Generation of Synthetic Transcriptome Data with Defined Statistical Properties for the Development and Testing of New Analysis Methods

Guillaume Brysbaert, Sebastian Noth, and Arndt Benecke*

Systems Epigenomics Group, Institut des Hautes Etudes Scientifiques/Institut de Recherches Interdisciplinaires, CNRS/INSERM, 91440 Bures sur Yvette, France.

We have previously developed a combined signal/variance distribution model that accounts for the particular statistical properties of datasets generated on the Applied Biosystems AB1700 transcriptome system. Here we show that this model can be efficiently used to generate synthetic datasets with statistical properties virtually identical to those of the actual data by aid of the JAVA application *ace.map creator 1.0* that we have developed. The fundamentally different structure of AB1700 transcriptome profiles requires re-evaluation, adaptation, or even redevelopment of many of the standard microarray analysis methods in order to avoid misinterpretation of the data on the one hand, and to draw full benefit from their increased specificity and sensitivity on the other hand. Our composite data model and the *ace.map creator 1.0* application thereby not only present proof of the correctness of our parameter estimation, but also provide a tool for the generation of synthetic test data that will be useful for further development and testing of analysis methods.

Key words: transcriptome, microarray analysis, signal/variance distribution, distribution modeling, parameter approximation, synthetic data generation

Introduction

We have recently presented an analysis of the data structure and features of the AB1700 microarray technology (Applied Biosystems) and compared those results with the well established Affymetrix technology (*1*). We evidenced significant increases in the overall signal dynamic range and sensitivity of AB1700 data. A second independent lognormal signal distribution at the low end of the signal range was described. The composite signal distribution, which had not been reported before for microarray data, has fundamental implications for data analysis and biologic interpretation. In absence of biophysical or technical explanations for this dual distribution, an essentially different biologic mechanism leading to the second distribution has to exist. Thus not only thousands of weakly expressed additional transcripts can be detected, but also direct conclusions might be drawn as to the fundamental mechanisms of gene regulation (*1*, *2*). Although we did present a theoretical model based on stochastic chromatin "breathing", the origin of the second lognormal signal distribution still remains an unsolved problem, and is likely to spur future research into the governing principals of gene regulation, while today it has already lent further support to the observations concerning the stochasticity in gene transcription start site selection (*2*–*4*).

We have also derived, based on a representative set of original AB1700 data, a composite 18-parameter (18p) model that takes into account the signal, the signal variance, and the variance over the signal variance distributions, and hence accurately describes the global features of AB1700 data (*1*). Using this AB1700 data structure model, new avenues for microarray data quality control could be explored. Moreover, the particularity of AB1700 data warrants re-evaluation, adaptation, and redevelopment of statistical analysis methods, since existing approaches explicitly or implicitly rely on a single lognormal signal distribution hypothesis (*1*, *5*, *6*). We then wondered how we could further demonstrate the applicability of our model to the description of original

***Corresponding author.**
E-mail: arndt@ihes.fr

AB1700 data. As the development and testing of analytical methods require large testing and training datasets, and given the fact that the AB1700 technology is very recent and today few publicly available datasets exist in transcriptome databases, combined with the fact that transcriptome studies are resource, time, and money intensive, we speculated that we might use the statistical model to generate synthetic AB1700-like data that could be used for testing and training purposes. In fact, we have previously applied such a strategy successfully in the testing of the NeONORM inter-assay normalization method (*7*).

In this study, we show that the composite model can be used to generate synthetic AB1700-like data, which are in their global statistical properties indistinguishable from original microarray experimental data. For this purpose, we have also developed a JAVA application—*ace.map creator 1.0*. The ability to generate large sets of AB1700-like data at no cost thereby will certainly help to provide the datasets required for the development and testing of novel statistics approaches.

Results and Discussion

Strategy for generating synthetic AB1700-like transcriptome data

The strategy we chose to generate synthetic AB1700-like data is schematized in Figure 1. Briefly, preprocessed original microarray datasets are analyzed for their signal distributions and signal-variance distributions using the previously described AB1700 model. A random number generator then utilizes the parameter estimates obtained from the analysis of the original data to draw random signal and variance values, thereby generating a dataset of identical size. These synthetic data are then re-evaluated using the model and are shown to display virtually indistinguishable global statistical properties when compared with the original data.

Synthetic HGS V1.0-like data generated from the composite AB1700 data structure model

In the first step, we estimate the parameters for the two independent three-parameter (3p) lognormal signal distributions, for one of which x_0 is allowed to diverge from zero, using an original experimental dataset to reconstruct the corresponding signal distribution (see Materials and Methods) (*1*). Next, using the identical original dataset, we estimate the parameters for the composed Neonex function resulting in the signal-variance distribution (*1*). Finally, using the estimation procedure for probability density function (see Materials and Methods and the Supporting Online Material “05Pseudocode.pdf”), we decompose the signal distribution in the original data such that we can separately estimate the signal-dependent variance distribution for each variance sub-distribution. As discussed previously, the signal-dependent variance

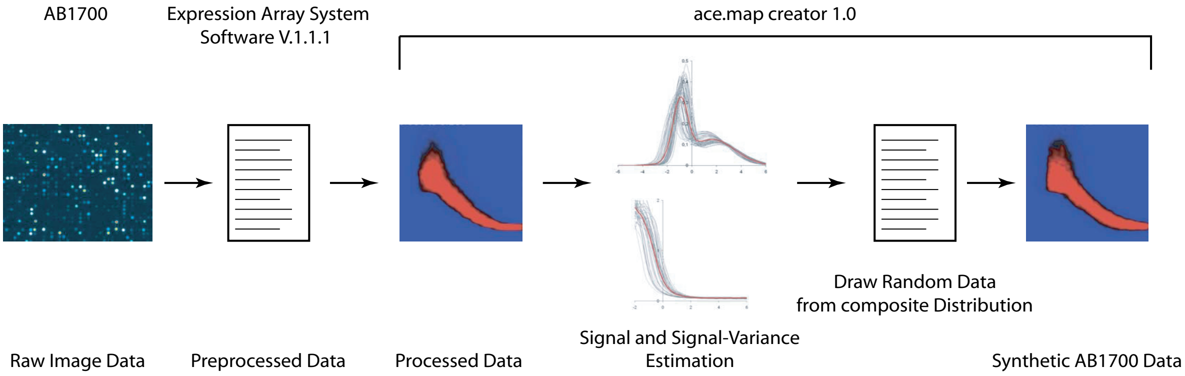

AB1700
Expression Array System Software V.1.1.1
ace.map creator 1.0
Draw Random Data from composite Distribution
Raw Image Data
Preprocessed Data
Processed Data
Signal and Signal-Variance Estimation
Synthetic AB1700 Data

Fig. 1 Overall strategy for generating synthetic AB1700-like data with statistical properties identical to those of the original data. Firstly, raw image data generated during data acquisition on the AB1700 platform are converted into preprocessed and median-normalized ASCII tables by the Expression Array System Software in its current version. Those preprocessed data are then read by the *ace.map creator 1.0* application, which analyses the statistical signal and signal-variance distributions, parameterizes our composite AB1700 data model, and finally draws a set of random data of identical size based on the model. These data are written as output and, due to the logics of the procedure implemented, share the statistical properties of the initial data file.

is lognormal distributed over the entire logarithmic signal range, and the parameters therefore can easily be determined on the decomposed original datasets. The two independent resulting density distributions are then combined together, resulting in the final logarithmic signal variance density over lognormal signal distribution. Taken together, a total of 18 seperate parameters need to be estimated from the original datasets (*1*).

We used the expectation maximization (EM) algorithm (http://www.cs.duke.edu/courses/spring04/cps196.1/handouts/EM/tomasiEM.pdf) (*8*) and the Gram-Schmidt orthogonalized gradient method (*9*) in order to efficiently estimate the above parameters (see Materials and Methods and the Supporting Online Material "05Pseudocode.pdf"). This procedure was followed for the entire set of fifty original AB1700 Human Genome Survey (HGS) V1.0 datasets (Supporting Online Material "04AllParam.pdf"). We next created fifty corresponding synthetic HGS V1.0-like datasets by drawing random logarithmic "signal" and "variance" numbers from the probability density distribution space defined through the 18p composite data structure model. Note that since the parameters are not independent of each other, for each synthetic dataset we always used one original dataset's set of parameter estimates. The obtained logarithmic signal values were de-logarithmized and written back together with the corresponding variances and probe_IDs to tab-delimited raw data ASCII files (see Materials and Methods and the Supporting Online Material "05Pseudocode.pdf" for procedural description). The generated fifty synthetic HGS V1.0-like data files are provided in the Supporting Online Material ("06SynHGS.zip").

Virtually identical properties for synthetic vs. original HGS V1.0 data

As a final validation of our model, we analyzed the fifty synthetic AB1700-like datasets using the same approaches employed throughout the previous study (*1*). First we calculated the virtual signal dynamic ranges according to the same criteria. The averaged results are shown in Table 1 (a complete list for all the original and synthetic data is available in the Supporting Online Material "00DynRange.pdf"). The numbers obtained closely resemble those for the original data, whereas a significant difference can only be noticed for the size of the signal dynamic range of the 98% interval when data are not filtered for a signal to noise ratio superior to three. This is explained by the absence of outliers in the synthetic data, since all values are drawn from within the probability density distribution. In consequence, the signal variance density distribution is curtailed for very low and very high signals, and breaks off earlier than the original data. However, the "signal dynamic range" of the synthetic data still significantly increased when compared with Affymetrix datasets (*1*), and the otherwise closely correlated values should also be noted.

Synthetic HGS V1.0-like data also display a dual lognormal signal distribution

We then plotted the logarithmic signal histograms and their approximations through the single and the mixture lognormal distribution model for the synthetic data (Figure 2A). Similarly, the different signal vari-

Table 1 Dynamic range estimation and comparison of original and synthetic AB1700 microarray data*

Data	Averages over 98% signal interval			Averages over 95% signal interval		
	No. of probes	Signal range (ln)	Variance	No. of probes	Signal range (ln)	Variance
Original	33,928	8.20	0.80	32,889	7.19	0.76
Original (S/N>3)	17,718	7.18	0.79	17,176	6.45	0.85
Synthetic	32,889	7.07	0.73	32,889	6.89	0.66
Synthetic (S/N>3)	17,214	7.15	0.82	17,214	6.14	0.78

*The dynamic ranges of the original AB1700 HGS V1.0 data and the synthetic HGS V1.0-like data generated in this study were estimated by calculating the average logarithmic signal range over fifty independent and heterogeneous datasets from each technology. The logarithmic signal range was determined for the 98% and 95% intervals in order to eliminate outliers at both ends of the dynamic range, and thus obtain more robust results. We also performed the calculation for only those signals that have a signal to noise ratio superior to three (S/N>3). The total average number of probes contributing to the signal range is also given. A complete list for all the original and synthetic data is available in the Supporting Online Material "00DynRange.pdf".

ance distribution plots were created for the fifty synthetic datasets (Figure 2B). Both sets of representations indicate absence of any significant difference between the synthetic data and the original data (*1*). This is further supported by the likelihood estimates that we calculated for the synthetic data using all of the three signal distribution models (Table 2; Supporting Online Material "01SignalDist.pdf" and "02Dualx0Param.pdf"). Note that the likelihood estimates are only marginally increased for the synthetic data when compared with the original data, which is a strong and very important indication that our composite model does not idealize the generated distributions. Finally, we generated 3D signal variance probability density plots (see Materials and Methods and the Supporting Online Material "05Pseudocode.pdf") for three randomly chosen original HGS V1.0 datasets (Figure 3A) as well as the three corresponding synthetic HGS V1.0-like datasets (Figure 3B). Albeit the complexities of the AB1700 data structure, the synthetic density maps closely resemble in the overall structure of their experimental counterparts.

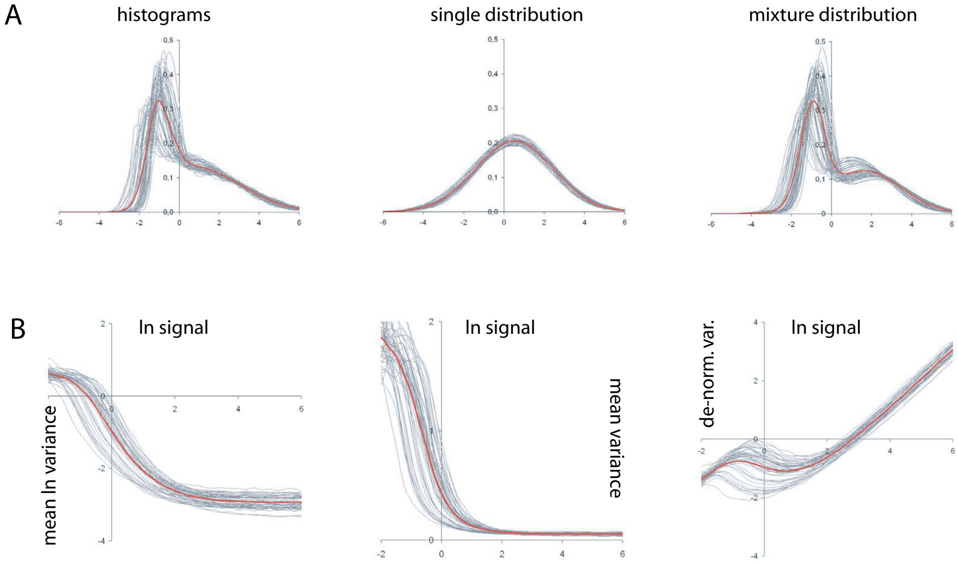


Fig. 2 Fifty synthetic HGS V1.0-like datasets generated using the composite model. We generated fifty synthetic HGS V1.0-like datasets using the parameter estimates from the fifty original HGS V1.0 experiments using the composite model described. The data structure for these synthetic experiments was assessed just as that for the original data. **A**. Left panel: histograms of the synthetic data lognormal signal distribution; Middle panel: approximation using the single lognormal distribution; Right panel: approximation using the mixture distribution model. Parameters can be found in Tables 1 and 2, as well as in the Supporting Online Material. **B**. The variance over lognormal signal distribution plots are shown for the fifty synthetic datasets.

Table 2 Likelihood estimates for different lognormal signal distribution models*

Data	Mean likelihood (L) estimates for signal distributions					
	L (single)	L (dual)	L (dual, x_0)	Single/Dual	Single/(Dual, x_0)	Gain
Original	−85885.00	−80233.11	−79468.81	1.0699	1.0802	8.02%
Synthetic	−85851.72	−80005.49	−79325.29	1.0725	1.0817	8.17%

*Likelihood measures were calculated and averaged for the original and the synthetic data according to three different lognormal distribution models. We considered a single 3p lognormal distribution, two combined 3p lognormal distributions where both x_0 are zero, and finally two combined 3p lognormal distributions where the x_0 of the second distribution diverges from zero. The likelihood averaged estimate ratios are given in order to calculate the gain in descriptive quality. The individual data for the 100 microarray experiments are available in the Supporting Online Material "01SignalDist.pdf".

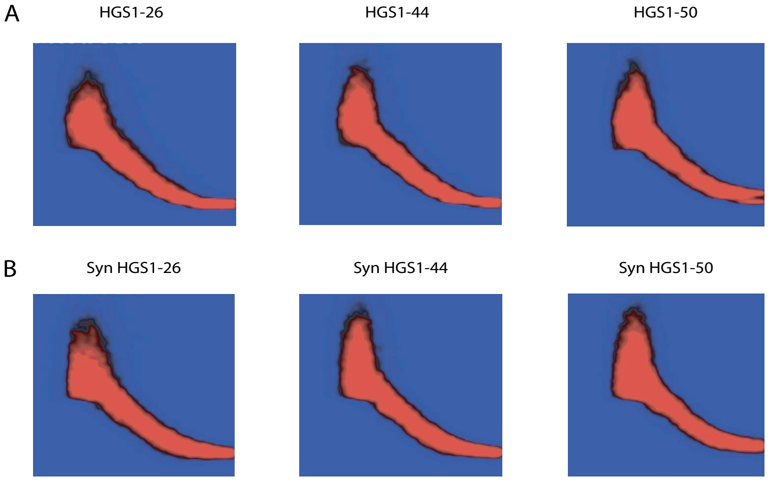


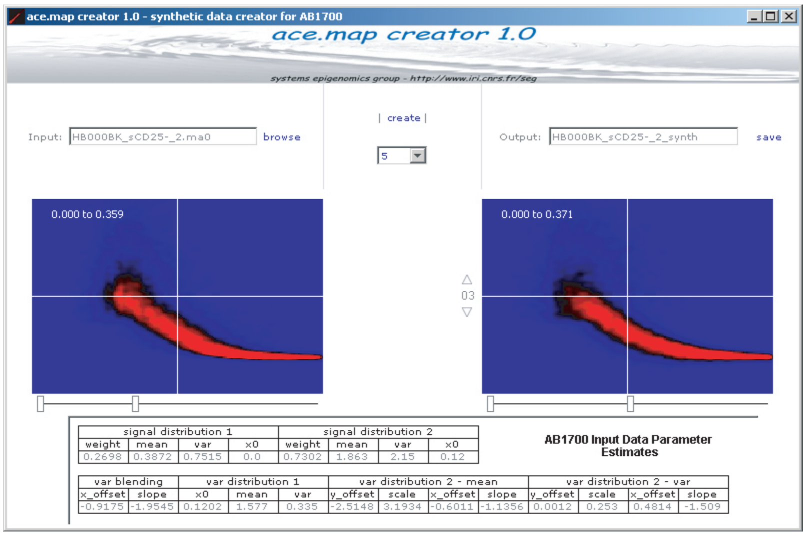
Fig. 3 Qualitative comparison of signal-dependent variance distributions and lognormal signal density plots for the original data and the corresponding synthetic data. **A**. The data density plots of signal-dependent logarithmic variance and logarithmic signal for three individual experimental HGS V1.0 datasets. **B**. The corresponding plots for the three synthetic HGS V1.0-like datasets generated using the parameter estimates from the original data. For all six plots the x-axis stands for ln(signal), with range from −4 to +6; the y-axis stands for ln(variance), with range from −2 to 2. Color scheme for density: blue (low)—black—red (high).

The *ace.map creator 1.0* application

Having shown that our model is of sufficient quality to allow the generation of synthetic HGS V1.0-like data, we decided to generate a software application for this purpose and make it available to researchers interested in further investigating the properties of the high-sensitivity AB1700 data. The idea is that large sets of statistically correctly structured data are necessary for the analysis method and algorithmic development in order to better understand, manipulate, and exploit the experimental advantages that the AB1700 system offers. However, such large sets are currently hard to come by. Since the technology is still very recent (first commercialized in 2005), very few published studies using the AB1700 system are available. Generating once one's data is certainly an option, but is very cost, time, labor, and resource intensive and not necessarily an option for researchers from the bioinformatics and statistics communities. Whereas original data are an absolute requirement for many statistics developments, many other methods can be tested, optimized, and trained even on synthetic data provided the synthetic data correctly reflect the data structure of *bona fide* experimental sets. The application that we sought to create is specifically tailored to such investigations. We present here the first version of *ace.map creator* that is capable of generating synthetic AB1700-like transcriptome data according to the model presented in the previous study (*1*). The software is a stand-alone JAVA application and executes, provided a recent JAVA virtual machine is preinstalled, under all standard operating systems. The user's guide containing a complete description of *ace.map creator 1.0* functionality is provided in the Supporting Online Material ("07acemapCreatorUsersGuide.pdf"). An executable of *ace.map creator 1.0* can be downloaded for non-commercial, public research from our website. Figure 4 is an exemplary screenshot of the running application.

Conclusion

By demonstrating that we can generate synthetic AB1700-like data files with their statistical properties indistinguishable from *bona fide* experimental data, we have further evidenced the correctness of the AB1700 data model we described previously (*1*). Due to the absence of a verified hypothesis to the origin of the mixture distribution observed in these data, our model only provides a statistically meaningful description of their structure without further insights into the underlying biology. Using the JAVA software *ace.map creator 1.0*, however, large synthetic datasets with defined statistical features can be generated in the future in the objective to further evaluate this technology. Therefore, for instance, statistically sound estimates to the number of required technical



ace.map creator 1.0 - synthetic data creator for AB1700

Input: HB000BK_sCD25-_2.ma0

5

Output: HB000BK_sCD25-_2_synth

AB1700 Input Data Parameter Estimates

signal distribution 1				signal distribution 2			
weight	mean	var	x0	weight	mean	var	x0
0.2698	0.3872	0.7515	0.0	0.7302	1.863	2.15	0.12

var blending		var distribution 1			var distribution 2 - mean				var distribution 2 - var			
x_offset	slope	x0	mean	var	y_offset	scale	x_offset	slope	y_offset	scale	x_offset	slope
-0.9175	-1.9545	0.1202	1.577	0.335	-2.5148	3.1934	-0.6011	-1.1356	0.0012	0.253	0.4814	-1.509

Fig. 4 A screenshot of the Java application *ace.map creator 1.0*. The application is freely available for non-commercial research from http://www.iri.cnrs.fr/seg.

and biological replicates for meeting preset significance thresholds could eventually be *a priori* calculated using a small number of actual preliminary experiments and a large set of derived synthetic data.

Another potential application of the 18p composite AB1700 data structure model is quality assessment of microarray experiments. Particularly, since this technology is very recent, researchers will find it quite difficult to obtain feedback on the quality of any single microarray experiment beyond the standard QC parameters determined by the analysis software (see Materials and Methods). Since these standard QC parameters do not capture the overall structure (for example, density distributions) of the transcriptome profile, they only provide some indications towards data quality. Using our model and the procedures described here, the 18 parameters can easily be estimated immediately after primary analysis of the data and can be compared with the average parameter values usually obtained, therefore provide a much better indication as to the quality of the individual measurement. Over time, with increasingly available AB1700 data in the public resources, these average parameter values will converge towards better defined ranges, and thus lead to ever more sensitive quality assessment estimates. Finally, we believe that the ability to generate synthetic AB1700 test data will be instrumental for the required statistical method evaluation and development. The particular structure of the AB1700 data has implications for data analysis that are not necessarily yet met by current methods of microarray analysis.

Materials and Methods

AB1700 microarray technology

The experimental data referred to as HGS V1.0 (Applied Biosystems, Foster City, USA; ProdNo: 4337467) used in this study were generated on two different AB1700 transcriptome platforms (ProdNo: 4338036) (http://www.appliedbiosystems.com). The dataset thereby is representative for the ensemble of data that we have so far generated and analyzed (*1*). Data preprocessing and primary analysis was performed using the Expression Array System Software v1.1.1 (ProdNo: 4364137) as previously described (*1*). Note that we renormalized the resulting data according to the median once more after having removed probes for which the AB1700 software has set flags equal to or greater than 2^{12}, indicating compromised measurements as well as the controls.

Parameter estimation for the 18p model

Most of the estimation has been described in detail in previous studies (*1*, *8*, *9*) and the Supporting Online Material. The estimation process is embedded into individual EM steps. Every step thereby re-estimates all parameters over the weighted sample data in the previous step (in logarithmic space). In our case, for every data point i [ln(Signal$_i$) | ln(Variance0.0$_i$)] and [ln(Signal$_i$) | ln(Variance0.34$_i$)] (hereinafter $[S_i|V_i]$), the weights $w_{1,i}$ and $w_{2,i}$ are calculated, which correspond to the combined probabilities:

$$p(\theta_n \mid [S_i|V_i]) \,/\, \big(p(\theta_1 \mid [S_i|V_i]) + p(\theta_2 \mid [S_i|V_i])\big)$$

These probabilities $p(\theta_{1,2} \mid [S_i|V_i])$ are the product of the *a priori* probability $p(\theta_n|S_i)$, and hence the mixture function, and also the probability that is determined over the lognormal probability density function at position V_i with the parameters for the corresponding S_i. The weights are being used for the calculation of weighted mean and weighted variance for the first lognormal distribution [$m_1(S_i)$ and $s_1(S_i)$]. They are also being used by the gradient method-based parameter estimation as factors for calculating the cumulative error, being minimized for the second lognormal distribution. After each EM estimation step, the mixture function is re-estimated using the new weights $w_{1,i}$. The EM algorithm terminates either after a preset number of steps is reached (negative abortion), or if the likelihood increase between two EM steps falls below a preset convergence threshold (positive abortion).

Expectation maximization

For the estimation of the lognormal distribution parameters modeling the signal distribution, a standard implementation was used. Weighted means and variances of each distribution were individually estimated/improved in each step:

$$P_{\mu,\sigma,x_0}(x) = \frac{e^{-(\ln(x-x_0)-m)^2/(2\sigma^2)}}{(x-x_0)\cdot\sigma\cdot\sqrt{2\pi}} \quad (1)$$

$$P_i(x) = P_{\mu_i,\sigma_i,x_{0i}}(x) \quad (2)$$

$$w_{i,j} = \frac{P_i(x_j)}{\sum_{k=1}^{2} P_k(x_j)} \quad (3)$$

$$\mu_i \Leftarrow \frac{\sum_{j=0}^{n} w_{i,j}\cdot\ln(x_j - x_{0j})}{\sum_{j=1}^{0} w_{i,j}} \quad (4)$$

$$\sigma_i^2 \Leftarrow \frac{\sum_{j=0}^{n} w_{i,j}\cdot\big(\ln(x_j - x_{0j}) - \mu_i\big)^2}{\sum_{j=1}^{0} w_{i,j}} \quad (5)$$

$$\pi_i \Leftarrow \frac{\sum_{j=1}^{0} w_{i,j}}{n} \quad (6)$$

The variance model is much more complicated but uses the same principle. The main difference is the calculation of π_i:

$$\pi_i = p(\theta_i) = \frac{\sum_j p(x_j|\theta_i)}{\sum_k \sum_j p(x_j|\theta_k)} \quad (7)$$

Gradient method

Gradient method was first used for estimating parameters for the Neonex function (*1*). As a conjugated gradient method, our method does not always use the gradient directly, but rather forms an orthonormal basis via the Gram-Schmidt orthogonalization method in succeeding steps to ensure to improve all parameters and avoid oscillations. The iterative search is thus subdivided into n orthogonalization steps, with n being the number of parameters of the function to find the minimum for:

$$\vec{g} = -\frac{\nabla_{f(\vec{x})}}{\|\nabla_{f(\vec{x})}\|}; \quad \vec{d} = \begin{cases} i = 1 : \vec{g} \\ i \neq 1 : \vec{g} - \sum_{j=1}^{i-1}(\vec{g}\cdot\vec{b}_j)\cdot\vec{b}_j \end{cases}$$

$$\vec{d} = \frac{\vec{d}}{\|\vec{d}\|}; \quad \text{and} \quad \vec{b_i} = \vec{d}$$

If a calculated $\vec{g}$ is a linear combination of $\vec{b}_1 \ldots \vec{b}_{i-1}$, it cannot contribute to the formation of an orthonormal base. If this is detected by the program, $\vec{d}$ is set to $\vec{g}$ and i is reset to 1.

Each scan consists either of a stepwise movement from the current parameter vector $\vec{x}$ into direction $\vec{d}$ using a predefined step-width until the error stops to decrease, or, if the first step already lead to a greater error, the step-width is divided by two until either the error decreases or a maximum number of divisions has been reached. In both cases, the errors of the last three sampled parameter points are used for quadratic interpolation to further improve the estimate. Depending on ε and/or c, a new $\vec{x}$ is calculated. If the corresponding error should be higher than that for the best scan estimate, it is replaced by the latter.

Generation of synthetic HGS V1.0-like data from the estimated distribution parameters

In the first step, one of the two signal distributions is randomly chosen with a given probability. This probability is dependent on the *a priori* probability for each signal distribution [$p(\theta_1)$ and $p(\theta_2)$]. A signal value S_i is then randomly drawn from the corresponding probability density. Using the logarithm of this signal value [$\ln(S_i)$], the *a priori* probabilities for the first and second variance distributions are determined [$p(\theta_n|S_i)$], and are used to randomly choose one of the variance distributions. For the logarithmic signal S_i, the corresponding parameters $m_n(S_i)$ and $s_n(S_i)$ are calculated, $x_{0,n}$ is chosen, and a random number is drawn from the resulting distribution. The number of such generated synthetic signal/variance pairs corresponds exactly to the number of probes in the original data file. Both the synthetic and the original data are signal rank sorted, and the synthetic data are attributed with the probe_ID corresponding to the same signal rank.

The *ace.map creator 1.0* application

Details on the *ace.map creator 1.0* software can be found in the accompanying user's guide (Supporting Online Material "07acemapCreatorUsersGuide.pdf"). The software, which executes on any standard operating system (Solaris, Linux, Windows, or Macintosh) equipped with the freely available SUN Microsystems JRE package, can be downloaded from http://www.iri.cnrs.fr/seg in the "web sources", "software" section.

Acknowledgements

We thank our collaborators Drs. B. Bell, J. Elion, M. Müller-Trutwein, L. Rogge, C. Zennaro, and their coworkers for sharing unpublished data, as well as their continuous support and encouragement throughout this study. All members of the systems epigenomics group are thanked for stimulating discussions. This work was supported by the European Hematology Association—José Carreras Foundation, the Institut des Hautes Etudes Scientifiques, the Institut de Recherche Interdisciplinaire, the Centre National de la Recherche Scientifique (CNRS), the Institut National de la Santé Et de la Recherche Médicale (INSERM), the Région Nord, and the French Ministry of Research through the "Complexité du Vivant—Action STICS-Santé" program (all to AB).

Authors' contributions

GB has implemented the *ace.map creator 1.0* application and has participated in the development of methods to generate synthetic AB1700 data. SN made the initial observation of a mixture lognormal signal distribution in AB1700 data, and has initially devised a method for generation of artificial data. AB has significantly participated in the mathematical formulations, the generation of artificial data, the statistical data analysis, and manuscript preparation. AB has designed and coordinated this study. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Noth, S., *et al.* 2006. High-sensitivity transcriptome data structure and implications for analysis and biologic interpretation. *Genomics Proteomics Bioinformatics* 4: 212-229.
2. Benecke, A. 2006. Chromatin code, local non-equilibrium dynamics, and the emergence of transcription regulatory programs. *Eur. Phys. J. E (Soft Matter)* 19: 353-366.
3. Kaern, M., *et al.* 2005. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* 6: 451-464.
4. Kurakin, A. 2005. Self-organization vs Watchmaker: stochastic gene expression and cell differentiation. *Dev. Genes Evol.* 215: 46-52.
5. Konishi, T. 2004. Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment. *BMC Bioinformatics* 5: 5.
6. Dean, N. and Raftery, A.E. 2005. Normal uniform mixture differential gene expression detection for cDNA microarrays. *BMC Bioinformatics* 6: 173.
7. Noth, S., *et al.* 2006. Normalization using weighted negative second order exponential error functions (NeONORM) provides robustness against asymmetries in comparative transcriptome profiles and avoids false calls. *Genomics Proteomics Bioinformatics* 4: 90-109.
8. Bilmes, J.A. 1998. A gentle tutorial of the EM algorithm and its appliaction to parameter estimation for Gaussian mixture and hidden Markov models. Technical report, ICSI TR-97-021, University of California at Berkley, USA.
9. Shewchuk, J.R. 1994. An introduction to the conjugate gradient method without the agonizing pain. Technical report, CS-94-125, Carnegie Mellon University, Pittsburgh, USA.

Supporting Online Material
http://www.iri.cnrs.fr/seg/synAB1700.zip