5 μg of fragmented biotinylated RNA at 65 °C overnight. For post-hybridization, the DNA microarray was washed twice in 0.12 M Tris–HCl/0.12 M NaCl/0.05% Tween 20 at 65 °C for 20 min, followed by washing in 0.12 M Tris–HCl/0.12 M NaCl for 10 min. The DNA microarray was then labeled with streptavidin–Cy5 (GE Healthcare Bio-Sciences, Tokyo). The fluorescent-labeled DNA microarray was washed for 5 min four times in 0.12 M Tris–HCl/0.12 M NaCl/0.05% Tween 20 at room temperature. Hybridization signal acquisition was performed using a DNA microarray reader adopting multi-beam excitation technology (Yokogawa Electric, Tokyo). The DNA microarray was scanned at multiple exposure times ranging from 0.1 to 40 s. The intensity values with the best exposure condition for each spot were then selected.

*GeneSQUARE*

The custom DNA microarray was designed using the GeneSQUARE Multiple Assay DNA Microarray platform. Alexa Fluor 555-labeled cDNA was prepared from total RNA cocktail (10 μg) by cDNA synthesis and in vitro transcription performed using the GeneSQUARE cDNA Direct Labeling System (Kurabo Industries) according to the manufacturer's instructions. Labeled cDNA was purified with a MinElute PCR Purification Kit (Qiagen) and added to hybridization buffer (5× sodium saline citrate [SSC, pH 7.0], 4× Denhardt's solution [Sigma–Aldrich, St. Louis, MO, USA], 1 μg of salmon sperm DNA [Life Technologies], and 0.5% [w/v] sodium dodecyl sulfate [SDS]). Hybridization was performed in a final volume of 8 μl per well on a GeneSQUARE Multiple Assay DNA Microarray (JMAC) in a hybridization chamber (Kurabo Industries) at 65 °C for 16 h in a water bath. After hybridization, the hybridized slides were washed by the following steps: (i) immersion in 1× SSC and 0.1% SDS solution for 5 min, (ii) immersion in 0.2× SSC and 0.1% SDS for 5 min, (iii) immersion in 0.2× SSC and 0.1% SDS at 55 °C for 5 min, (iv) rocking in 0.2× SSC, and (v) immersion in 0.05% SSC for 2 min. After they were dried by centrifuge, the slides were scanned with GenePix 4000B (Molecular Devices). Fluorescence intensities of scanned images were quantified with GenePix Pro 6.0 software (Molecular Devices).

*S-Bio*

The custom microarray was designed using the S-Bio plastic slide platform (Sumitomo Bakelite) and oligonucleotide probes spotted with ProbeBank [17,13] and EC amino linker [18]. The total amount of RNA required can be reduced by mechanical spotting using the GENESHOT Spotting Device (NGK Insulators, Nagoya, Japan).

Total RNA cocktail was amplified using the MessageAmp II Biotin-Enhanced Single Round aRNA Amplification Kit (Life Technologies). Briefly, the total RNA cocktail of each sample (1 μg each) was transcribed into double-stranded T7 RNA polymerase promoter-tagged cDNA and then amplified into single-stranded biotin-labeled aRNA by T7 polymerase. aRNA (3 μg) was fragmented at 94 °C for 15 min and hybridized on the microarray in the presence of formamide (final concentration 10%, v/v) at 37 °C for 16 h. The microarray was washed at room temperature for 5 min in 0.1× SSC and 0.1% SDS, followed by another 5 min wash in 0.05× SSC and 0.1% SDS at 43 °C. Finally, the microarray was rinsed in 0.05× SSC before drying by low-speed centrifugation. For staining, the microarray was immersed in a phosphate-buffered saline (PBS) solution containing 10 μg/ml streptavidin, R-phycoerythrin conjugate (Life Technologies), Tween 20 (0.05%, v/v), and bovine serum albumin (2 mg/ml) for 30 min. Washing was performed to remove the additional stain at room temperature in PBS buffer for 5 min, followed by another wash in a similar buffer prepared separately for 30 s. The microarray was rinsed in 0.05× SSC at room temperature before drying by low-speed centrifugation.

The microarray was scanned using an Agilent DNA Microarray platform scanner (Agilent Technologies, cat. no. G2565BA) at photo-multiplier tube 800 in a resolution of 10 μm. The intensity values of each feature of the scanned image were quantitated using Feature Extraction software (version 9.1, Agilent Technologies).

*NimbleGen*

The custom microarray was designed using the NimbleGen platform (Roche NimbleGen). Total RNA cocktail (0.5 μg) was used as a starting material to prepare double-stranded cDNA using a SuperScript Double-Stranded cDNA Synthesis Kit (Life Technologies). cDNA was labeled using the random priming method with Cy3-labeled random nonamer primers and Klenow DNA polymerase at 37 °C for 2 h using a NimbleGen One-Color DNA Labeling Kit (Roche NimbleGen.). The Cy3-labeled cDNA (4 μg) was hybridized to the DNA microarray using a NimbleGen Hybridization Kit (Roche NimbleGen) for 17 h at 42 °C. The microarray platform slides were washed and scanned with a NimbleGen MS200 microarray platform scanner (Roche NimbleGen).

*Next-generation sequencer GAII*

The RNA samples with standard RNA were prepared with an mRNA-Seq Sample Prep Kit (Illumina, San Diego, CA, USA). A 75-base single run was performed on the next-generation sequencer using a Single-Read Cluster Generation Kit (version 4) and TruSeq SBS v5-GA (Illumina GAII) with two samples. The read sequences were aligned against the human genome and the standard RNA sequences using a BLAST program (http://blast.ncbi.nlm.nih.gov/Blast.cgi) [19].

*Taqman RT–PCR*

A High-Capacity cDNA Reverse Transcription Kit (Life Technologies) was used for cDNA synthesis. The messenger RNA (mRNA) level was monitored with the 7500 Fast Real-Time PCR System (Life Technologies) and TaqMan Fast Universal PCR Master Mix (Life Technologies) following the manufacturer's instructions. TaqMan primer probes for standard combination number 183 except external standard 1000_4 were custom designed as follows: 500_2, 500_4, 1000_3, 500_1.

In addition, Taqman primer probes for ACTB (TaqMan Probe ID: Hs03023880_g1), B2M (TaqMan Probe ID: Hs00984239_m1), GAPDH (TaqMan Probe ID: Hs99999905_m1), GUSB (TaqMan Probe ID: Hs99999908_m1), HPRT1 (TaqMan Probe ID: Hs99999909_m1), PGK1 (TaqMan Probe ID: Hs00943178_g1), PPIA (TaqMan Probe ID: Hs99999904_m1), RPLP0 (TaqMan Probe ID: Hs99999902_m1), TBP (TaqMan Probe ID: Hs00920497_m1), and YWHAZ (TaqMan Probe ID: Hs01122451_m1) were purchased (Life Technologies). The RNA copy numbers were normalized to those of internal ACTB.

## Results

*Signal correction*

In DNA microarray analyses, signal intensities may vary significantly between different platforms, even if measuring the same sample, making it difficult to compare cross-platform data sets. Therefore, signal normalization using internal standards was performed to account for this variation. Specifically, correction was performed by dividing the values obtained by subtracting background levels and by the median values of positive controls (Eq. (1)):

$$\frac{[500\_1\_P1]_{\text{signal-BG}}}{[ACTB, B2M, GAPDH, GUSB, HPRT1, PGK1, PPIA, RPLP0, TBP, YWHAZ]_{\text{median}}}. \quad (1)$$

$[\text{geneX}]_{\text{signal-BG}}$ = Background (BG) subtracted signal intensity of gene X probe
$[\text{geneA, B, C,...}]_{\text{median}}$ = median of BG subtracted signal intensities of gene A, B, C,... probe

Background levels were calculated using a standard procedure for each microarray. In this correction procedure, 10 positive controls—ACTB, B2M, GAPDH, GUSB, HPRT1, PGK1, PPIA, RPLP0, TBP, and YWHAZ—were treated as internal standards. Although only a single RNA standard is used in typical chemical analyses, in DNA microarray platforms, where expression of the positive control results in considerably different signal intensities between platforms, multiple positive controls based on a large number of quantification targets are required. Their median value is then used for making corrections.

*Confirmation of concentration dependency*

First, confirmation was sought as to whether or not each microarray demonstrates concentration dependency. Here, a linear evaluation was conducted on four types of probe for each of 10 types of standard in order to select suitable combinations (1000_1–5, 500_1–5). The concentration of the standards ranged from 10 to 100,000 zmol, and serial dilutions were prepared in 10-fold increments. Measurements were repeated at least two times on the same solution.

Signal intensities obtained by hybridization using each DNA microarray platform were corrected using the median values of the positive controls (Eq. (1)). The relationship between those values and RNA concentration was plotted, linear regression was performed for each of the standards and probes (P1–P4), and their slopes and correlation coefficients were calculated. An example indicating the concentration dependency of standard substance candidate 500_1 is shown in Fig. 3.

Although no correlation was observed between concentration increases in the standards and the corrected signal values at a low amount (10 zmol), by selecting a higher amount (1 nmol) range correlation coefficients of 0.97 and above were obtained, thereby confirming concentration dependency. The observation of a direct association between signal and concentration strongly suggests that our probes are detecting our external RNA standards specifically.

*Selection of standard substance candidate–probe combinations*

The most linear concentration ranges were selected and linear regression was performed for each standard, each platform, and each probe using the results described in the previous section. Next, the slopes and correlation coefficients were extracted for those ranges. Microarray data were summarized for each probe, and average values for slopes and correlation coefficients were determined as shown in Table 1 and Supplementary Table S4 for all data:

$$\frac{|[\text{P1}]_{\text{Ave}} - [\text{P1, P2, P3, P4}]_{\text{Ave}}|}{\{[\text{P1, P2, P3, P4}]_{75\%} - [\text{P1, P2, P3, P4}]_{25\%}\} * 0.7413}. \quad (2)$$

Next, $Z$ scores were calculated for each standard–probe combination (Eq. (2)), and combinations of standard and probe were selected by extracting and excluding outlier probes (probes having scores $\geqslant 2$). In other words, those probes that demonstrated bias were excluded from the P1 to P4 probes prepared for each standard.

Considering $Z$ score and average and standard deviation of slope in each platform, the selected combinations of standard and probes consisted of the following nine types: 1000_2_P2, 1000_2_P3, 1000_3_P3, 1000_4_P2, 500_1_P2, 500_1_P3, 500_1_P4, 500_2_P4, and 500_4_P2.

*Approach used to prepare calibration curves*

Calibration curves indicate the relationship between signal and concentration as a result of preparing serially diluted standard solutions and measuring those solutions under the same conditions as samples.

DNA microarray platforms are characterized by being able to acquire a large amount of data using a single microarray. Thus, we devised a method in which mixtures of multiple standards were used and calibration curves were produced using a single DNA microarray by changing the concentration level of each. A graphical representation of this approach is indicated in Fig. 4. In general, although standards are prepared by sequentially diluting one type of standard to produce a single quantification target, in this study five different types of standard (RNA standards E1–E5) were prepared while changing their respective concentration levels. This results in a "relative" quantification of expression level instead of an "absolute" quantification.

*Analysis of calibration curves*

Combinations of the nine types of standard–probe combinations selected (see "Selection of standard substance candidate–probe combinations" section above) and at five concentration levels (10, 100, 1000, 10,000, and 100,000 zmol) were prepared based on the approach depicted in Fig. 4, and combinations having the lowest levels of error between linearity and microarray were selected.

Corrected values for the nine types of standard–probe pairs were calculated for each of the microarray platforms A to F and plotted. In those plots, a line corresponding to a slope of 1 and a line passing near the center of each concentration were drawn. In cases where all six microarray platforms fell between these two lines, the data for those microarray platforms were used and points falling outside these lines were excluded. A round-robin system was used until a standard probe pair that met the above criteria was found. A total of 186 combinations were evaluated.

Signals were corrected as defined above for each microarray platform A to F, for each combination of standard–probe pair, and for each concentration level, and linear regression was performed. Those combinations that demonstrated an average

$[\text{P1-PL-X}]_{\text{slope}}$ : slope of probe P1 in platform X extracted for the most linear concentration ranges
$[\text{P1}]_{\text{Ave}}$ : Average slope of $[\text{P1-PL-A}]_{\text{slope}}, [\text{P1-PL-B}]_{\text{slope}}, [\text{P1-PL-C}]_{\text{slope}}, [\text{P1-PL-D}]_{\text{slope}}, [\text{P1-PL-E}]_{\text{slope}}, [\text{P1-PL-F}]_{\text{slope}}$
$[\text{P1, P2, P3, P4}]25\%$ : 25% quintile of $[\text{P1}]_{\text{Ave}}, [\text{P2}]_{\text{Ave}}, [\text{P3}]_{\text{Ave}}$ and $[\text{P4}]_{\text{Ave}}$
$[\text{P1, P2, P3, P4}]75\%$ : 75% quintile of $[\text{P1}]_{\text{Ave}}, [\text{P2}]_{\text{Ave}}, [\text{P3}]_{\text{Ave}}$ and $[\text{P4}]_{\text{Ave}}$

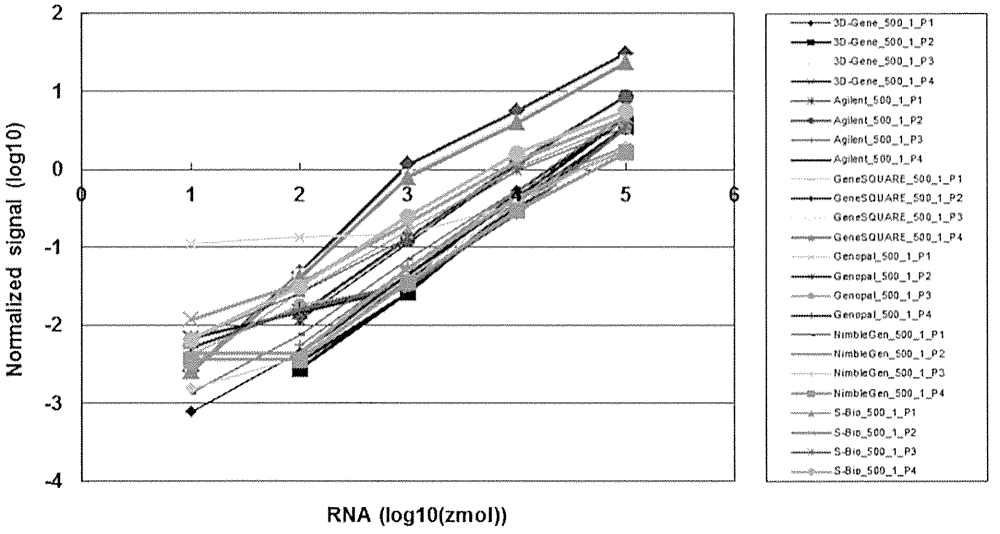



**Fig.3.** Procedures for selecting standard substance candidate–probe combinations using *Z* scores. The relationship between normalized signal intensities and RNA concentration of standard substance candidate 500_1 was plotted. Linear regression was performed for each of the standards and probes (P1–P4), and their slopes and correlation coefficients were calculated.

**Table 1**
Averages of slopes and correlation coefficients of six microarray platforms for P1 of standard substance candidate 500_1.

| RNA | Probe name | Platform | RNA (log10 (zmol)) | | Slope | $r^2$ |
|---|---|---|---|---|---|---|
| | | | Max | Min | | |
| 500_1 | 500_1_P1 | 3D-Gene | 5 | 2 | 1.020 | 0.998 |
| | | Agilent | 5 | 2 | 0.965 | 0.999 |
| | | GeneSQUARE | 4 | 1 | 1.032 | 0.996 |
| | | Genopal | 5 | 3 | 0.769 | 0.972 |
| | | NimbleGen | 4 | 1 | 0.884 | 0.993 |
| | | S-Bio | 5 | 1 | 0.777 | 0.994 |
| | | Average | | | 0.908 | 0.992 |
| | | SD | | | 0.117 | 0.010 |
| | | *Z* score | | | 1.429 | 1.587 |
| | 500_1_P2 | 3D-Gene | 5 | 2 | 1.047 | 0.999 |
| | | Agilent | 5 | 2 | 0.935 | 0.995 |
| | | GeneSQUARE | 4 | 1 | 1.071 | 0.998 |
| | | Genopal | 5 | 3 | 0.993 | 0.999 |
| | | NimbleGen | 5 | 2 | 0.881 | 0.915 |
| | | S-Bio | 5 | 2 | 0.722 | 0.996 |
| | | Average | | | 0.941 | 0.984 |
| | | SD | | | 0.128 | 0.034 |
| | | *Z* score | | | 2.137 | 1.189 |
| | 500_1_P3 | 3D-Gene | 5 | 2 | 1.026 | 0.999 |
| | | Agilent | 5 | 2 | 0.993 | 0.992 |
| | | GeneSQUARE | 5 | 2 | 0.883 | 0.999 |
| | | Genopal | 5 | 3 | 0.977 | 1.000 |
| | | NimbleGen | 5 | 2 | 0.908 | 0.929 |
| | | S-Bio | 5 | 2 | 0.723 | 0.997 |
| | | Average | | | 0.918 | 0.986 |
| | | SD | | | 0.110 | 0.028 |
| | | *Z* score | | | 0.305 | 0.437 |
| | 500_1_P4 | 3D-Gene | 5 | 2 | 1.023 | 0.998 |
| | | Agilent | 5 | 2 | 1.002 | 1.000 |
| | | GeneSQUARE | 5 | 2 | 0.888 | 0.997 |
| | | Genopal | 5 | 3 | 0.981 | 0.998 |
| | | NimbleGen | 5 | 2 | 0.889 | 0.942 |
| | | S-Bio | 5 | 1 | 0.763 | 0.997 |
| | | Average | | | 0.924 | 0.989 |
| | | SD | | | 0.098 | 0.023 |
| | | *Z* score | | | 0.305 | 0.437 |

*Note:* SD, standard deviation.

slope of microarray platforms A to F of 0.8 or higher, an average correlation coefficient of 0.95 or higher, and variability in the correlation coefficients of microarray platforms A to F of 5% or less were selected.

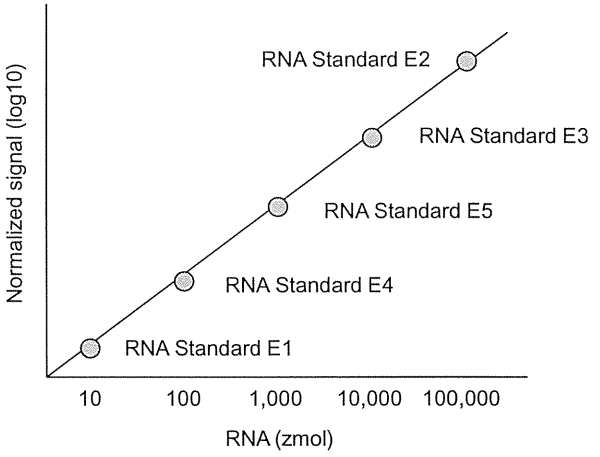



**Fig.4.** Approach used to prepare calibration curves: Evaluation of assay linearity at input levels for spike-in RNAs to the microarray platform. Linearity was assessed based on a five-point dilution series, ranging four orders of magnitude, of the spiked external RNA E1 to RNA E5. Each dilution series was measured, and the linearity of each of the 186 combinations was estimated by calculating the A round-robin system. The standard probe combination was determined to the extent that linearity was obtained.

Next, the variability of the signal correction values at each concentration was determined, and those combinations for which the variability between microarray platforms was 25% or less at three or more of the five concentration levels were selected (nos. 16, 33, and 183). The selected combinations are shown in Supplementary Table S5. A reproducibility experiment was conducted to confirm whether or not linearity is reproducible (Supplementary Table S6).

*Best combination of probe sets for standard RNAs*

Next, the standard combination numbers 16, 33, and 183 were added to HURR and HBRR, and microarray expression analyses were conducted using each platform. Standard combination number 183 demonstrated the best concentration linearity and reproducibility among all six platforms (Fig. 5). This consisted of the standard–probe combinations of 500_2_P4, 500_4_P2,

1000_3_P3, 500_1_P3, and 1000_4_P2. Thus, the use of this combination of standard substances and probes can ensure data compatibility between DNA microarray platforms. Although the dynamic range of compatibility varies according to each platform, it is thought to be at least three orders of magnitude.

*Digital expression profiling (RNA-Seq)*

Next, an RNA-Seq expression analysis was conducted using the model GAII manufactured by Illumina on the standard RNA combination number 183. The results were then compared with expression data obtained with DNA microarray in order to investigate the correlation with next-generation sequencers. In addition, RT–PCR was performed by using TaqMan primer probes for standard combination number 183. The results demonstrate a close correlation with respect to dynamic range, linearity, and the like (Fig. 6). We suggest that these results indicate that the use of standards can be expanded not only to include compatibility between DNA microarray platforms but also to evaluate the data obtained using next-generation sequencers.

*Results of quantitative PCR comparisons*

Quantitative PCR was conducted on four RNA external standards (500_2, 500_4, 1000_3, and 500_1) and 10 types of genes (ACTB, B2M, GAPDH, GUSB, HPRT1, PGK1, PPIA, RPLP0, TBP, and YWHAZ) for each of the HURR and HBRR samples to which a standard combination number 183 was added. Expression ratios for each sample were calculated using the ddCt method. When these results were compared with the expression ratio data obtained by three DNA microarray platforms and next-generation sequencer, significant correlations were observed for each DNA microarray platform and next-generation sequencer (Fig. 7). We conclude that we are able to obtain equivalent data across microarray platforms for both standard and real-time applications provided that the expression level is above a certain threshold level.

## Discussion

Attempts to compare DNA microarray data across platforms have been made in the past, and data compatibility has been maintained by comparing expression fluctuation ratios. Signal intensity is not generally considered to be comparable directly across platforms due to differences in various factors such as the target preparation method, probe sequence, and detection method.

In this study, we were able to select standard–probe combinations that can be used across multiple platforms by developing methods for optimizing both the standard and the detection probe. When detection signal intensities from each platform were corrected using our RNA standards, the correlation with quantitative RT–PCR, considered to be the "gold standard," was shown to improve. We suggest that a calibration method based on our standards is effective and can contribute to improvements in data reliability. Moreover, due to the high level of correlation in signal

#183/HURR

| probe | RNA (LOG10) | Normalized Signal (Log10) | | | | | | max | min | ave | σ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 3D-Gene | Agilent | GAII | GeneSQUARE | NimbleGen | S-Bio | | | | |
| 500_2_4 | 5 | 0.628 | 0.714 | 0.664 | 0.308 | 0.182 | 0.841 | 0.841 | 0.182 | 0.556 | 0.255 |
| 500_4_2 | 4 | -0.368 | -0.036 | -0.437 | -0.684 | -0.390 | 0.178 | 0.178 | -0.684 | -0.289 | 0.309 |
| 1000_3_3 | 3 | -1.204 | -1.684 | -1.333 | -1.567 | -0.927 | -0.711 | -0.711 | -1.684 | -1.238 | 0.372 |
| 500_1_3 | 2 | -2.347 | -2.024 | -2.410 | -2.785 | -2.275 | -1.607 | -1.607 | -2.785 | -2.241 | 0.397 |
| 1000_4_2 | 1 | -2.903 | -2.751 | -3.586 | -3.433 | -2.884 | -2.209 | -2.209 | -3.586 | -2.961 | 0.497 |

#183/HBRR

| probe | RNA (LOG10) | Normalized Signal (Log10) | | | | | | max | min | ave | σ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 3D-Gene | Agilent | GAII | GeneSQUARE | NimbleGen | S-Bio | | | | |
| 500_2_4 | 5 | 0.931 | 1.191 | 0.562 | 0.782 | 0.471 | 1.058 | 1.191 | 0.471 | 0.832 | 0.281 |
| 500_4_2 | 4 | 0.019 | 0.466 | -0.504 | -0.251 | -0.071 | 0.356 | 0.466 | -0.504 | 0.002 | 0.365 |
| 1000_3_3 | 3 | -0.828 | -1.126 | -1.366 | -1.157 | -0.637 | -0.485 | -0.485 | -1.366 | -0.933 | 0.339 |
| 500_1_3 | 2 | -2.003 | -1.532 | -2.359 | -2.572 | -1.868 | -1.370 | -1.370 | -2.572 | -1.951 | 0.464 |
| 1000_4_2 | 1 | -2.593 | -2.268 | -3.262 | -3.047 | -2.537 | -2.392 | -2.268 | -3.262 | -2.683 | 0.388 |

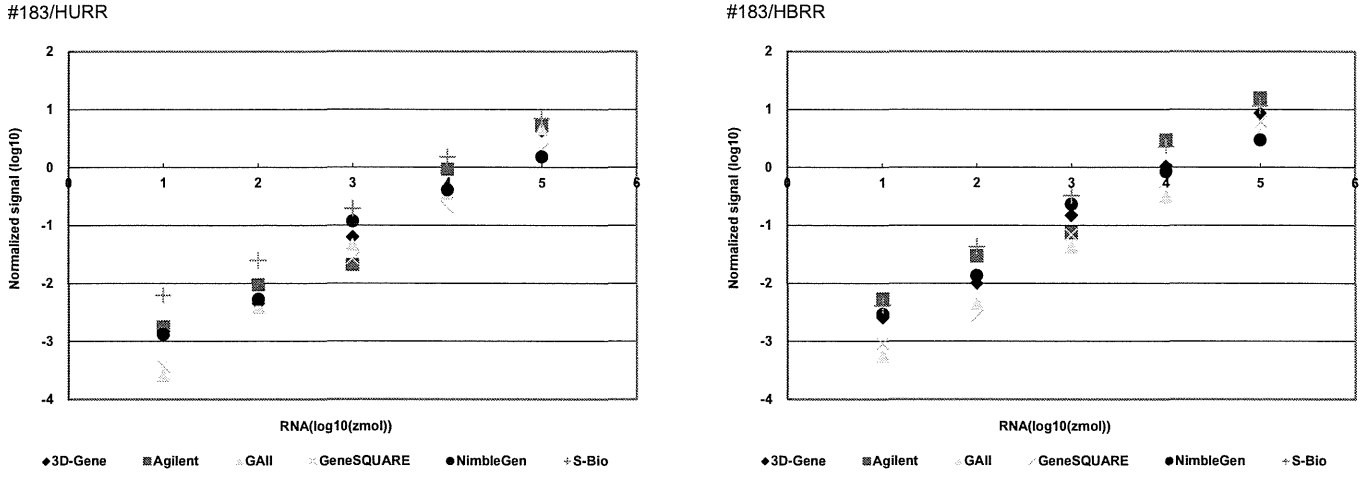



**Fig.5.** Linearity and reproducibility among all six microarray platforms of standard combination number 183. Among the three combinations, number 183 demonstrated the best concentration linearity and reproducibility (lowest standard deviation value) in all six platforms.

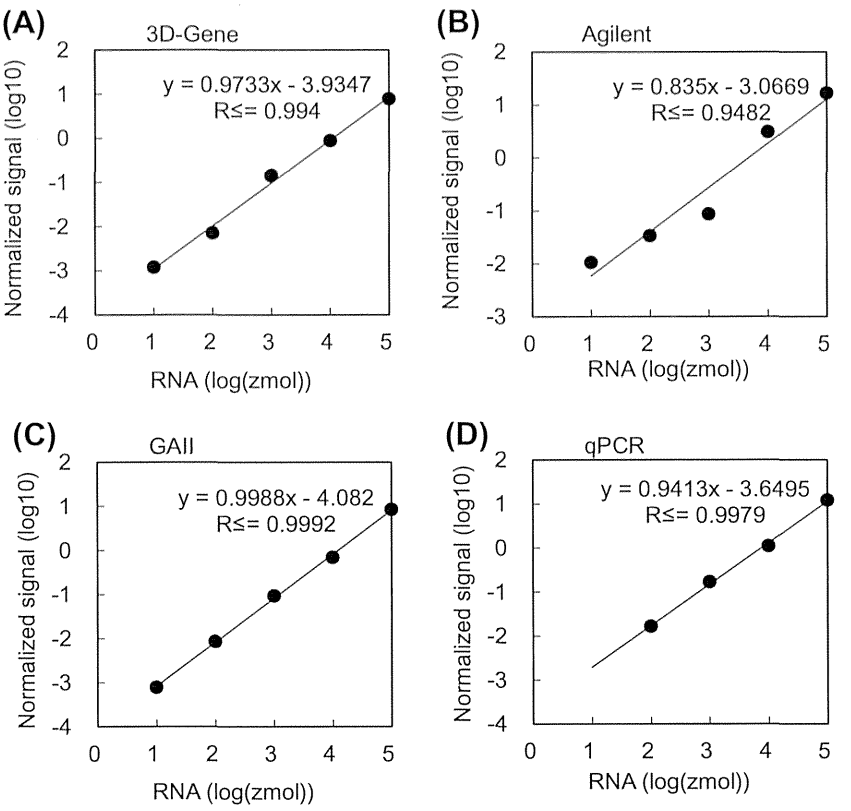



**Fig.6.** Comparison of quantitative RT–PCR data derived from microarray platforms and a next-generation sequencer: Assessment of the assay linearity between and microarray platforms and a next-generation sequencer. Using the standard RNA combination number 183 contents enabled us to obtain the high correlation with respect to dynamic range, linearity between microarray platforms, and a next-generation sequencer. qPCR, quantitative RT–PCR.

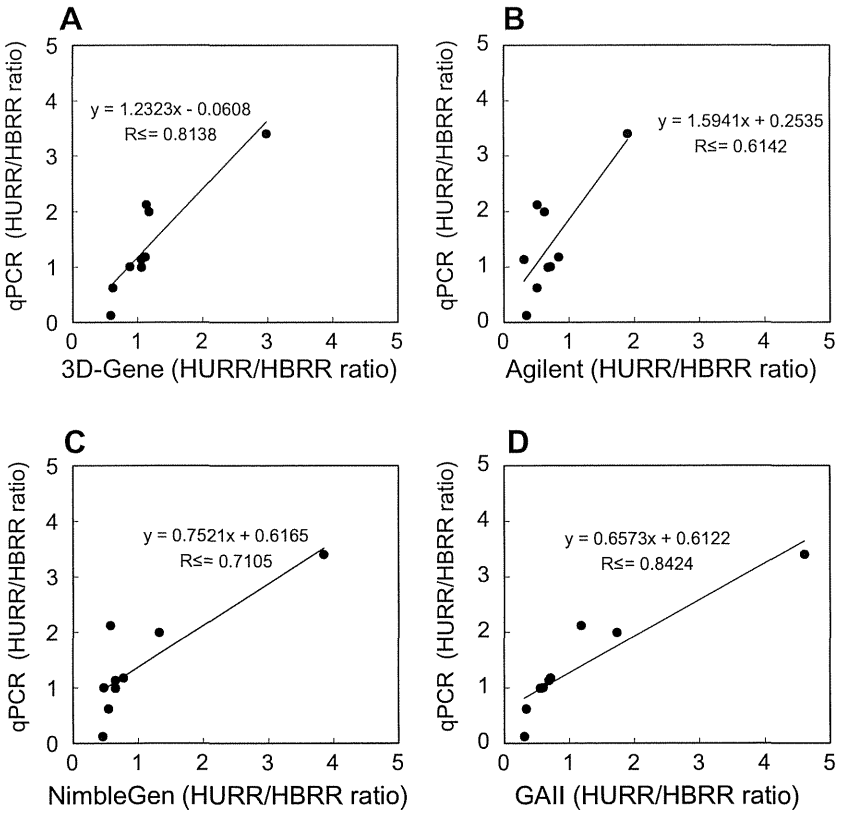



**Fig.7.** Comparison of fold change ratio (HURR/HBRR) among quantitative RT–PCR, microarray platforms, and next-generation sequencer. (A–C) Comparison of fold change ratio (HURR/HBRR) between microarrays and quantitative RT–PCR. (D) Comparison of fold change ratio (HURR/HBRR) between next-generation sequencer and quantitative RT–PCR. The dot plot indicates the fold change ratio (HURR/HBRR). Nine genes indicated that there is an expression difference. qPCR, quantitative RT–PCR.

intensity, the use of this calibration method makes it possible to directly compare detection across platforms.

Because this calibration method was also effective with next-generation sequencers employing different detection principles, we further suggest that our calibration method can be applied to various gene expression analysis techniques. In addition, we suggest that the standard selection method and calibration method we have developed is effective for detection methods other than DNA microarray platforms.

The Affymetrix DNA microarray was not assessed in this study because it is not possible to produce a custom microarray with this platform. The current pace of progress in the field of genetic diagnostics has resulted in DNA microarray platforms being increasingly used in patient management such as MammaPrint in breast cancer prognosis prediction. The calibration method we have developed enables evaluation of intra-run, inter-run, and cross-platform DNA microarray detection data, thereby making it possible to improve and maintain reliability [20–22].

In conclusion, we suggest that our set of validated nucleotide standards will enable direct comparison of data produced using multiple DNA microarray platforms provided with identical clinical samples, thereby ensuring the compatibility of detection results, inter-laboratory communication, and diagnoses. Moreover, because external RNA standards enable identification of failed steps during the assay process, it is possible to improve reliability and ensure compatibility between data sets, suggesting that similar results can be obtained in clinical diagnostic testing independent of the specific platform used.

## Acknowledgments

We thank Ayumi Hasegawa (Sumika Chemical Analysis Service) for the design of data analysis and Tatsunobu Fukushima (Mitsubishi Chemical Holdings) for helpful discussion during this work. We also acknowledge support from a New Energy and Industrial Technology Development Organization (NEDO) grant (NEDO 10001779-0) funded by the Ministry of Economy, Trade and Industry – Japan. We also thank Dr. Keith Hart, Cardiff University and members of Japan MicroArray Consortium (JMAC) for helpful discussion in this work.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ab.2014.11.012.

## References

[1] J. de Leon, M.T. Susce, E. Murray-Carmichael, The AmpliChip CYP450 Genotyping Test: integrating a new clinical tool, Mol. Diagn. Ther. 10 (2006) 135–151.

[2] T. Satoh, K. Matsumoto, T. Fujii, O. Sato, N. Gemma, M. Onuki, H. Saito, D. Aoki, Y. Hirai, H. Yoshikawa, Rapid genotyping of carcinogenic human papillomavirus by loop-mediated isothermal amplification using a new automated DNA test (CliniChip HPV), J. Virol. Methods 188 (2013) 83–93.

[3] C.I. Dumur, C.E. Fuller, T.L. Blevins, J.C. Schaum, D.S. Wilkinson, C.T. Garrett, C.N. Powers, Clinical verification of the performance of the pathwork tissue of origin test: utility and limitations, Am. J. Clin. Pathol. 136 (2011) 924–933.

[4] Z. Wang, A.P. Malanoski, B. Lin, C. Kidd, N.C. Long, K.M. Blaney, D.C. Thach, C. Tibbetts, D.A. Stenger, Resequencing microarray probe design for typing genetically diverse viruses: human rhinoviruses and enteroviruses, BMC Genomics 9 (2008) 577.

[5] M.J. van de Vijver, Y.D. He, L.J. van't Veer, H. Dai, A.A. Hart, D.W. Voskuil, G.J. Schreiber, J.L. Peterse, C. Roberts, M.J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E.T. Rutgers, S.H. Friend, R. Bernards, A gene-expression signature as a predictor of survival in breast cancer, N. Engl. J. Med. 347 (2002) 1999–2009.

[6] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F.L. Baehner, M.G. Walker, D. Watson, T. Park, W. Hiller, E.R. Fisher, D.L. Wickerham, J. Bryant, N. Wolmark, A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer, N. Engl. J. Med. 351 (2004) 2817–2826.

[7] R. Salazar, P. Roepman, G. Capella, V. Moreno, I. Simon, C. Dreezen, A. Lopez-Doriga, C. Santos, C. Marijnen, J. Westerga, S. Bruin, D. Kerr, P. Kuppen, C. van de Velde, H. Morreau, L. van Velthuysen, A.M. Glas, L.J. van't Veer, R. Tollenaar, Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer, J. Clin. Oncol. 29 (2011) 17–24.

[8] J. Ambroise, B. Bearzatto, A. Robert, B. Govaerts, B. Macq, J.L. Gala, Impact of the spotted microarray preprocessing method on fold-change compression and variance stability, BMC Bioinformatics 12 (2011) 413.

[9] S. Mao, C. Wang, G. Dong, Evaluation of inter-laboratory and cross-platform concordance of DNA microarrays through discriminating genes and classifier transferability, J. Bioinf. Comput. Biol. 7 (2009) 157–173.

[10] S. Schwers, E. Reifenberger, M. Gehrmann, A. Izmailov, K. Bohmann, A high-sensitivity, medium-density, and target amplification-free planar waveguide microarray system for gene expression analysis of formalin-fixed and paraffin-embedded tissue, Clin. Chem. 55 (2009) 1995–2003.

[11] W.R. Pearson, Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms, Genomics 11 (1991) 635–650.

[12] N. Markham, M. Zuker, UNAFold: Software for nucleic acid folding and hybridization, Methods Mol. Biol. 453 (2008) 3–31.

[13] T. Yamada, H. Soma, S. Morishita, PrimerStation: a highly specific multiplex genomic PCR primer design server for the human genome, Nucleic Acids Res. 34 (2006) W665–W669.

[14] J. SantaLucia Jr., A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics, Proc. Natl. Acad. Sci. USA 95 (1998) 1460–1465.

[15] K. Nagino, O. Nomura, Y. Takii, A. Myomoto, M. Ichikawa, F. Nakamura, M. Higasa, H. Akiyama, H. Nobumasa, S. Shiojima, G. Tsujimoto, Ultrasensitive DNA chip: gene expression profile analysis without RNA amplification, J. Biochem. 139 (2006) 697–703.

[16] D. Okuzaki, T. Fukushima, T. Tougan, T. Ishii, S. Kobayashi, K. Yoshizaki, T. Akita, H. Nojima, "Genopal: a novel hollow fiber array for focused microarray analysis, DNA Res. 8 (2010) 1–11.

[17] T. Yamada, S. Morishita, Computing highly specific and noise tolerant oligomers efficiently, J. Bioinf. Comput. Biol. 2 (2004) 21–46.

[18] Y. Komatsu, N. Kojima, M. Sugino, A. Mikami, K. Nonaka, Y. Fujinawa, T. Sugimoto, K. Sato, K. Matsubara, E. Ohtsuka, Novel amino linkers enabling efficient labeling and convenient purification of amino-modified oligonucleotides, Bioorg. Med. Chem. 16 (2008) 941–949.

[19] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[20] MAQC Consortium, The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements, Nat. Biotechnol. 24 (2006) 1151–1161.

[21] MAQC Consortium, The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models, Nat. Biotechnol. 28 (2010) 827–838.

[22] T.A. Patterson, E.K. Lobenhofer, S.B. Fulmer-Smentek, P.J. Collins, T.-M. Chu, W. Bao, H. Fang, E.S. Kawasaki, J. Hager, I.R. Tikhonova, S.J. Walker, L. Zhang, P. Hurban, F. de Longueville, J.C. Fuscoe, W. Tong, L. Shi, R.D. Wolfinger, Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project, Nat. Biotechnol. 24 (2006) 1140–1150.

# 3　iPS細胞からのビッグデータの情報セキュリティと創薬，医療への活用

藤渕　航*

## 3.1　はじめに

2006年にマウス，2007年にヒトのiPS細胞の作製が発表されて以来，目覚ましい勢いでiPS細胞やiPS細胞由来の細胞の医療や創薬への応用への準備が進んでいる。iPS細胞は正に夢の生体ツールとして大きな期待が膨らんでおり，将来には総国民iPS細胞ストック時代を迎える可能性も秘めている。本節執筆の現時点では，コスト的にも技術的にも難しく総国民iPS細胞ストックには否定的な見解も多いが，既に2013年度よりJSTから「再生医療実現拠点ネットワークプログラム」が開始され，その中核拠点に選出された京都大学iPS細胞研究所で，10年の歳月をかけて日本人の９割をカバーする推定約140名の組織適合性遺伝子HLAがA，B，DR三座ホモドナーのiPS細胞化ストック構想が始まった。少なくとも本構想は，未来においてiPS細胞を医療の中核に据えるための基本科学技術を発展させることは間違いなく，10年間で質の高いiPS細胞を低コストで作製できる時代を到来させるであろう。思い起こすことは，米国NIHで2003年に1,000ドルでヒトゲノムを解読する構想のロードマップが提示されてから，僅か10年以内の2012年に1,000ドルゲノムを達成したと報道が出るまでになった。これは当時，ヒトゲノム１人分の解読に100万ドル以上かかっていた事実からすれば，コストが1,000分の１になるとは誰も想像がつかなかったことである。このような経験を踏まえて本節では，現在から10年後の未来にわたってiPS細胞から生じるいわゆる「ビッグデータ」について，現状を踏まえた上でできる限り正確なこれからの動向について予測を含め，記載したいと思う。

## 3.2　iPS細胞がもたらすビッグデータ

### 3.2.1　iPS細胞の品質管理

１個人の体細胞からiPS細胞を作るのに，熟練した技術者なら数ヶ月程度しかかからない。簡易な実験用iPS細胞であればこのようなiPS細胞でも問題ないが，医療や創薬用となると厳重な品質管理が必要とされ，この期間は最低でも６ヶ月はかかると考えられる。例えば，表１に示したように，iPS細胞の特性は，ドナー自身が持つ内在的な性質であるbiological characteristicsとiPS細胞の作製過程で生じる様々な技術的な性質であるtechnical characteristicsに大別される。biological characteristicsの方は，性別や血液型や年齢などの基本情報の他に，このドナー由来のiPS細胞は免疫寛容性が高いか，また，遺伝的疾患を持っていないか，AIDSウイルスなど後天的要因の病気に罹患していないか，など他者の再生医療に用いることが可能かを判定する基準となる。一方，technical characteristicsの方はiPS細胞を生成した時の完成度であり，幹細胞マーカーの発現，細胞分裂能力，分化能力などの他に，作製過程で用いた外来遺伝子の残存度など作製方法にも依存するものである。

---

＊　Wataru Fujibuchi　京都大学　iPS細胞研究所　増殖分化機構研究部門　教授

表1　iPS細胞の品質評価で必要とされる情報例（文献1）より改修）

| 特徴分類 | カテゴリー | 例 |
|---|---|---|
| biological characteristics | 基礎情報 | 性別，血液型，年齢，人種など |
| | 免疫情報 | HLAタイプなど |
| | 先天的疾患 | 既往症，遺伝病など |
| | 後天的疾患 | HIV，A/C型肝炎，HTLVなど |
| | 由来組織 | 皮膚，脂肪，歯髄，血液など |
| technical characteristics | 細胞形態 | コロニー形状，数，密度など |
| | 幹細胞マーカー | Nanog，Oct3/4など |
| | 増殖能力 | 分裂速度，分裂形式など |
| | 分化能力 | 外，内，中胚葉層形成など |
| | 外来遺伝子残存度 | 導入遺伝子検査 |
| | 体細胞突然変異 | SNPs，CNVsなど |
| | 細胞同一性 | STR解析など |
| | 微生物汚染 | マイコプラズマ検査など |
| | ゲノム安定性 | 核型解析など |
| | エピゲノム | アレイ，シーケンサー解析など |

### 3.2.2　遺伝的リスク

特に，今後は，biological characteristicsであるSNPs情報からわかる遺伝的なリスクが重用視されると考えられる。既に世界的に有名なGWAS研究などによるSNPsとフェノタイプを結ぶdbGaPなど情報データベースが蓄積されている。近年，SNPsは親由来だけでなく，卵細胞から成体に至るまでに高度に蓄積された体細胞突然変異が原因となっていることもわかっており，これがさらにiPS細胞作製過程でも生じる変異と合わせて最終的に生じたiPS細胞での遺伝的リスクを調べる必要がある。表2に主なSNPsやCNVsなどゲノム変異に関わる情報データベースを記した。

現在，このようなゲノム上でわかっている疾患関連サイトについて，潜在的な遺伝的リスクを調べる高性能なツールの開発も世界中で進められていると考えられる。例えば，dbSNPなどのデータベースを用いてSNPsを検索してアノテーションするSNPdat，SNPnexus，Snap，SNP Function Portal，SNPper，Fans，FunctSNP，Annovarなどがある。一方，データベースにないrare variantsの場合には既知の疾病パスウェイや遺伝子発現やデータなどとassociateさせるVAASTやBioBinなどの様々なソフトウェアが使用されている。

## 3.3　ゲノム情報産業と我が国での個人情報保護

### 3.3.1　ゲノム情報を活用した産業

個人から得られるゲノム配列には膨大な情報が含まれている。これを利用して，個人の疾病リスクの予想のみならず，性格や適正などまで検査する産業が発展してきている。既に欧米では，

表2　代表的なゲノム変異関連のデータベース例

| 名称 | 開発機関 | 主な特徴 | アドレス |
|---|---|---|---|
| dbSNP | 米国National Center for Biotechnology Information/NIH | 1塩基置換および短い欠失・挿入，レトロポゾン，STR多型のデータレポジトリ | http://www.ncbi.nlm.nih.gov/snp |
| dbVar | 米国National Center for Biotechnology Information/NIH | 1kb以上の逆位，転座，欠失，挿入などゲノム構造変異のデータレポジトリ | http://www.ncbi.nlm.nih.gov/dbvar |
| dbGaP | 米国National Center for Biotechnology Information/NIH | 医療以外のフェノタイプを含むジェノタイプとの関連データベース／アクセス制限データ有り | http://www.ncbi.nlm.nih.gov/gap |
| ClinVar | 米国National Center for Biotechnology Information/NIH | 疾病に関係する塩基やアミノ酸変異のデータベース | http://www.ncbi.nlm.nih.gov/clinvar |
| DGV | カナダThe Centre for Applied Genomics | 健康人における50b以上の逆位，転座，欠失，挿入などゲノム構造変異のデータレポジトリ | http://dgv.tcag.ca/dgv/app/home |
| GWAScentral/HGVBase | スウェーデンKarolinska Institute，英国European Bioinformatics Institute，ドイツEuropean Molecular Biology Laboratory | ヒトGWAS研究のレポジトリ | http://www.gwascentral.org |
| PharmGKB | 米国Stanford University | ゲノム変異と薬物反応に関するデータベース | http://www.pharmgkb.org |
| HGMD | 英国Cardiff University | 遺伝子欠失と遺伝病のデータベース | http://www.hgmd.cf.ac.uk |
| JSNP | 東京大学医科学研究所，JST | 日本人のSNPsデータベース | http://snp.ims.u-tokyo.ac.jp |
| DECIPHER | 英国ウェルカム・トラスト　サンガー研究所 | Ensemblを利用したヒト染色体不均衡とフェノタイプデータベース | http://decipher.sanger.ac.uk |

数年前から民間で有料の遺伝子検査サービスが始まっており，脳梗塞，心筋梗塞，糖尿病，乳がんリスク，などのDNA多型情報に基づく情報を蓄積している。2012年3月には，米国立衛生研究所（NIH）から，企業などが提供した遺伝子検査情報を提供する遺伝子検査レジストリ（GTR）というデータベースが公開された（図1）。このデータベースの目的は，乱立する遺伝子検査企業に透明性を持たせ，どこでどのような検査を行っているか，また，遺伝子データを研究者が共有する仕組みを提供して科学的研究を促進させようとするものである。

一方，これに伴う個人情報の保護が問題となってきている。欧米では個人情報保護は随分と進んでおり，1980年のOECD（経済協力開発機構）によるプライバシーガイドラインを受けて，早くから強固なデータ保護やプライバシー保護の法律ができ，DPA(欧州Data Protection Authority)やFTC（米連邦取引委員会）などの公的組織が情報保護の取り締まりを行っている。また，両者の間では2001年のSafe Harbor合意に基づく国際間での情報保護の取り扱いまで法的に整備されている。2013年6月に米国立衛生研究所（NIH）のNCBIを訪問し，実際にどのようにしてプラ

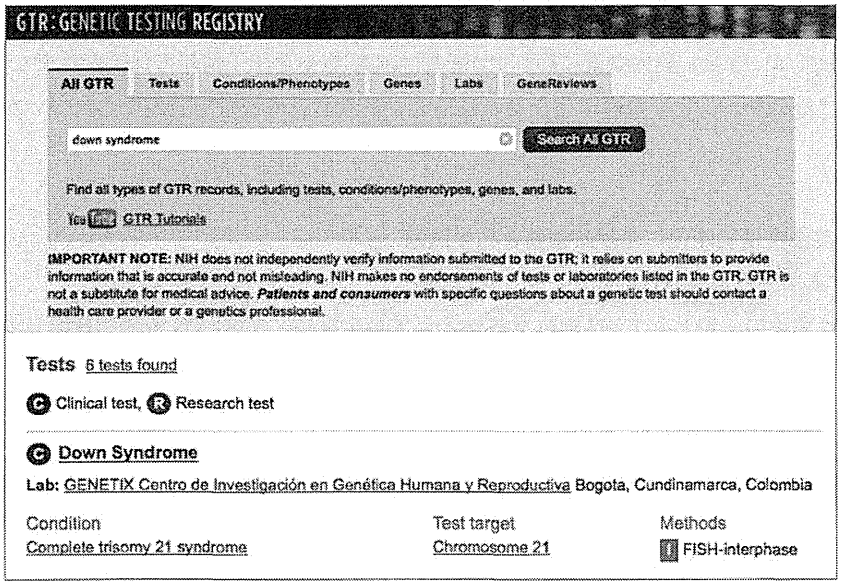



図１　遺伝子検査レジストリ（GTR）
http://www.ncbi.nlm.nih.gov/gtr/

イバシーと健康情報提供を両立しているのか調査した。その結果，①個人を特定できるDNA情報についてはインフォームドコンセントなどを含む規定の整備，②情報漏洩を想定したデータの暗号化や自動削除技術開発，③ユーザの階層的利用権限化などを行って保護する方法の開発を進行中であるとの情報を得ている。

### 3.3.2　日本における個人情報保護

我が国は個人情報保護においては欧米に比べてやや後進国である。日本では2003年にようやく個人情報保護法に基づく規制が制定されたばかりで，急増する個人の遺伝情報の取り扱いについて国内の法整備は遅れている。また，日本人から得られた遺伝情報を国外へ持ち出すことについて何ら法的規制がないのが現状である。本節の執筆中に「特定秘密保護法」が国会で可決されたが，今後，これがどのように日本人の遺伝情報の漏洩に影響するのかはまだ予想がつかない段階である。我が国でも国内向けに遺伝子検査の企業も増加している中，国立バイオサイエンスデータベースセンター（NBDC）では，2013年４月に「ヒトデータ共有ガイドライン」および「ヒトデータ取り扱いセキュリティガイドライン」の策定を始めている。また，遺伝子から得られる個人情報保護の規定や法令については，各省庁で様々に検討が開始されつつある。どのように統廃合され，国全体として整備されるのかにはまだ少し時間がかかるであろう。

さらに，バイオ関連情報漏洩をさせないためのセキュリティ技術としても，化合物の秘密検索（産総研＆東大 2011年），臨床データの秘密計算（日本成人白血病治療共同研究グループ&NTT 2012年），クラウド上遺伝子情報データでの秘密検索（日立ソリューションズ 2012年）などと，

いずれも計算機上のデータを暗号化したまま計算が可能な技術が報告されている。しかしながら全体像としては，今後，莫大な量に膨れ上がる個人の細胞や遺伝子情報の保護だけでなく，転送，利用におけるセキュリティの標準化などの包括的な取り組みや技術開発がなされていないままである。

### 3.4　高度医療情報時代における創薬と再生医療

#### 3.4.1　高度医療情報時代の到来

我が国は国策の一環としてiPS細胞を基軸として再生医療・医療情報大国を目指している。この過程においてこれまで見過ごされてきた個人情報の問題が浮き彫りになってくるのは必定である。今後は，全国の病院・研究所などからiPS細胞情報データベースにアクセス可能であるが，同時に高いセキュリティを維持するシステムが必要である。その上で，国民がiPS細胞を個人で作製する上で不安となるプライバシーの問題を解決しつつ医療上重要な遺伝情報を提供するという，真に医療情報大国として発展するために必要なインフラを構築する必要がある。具体的には，国内の病院やデータセンターなどの拠点間でiPS細胞の異なる情報を相互交換するため，セキュアなデータ検索・表示とデータのダウンロードアップロードのための計算機技術が必要である。基本技術として，一部先述したが，データの暗号化や第三者に情報が漏洩した場合の自動削除，ユーザの階層的利用権限化，分散データなどが必要な開発項目と想定される。

#### 3.4.2　iPS細胞の創薬・毒性評価からの情報

これまでマウス，ラットなどで行われていた創薬や毒性評価の研究が急激な勢いでヒトiPS細胞を用いたシステムに置き換わりつつある。これは，サリドマイドなどに代表されるようにマウスで無毒性が保障された化合物でもヒトでは毒性を示すなどヒト特有の効果や副作用を示すことが明らかになったからである。例えば，ヒトES細胞から神経細胞を誘導する過程でメチル水銀を投与すると，ヒトでは神経樹状突起が縮退するが，マウスでは縮退せず，替わりに細胞死が観察された[2]。このように，より実態を反映したヒトiPS細胞システムが今後は創薬や毒性評価の主力となると考えられる。また，米国NIHのNCATS（National Center for Advancing Translational Sciences）では，iPS細胞チップの開発も行っており，これまで臨床試験でドロップアウトした化合物を再利用できないか研究する計画もある。今後はこのような化合物の投与データも多く蓄積するが，そこから有益な情報を取り出して活用することも必要となる。例えば，我々は少ない毒性のデータを学習させ，新規化合物の毒性を予測するための手法を開発し報告した[3]。そこでは，遺伝子発現情報から遺伝子ネットワークを構築し，これを従来の遺伝子発現データだけのサポートベクターマシン予測と比較するもので，遺伝子ネットワークを学習に加えると予測精度が向上する結果が得られている（図2）。

#### 3.4.3　再生医療情報のデータマイニング

今後，iPS細胞を含め，蓄積する細胞ビッグデータから有益な情報を引き出すためのツール開発を我々の研究室では進めてきた。例えば，高速類似細胞検索CellMontage[4]，網羅的遺伝子モジュ

ール探索システムSAMURAI[5]などがある（図３）。CellMontageでは，手持ちの細胞や人工作製した細胞がどのタイプの細胞に近いかを瞬時に検索することが可能である。現在，開発されている計算システムでは１秒間に数千件の細胞データを検索することができる[6]。例では膵臓の細胞

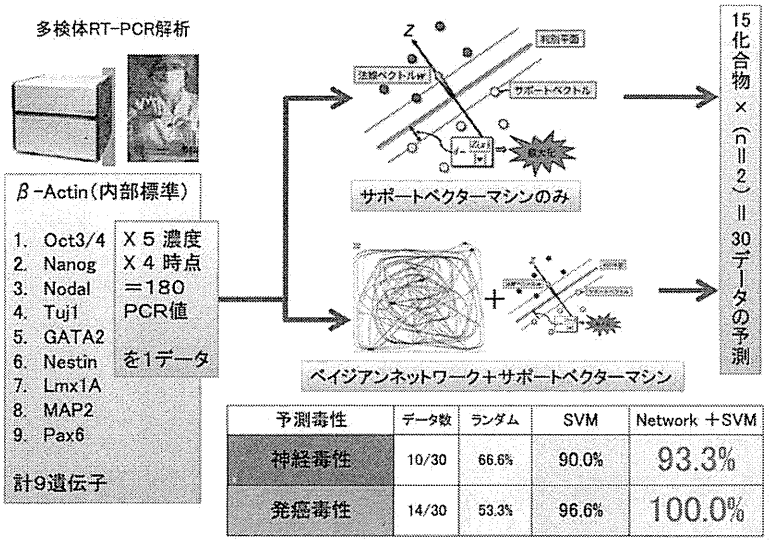
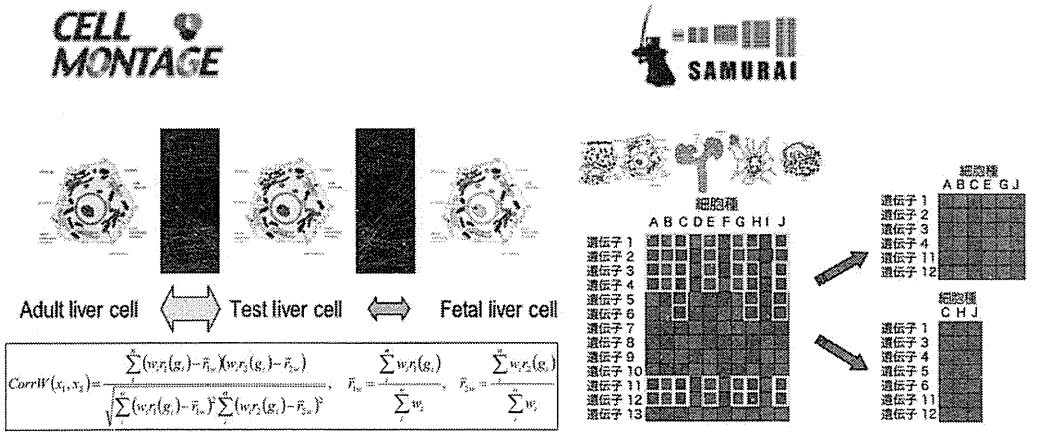



図２　幹細胞を用いた毒性化合物予測
（データ提供：平成21～23年度厚生労働科研費「化学物質リスク事業：大迫班」による）




重み付き順位相関係数に基づく高速検索アルゴリズムRaPiDS

Horton *et al.*, *Genome Informatics* (2006)
Fujibuchi *et al.*, *Bioinformatics* (2007)

飽和アイテム集合マイニングとノイズを許したモジュール合成

Okada and Fujibuchi, *CAMDA* (2007)
Fujibuchi *et al.*, *Methods Mol. Biol.* (2009)

図３　高速類似細胞検索CellMontage（左）と網羅的遺伝子モジュール探索システムSAMURAI（右）

Query >NR7f5 IVALUEIIAFLPIsingleIIAFLPIHomo sapiensladult endocrine/exocrine pancreas pancreas NR7f5 . (total genes)
Platform: GPL999999(129 entries), sampling genes: 32512(GO-bound all), Probability: 0.01, Correlation: 0.0
Found: 100 entries. Start: Tue Sep 26 18:22:46 2006 End: Tue Sep 26 18:22:47 2006

| Top100 | Sample | DataSet | Platform | Type | Channel | Organism | Description | Probability(Correlation,#Genes) | Data |
|---|---|---|---|---|---|---|---|---|---|
| 1 | IVALUEIIAFLPIsingleIIAFLPIHomo sapiensladult endocrine/exocrine pancreas pancreas NR7f5 | iAFLP | iAFLP | single VALUE | Homo sapiens | adult endocrine/exocrine pancreas pancreas NR7f5 | 0.000e+00(1.00,32512) | |
| 2 | IVALUEIIAFLPIsingleIIAFLPIHomo sapiensladult endocrine/exocrine pancreas pancreas(caudal) NR8d5 | iAFLP | iAFLP | single VALUE | Homo sapiens | adult endocrine/exocrine pancreas pancreas(caudal) NR8d5 | 2.413e-820(0.33,32512) | |
| 3 | IVALUEIIAFLPIsingleIIAFLPIHomo sapiensladult endocrine/exocrine pancreas pancreas(caput) NR8c5 | iAFLP | iAFLP | single VALUE | Homo sapiens | adult endocrine/exocrine pancreas pancreas(caput) NR8c5 | 2.085e-610(0.29,32512) | |
| 4 | IVALUEIIAFLPIsingleIIAFLPIHomo sapiensladult endocrine/exocrine pancreas pancreas(caput) NR8b5 | iAFLP | iAFLP | single VALUE | Homo sapiens | adult endocrine/exocrine pancreas pancreas(caput) NR8b5 | 7.868e-583(0.28,32512) | |
| 5 | IVALUEIIAFLPIsingleIIAFLPIHomo sapiensladult stomach/colon stomach stomach NR7f4 | iAFLP | iAFLP | single VALUE | Homo sapiens | adult stomach/colon stomach stomach NR7f4 | 4.721e-439(0.24,32512) | |
| 6 | IVALUEIIAFLPIsingleIIAFLPIHomo sapiensladult endocrine/exocrine pancreas pancreas(caudal) NR8f5 | iAFLP | iAFLP | single VALUE | Homo sapiens | adult endocrine/exocrine pancreas pancreas(caudal) NR8f5 | 1.410e-438(0.24,32512) | |
| 7 | IVALUEIIAFLPIsingleIIAFLPIHomo sapiensladult endocrine/exocrine pancreas pancreas(corpus) NR8e5 | iAFLP | iAFLP | single VALUE | Homo sapiens | adult endocrine/exocrine pancreas pancreas(corpus) NR8e5 | 3.498e-300(0.20,32512) | |
| 8 | IVALUEIIAFLPIsingleIIAFLPIHomo sapiensladult stomach/colon small_intestine duodenum NR7d5 | iAFLP | iAFLP | single VALUE | Homo sapiens | adult stomach/colon small_intestine duodenum NR7d5 | 4.833e-266(0.19,32512) | |
| 9 | IVALUEIIAFLPIsingleIIAFLPIHomo sapiensladult liver liver liver NR7c3 | iAFLP | iAFLP | single VALUE | Homo sapiens | adult liver liver liver NR7c3 | 5.979e-204(0.17,32512) | |

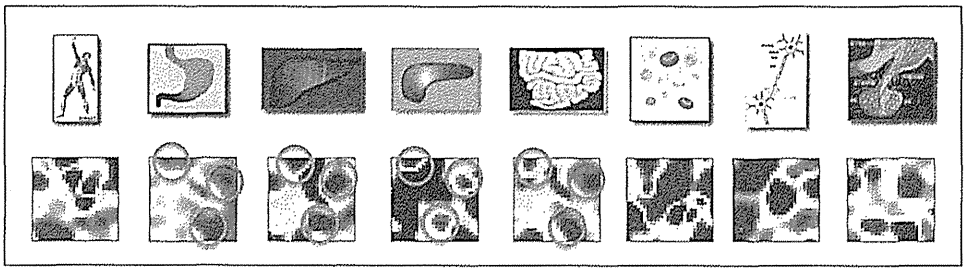


図4　CellMontageで膵臓の細胞をクエリーとして検索

をクエリーとして検索しているが，膵臓に近い細胞は胃や肝臓などである（図4）。これらの臓器は全て内胚葉性であり，発生上，近い関係にある。また，機械学習の手法と組み合わせることによって人工作製したiPS細胞の質も検索可能である。例えば，iPS細胞がどの細胞に由来していたかについての試験的な結果では，10種の由来細胞を含む73マイクロアレイデータ検索の順番間違いがわずか6％と大変に有効な結果が得られている。また，iPS細胞コロニーの写真画像から良質なiPS細胞を推定する研究では，17日目以降の初期化をほぼ完了したiPS細胞においては100％の正確さで判定できた。

また，SAMURAIシステムでは，買い物をする時のassociation ruleという考えから生まれたLCM：Linear Time Closed Itemset Minerと呼ばれるアルゴリズムを用いて，大量の細胞データに共通する遺伝子の組み合わせ（飽和アイテム集合）を網羅的に取り出すものである（図5）。現在，2,912件の多様なヒト細胞マイクロアレイデータから遺伝子モジュールの辞書を作成する研究が当ラボで進行中である。これから将来においても，益々，大量の細胞関連データからのデータマイニングツールのアイデアが必要となるであろう。

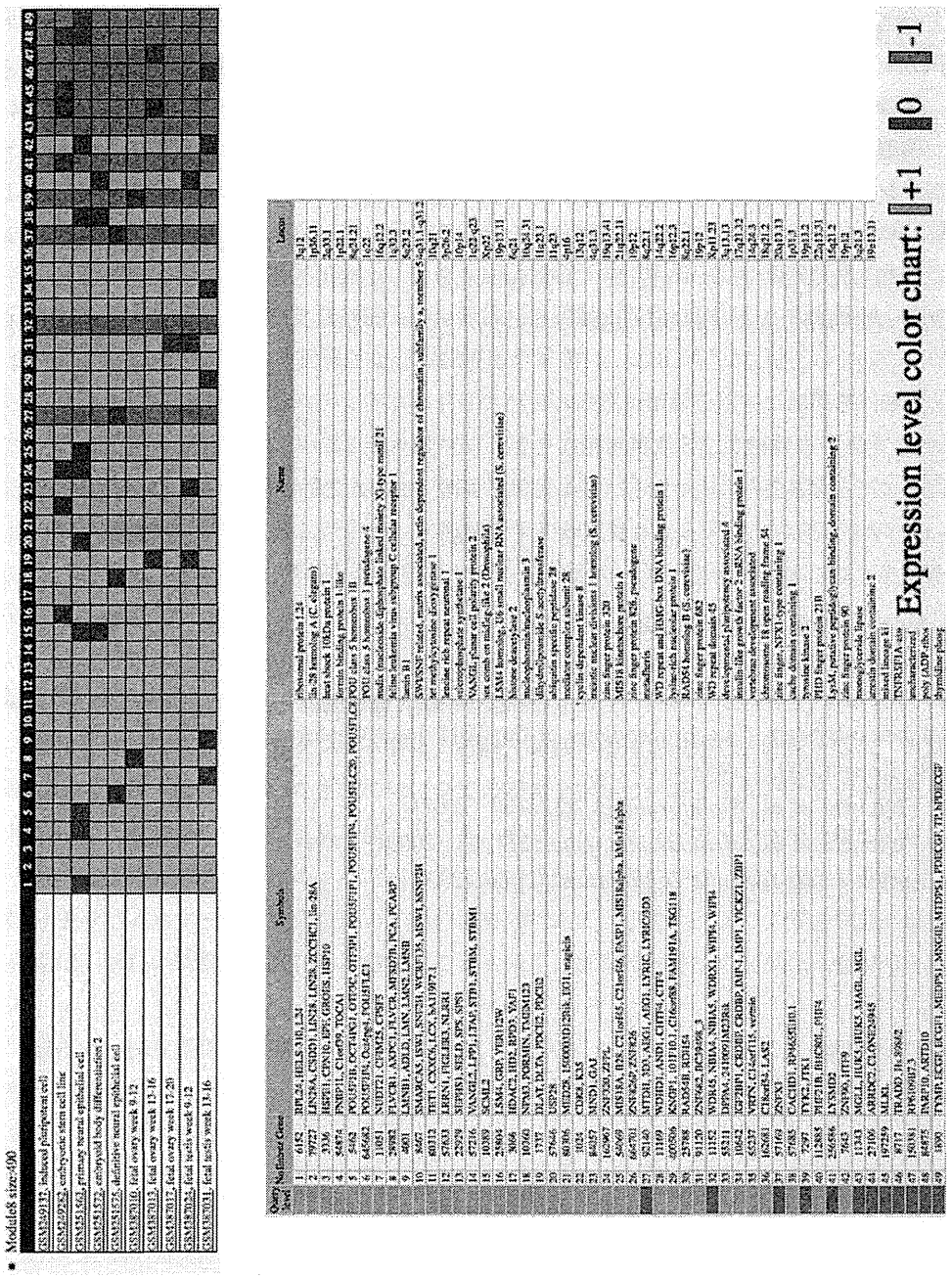



図５　iPS細胞，ES細胞，胚葉体，神経表皮，胎児の生殖細胞に共通な遺伝子モジュールの例
POU5F1，LIN28などの49遺伝子が抽出されている。

### 3.5　今後必要とされる解析技術について

今後はiPS細胞の情報だけでなく，ヒト細胞全てについての最新情報を全国の医療関連機関を結んで相互に共有できることが，未来の細胞医療には必要となってくると考えられる。その時に

大量のデータを首尾よく保存し，検索可能にするには，従来のようなデータベーススキーマに基づく関係データベースのような変化に弱いシステムでは追い付かないであろう。今後必要とされるのは，乱雑な大量のデータを記憶装置上にただ単に置いただけで，計算機が自動的にデータを整理し，関連付けを行い，検索やデータマイニング，機械学習を行うことを可能にする，より一般性の高いシステムを開発することが必要である。特に関連する報告として特筆したいのは，IBMのWatsonシステムが百科事典を機械学習し，「Jeopardy!」という米国の人気クイズ番組で人間の優勝者に勝利したことである。今後は，このWatsonをさらに進化させて，自動整理，自動学習するシステムを開発することがiPS細胞からのビッグデータマイニングには必要であるかも知れない。

## 文　　献


1) G. Stacey, *Prog. Brain Res.*, **200**, 41 (2012)
2) X. He *et al.*, *Toxicol. Lett.*, **212**, 1 (2012)
3) W. Fujibuchi *et al.*, Prediction of Chemical Toxicity by Network-based SVM on ES-cell Validation System, The Proceedings of the 2011 Joint Conference of CBI-Society and JSBi, Kobe (2011)
4) W. Fujibuchi *et al.*, *Bioinformatics*, **23**, 3103 (2007)
5) W. Fujibuchi *et al.*, *Methods Mol. Biol.*, **577**, 55 (2009)
6) P. Horton *et al.*, *Genome Informatics*, **17**, 67 (2006)

第 7 章　ビッグデータの活用事例と今後の狙いどころ

第 2 節　新規事業，研究開発テーマ創出への活用

# ［3］ヒト細胞からのビッグデータの情報管理と情報解析技術

加藤 有己　　京都大学　iPS 細胞研究所　特定拠点助教

桜井 都衣　　京都大学　iPS 細胞研究所　研究員

藤渕 航　　京都大学　iPS 細胞研究所　教授

(株)技術情報協会
「ビッグデータの収集、調査、分析と活用事例」

# [3] ヒト細胞からのビッグデータの情報管理と情報解析技術

京都大学　加藤有己，桜井都衣，藤渕航

はじめに

ライフサイエンスの分野では，近年の目覚ましいシークエンシング技術の進展により，膨大な個数の配列データを代表とするビッグデータが産生されている。ビッグデータが与えられたとき，大量の個々の事象を解析できるだけでなく，データ全体を包括的に取り扱うことが可能となる。ビッグデータを有効利用するためには，それが誰にでも使える状態として保持されている必要があり，データベースが果たす役割は非常に大きい。また，ビッグデータから有用な情報を取り出すために，コンピューターを用いた情報解析技術が必要不可欠である。本節では，ヒト細胞からのビッグデータに焦点を当て，主として研究が進展している遺伝子発現データと，それを利用した細胞解析について言及する。

## 1. ライフサイエンス研究分野におけるビッグデータ産出の現状

1985 年にアメリカのエネルギー省によって開始されたヒトゲノムプロジェクトは，多くの国際的な研究機関の協力の下行われ，2003 年にアメリカのホワイトハウスにてヒトゲノムの全塩基配列解読完了の報告がなされた。このときに用いられた技術がシークエンシング技術であり，現在では各種のゲノムプロジェクトによって，さらに多くのモデル生物種のゲノム配列解読が完了，もしくは進行中である。ゲノム配列データはアメリカ，ヨーロッパ，そして日本を拠点として運営されている国際DNAバンクに登録されている。2012年の時点でそのデータサイズは約540 GBにまで上り，さらに指数的な増加を続けている。また，2000 年以降はゲノムプロジェクトの成果として生命システムの構成要素が明らかになってきたことで，それまでに行われていた断片的な材料を用いて現象の一部のみを観察する方法よりも，生命システム全体を包括的に解析する方法（オミクス解析）の有用性が重視された研究が多くなったのではないだろうか。そこで以下では，現在のライフサイエンス分野研究におけるビッグデータの産出技術の一例と，それらを取り巻く状況について述べる。

### 1.1 現代のライフサイエンス研究とオミクス解析技術

ゲノム配列情報の取得は生命現象の研究において非常に重要なものであるが，生物の複雑かつ精密な機能制御の理解のためには，ゲノム配列解読だけでは不十分である。現在までの研究で体内位置情報を区別して計算すると約 2,600 種類あるとされているヒト細胞[1]のうち，細胞 1 種類の機能解明のためにも，ゲノム配列情報に加えて以下のような詳細な階層，構成成分の網羅的な解析が必要になる。

・分子単体レベルでの情報（遺伝子・タンパク質発現情報：配列・構造・時空間的制御）
・分子ネットワークレベルでの機能情報（分子間相互作用，パスウェイ，リアクトーム）
・細胞レベルの機能情報（代謝制御，フェノタイプ発現）
・組織・器官の一部としての機能

これらのデータを効率よく取得するため，現在ではシークエンシング技術だけではなく，オミクス研究のための様々な技術の開発がなされている。また，検出やデータ解析技術のハイテク化，アメリカの国家プロジェクト「1000 ドルゲノム計画」に代表されるように，ハイスループットなゲノムシークエンシング技術が安価になっており、ビッグデータを産出する研究が以前よりも盛んに行われる環境になっている。表 1 において，細胞の機能解析に用いられる手法の代表的なものを示す。これらを組み合わせることで，研究対象となる細胞固有のシステムや生体におけるその機能の網羅的な解析が可能となる。表 1 の中でも，近年は特に，RNA シークエンシング（RNA-seq）やバイサルファイトシークエンシング等の技術を用いて，個々の細胞の持つ疾病関連状況や，iPS 細胞，幹細胞の分化の程度を網羅的に調べる研究が盛んに行われている。

表 1　細胞機能解析に用いられる手法の一例

| 実験デザイン | 解析技術の一例 |
|---|---|
| ゲノム解析 | ゲノムシークエンシング |
| エピゲノム解析 | マイクロアレイ，ゲノムシークエンシング，ChIP シークエンシング |
| トランスクリプトーム解析 | マイクロアレイ，RNA シークエンシング |
| プロテオーム解析 | タンパク質マイクロアレイ，2D-MS，LC-MS |
| インタラクトーム解析 | LC-MS，酵母 two hybrid 法，FRET/BRET |
| メタボローム解析 | イメージング質量分析法 |
| フェノーム解析 | イメージング |

## 1.2　ライフサイエンス研究ビッグデータと情報統合技術

ハイスループットな解析技術は，以前まで特定の研究機関や創薬研究開発分野など，限られた施設でのみ行うことが可能であったが，現在はそれらの技術の確立やコスト面での問題が解決されつつあるため，様々な研究で用いられるようになった。そのため，ライフサイエンス研究分野全体のアウトプットとしてビッグデータの産出が一般的になりつつある。また，データ量だけでなく，例えばトランスクリプトーム解析のように RNA-seq とマイクロアレイなど，データを得るために用いる実験，計測技術が異なる場合もあるため，これらの実験デザインの情報も保持した上で，ビッグデータを効率的に取り扱うツールをどう設計するかが現在の大きな課題である。ライフサイエンス研究分野におけるデータベースシステムはまさに，あるカテゴリーの実験系で得られるビッグデータの収集，管理，情報の検索，そしてその利用という作業を容易に実現させるためのツールであり、そのビッグデータ処理における重要性は日に日に高まっている。例えば，遺伝子発現解析データのための公共データリポジトリ・データベースとして，NCBI の Gene Expression Omnibus、EBI の ArrayExpress，国立遺伝学研究所の Omics Archive があり，これらの登録実験データ数は合計で 140 万件以上にものぼる（2014 年 5 月現在）。これら以外にも，遺伝子発現解析研究に関する大量のデータを取り扱うものに限定しても，世界中で数多くのデータベースが開発されている（表 2 参照）。

表 2　ヒト細胞における遺伝子発現データを取り扱うデータベースの一例

| 名称 | 公開状況 | 開発機関 | URL |
|---|---|---|---|
| Gene Expression Omnibus | Public | NIH | http://www.ncbi.nlm.nih.gov/geo/ |
| ArrayTrack | Public and private | FDA | http://www.fda.gov/ScienceResearch/BioinformaticsTools/Arraytrack/ |
| ArrayExpress | Public | EMBL-EBI | https://www.ebi.ac.uk/arrayexpress/ |
| Genenvestigator | Private | NEBION | https://www.genevestigator.com/gv/ |
| Stanford Microarray Database | Public and private | Stanford University School of Medicine | http://smd.princeton.edu/index.shtml |
| GeneNetwork system | Public | University of Tennessee | http://www.genenetwork.org/webqtl/main.py |
| GeneCards | Public | Weizmann Institute of Science | http://www.genecards.org |
| DDBJ Omics Archive | Public | 国立遺伝学研究所 | http://trace.ddbj.nig.ac.jp/dor/index.html |
| RefExA | Public | 東京大学先端科学技術研究センター | http://157.82.78.238/refexa/main_search.jsp |

しかしながら，多くのデータベースは異なる組織によって運営されているため，登録されているプロジェクトデータの重複の有無がある。また，データベースごとに異なる ID やフォーマットを用いてデータを登録するという方式のため，登録項目や使用されている語彙が統一されていないという場合が多い。このようなことから，異なるデータベース間での内容の比較やデータ解析が困難という問題も生じている。この問題を解決し，データの共有や再利用を促進し付加的な情報を引き出すために，上記 3 つの遺伝子発現解析データベースを始めとした様々なデータベースの統合プロジェクトが，既に国家規模で立ち上げられ取り組まれている。さらに，データベースの統合化の必要性が増すにつれて，実験データ報告における必要最小限の情報項目チェックリスト（The minimum information standards, MI standards）[2] や制御さ

れた語彙を用いるオントロジー分野も注目されるようになってきた。

MI standards は，実験の全体像の形式的な記述や報告を可能にする，また実験条件を形式的に管理し実験データの再現性を高める必要最小限の情報を統一する，という目的で提案された報告用チェックリストである。最初の MI standards の例としては、マイクロアレイデータの報告のために作成された Minimum Information about a Microarray Experiment（MIAME）チェックリスト[3]があり，異なる研究機関で行われた実験データやメタデータを統合する際のテンプレートとして利用されている。また，MIAME を始めとした MI standards に基づいた形式でのデータ，メタデータの投稿を推奨する学術系雑誌もある[4]。2008 年には生命医科学や生物学の分野で必要最小限の情報を統一する Minimum Information about a Biomedical or Biological Investigation（MIBBI）プロジェクト（http://www.biosharing.org/standards/mibbi）が国際協力の下発足され，現在までに様々な MI standards が作成されている。MI standards では実験データに付随するメタデータの報告項目を対象となる研究分野や実験技術ごとに定めており，特にビッグデータを産出するハイスループット実験技術を用いて行う研究の情報管理に有用であるとされている。さらに，MI standards はある特定の研究分野や手法でデータを得るための重要な要素を項目として含むので、データベースのフレームワーク構築にも有用であると考えられている。

オントロジーは，ある知識領域における様々な概念を構成する語彙の体系的な分類や，それらの語彙の共有を可能とする。例えば，以前まで新たに発見された遺伝子は，シークエンシングの結果で得られた配列や構造の類似性から命名または分類されることが多かった。しかし，ゲノムプロジェクト以降，研究開発技術の発達に伴い膨大な数の遺伝子配列およびその機能が明らかになり，その蓄積された知識に基づいて遺伝子を分類し直すために構築されたものが，よく知られている Gene Ontology（GO）である[5]。すなわち，GO は様々な生物種の遺伝子機能に関する知識やその記述に必要な共通語彙（GO term）を統一された形式で分類し有している，遺伝子機能の辞書のようなものである。さらに、現在のライフサイエンス分野のデータ解析ソフトウェアやデータベースには，共通語彙を用いることで異なるツール間の知識表現の統一化や比較化を容易にするという目的で，解析や検索のアウトプットに GO term を使用するものも多い。このことから，オントロジーの有用性に期待し，近年では他のライフサイエンス研究分野のオントロジーの作成が進められている。

ビッグデータは活用されてこそ価値がある。今後は次々と新しいデータを生み出していくだけではなく，既にあるビッグデータを生物学的な意味のある知識として蓄積し，応用研究や新たな基礎研究のアイディア取得のためのリソースとしてリサイクルしていくことが重要である。そのためには，情報や知識が統一された形式で表示され、誰もが共有できるシステムが必要である。現在は、上記の MI standards やオントロジーを組み合わせた MAGE-TAB（http://www.mged.org/mage-tab/）や ISA-TAB（http://isatab.sourceforge.net/index.html）のようなアッセイアノテーションフォーマットも作成されている。このような環境を整備し利用することができれば，多くの研究グループにより膨大な時間とコストをかけて産出されたビッグデータおよび大量の知識が様々な研究グループにより再度活用されることが可能になり，ライフサイエンス研究の促進的な発展が期待できる。

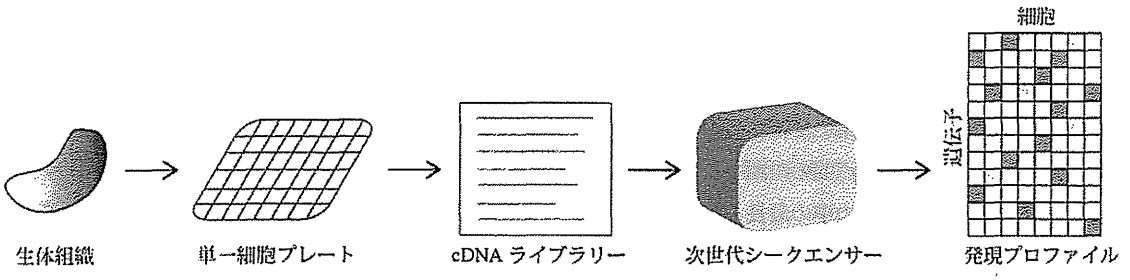
## 2. 細胞の自動分類技術

前述のように，ヒト細胞からはトランスクリプトームなどの大量のデータが得られる。これらのビッグデータから意味のある情報を抽出し，細胞から構成される生命の謎に迫るためには，情報科学，統計科学などに裏打ちされた方法論およびコンピューターの能力を駆使して網羅的な解析を行うことが重要である。以下では，近年注目を浴び始めている単一細胞解析を用いた細胞分類問題に焦点を当て，用いられている実験技術および情報技術の具体例を紹介し，今後の展開について述べる。

### 2.1 RNA-seq を用いた単一細胞解析

近年では細胞解析を行う際，細胞の集団レベルで遺伝子発現量を考察するよりも，より解像度の高い単一細胞レベルで遺伝子発現量解析を行うことが有効であると考えられている。生体の組織から単一細胞レベルでトランスクリプトー

ム解析を行う手法として，RNA-seq[6] を利用するアプローチが知られている（図 1 参照）。まず，対象とする生体の組織から，FACS（fluorescence-activated cell sorting）などで知られる細胞分取法を用い，組織の構成要素である単一細胞に分離する。次に，分離された各細胞から発現している RNA を RNA-seq により配列解析するが，複数の単一細胞からの RNA を同時に大量解析するため，どの細胞から得られた RNA なのかを記憶するために，核酸配列にバーコードを埋め込むなどの工夫が必要である[7]。RNA-seq で転写配列断片の定量的情報により得られた遺伝子発現の全体像を発現プロファイルと呼んでいる。

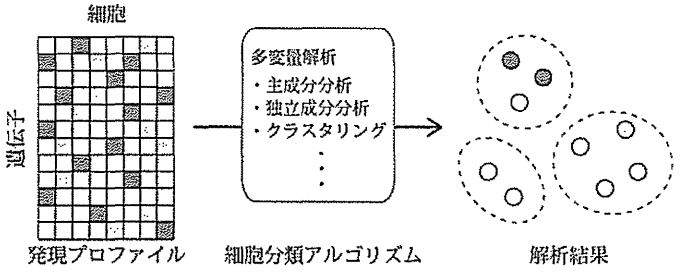



図の cDNA ライブラリーの各横線は各細胞内で発現している RNA から作成した cDNA を表す。遺伝子発現プロファイルでは，行と列がそれぞれ遺伝子と細胞を表し，各マスの色が濃く塗りつぶされているほど発現量が多いことを示す。

図 1　単一細胞 RNA-seq 解析

## 2.2　発現プロファイルからの細胞の自動分類

発現プロファイルにおいて，各細胞は遺伝子の個数だけ次元を持つ実数値ベクトル（多変量）とみなすことができる。例えばヒトの場合，約 22,000 個の遺伝子があると推定されているため，その全てを考慮した場合の発現空間における各細胞は 22,000 次元のデータとなる。したがって，細胞情報解析を行うに当たって，高次元データの集合からデータ間の隠れた構造の導出をめざす多変量解析が有効であると考えられる（図 2 参照）。ただし，細胞（入力）に対する出力値（教師出力）が与えられていないことが多いため，入力群から必要な情報だけを抽出する必要がある。このとき，低次元の実数値ベクトルだけを出力として考えると良い場合があり，代表的な手法として主成分分析がある。また，出力としてラベルなどの離散値を考えて分類するクラスタリングもしばしば行われる。しかしながら，現実には入力データには様々な因子が混在しており，主成分分析などのみで隠れた構造をあぶり出せることは稀である。以下では，ごく最近発表された 2 件の論文で行われている組織特異的な細胞分類手法を紹介する。両者は，解析の最初の段階で細胞データの次元圧縮を行って扱いやすくしている点において共通しているが，その後の処理において，前者は確率モデルを扱い，後者は離散的なグラフモデルを扱っている点が異なる。




解析結果の各丸は細胞を表し，外側の破線の円は分類後のクラスを表す。

図 2　細胞の自動分類

### 2.2.1　脾臓における細胞の分類

Jaitin らは大量並列単一細胞 RNA-seq 解析と教師なし分類アルゴリズムを利用して，マウスの脾臓中の血液細胞をその種類によって分類することに成功している[8]。分類アルゴリズムでは，発現プロファイルから発現量の分散の大きな