorithms can exploit to create seemingly convincing classifiers. As shown in Figure 1, many pseudo classifiers achieved AUCs well in excess of 0.5, the customary value of random classifiers using prior probabilities. However, as shown in Figure 1, only 17 of 1000 pseudo AUCs exceeded the true AUC, yielding[36] a P-value of $(17+1)/(1000+1) = 0.018$. Alternatively, fitting a beta distribution to a histogram of pseudo AUCs using EasyFit (MathWave, Dnepropetrovsk, Ukraine) led to an estimated P-value $= 0.012$.

Prescription medication use is common in subjects at high-risk of psychosis[43] (Table 1) and may affect leukocytic miRNA expression. To investigate the effects, the 5-miRNA sum (equation (1)) was applied to 1000 random selections of 25 samples from the set of nonprogressed and 25 samples from the set of progressed subjects. The average AUC was 0.86 with s.d. = 0.029. We then selected 1000 times random subsets in 2 ways: to maximize or minimize the number of treated subjects plus a number of untreated subjects needed to make a total of 25. The resulting AUCs were in [0.83, 0.87]. This experiment therefore provided evidence of little influence of medications on the performance of the classifier (equation (1)) (Supplementary Materials and Methods, Supplementary Table S1).

### miRNA–miRNA correlation networks within groups

We next explored the degree of co-regulation among 136 robustly expressed miRNAs. We randomly selected 1000 times sets of 25 subjects from each of the 3 groups. From each selection, Pearson correlations of all 9180 distinct pairs of 136 miRNAs were
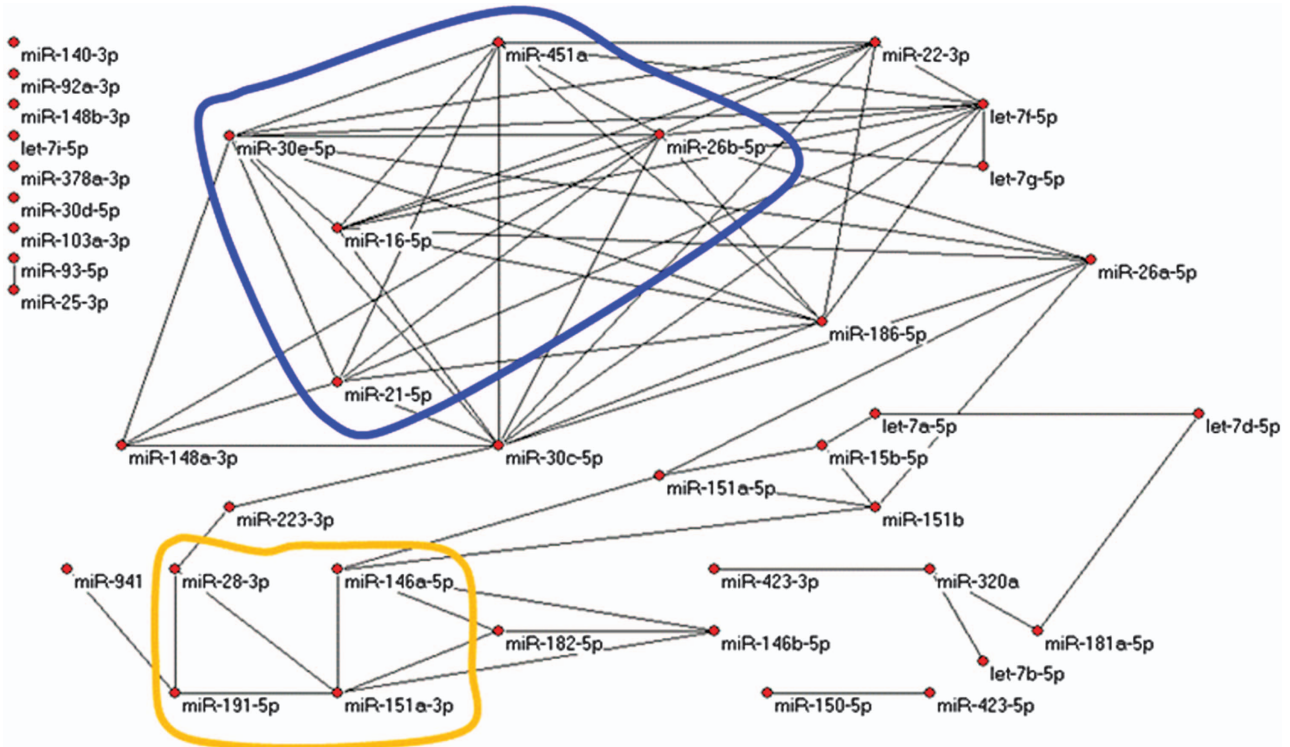
**Figure 3.** Graphs from miRNA–miRNA correlations. Edges represent strong correlations. Looped regions are common subgraphs. Networks shown are strongly correlated miRNAs among unaffected controls. miRNA, microRNA.
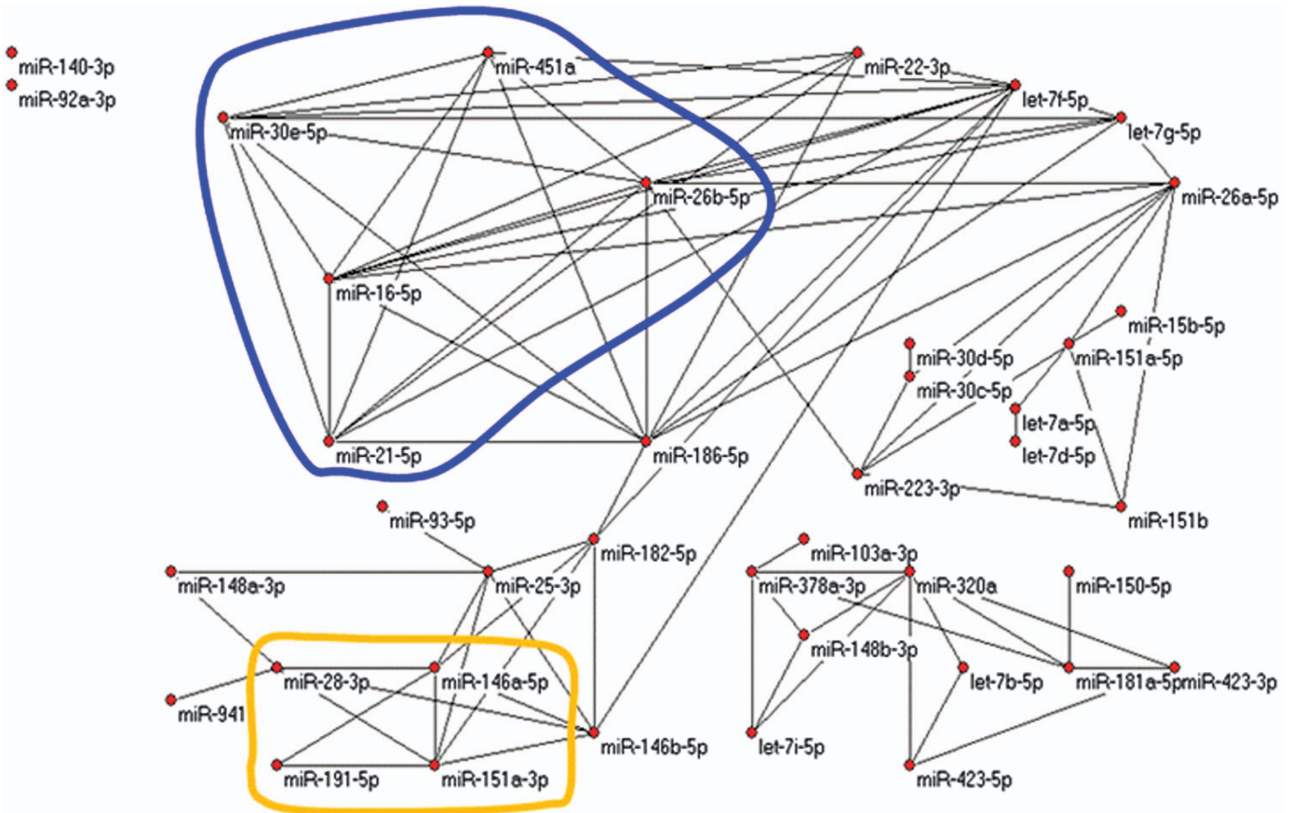


**Figure 4.** Graphs from miRNA–miRNA correlation with the same criteria as in Figure 3 but for nonprogressed subjects. This graph is similar to that in Figure 3. miRNA, microRNA.
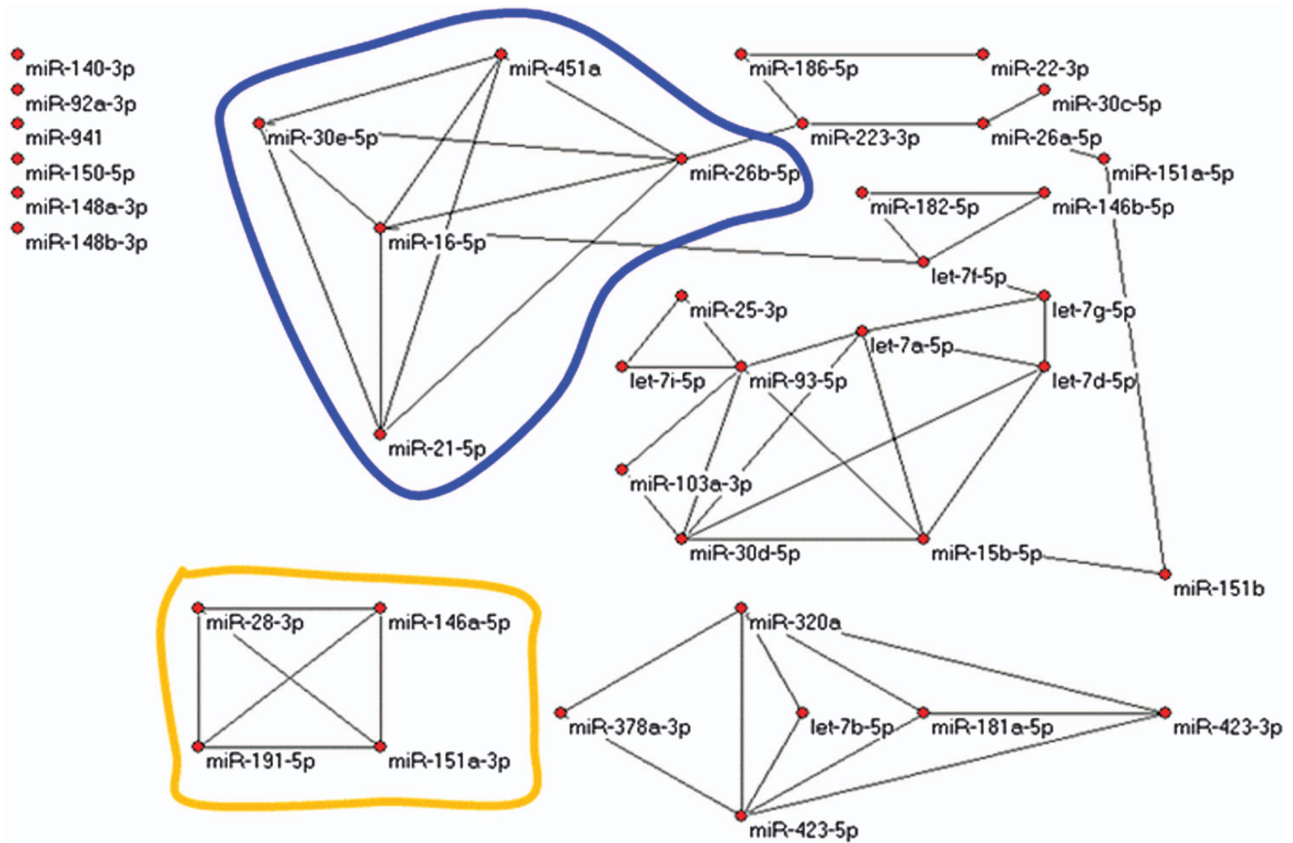
**Figure 5.** Graphs from miRNA–miRNA correlations with the same criteria as in Figures 3 and 4 but for progressed subjects. Evidently much organisation of miRNA networks is lost in subjects who eventually progressed to psychosis. miRNA, microRNA.

calculated. Surprised by the consistent differences group vs. group, we redid all calculations with random subsets of 23 and 21 subjects instead of 25. This yielded essentially the same patterns (Supplementary Materials and Methods). Correlation graphs of miRNA data have been reported elsewhere in neuroscience.[44]

Restricting attention to the 40 most robustly expressed miRNAs (that include 3 miRNAs from equation (1), namely, miR-941, miR-103a-3p, and miR-92a-3p), we contrasted miRNA–miRNA correlation networks over the 3 groups by randomly selecting 1000 times subsets of 25 samples from each group and calculating all 780 Pearson correlations of pairs (from 40 miRNAs represented by 25-dimensional vectors). In each group and for each pair of miRNAs, we tabulated the number of times of 1000 possible that the correlation exceeded 0.5878 ($P$-value $= \sim 1.00\mathrm{E}-3$ for normally distributed 25-dimensional vectors, hence expecting 0.78 times among 780 random pairs to exceed that threshold). Such correlations were much more frequent than chance; those occurring in $>500$ of 1000 trials became edges in graphs in Figures 3–5 (drawn with Pajek http://pajek.imfm.si/doku.php). Clearly, highly correlated miRNAs were more numerous in nonprogressed and unaffected subjects than in progressed subjects. Notably, miR-941 and miR-103a-3p were both included in the correlation networks for unaffected and nonprogressed subjects but absent from the network for progressed subjects.

Bioinformatic analyses
Although the seed region (nucleotides 2 through 8, numbered from 5′ end) was originally proposed to define the agency of miRNA targeting,[45,46] recent reports suggest that all of the mature miRNA sequence may be involved. Other types of targeting include 'offset' (starting at base 3), 'supplementary' (additional binding in a second region) and 'compensatory' (supplementary targeting that tolerates limited mismatches in the seed).[47] Also, 'centered' sites were found as a class of miRNA target sites that lack both perfect seed pairing and 3′-compensatory pairing and instead have 11–12 contiguous Watson–Crick pairs near the center of the canonical miRNA.[48] Regarding prevalence of non-seed binding, a transcriptome-wide survey for miR-155 found that ~ 40% of miR-155-dependent Argonaute (Ago) binding occurs at sites without perfect seed matches.[49] In mouse brain, G-bulge sites (positions 5 or 6 in the seed) were found often bound and regulated by miR-124, and more generally, bulged sites comprise ⩾ 15% of all Ago-miRNA interactions.[50] An analysis of 18 000 high-confidence miRNA–mRNA interactions found ~ 60% of seed interactions to be noncanonical, containing bulged or mismatched nucleotides.[51] In summary, evidence suggests miRNA targeting is not necessarily a function of base pairing of the seed region.

Moreover, many canonical miRNAs are very similar as sequences. To filter our list of 136, we used Ingenuity (QIAGEN). The Ingenuity list typically represents sets of miRNAs with very similar sequences by just one from the set. We noted that consequently only 92 of the 136 robustly expressed miRNAs were in the Ingenuity miRNA targeting database (for example, the nine let-7 species in our 136 were represented by let-7a-5p). There are 5 miRNAs in equation (1), so 10 pairs. We calculated the Smith–Waterman sequence alignment score[52] for all 10 pairs using weights: match +1, mismatch − 1, gap − 1. Pairs of the five miRNAs in equation (1) had strong similarities. For example, the score $11 - 0 - 2 = 9$ is reached for miR-199a-3p 5′-ACAGUAGUCUGCA CAU_UGGUUA-3' and miR-941 5'-CACCCGGCUGUGUGCACAUG

UGC-3′ because they contain the common subsequence GU-UGCACAU-UG (Supplementary Materials and Methods).

The average of the 10 alignment scores was 6.4. Then we selected 1000 times a random set of 5 miRNAs from the 87 in Ingenuity excluding the 5 in equation (1) and calculated the 10 Smith–Waterman scores and their 1000 averages. The average was 5.1. A total of 19 times of 1000 the random averages were ⩾ 6.4. Thus, the Monte Carlo[36] estimated P-value that the true sequence similarities are due to chance is 0.02 (Supplementary Materials and Methods). Other Smith–Waterman weight choices including extreme choices with mismatch or gap equal to − 100 (so exact match subsequences) led to the same conclusions (data not shown). In summary, taken as full, canonical sequences, the five miRNAs selected in equation (1) from nonprogressed vs progressed analysis were as sequences more similar than would be expected by chance.

## DISCUSSION

This study suggests that expression patterns of small, regulatory miRNAs in leukocytes differentiate persons at clinical high risk for psychosis who subsequently develop psychosis from those who do not. While no single miRNA exhibits statistically significant predictive power, we found that a sum of five abundantly expressed miRNAs produced a risk classifier that survives randomization testing. That is, stringent randomization tests have implied: original data likely contained miRNA information that distinguished nonprogressed from progressed; the normalization procedure we used likely did not obliterate that information; and the greedy algorithm (CALF) found a classifier function with AUC performance better than chance would readily allow. Randomization tests might seem obviously necessary, but many reports of classifier constructions in the scientific literature do not include them.

Our results are consistent with a study conducted by Gardiner et al.[53] that compared miRNA expression in 112 schizophrenia subjects to that of 76 unaffected comparison subjects. They reported 83 miRNAs as downregulated, 30 of which were robustly expressed and thus considered in our analyses; remarkably, miR-92a-3p, miR-199a-3p, miR-31-5p in equation (1) were among these. There was, however, no apparent overlap between our study and a second investigation of monocyte miRNA expression in schizophrenia.[54] Certain other studies did not consider the five miRNAs in equation (1) in their analyses.[55–58]

miRNA regulation of gene expression in humans is based on imperfect base-pair binding of the mature miRNA to a targeted mRNA. The canonical sequences of the five miRNAs selected from nonprogressed vs. progressed analysis in equation (1) are significantly more similar than expected. This finding implies that these five miRNAs may in some way co-regulate gene expression and may be in themselves co-regulated. Finally, the remarkably different miRNA–miRNA correlation networks in Figure 3 suggest a shift in network orchestration in persons who progressed to psychosis.

Our findings require further investigation in terms of affected genes and cellular products. Most informative would be mRNA: miRNA co-expression analyses. Assuming correct and representative sampling and reproducibility of lab assays, and noting favorable outcomes of randomization tests, the classifiers and regulatory network patterns herein are unlikely to be due to chance. However, verification is needed with additional samples, possibly using an alternative assay technology such as HT-PCR.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## REFERENCES

1 Perala J, Suvisaari J, Saarni SI, Kuoppasalmi K, Isometsa E, Pirkola S et al. Lifetime prevalence of psychotic and bipolar I disorders in a general population. Arch Gen Psychiatry 2007; 64: 19–28.

2 Jaaskelainen E, Juola P, Hirvonen N, McGrath JJ, Saha S, Isohanni M et al. A systematic review and meta-analysis of recovery in schizophrenia. Schizophr Bull 2013; 39: 1296–1306.

3 Perkins DO, Gu H, Boteva K, Lieberman JA. Relationship between duration of untreated psychosis and outcome in first-episode schizophrenia: a critical review and meta-analysis. Am J Psychiatry 2005; 162: 1785–1804.

4 Fusar-Poli P, Borgwardt S, Bechdolf A, Addington J, Riecher-Rossler A, Schultze-Lutter F et al. The psychosis high-risk state: a comprehensive state-of-the-art review. JAMA Psychiatry 2013; 70: 107–120.

5 Ziermans TB, Schothorst PF, Sprong M, van Engeland H. Transition and remission in adolescents at ultra-high risk for psychosis. Schizophr Res 2011; 126: 58–64.

6 Fusar-Poli P, Bonoldi I, Yung AR, Borgwardt S, Kempton MJ, Valmaggia L et al. Predicting psychosis: meta-analysis of transition outcomes in individuals at high clinical risk. Arch Gen Psychiatry 2012; 69: 220–229.

7 Katsura M, Ohmuro N, Obara C, Kikuchi T, Ito F, Miyakoshi T et al. A naturalistic longitudinal study of at-risk mental state with a 2.4 year follow-up at a specialized clinic setting in Japan. Schizophr Res 2014; 158: 32–38.

8 Demjaha A, Valmaggia L, Stahl D, Byrne M, McGuire P. Disorganization/cognitive and negative symptom dimensions in the at-risk mental state predict subsequent transition to psychosis. Schizophr Bull 2012; 38: 351–359.

9 Ruhrmann S, Schultze-Lutter F, Salokangas RK, Heinimaa M, Linszen D, Dingemans P et al. Prediction of psychosis in adolescents and young adults at high risk: results from the prospective European prediction of psychosis study. Arch Gen Psychiatry 2010; 67: 241–251.

10 DeVylder JE, Muchomba FM, Gill KE, Ben-David S, Walder DJ, Malaspina D et al. Symptom trajectories and psychosis onset in a clinical high-risk cohort: the relevance of subthreshold thought disorder. Schizophr Res 2014; 159: 278–283.

11 Nelson B, Yuen HP, Wood SJ, Lin A, Spiliotacopoulos D, Bruxner A et al. Long-term follow-up of a group at ultra high risk ("prodromal") for psychosis: the PACE 400 study. JAMA Psychiatry 2013; 70: 793–802.

12 Aberg KA, McClay JL, Nerella S, Clark S, Kumar G, Chen W et al. Methylome-wide association study of schizophrenia: identifying blood biomarker signatures of environmental insults. JAMA Psychiatry 2014; 71: 255–264.

13 Sullivan PF, Fan C, Perou CM. Evaluating the comparability of gene expression in blood and brain. Am J Med Genet B Neuropsychiatr Genet 2006; 141B: 261–268.

14 Schwartz M, Kipnis J, Rivest S, Prat A. How do immune cells support and shape the brain in health, disease, and aging? J Neurosci 2013; 33: 17587–17596.

15 Reader BF, Jarrett BL, McKim DB, Wohleb ES, Godbout JP, Sheridan JF. Peripheral and central effects of repeated social defeat stress: monocyte trafficking, microglial activation, and anxiety. Neuroscience 2015; 289: 429–442.

16 Baruch K, Ron-Harel N, Gal H, Deczkowska A, Shifrut E, Ndifon W et al. CNS-specific immunity at the choroid plexus shifts toward destructive Th2 inflammation in brain aging. Proc Natl Acad Sci USA 2013; 110: 2264–2269.

17 Marin I, Kipnis J. Learning and memory and the immune system. Learn Memory 2013; 20: 601–606.

18 Ziv Y, Ron N, Butovsky O, Landa G, Sudai E, Greenberg N et al. Immune cells contribute to the maintenance of neurogenesis and spatial learning abilities in adulthood. Nat Neurosci 2006; 9: 268–275.

19 Wolf SA, Steiner B, Akpinarli A, Kammertoens T, Nassenstein C, Braun A et al. CD4-positive T lymphocytes provide a neuroimmunological link in the control of adult hippocampal neurogenesis. J Immunol 2009; 182: 3979–3984.

20 Horvath S, Mirnics K. Immune system disturbances in schizophrenia. Biol Psychiatry 2014; 75: 316–323.

21 Miller BJ, Buckley P, Seabolt W, Mellor A, Kirkpatrick B. Meta-analysis of cytokine alterations in schizophrenia: clinical status and antipsychotic effects. *Biol Psychiatry* 2011; **70**: 663–671.

22 Upthegrove R, Manzanares-Teson N, Barnes NM. Cytokine function in medication-naive first episode psychosis: a systematic review and meta-analysis. *Schizophr Res* 2014; **155**: 101–108.

23 Miller BJ, Gassama B, Sebastian D, Buckley P, Mellor A. Meta-analysis of lymphocytes in schizophrenia: clinical status and antipsychotic effects. *Biol Psychiatry* 2013; **73**: 993–999.

24 Perkins DO, Jeffries CD, Addington J, Bearden CE, Cadenhead KS, Cannon TD et al. Towards a psychosis risk blood diagnostic for persons experiencing high-risk symptoms: preliminary results from the NAPLS project. *Schizophr Bull* 2015; **41**: 419–428.

25 Chan MK, Krebs MO, Cox D, Guest PC, Yolken RH, Rahmoune H et al. Development of a blood-based molecular biomarker test for identification of schizophrenia before disease onset. *Transl Psychiatry* 2015; **5**: e601.

26 Ha M, Kim VN. Regulation of microRNA biogenesis. *Nat Rev Mol Cell Biol* 2014; **15**: 509–524.

27 Hammond SM. Dicing and slicing: the core machinery of the RNA interference pathway. *FEBS Lett* 2005; **579**: 5822–5829.

28 Hammond SM. MicroRNA therapeutics: a new niche for antisense nucleic acids. *Trends Mol Med* 2006; **12**: 99–101.

29 Addington J, Cadenhead KS, Cornblatt BA, Mathalon DH, McGlashan TH, Perkins DO et al. North American Prodrome Longitudinal Study (NAPLS 2): overview and recruitment. *Schizophr Res* 2012; **142**: 77–82.

30 Miller TJ, McGlashan TH, Rosen JL, Somjee L, Markovich PJ, Stein K et al. Prospective diagnosis of the initial prodrome for schizophrenia based on the Structured Interview for Prodromal Syndromes: preliminary evidence of interrater reliability and predictive validity. *Am J Psychiatry* 2002; **159**: 863–865.

31 First MB, Spitzer RL, Givvon M, Williams JBW. *Structured Clinical Interview for DSM-IV TR Axis I Disorders, Non-patient Edition (SCID-I/NP)*. Biometrics Research, New York State Psychiatric Institute: New York, NY, USA, 2002.

32 Glatt SJ, Tsuang MT, Winn M, Chandler SD, Collins M, Lopez L et al. Blood-based gene expression signatures of infants and toddlers with autism. *J Am Acad Child Adolesc Psychiatry* 2012; **51**: 934–44 e2.

33 Zhang W, Gao S, Zhou X, Xia J, Chellappan P, Zhou X et al. Multiple distinct small RNAs originate from the same microRNA precursors. *Genome Biol* 2010; **11**: R81.

34 Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*. 2014; **42**: D68–D73.

35 Zhang W, Zeng T, Chen L. EdgeMarker: Identifying differentially correlated molecule pairs as edge-biomarkers. *J Theor Biol* 2014; **362**: 35–43.

36 North BV, Curtis D, Sham PC. A note on calculation of empirical P values from Monte Carlo procedure. *Am J Human Genet* 2003; **72**: 498–499.

37 Lindgren F, Hansen B, W K, Sjostrom M, L E. Model validation by permutation tests. *J Chemometrics* 1996; **10**: 521–532.

38 Rucker CL, Rucker G, Meringer M. y-Randomization and its variants in QSPR/QSAR. *J Chem Inf Model* 2007; **47**: 2345–2357.

39 Smit S, Hoefsloot HC, Smilde AK. Statistical data processing in clinical proteomics. *J Chromatogr B Analyt Technol Biomed Life Sci* 2008; **866**: 77–88.

40 Tropsha A. Best practices for QSAR model development, validation, and exploitation. *Mol Informatics* 2010; **29**: 476–488.

41 Buzkova PL, Lumley T, Rice K. Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions. *Ann Hum Genet* 2011; **75**: 36–45.

42 Fredrickson BL, Grewen KM, Coffey KA, Algoe SB, Firestine AM, Arevalo JM et al. A functional genomic perspective on human well-being. *Proc Natl Acad Sci USA* 2013; **110**: 13684–13689.

43 Woods SW, Addington J, Bearden CE, Cadenhead KS, Cannon TD, Cornblatt BA et al. Psychotropic medication use in youth at high risk for psychosis: comparison of baseline data from two research cohorts 1998-2005 and 2008-2011. *Schizophr Res* 2013; **148**: 99–104.

44 Smalheiser NR, Lugli G, Rizavi HS, Torvik VI, Turecki G, Dwivedi Y. MicroRNA expression is downregulated and reorganized in prefrontal cortex of depressed suicide subjects. *PLoS One* 2012; **7**: e33201.

45 Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004; **116**: 281–297.

46 Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 2005; **120**: 15–20.

47 Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell* 2009; **136**: 215–233.

48 Shin C, Nam JW, Farh KK, Chiang HR, Shkumatava A, Bartel DP. Expanding the microRNA targeting code: functional sites with centered pairing. *Mol Cell* 2010; **38**: 789–802.

49 Loeb GB, Khan AA, Canner D, Hiatt JB, Shendure J, Darnell RB et al. Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting. *Mol Cell* 2012; **48**: 760–770.

50 Chi SW, Hannon GJ, Darnell RB. An alternative mode of microRNA target recognition. *Nat Struct Mol Biol* 2012; **19**: 321–327.

51 Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 2013; **153**: 654–665.

52 Smith TF, Waterman MS, Fitch WM. Comparative biosequence metrics. *J Mol Evol* 1981; **18**: 38–46.

53 Gardiner E, Beveridge NJ, Wu JQ, Carr V, Scott RJ, Tooney PA et al. Imprinted DLK1-DIO3 region of 14q32 defines a schizophrenia-associated miRNA signature in peripheral blood mononuclear cells. *Mol Psychiatry* 2012; **17**: 827–840.

54 Lai CY, Yu SL, Hsieh MH, Chen CH, Chen HY, Wen CC et al. MicroRNA expression aberration as potential peripheral blood biomarkers for schizophrenia. *PLoS One* 2011; **6**: e21635.

55 Fan HM, Sun XY, Niu W, Zhao L, Zhang QL, Li WS et al. Altered microRNA expression in peripheral blood mononuclear cells from young patients with schizophrenia. *J Mol Neurosci* 2015; **56**: 562–571.

56 Song HT, Sun XY, Zhang L, Zhao L, Guo ZM, Fan HM et al. A preliminary analysis of association between the downregulation of microRNA-181b expression and symptomatology improvement in schizophrenia patients before and after antipsychotic treatment. *J Psychiatr Res* 2014; **54**: 134–140.

57 Sun XY, Lu J, Zhang L, Song HT, Zhao L, Fan HM et al. Aberrant microRNA expression in peripheral plasma and mononuclear cells as specific blood-based biomarkers in schizophrenia patients. *J Clin Neurosci* 2015; **22**: 570–574.

58 Yu HC, Wu J, Zhang HX, Zhang GL, Sui J, Tong WW et al. Alterations of miR-132 are novel diagnostic biomarkers in peripheral blood of schizophrenia patients. *Prog Neuropsychopharmacol Biol Psychiatry* 2015; **63**: 23–29.

Supplementary Information accompanies the paper on the Translational Psychiatry website (http://www.nature.com/tp)