ISSN 1750-9823 (print)
International Journal of Sports Science and Engineering
Vol. 03 (2009) No. 01, pp. 011-016

DNA Genome-based Diagnostics for Athletes' Physiological Conditions Using Cluster Analysis

Ling-Hong Tseng [1,+] , Ilene Chen [2, 3, +] , Hiroto Homma [3], Hong Yan [2, 4], Chyi-Long Lee [1]

[1] Department of Obstetrics and Gynecology, Chang Gung Memorial Hospital and University of Chang Gung School of Medicine, Taiwan, China

[2] School of Electrical and Information Engineering, University of Sydney, Australia

[3] Hiroto Homma Swim School, NSW, Australia

[4] Department of Electronic Engineering, City University of Hong Kong, Hong Kong, China

(*Received July 18, 2008, accepted November 30, 2008*)

Abstract. A system of cluster analyses for gene expression signatures from cDNA microarrays is described to diagnose athletes' physiological conditions in response to training loads, which uses standard statistical algorithms to arrange altered genes according to similarity in the patterns of gene expression. The output is displayed graphically using heat maps and dendrograms, conveying the clustering and the underlying patterns of gene expression in a form intuitive for coaches. A "39-gene" model is developed to diagnose athletes' physiological conditions with cDNA microarrays. Since the pattern seen in gene expression signatures indicates the status of cellular processes, our results suggest a strategy to "see" training -induced cellular processes based on all of the 39 altered genes with cDNA microarrays.

Key words: cluster analysis, physiological conditions, gene expression signatures, cDNA microarrays

1. Introduction

During the last three decades there has been a movement away from judging the worth of training by the physical challenge it presents to judging training by its effect on the physiological mechanisms of the human body [1]. Training was designed to place maximum stress on high performance athletes [1]. Coaches designed various programs around the idea of pushing athletes to the limit of their pain tolerance and then motivating them to go beyond it [1]. Programs had athletes swimming faster, farther, or swimming with fewer rests during training than they or their competitors had ever swum before [1]. These programs involved targeting each of the major phases of energy metabolism and other aspects of physical conditioning. Training loads are the coaches' responsibility, directly affecting the athletes' health [2].

At present, highly trained swimmers at the Australian Institute of Sport (AIS) are still tested for blood lactate, oxygen consumption and heart rate using traditional medical instrumentation [2]. To do this, for example, a sample of blood is collected from the ear or fingertip from each athlete after each swim in order to measure the amount of lactate acid [1]. Such blood testing involves swimming a series of repeated at progressively faster speeds during the training of competitive swimmers [1]. Indeed, athletes may not like such blood testing since it can increase the risk of transferring HIV/AIDS, hepatitis B, hepatitis C and many other blood-borne infectious diseases from one person to another [1]. Also, it would be unnecessary to do so since the quantities that are reported are the merely the amount of lactate acid; consequently, an integrated knowledge of the athletes' physiological condition being studied would be quite superficial. We need to develop some better fundamental tools to diagnose athletes' physiological conditions in response to training.

The Human Genome Project, completed in April 1953, has driven the development of technology to diagnose athletes' physiological conditions in response to training in new ways. The new molecular tools center on gene cloning and sequencing technology and molecular probing [2], which amount to hundreds of data points for thousands or tens of thousands of genes, can be used to recognize different cells and cellular

[+] Ling-Hong Tseng and Ilene Chen contributed equally to this paper.
Correspondence should be addressed to Ilene Chen ilene@ee.usyd.edu.au TEL: +61 2 93517221 FAX: +61 2 93513847

structures. DNA molecules can now be attached to small squares of glass that, by analogy with computer circuits, are called cDNA microarrays or gene chips [3-5]. Most human genes exist in several forms that differ very slightly in their nucleotide sequences, and some of these altered genes contribute to training-induced stress [3-5]. Gene chips can be designed that will reveal these altered genes and therefore examine an athlete's physiological conditions whose underlying gene alternations are known [3-5].

In our previous work [6], we have shown how fixed effect logistic regression model can be used to identify training-induced cellular reactions with gene expression signatures from cDNA microarrays. In the present work, we aim to show how cluster analysis can be used to "see" athletes' physiological conditions, using standard statistical algorithms to arrange altered genes according to similarity in patterns of gene expression. A "39-gene" model has been developed to diagnose athletes' physiological conditions with cDNA microarrays. Our results indicate that patterns of gene expression exhibit the status of cellular processes, which may not be observed easily otherwise. The technique as presented is referred to as DNA genome-based diagnostics for athletes' physiological conditions.

2. Methods

Source of Experimental Data. The data used here was collected from Gene Expression Omnibus under platform accession no GDS1432; its web address is http://www.ncbi.nlm.nih.gov/geo/gds/gds_browse.cgi?gds=1432. In this analysis, we re-evaluate the gene expression based RNA/sported DNA/cDNA array by Fluck et al. [6] on vastus lateralis muscles before and after a 6-week endurance training course, respectively. Expression patterns of vastus lateralis muscles were examined repeatedly 0, 1, 8 and 24 hours following a 30-minute endurance training course. A detailed description relating to its experimental procedure can be found in Fluck et al. [6].

Hierarchical Clustering. Cluster analysis is a generic term that attempts to determine whether or not a given dataset contains distinct groups, and, if so, to find the groups [8]. Two general references on the topic of cluster analysis are given by Hartigan [9] and Gordon [10]. This method starts from a similarity matrix calculation between the altered genes to be clustered on the basis of their expression patterns. The similarity metric used here is the Euclidean distance, defined as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^{d} \left(X_{ik} - X_{jk} \right)^2}$$

where X_{ik} and X_{ik} are the measured levels of gene expression for test samples i and j. In this analysis, a hierarchical cluster algorithm is used, which leads to a series of hierarchical groups, reported by tree-like dendrograms [8]. The similarity between two clusters is based on average linkage clustering, where the inter-cluster distance are defined as the average of the distance between all pairs of data points and members of a pair which are in distinct clusters [8].

Displays. Here we use heat maps to plot the measured levels of gene expression on the 2-dimensional grids and use the rank-by-feature color scale scheme to represent the magnitude of gene expression on the basis of their expression patterns. In this analysis, medium data points are colored black, data points with increasing ranks with reds of increasing intensity and decreasing ranks with greens of decreasing intensity [11]. Then, outputs from average linkage hierarchical clustering are reported using tree-like dendrograms that are appended to the corresponding heat maps to represent the nature of the computed relationships among altered genes using any ordering [11].

3. Results

We have used average linkage hierarchical clustering to study the gene expression model relating to endurance training-induced vastus lateralis muscle contraction (Table 1, Fig. 1). Given the measured levels of gene expression within vastus lateralis muscle cells 0, 1, 8 and 24 hours, respectively, following a 30-minute endurance training course, the case and control were characterized at 6-week intervals before and after undergoing the training course, respectively.

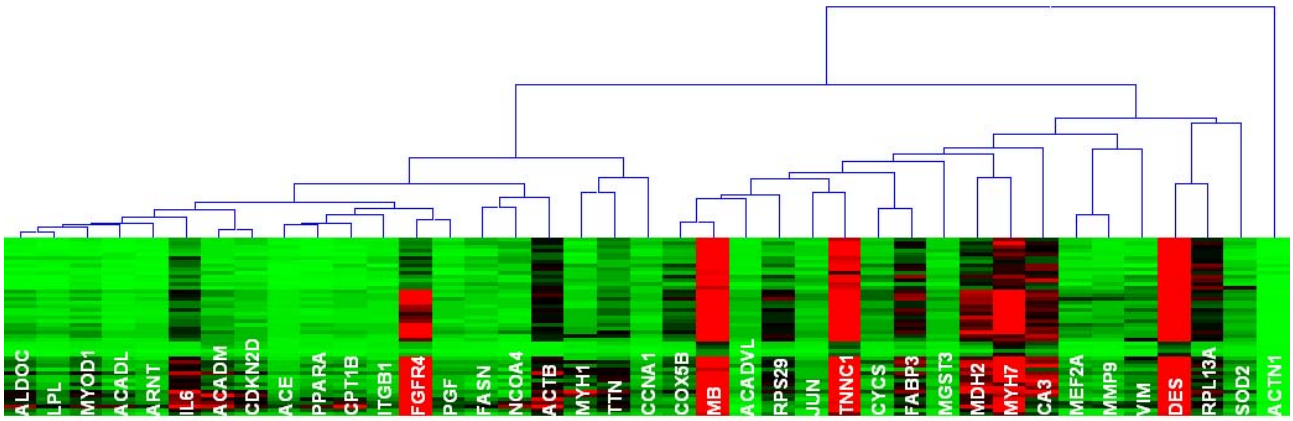

Fig. 1. Clustered display of altered gene expression data of endurance training –induced muscle contraction.

Table 1. Altered genes and their ontology in response to endurance training –induced muscle contraction.

Gene symbol	**Gene Ontology**
ACADL	lipid metabolism; electron transport
ACADM	lipid metabolism; electron transport
ACADVL	lipid metabolism; electron transport
ACE	angiotensin catabolic process in blood
ACT	cell motility
ACTN1	cell motility
ALDOC	glycolysis
ARNT	regulation of glycolysis
CA3	energy metabolism
CCNA1	cell cycle
CDKN2D	cell cycle
COX5B	respiratory gaseous exchange
CPT1B	lipid metabolism
CYCS	apoptosis
DES	muscle contraction
FABP3	phosphatidylcholine anabolism
FASN	lipid biosynthesis
FGFR2	amino acid phosphorylation
IL6	inflammatory response
ITGB1	leukocyte adhesion
JUN	leading edge cell differentiation;
LPL	lipid metabolism
MB	oxygen transport
MDH2	carbohydrate metabolism; glycolysis
MEF2A	muscle development
MGST3	lipid metabolism
MMP9	macrophage differentiation
MYH1	striated muscle contraction
MYH7	striated muscle contraction
MYOD1	striated muscle development
NCOA4	transcription
PGF	angiogenesis
PPARalpha	lipid metabolism; regulation of energy homeostasis
PPL13A	protein biosynthesis (ribosomal protein L13a)
RPS29	protein biosynthesis (ribosomal protein S29)

SOD2	response to oxidative stress/reactive oxygen species excess
TNNC1	regulation of muscle contraction
TTN	carbohydrate metabolism
VIM	cell motility

The results were very interesting. Based on the resultant expression patterns of the 39 altered genes observed in Figure 1, it is our proposition that this "39-gene" model can be used to predict training-induced physiological mechanisms at the molecular level with cDNA microarrays. Firstly, MDF2A and MYOD are found to be responsible for muscle development. MYH1 is responsible for skeletal muscle contraction, whereas MYH7 is responsible for cardiac muscle contraction. This muscle contraction is regulated by TNNC1.

Secondly, complex biosynthesis and metabolic processes, including energy, carbohydrate and lipid metabolisms, were found to have strong relationship with endurance training–induced muscle contraction: CA3 for energy metabolisms, MDH2 and TTN for carbohydrate metabolism, and ADADL, ACADM, ACADVL, CPT1B, FASN, LPL, MGST3, and PPARalpha for lipid metabolisms. This muscle contraction requires energy, which is made available from the following two ribosomal molecules: PPL13A and PPS29. This energy homeostasis is regulated by PPARalpha pathway of signal transduction. Through the PPAR/RXR heterodimetric complex, PPARalpha promotes the transport of long-chain fatty acids and retinoids X acids into mitochondrial that in turn stimulates the beta-oxidative degradation of fatty acids by the carnitine O-palmitoyltransferase I (CPTI) system [12, 13]. While the mitochondria devote organelles in the cells and the sites of metabolic enzyme-specific reactions in the respiratory chains, it is also the site where oxygen acts as the final electron receptor in an electron transport chain [12, 13]. In the presence of inflammatory muscle cells and tissues (IL6), we have found that this endurance training course has triggered mitochondrial oxidative stress (SOC2) causing mitochondrial dysfunction, reducing energy production and triggering the apoptosis programs (CYCS).

Finally, there is an indication of reactive oxygen species (ROS) excess triggers apoptosis (CYCS), leukocyte endothelial migration (ITGB1), leading edge cell differentiation (JUN), macrophage differentiation (MMP9), angiotensin catabolic process (ACE), and angiogenesis (PGF). To date, it is well known that ROS are a family of molecules and its derivatives are produced in all aerobic cells yielded from the metabolisms of molecular oxygen [14]. On the basis of our observations here, this training course has led to the deleterious effects of oxygen from the metabolic reduction of the highly reactive and toxic species, promoting endothelial damage or dysfunction and atherosclerosis [14]

4. Discussion

The Australian Institute of Sport (AIS) is a major center for swimming in Australia and trains approximately one-third of the swimmers for the Australian swimming team [2]. Nonetheless, considerable evidence suggests that athletes at AIS have suffered from a higher risk of upper respiratory, gastrointestinal and immune illnesses [15-31]. Therefore, many swimmers and coaches have expressed concern that high-level training loads increase risk of illness and infection [2]. Although research in the sport science and medicine at AIS develops at a slow rate [2], our approach presented here has made available the diagnostics of athletes' physiological conditions in response to training based on the 39 altered genes with cDNA microarrays.

A study of the gene expression model on endurance training–induced muscle contraction exhibits the presence of muscle development and contraction, the presence of energy, carbohydrate and lipid metabolisms, as well as the presence of ROS excess triggering endothelial damage or dysfunction and atherosclerosis. Nowadays, ROS excess is well known to be cytotoxic and has been implicated in the etiology of a wide array of human diseases, including cancer [14]. Various carcinogens may also partly exert their effect by generating ROS during their metabolism [14]. Oxidative damage to cellular DNA can lead to mutations and may, therefore, play an important role in the initiation and progression of multistage carcinogenesis [14]. The changes in DNA such as base modification, rearrangement of DNA sequence, miscoding of DNA lesion, gene duplication and the activation of oncogenes may be involved in the initiation of various cancers [14]. Elevated levels of ROS and down regulation of ROS scavengers and antioxidant enzymes are associated with various human diseases including various cancers [14]. ROS are also implicated in diabtes and neurodegenerative diseases [14]. Understanding the role of ROS as key mediators in signaling cascades may provide various opportunities for pharmacological intervention [14].

On the other hand, if we can scan an athlete's entire genome looking for all of the 39 altered genes that

predispose toward ROS excess, our approach as described has provided a comprehensive approach which accounts for the molecular events leading to the solution of the roots of training-induced ROS excess in the variation in the genome. With the genome sequence in hand, we can use genome-based diagnostics for athletes' physiological conditions, examine training-induced cellular reactions at the genetic level and, in time, develop treatments or cures for ROS excess on the basis of this understanding. In any case, the prospect of conquering the deleterious effect of oxygen from the metabolic reduction of the highly reactive and toxic species can now be entertained.

Overall, microarray-based gene expression signatures hold great promise to improve diagnostics of athletes' physiological conditions in response to training at the molecular levels. The success of this computational approach has given us confidence to "see" training-induced cellular processes with cDNA microarrays based on coaching necessities.

5. References

[1] Maglischo, E.W. *Swimming Fastest*. Illinois: Human Kinetics Press. 2002

[2] Colwin, C.M. *Breakthrough Swimming*. Illinois: Human Kinetics Press. 2002

[3] J.D. Watson. The Human Genome Project: Past, present, and future. *Science* 1990 **248:** 44-49.

[4] "Human Genome Report Press Release", International Consortium Completes Human Genome Project, [On-line], URL: http://www.ornl.gov/TechResources/Human_Genome/project/50yr.html

[5] NCBI. "Human Genome Resources," [On-line], URL: http://www.ncbi.nlm.nih.gov/genome/guide/human/

[6] I. Chen, L. Tseng, H. Homma, H Yan, and L Keith. Monitoring athletes' physiological response to endurance training with genome-wide expression data. *International Journal of Sport Science and Engineering*. **1:** 147-156

[7] M. Fluck, C. Dapp, S. Schmutz, E. Wit, and H. Hoppeler. Transcriptional profiling of tissue plasticity: role of shifts in genes expression and technical limitations. *Journal of Applied Physiology*. 2005, **99:** 397-413

[8] Everitt. B.S. *Statistical analysis using S-Plus.* Chapman and Hall, 1994

[9] Hartigan. J.A. *Clustering algorithm*. John Wiley and Sons, 1975

[10] Gordon. A.d. *Classification: methods for the exploratory analysis of multivariate data.* Chapman and Hall, 1981

[11] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Postein, Cluster analysis and display of genome-wide expression patterns. *Processing to National Academic of Science*. 1998, **95:** 14863-14868

[12] I. Takada, and S. Koto. PPARs target genes. *Nippon Rinsho.* 2005, **63:** 573-577

[13] N. Takahashi, T. Goto, T. Kusudo, T. Moriyama, and T. Kawada. The structures and functions of peroxisome proliferator-activated receptors (PPARs). *Nippon Rinsho*. 2005, **63:** 557-564

[14] W. Waris, and H. Ahsah. Reactive oxygen species: role in the development of cancer and various chronic conditions. *Carcinogenesis.* 2006, **5:** 1186-1192

[15] C. Hore. Important unusual infections in Australia: a critical care perspective. *Critical care and resuscitation: journal of the Australasian Academy of Critical Care Medicine.* 2001, **3:** 262-272

[16] M. Cleeson, D.B. Pyne, W.A. McDonald, S.J. Bowe, R.L. Clancy, and P.A. Fricker. In-vivo cell mediated immunity in elite swimmers in response to training. *Journal of science and medicine in sport / Sports Medicine Australia.* 2004, **7:** 38-46

[17] M. Gleeson, D.B. Pyne, J.P. Austin, J.L. Francis, R.L. Clancy, W.A. Mcdonald, and P.A. Fricker. Epstein-Barr virus reactivation and upper-respiratory illness in elite swimmers. *Medicine and science in sports and exercise.* 2002**, 34:** 411-417

[18] D.B. Pyne, W.A. McDonald, M. Gleeson, A. Flanagan, R.L. Clancy, and P.A. Fricker. Mucosal immunity, respiratory illness, and competitive performance in elite swimmers. *Medicine and science in sports and exercise*. 2001, **33:** 348-353

[19] M. Gleeson, D.B. Pyne. Special feature for the Olympics: effects of exercise on the immune system: exercise effects on mucosal immunity. *Immunology and cell biology.* 2000, **78:** 536-544

[20] M. Gleeson. Special feature for the Olympics: effects of exercise on the immune system. Overview: exercise immunology. *Immunology and cell biology*. 2000, **78:** 483-484

[21] M. Gleeson. Mucosal immunity and respiratory illness in elite athletes. *International journal of sports medicine.* 2000, **21(Suppl 1):** S33-S43

[22] M. Gleeson, S.T. Hall, W.A. McDonald, A.J. Flanagan and R.L. Clancy. Salivary IgA subclasses and infection risk in elite swimmers. *Immunology and cell biology.* 1999, **77:** 351-355

[23] P.A. Fricker, W.A. McDonald, M. Gleeson, and R.L. Clancy. Exercise-associated hypogammaglobulinemia. *Clinical journal of sport medicine : official journal of the Canadian Academy of Sport Medicine*. 1999, **9:** 46-48.

[24] M. Gleeson, W.A. McDonald, D.B. Pyne, A.W. Cripps, J.L. Francis, P.A. Fricker, and R.L. Clancy. Salivary IgA levels and infection risk in elite swimmers. *Medicine and science in sports and exercise.* 1999, **31:** 67-73

[25] L. Spence, W.J. Brown, D.B. Pyne, M.D. Nissen, T.P. Sloots, J.G. McCormack, A.S. Locke, and P.A. Fricker. Incidence, etiology, and symptomatology of upper respiratory illness in elite athletes. *Medicine and science in sports and exercise.*2007, **39:** 577-584

[26] M. Gleeson, W.A. McDonale, A.W. Cripps, D.B. Pyne, R.L. Clancy, and P.A. Fricker. The effect on immunity of long-term intensive training in elite swimmers. *Clinical and experimental immunology.* 1995, **102:** 210-216

[27] D.B. Pyne, M.S. Baker, P.A. Fricker, W.A. McDonald, R.D. Telford, and M.J. Weidemann. Effects of an intensive 12-wk training program by elite swimmers on neutrophil oxidative activity. *Medicine and science in sports and exercise.* 1995, **27:** 536-542

[28] M. Gleeson, W.A. McDonald, A.W. Cripps, D.B. Pyne, R.L. Clancy, P.A. Fricker, and J.H. Wlodarczyk. Exercise, stress and mucosal immunity in elite swimmers. *Advances in experimental medicine and biology.* 1995, **371A:** 571-574

[29] M. Gleeson, W.A. McDonald, D.B. Pyne, R.L. Clancy, A.W. Cripps, J.L. Francis, and P.A. Ficker. Immune status and respiratory illness for elite swimmers during a 12-week training cycle. *International journal of sports medicine*. 2000, **21:** 302-307

[30] M. Gleeson, S.T. Hall, W.A. McDonald, A.J. Flanagan, and R.L. Clancy. Salivary IgA subclasses and infection risk in elite swimmers. *Immunology and cell biology*. 1999, **77**: 351-355

[31] M. Gleeson, S.T. Hall, W.A. McDonald, A.J. Flanagan, and R.L. Clancy. Salivary IgA levels and infection risk in elite swimmers. *Medicine and science in sports and exercise.* 1999, **31:** 67-73