Applications of Extended Kramers Model for the Rate of Chemical Reactions to Noise Models for Single-Cell Transcriptomics, different DNA analyses, and for Topographic and Retinotopic Mapping.

Michael Fundator

Division of Behavioral and Social Sciences and Education of the National Academies of Sciences, Engineering, and Medicine, USA.

Extended Kramers Model based on Fokker-Plank Stochastic Differential Eequation for Velosity of Chemical Reactions can be applied to noise models, such as Brownian noise due to the Brownian motion of diaphragm in ultrasensitive pressure sensors that are extremely sensitive to noise caused by the thermal agitation of the molecules of the fluid medium surrounding the diaphragm and eventually produce incorrect pressure values. The model based on Cumulant Analysis can be extended to uncertainty quantification of complex computational models in different areas of sciences that can require calculation of 5-fold Boltzman integrals and the like. Along with chemical problems based on differences between single and multi cell analysis there are statistical problems related to noise models for single-cell transcriptomics and to challenging problems to distinguish genuine from technical stochastic allelic expression to challenging problems to distinguish genuine from technical stochastic allelic expression that is important in such questions as decomposition of tissues, and different DNA analyses, such as Cytoplasmic membrane-associated DNA (cmDNA) and plasmmid DNA along with Neuroscientific analysis for Topographic and Retinotopic Mapping. These analyses have numerious applications in Biopharmatheutical and Chemical Industies.

Keywords: transcriptomics, retinotop1c map, stochastic allelic expression

The commemoration of I 00 years since introduction of Fokker-Planck Stochastic Differential Equation, and 65 years since the introduction of Hodgkin and Huxley model to Biochemistry and Neuroscience coincides with the commemoration of 250 years of history of Rutgers University.

1. Introduction.

More than 100 years of historical development of research in the relation of cancer to immune system that were accompanied with controversy mainly because of little knowledge of cellular structure led to the development of the field of transcriptomics.(Robert D. Schreiber et al The three es of cancer immunoediting Annu. Rev. Immunol. 329–60 doi: 10.1146/annurev.immunol.22.012703.104803)

"Transcriptomics is the study of the complete set of RNAs (transcriptome) produced by the genome of a specific cell or organism at a specific time or under specific circumstances or in a specific cell using high throughput methods, e.g. MRI"(www.nature.com › subjects). "Another definition of the transcriptome as all the RNA molecules — which includes a wide variety of untranslated, nonprotein encoding RNA transcribed from the DNA of the genome. It is now thought that 76% of our DNA is transcribed into RNA although only 1.5% of this is messenger RNA for protein synthesis."(http://www.biology-pages.info/E/ESTs.html)

DeoxyriboNucleicAcid(DNA) is 1)double- stranded molecule with a 2)long chain of nucleotides with 3)a deoxyribose and phosphate backbone and 4)four different bases: adenine(A), guanine(G), cytosine(C), and thymine(T), versus RiboNucleicAcid(RNA)that is a single-stranded molecule with a shorter chain of nucleotides with a 3)ribose and phosphate backbone and 4)four different bases: adenine(A), guanine(G), cytosine(C), and uracil(U).(https://www.saylor.org/site/wp-content/uploads/2010/11/BIO101-DNA-vs-RNA.pdf)

An allele is a variant form of a gene. Some genes have a variety of different forms, which are located at the same position, or genetic locus, on a chromosome. Humans are called diploid organisms because they have two alleles at each genetic locus, with one allele inherited from each parent. (http://www.nature.com/scitable/definition/allele-48)

Equilibrium

"The next step in cancer immunoediting proceeds to the equilibrium phase in which a continuous sculpting of tumour cells produces cells resistant to immune effector cells. This process leads to the immune selection of tumour cells with reduced immunogenicity. These cells are more capable of surviving in an immunocompetent host, which explains the apparent paradox of tumour formation in immunologically intact individuals." (Kim R, et al Cancer immunoediting from immune surveillance to immune escape. Immunology. 2007;121(1):1-14. doi:10.1111/j.1365-2567.2007.02587.x.)

Sigmoid semilogarithmic functions with shape of Boltzmann equations are very much in use for description of diverse biological situations that can be applied to wide variety of scientific fields ranging from the behavior of proteins involved in molecular biology and physiology to different calculations related to the equilibrium potential of ion species that is permeant through membrane channels.

The mathematical structure of the models of Boltzmann type kinetic equations for reacting gas mixtures for particles undergoing inelastic interactions with reactions of bimolecular and dissociation-recombination type is very complicated, because of the collisional operators that usually in the full Boltzmann equations, are expressed by 5-fold integrals.

Consequently direct numerical applications of these models present several computational difficulties.

The search for the simpler solution had its long way till the introduction of the equation for the Brownian motion by Albert Einstein.

With the application of

Transition State Theory to Arrhenius equation,

R = C $e^{-\frac{E_b}{kT}}$

where R is the rate of chemical reaction, Eb the activation energy barrier , k is the Boltzmann constant, T is the temperature, and C is a constant transforms C to

C = kT/h ,

where h is Plank's constant, this however, does not consider the state of equilibrium of the reactants.

In the theory of the velocity of chemical reactions the problem of study by Kramers was based on empirical knowledge that the reactants are in the state of equilibrium. His introduction of diffusion equation is given in the following form of Fokker-Plank equation,

m$\frac{\partial^2 x}{\partial t^2}$ = - $(\partial U(x))/\partial x$ - $\gamma m\, \partial x/\partial t$ + F(t), where m is reduced the mass in the potential of mean force U, and F is a noise of a random fluctuating force, originating from the thermal motion, γ is a viscosity.

It was based on the assumptions about a particle that moves in an external field of force and additionally is subject to the irregular forces of a surrounding medium in temperature equilibrium , which he called Brownian motion. The conditions are such that the particle is thought of as caught in a potential hole but may escape in the course of time by passing over a potential barrier. The problem is to calculate the probability of escape in its dependency on temperature and viscosity of the medium.

2. Applications to different models of Biochemistry and Molecular Biology.

In electrophysiology an introduction of the Hodgkin and Huxley (H&H) model for ion Gating dynamics forms basis for the description of dynamics of the open or closed state region of an ion channel by a probability density p(x, t) which satisfies a Fokker-Planck equation.

Hodgkin and Huxley work was based on intuition, the existence of cellular plasma membranes has not been proven, and there was no knowledge about the nature of macromolecules associated with nerve excitation. Only after 20 years the development in Neuroscience offered proof of predicted H&H model.

The spatial order of neurons in the order of their axonal connections, is related to Topographic maps that are found in many parts of the nervous system. As one of the simplest examples are the Retinotopic maps for communication between retinal axons and lamina neurons.

3. Applications to surface acoustic wave pressure sensors and biosensors.

Aluminum Nitride (AlN) based surface acoustic wave (SAW) pressure sensors for harsh environment applications are of great interest in recent years. Such sensor employs a thick diaphragm (~50 μm) to endure the high pressure, but this seriously limits the sensitivity of these devices, when employed in high temperature environments.

4. Applications to riboswitches

As a counterpart in molecular biology riboswitches that were introduced quite recently and developed in span of the last 20 years are the fragments of a messenger RNA that bind a small RNA, and in this way mRNAs bind metabolites that involves untranslated regions (UTR) of RNA (Author's unpublished manuscript)

The statistical properties and spectral characteristics of the noise, due to Brownian motion of diaphragm in ultrasensitive solid-state capacitive and piezoresistive pressure sensors operating at sub-millimeters of mercury pressures in a gaseous ambient, are obtained as functions of the diaphragm dimensions, temperature, and applied pressure.(Wise K D et al (987) Noise Due to Brownian Motion in Ultrasensitive Solid-State Pressure Sensors DOI 1011091'-ED 19872300:z}

It comes out of the study of diaphragms that using the notions of equilibrium and escape barrier potential, we can apply the Kramers theory to diaphragm sensors also. After realization of this result, it can be applied to the mathematics of Neuroscience.

The spatial order of neurons in the order of their axonal connections, is related to Topographic maps that are found in many parts of the nervous system. As one of the simplest examples are the Retinotopic maps for communication between retinal axons and lamina neurons. "Topographic maps, which maintain the spatial organization of neurons in the order of their axonal connections, are found throughout the nervous system" (Udin and Fawcett, 1988). "A notable model of topographic map formation is found in the visual system, in which a relay of visual information is arranged in a spatially ordered manner from the retina to the visual center in the brain. This topographic map is termed a retinotopic map".(Karl Friedrich Fischbach et al Recognition of pre- and postsynaptic neurons vianephrin/NEPH1 homologs is a basis for the formation of the Drosophila retinotopic map) As in H&H was purely analyticalHodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol.1952;117:500-44.

5. Considerations for gene expressions.

Single nucleotide polymorphisms, frequently called SNPs (pronounced “snips”), are the most common type of genetic variation among people. Each SNP represents a difference in a single DNA building block, called a nucleotide. For example, a SNP may replace the nucleotide cytosine (C) with the nucleotide thymine (T) in a certain stretch of DNA.

SNPs occur normally throughout a person’s DNA. They occur once in every 300 nucleotides on average, which means there are roughly 10 million SNPs in the human genome. Most commonly, these variations are found in the DNA between genes. They can act as biological markers, helping scientists locate genes that are associated with disease. When SNPs occur within a gene or in a regulatory region near a gene, they may play a more direct role in disease by affecting the gene’s function.(https://ghr.nlm.nih.gov/primer/genomicresearch/snp)

As an example of the above discussion:

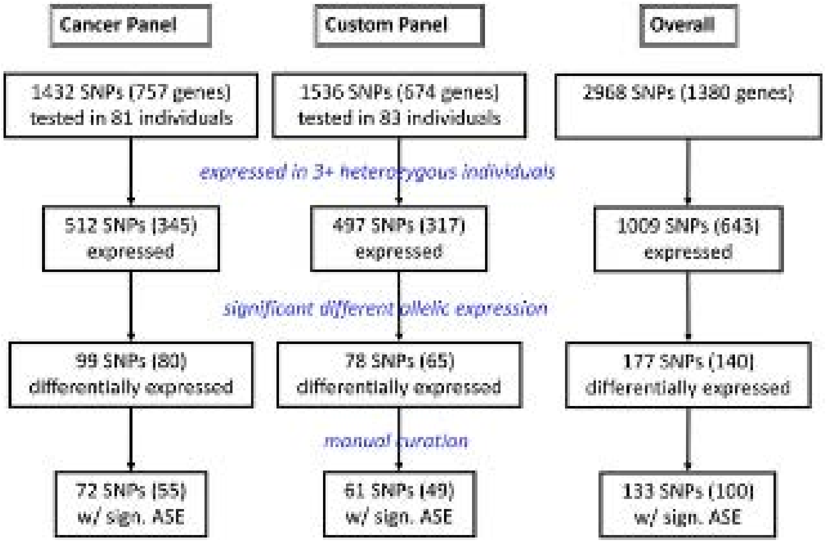
Cancer Panel
Custom Panel
Overall
1432 SNPs (757 genes) tested in 81 individuals
1536 SNPs (674 genes) tested in 83 individuals
2968 SNPs (1380 genes)
expressed in 3+ heterozygous individuals
512 SNPs (345) expressed
497 SNPs (317) expressed
1009 SNPs (643) expressed
significant different allelic expression
99 SNPs (80) differentially expressed
78 SNPs (65) differentially expressed
177 SNPs (140) differentially expressed
manual curation
72 SNPs (55) w/ sign. ASE
61 SNPs (49) w/ sign. ASE
133 SNPs (100) w/ sign. ASE

Figure 1. Experiment design and results obtained for the two panels used in the study.The Overall column corresponds to the combination of the two panels. David Serre et al Differential Allelic Expression in the Human Genome: A Robust Approach To Identify Genetic and Epigenetic Cis-Acting Mechanisms Regulating Gene Expression https://doi.org/10.1371/journal.pgen.1000006.g001

6. Insights into near-Gaussian distributions.

Before considering any discussion about the above phenomenon or any possible approach to analyze or investigate it, it seems appropriate to quote Karl Pearson, who wrote 110 years ago on p. 189 “My custom of terming the curve the Gauss–Laplacian or normal curve saves us from proportioning the merit of discovery between the two great astronomer mathematicians." One of the definitions of Peirce of "normal" as of what would, in the long run, occur under certain circumstances, clearly implies Principle of prediction and LLN.

“It is undeniable that, in a large number of important applications, we meet distributions which are at least approximately normal. Such is the case, e.g., with the distributions of errors of physical and astronomical measurements, a great number of demographical and biological distributions, etc.” Cramer.

The first investigation of slightly non-Gaussian distributions was undertaken by Chebyshev around a century and a half ago, who studied in detail a family of orthogonal polynomials which form a natural basis for the expansions of these distributions. A few years later the same polynomials were also investigated by Hermite and they are called Chebyshev-Hermite or simply Hermite polynomials, their definition was first given by Laplace.

These methods use Edgeworth’s form that is equivalent to the Gram-Charlier Type A series with use cumulant analysis for the representation of the distribution function in terms of different types of sums of functions of Gaussian processes.

A standard method of exploring high-dimensional datasets is to examine various low-dimensional projections thereof. In fact, many statistical procedures are based explicitly or implicitly on a projection pursuit. Under weak regularity conditions on a distribution P = P(n) on Rn, most d-dimensional orthonormal projections of P are similar (in the weak topology) to a mixture of centered, spherically symmetric Gaussian distributions on Rd if n tends to infinity while d is fixed.

To strengthen this notion consider cumulants properties for time series analysis that provide measure of Gaussianity. If r.v. X is normal, then cumk{X} = 0 for k > 2, where cumk denotes the joint cumulants of X with itself k times.

For simplicity consider seq of iid Xi with all moments and E{Xi} = 0 and var{Xi} = 1, then for

Sn = ΣXi/$\sqrt{n}$ cum_k\{ Sn \} = n cum_k\{X\}/$n^{k/2}$ that tends to 0 for k > 2, as n tends to infinity, so Sn has a limiting normal distribution.

And for time series analysis the moment function

E{X(t+u1)… X(t+uk-1)X(t)} would not depend on t, and on the short time interval centered at point of time t can be approximated by normal distribution.

7. DNA and RNA sequencing. Tests ofχ2, Poisson, and other assumptions.

Different fields of research in Neuroscience are related to various types of genetical analysis, e.g. cytoplasmic membrane-associated deoxyribonucleic acid (cmDNA), plasma cell membrane DNA (plasmidDNA) , etc…

Statistical problems related to noise models for single-cell transcriptomics and to challenging problems to distinguish genuine from technical stochastic allelic expression that is important in such questions as decomposition of tissues into cell types. In general DNA analysis data from next generation sequencing (NGS) technologies posed new computational and statistical challenges because of their massive size, complicated structure (large number of genes), limited information, and discreteness. NGS data are more complex than data from previous high throughput technologies such as microarrays that were increasingly utilized for genetic testing of individuals with unexplained developmental disorders. From a statistical perspective, the theory of multiple hypotheses testing (Efron et al., 2001), and the use of false discovery rates (FDR) for multiple testing problems (Benjamini and Hochberg, 1995; Storey, 2003; Efron, 2010), which were motivated by microarray data, also apply to NGS data analysis with modifications to acknowledge the data specific features (Oshlack et al., 2010)

Let ngstbe the observed count of gene g in the sample s with treatment t, and θgt is the expected value of ngst. The approach is similar to that of microarray analysis, to obtain test statistics or posterior distributions, for testing gene-wise differential expression, μg1 −μg2between the two treatment groups. Such approach has its important application in clinical trials for the mentioned above cluster randomized trialsor mixed models repeated measures models that are closely related to such fields as epidemiology and immunology and are discussed later. However, the sampling distribution of the test statistic for NGS data analysis is difficult to obtain without restrictive distributional assumptions.

When analyzing microarray data it is quite common to assume the data follow Gaussian distribution after normalization and log transformation. As such, for each gene, the test statistic for differential expression follows a t-distribution with (2S-2) degrees of freedom (Efron et al.,2001).

However, in the last 4-5 years many works used FDR and multiple TSH application to determine if gene differential expression follows Poisson distribution and χ2, or Negative Binomial assumptions that are often used to model reads from RNA-seq data. Although a Poisson model may be appropriate for technical replicates when variability is lower, with higher variability in biological replicates or surrogate replicates, the Poisson model does not control Type-I error and underestimates the variability, as overdispersion can be observed. Because multiple tests are being performed across the set of genes, q-values are reported that control the false discovery rate using the Benjamini–Hochberg procedure [51].Data were gathered from 75 RNA-seq experiments conducted in five different bacteria: E. coli, N.gonorrhoeae, S. enterica, S. pyogenesand, X. nematophila.Altogether, the RNA-seq experiments yielded over two billion sequencing reads corresponding to 189 billion nt. Without application of TSH it has been deemed impossible to observe that real data for gene sequencing does not follow assumptions of χ_1^2, Poisson, and other distributions. It can be even guessed from "Canonical model for noise gene expression, where the processes of DNA activation, transcription and translation are all represented as Poisson processes with given rates, gives a master equation which may be solved exactly (with generating functions) under various assumptions or approximated with stochastic tools like Van Kampen's system size expansion." (Stochastic switching in biology: from genotype to phenotype Paul C Bressloff)

The following is the author's idea.

Since RRKM theory is based on the following assumptions:

A molecule is considered as a collection of s coupled harmonic oscillators.

The intermolecular distribution of the excess energy (IVR) occurs faster than the unimolecular decomposition of the activated complex back to reactants (referred to as the "ergodic assumption").

The same approach can be used for DNA and RNA analysis along with mathematical analysis of different equilibrium calculations for cell membranes.

References.

1. L. Dumbgen, P. D. Conte-Zerial “On low-dimensional projections of high-dimensional distributions”.

2. M. P. BianchiA BGK-type model for a gas mixture undergoing reversible reaction.

3. H. A. Kramers “Brownian motion in a field of force and the diffusion model of chemical reactions.” Physica, 7, 4, 284-304 (1940)

4. R. von Mises “Probability, Statistics and Truth”

5. Straube, A. V. et al(2011), "How accurate are the nonlinear chemical Fokker-Planck and chemical Langevin equations?", The Journal of Chemical Physics, 135:084103

6. P. Morters “Lecture notes.”

7. D.R. Brillinger “Moments, cumulants, and some applications to stationary random processes”.

8. Y. L. Tong, “Relationship between stochastic inequalities and some classical mathematical inequalities,” Journal of Inequalities and Applications, vol. 1, no. 1, pp. 85-98, 1997.

9. Alicia Oshlack, Mark D Robinson Matthew D Young“From RNA-seq reads to differential expression results.”Genome Biology November/10.

10. By Marie-Josée Fortin, Mark R. T. Dale “Spatial Analysis: A Guide for Ecologists” CUP

11. Gerald Paul, H Eugene Stanley “Partial test of the Universality Hypothesis: The case of different coupling strengths in different lattice directions. “

12. D. R. Anderson, K.P. Burnham “Kullback-Leibler Information as a basis for strong inference in ecological studies” Wildlife Research, 28, 111-119

13. G.Polenta, D.Marintucci, A.Balbi, P.de Bernardis, E.Hivon, S.Masi, P.Natoli, N.Vittorio “Unbiased Estimation of an Angular Power Spectrum”Astrophysics

14. P. J. Marshall, M. P. Hobson, S. F. Gull, andS. L. Bridle “Maximum-entropy weak lens reconstruction: improved methods and application to data”Monthly Notices of the Royal Astronomical Society Volume 335, Issue 4, pages 1037–1048, October/02

15. Bradley Efron“Large-Scale Inference:Empirical Bayes Methods for Estimation, Testing and Prediction”

16. Jelle J. Goemanan, Aldo Solarib“Tutorial in biostatistics: multiple hypothesis testing in genomics” Statist. Med. 2012, 00 1-27.

17. Kenneth F. Manly, Dan Nettleton and J.T. Gene Hwang “Hypotheses Genomics, Prior Probability, and Statistical Tests of Multiple Hypotheses.”

18. Jeffrey T. Leek, John D. Storey “A general framework for multiple testing dependence.”

19. T. A. B. Snijders“Hypothesis Testing: Methodology and Limitations.”

20. Ullah, A. & D.E.A. Giles. “The positive-part Stein-rule estimator and tests of hypotheses.” Economics Letters 26 (1988): 49-52.

21. H. Spohn et al “Numerical test of hydrodynamic fluctuation theory in the Fermi-Pasta-Ulam chain.” Physical review E 90, 012124 (2014)

22. H.M. Hudson “A Natural Identity for exponential families with applications in multiparameter estimation” The Annals of Statistics V.6, N 3

23. Martin Vetterli et al “Wavelet-Based Texture Retrieval Using Generalized Gaussian Density and Kullback–Leibler Distance” IEEE

24. J.D Esary, F Proschan, D.W Walkup “Association of random variables with applications”

Ann. Math. Statist., 38 (1976), pp. 1466–1474

25. Paul Richard Halmos “Lectures on Boolean Algebras” Van Nostrand

26. D. E. Caldwell, S. H. Lai and J. M. Tiedje “A Two-Dimensional Steady-State Diffusion Gradient for Ecological Studies” Bulletins from the Ecological Research Committee/NFR No. 17, Modern Methods in the Study of Microbial Ecology (1973), pp. 151-158.

27, Modern Methods in the Study of Microbial Ecology (1973), pp. I 51-158.

28. David Case Hydrodynamics - David Case casegroup.rutgers.edu/lnotes/hydrodynamics.pdf

29. CR Guest Dynamics of Mismatched Base Pairs in DNA Biochemistry 91(30),3271-3279

30. Van Oudenaarden, A et al (2008). "Nature, nurture, or chance: stochastic gene expression

and its consequences". Cell. 135 (2): 216-26. doi: I 0.1016/j.cell.2008.09.050. PMC 3118044 31.

Golding, I;

31 . Cox, EC et al (2005). "Real-time kinetics of gene activity in individual

bacteria". Cell. 123 (6): 1025-36. doi: 10. 1016/j.cell.2005.09.031. PMID 16360033.

32. Singer, RH et al (2006). "Transcriptional pulsing of a developmental gene". Current Biology.

16 (10): 1018-25. doi:10.1016/j.cub.2006.03.092. PMID 16713960.

33. Tyagi, Set al (2006). "Stochastic mRNA synthesis in mammalian cells". PLoS Biology. 4 (10):

e309. doi:10.1371/journal.pbio.0040309. PMC 1563489 Freely accessible. PMID 17048983.

34. Paulsson, J. et al (2010). "Non-genetic heterogeneity from stochastic partitioning at cell

division". Nature Genetics. 43 (2): 95- 100. doi:10.1038/ng.729. PMC 3208402.

35. Robert A. Meyers Mathematics of Complexity and Dynamical Systems

36. Hodgkin, A. L., and A. F. Huxley. 1952. A quantitative description of membrane

current and its application to conduction and excitation in nerve. J. Physiol.

{Lond.). 117:500 -544.

37. Elf, J. and Ehrenberg, M. (2003) "Fast Evaluation of Fluctuations in Biochemical Networks

With the Linear Noise Approximation", Genome Research, 13 :2475- 2484.

38. Pawitan, Y., Seng, K. C. & Magnusson, P. K. E. How many genetic variants remain to be discovered? PLoS ONE 4, e7969 (2009).

39. Ioannidis, J. P A Genetic associations: false ortrue? Trends Mot. Med. 9, 135- 138 (2003). McCarthy, M. l. et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Rev. Genet. 9, 356-369 (2008).

40. Hayot, F. and Jayaprakash, C. (2004), "The linear noise approximation for molecular fluctuations within cells", Physical Biology, I :205

41. Grima, R. and Thomas, P. and Straube, A. V. (2011), "How accurate are the nonlinear chemical Fokker-Planck and chemical Langevin equations?", The Journal of Chemical Physics, 135:084103

37. M. Fundator Applications of Multidimensional Time Model for Probability Cumulative Function for Parameter and Risk Reduction. In JSM Proceedings Health Policy Statistics Section Alexandria, VA: American Statistical Association. 433-441.

38. M. Fundator Multidimensional Time Model for Probability Cumulative Function. In JSM Proceedings Health Policy Statistics Section. 4029-4039.

39. M. Fundator Testing Statistical Hypothesis in Light of Mathematical Aspects in Analysis of Probability doi :10. 20944/preprints201607 .0069. vl

40. Elf, J. and Ehrenberg, M. (2003) "Fast Evaluation of Fluctuations in Biochemical Networks With the Linear Noise Approximation", Genome Research, 13 :2475- 2484.

41. Pawitan, Y., Seng, K. C. & Magnusson, P. K. E. How many genetic variants remain to be discovered? PLoS ONE 4, e7969 (2009).

42. Hayot, F. and Jayaprakash, C. (2004), "The linear noise approximation for molecular fluctuations within cells", Physical Biology, I :205

43. Ioannidis, J. P A Genetic associations: false or true? Trends Mot. Med. 9, 135- 138 (2003).

44. McCarthy, M. l. et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Rev. Genet. 9, 356-369 (2008).