

REVIEW ARTICLE

Applying computation biology and “big data” to develop multiplex diagnostics for complex chronic diseases such as osteoarthritis

Guomin Ren¹ and Roman Krawetz^{1,2,3}

¹McCaig Institute for Bone & Joint Health, University of Calgary, Calgary, AB, Canada, ²Department of Surgery, University of Calgary, Calgary, AB, Canada, and ³Department of Anatomy and Cell Biology, University of Calgary, Calgary, AB, Canada

Abstract

The data explosion in the last decade is revolutionizing diagnostics research and the healthcare industry, offering both opportunities and challenges. These high-throughput “omics” techniques have generated more scientific data in the last few years than in the entire history of mankind. Here we present a brief summary of how “big data” have influenced early diagnosis of complex diseases. We will also review some of the most commonly used “omics” techniques and their applications in diagnostics. Finally, we will discuss the issues brought by these new techniques when translating laboratory discoveries to clinical practice.

Keywords

Chronic disease, computational biology, diagnostics, osteoarthritis

History

Received 15 December 2014
Revised 18 April 2015
Accepted 18 April 2015
Published online 10 December 2015

Computational techniques in early diagnosis

The ability to provide effective treatments in the early stages of a disease tends to lead to significantly better outcomes for the patient when compared with providing the same treatment at a significantly later stage of progression. This is particularly true for a number of diseases such as cancer and cardiovascular diseases, where any time lost can be a matter of life or death. However, early diagnosis of these diseases may be difficult using traditional biochemical methods due to their asymptomatic nature and the lack of efficient detective technologies. In the last decade, an exponential increase in the amount of data has been produced by various high-throughput “omic” technologies and we have now effectively entered the era of “big data”. Although requiring massive computational resources and advanced data processing and analysis methods, “big data” approaches to diagnostic medicine and biomarker development have been successfully applied to the early detection of complex chronic diseases. In some cases, this has given us a deeper understanding of the molecular pathogenesis of disease. In this review, we will discuss the current state of “big data” and computational techniques for early-stage disease diagnosis and how advances in these techniques may promote a better understanding of complex diseases.

“Big data” in disease diagnosis

Although great progress has been made within the last few decades, classical biomedical research methodology is still facing a challenge with diagnosis of complex diseases. These are typically associated with the effects of multiple genes in combination with lifestyle and environmental factors. One of the reasons for this difficulty in early diagnosis (or prediction) is that changes in traditional biomarkers can be too subtle at the asymptomatic stage to efficiently distinguish patients from normal individuals (Chen et al., 2012), and useful information can often be masked by the “noise” generated from naturally occurring variation within a given population. Therefore, many groups have suggested that diagnosis should be considered in a more comprehensive manner. Hampel et al. (2011) suggested that a combination of multiple biomarkers as well as genetic predisposition and environmental factors should all be taken into account for early diagnosis and personalized therapies of complex diseases such as Alzheimer’s disease. However, such studies require large-scale measurements on a large number of individuals to eliminate over-fitting of predictive models. With the development of high-throughput “omics” techniques and the reduction in prices per sample, these types of analyzes are now a reality. An enormous number of data have been generated, providing a global view with rich information on diseases and their diagnosis.

One of the largest projects is The Cancer Genome Atlas (TCGA, <http://cancergenome.nih.gov/>), which contains clinical information, histopathology slide images and molecular information from over 8000 tissue samples of 34 types of cancer. The goal of TCGA is to improve early detection of cancer and treatment by understanding how DNA mutations interact to drive cancers. However, the interpretation of such

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Address for correspondence: Dr. Roman Krawetz, McCaig Institute for Bone and Joint Health, Faculty of Medicine, University of Calgary, 3330 Hospital Drive NW, Calgary, AB, Canada T2N 4N1. Tel: +1 403 210 6268. E-mail: rkrawetz@ucalgary.ca

rich information seems to be a “big data problem”. Big data is a concept that varies in different fields. In biomedical research, “big data” essentially refers to computational analyses that help scientists make sense of the chaos of extremely large experimental and clinical data sets. Conceivably, big data are already impacting disease diagnosis. For example, by studying a large sample set, Chen et al. (2011) achieved considerable high specificity (98.9% and 91.9%) for non-invasive prenatal diagnosis of trisomy 13 and trisomy 18 using maternal plasma DNA sequencing.

Big data in disease diagnosis shares the same IT challenges as in other fields, including data storage, transfer, access control, and management (Marx, 2013; Schadt et al., 2010). Another challenge is the computational modeling of complex biology systems. Due to the large scale and diversity of the data, non-optimized models may fall into Non-deterministic Polynomial-time hard (NP hard) problems whose time complexity increases super-exponentially (Schadt et al., 2010). Moreover, sampling bias should not be neglected. According to the study of Kaplan et al. (2014), bigger data are not always better, since large sample studies sometimes can magnify biases associated with error resulting from sampling or study design.

Computational “omics” techniques

Diseases with an identifiable genetic component play a role in nine of the ten leading causes of death in the United States (Hoyert & Xu, 2012). A positive association between genetic variation and disease may not only help diagnose diseases at an early stage but also predict disease onset before the initiation of pathogenesis. Genome-wide association study (GWAS) is one of the most common statistical approaches that involves rapidly scanning millions of markers (single-nucleotide polymorphisms, SNPs) at the same time across genomes to find genetic variations associated with a common complex disease (Visscher et al., 2012; Wellcome Trust Case Control, 2007). Liu et al. (2014) reported that the inclusion of the GWAS genetic variants data significantly improved their breast cancer naïve Bayes diagnostic model. As technological improvements continue to decrease DNA sequencing costs, whole genome sequencing (WGS) or whole exome sequencing (WES, sequence protein-coding genes only) becomes more practical for clinical applications and might be a potential alternative to GWAS as it provides more information on whole genomes (Berg et al., 2011). However, WGS/WES generates large quantities of data that require tremendous computational capacity for analysis such as sequence alignment, variant calling, filtering, and identifying disease susceptibility genes. In fact, sequence data are produced significantly faster than current computational resources can handle (Stein, 2010). Thus, more efficient algorithms and/or more powerful hardware need to be developed in the future (Ding et al., 2014). However, this may lead to an “arms race” between hardware and software resulting in increased rates of obsolescence in the field. Therefore, it is clear that data acquisition (hardware) and analysis (software) cannot be pursued independently of each other.

Gene expression (transcriptomics) profiling provides an opportunity for accurate, definitive diagnosis (Wen et al., 2013; Wiseman et al., 2013). High-throughput mRNA sequencing (RNA-Seq) is one of the most popular techniques in

transcriptomics since this technology allows for investigating both known transcripts and uncovering new ones. Since transcripts (RNAs) need to be converted to cDNA and then sequenced, RNA sequence assembly algorithms for short, low-quality reads without references are required (Martin & Wang, 2011). While microarray suffers from a number of limitations compared with RNA-Seq (e.g. unbiased detection of transcripts, increased dynamic range, increased specificity/sensitivity, and increased detection of rare/low-abundance transcripts), it can be used to measure large numbers of gene expression levels simultaneously. In addition to regular clinical diagnosis, many recent articles reported the success of applying microarray in prenatal diagnosis (Shaffer et al., 2012; Wapner et al., 2012).

Proteomics can also be used for the biomarker detection of early-stage disease such as cancer (Mehrotra & Gupta, 2011; Rahman et al., 2011), cardiovascular disease (Delles et al., 2010; Gerszten et al., 2011), Alzheimer’s disease (Craig-Schapiro et al., 2011), and other chronic diseases (Good et al., 2010; Zurbig et al., 2012). Mass spectrometry (MS)-based proteomics can help identify all differentially expressed proteins and their post-translational modifications during disease progression that can be used as biomarkers for early diagnosis and monitoring disease treatment (Colinge & Bennett, 2007). The data process of MS relies heavily on open access public proteomics databases. Both our own group and others in the field have employed the use of high-throughput ELISA technology such as Luminex and Meso Scale to examine panels of proteins (typically numbering between 20 and 60) in chronic diseases such as osteoarthritis (Heard et al., 2013) and traumatic injuries (Helmy et al., 2012).

Metabolomics, while a younger field than the rest, is rapidly expanding in the diagnostics field in “post-genomic era”. Metabolic characteristics and changes in patients are influenced not only by which genes are transcribed, but also the composition of material that the cells obtain from their micro-environment. Many reviews have discussed the application of metabolomics in diagnostics using high-throughput techniques such as nuclear magnetic resonance spectroscopy (NMR) and MS. Madsen et al. (2010) made a comprehensive summary of metabolomics in cancer, diabetes, cardiovascular, and other complex disease diagnosis. Zhang et al. (2012) pointed out that saliva metabolomics is a potential method for personalized therapy and treatment monitoring. However, the type of data analysis is crucial for metabolomics-based diagnosis: in some cases, one single marker from the metabolic profile might be sufficient to detect the disease specifically, in most cases, machine learning techniques are applied to recognize and classify metabolic profiles or fingerprints between normal and disease states. The most widely used are linear discriminant analysis (LDA), artificial neural networks (ANN), and support vector machines (SVM). Principal component analysis (PCA) is often employed for data dimension reduction before model training in order to lower the chance of over-fitting the model. Another way to avoid model over-fitting is to apply cross-validation techniques at the model training step.

System biology

Not until the completion of the Human Genome Project was it realized that gene sequence alone was insufficient to identify

all the biologic origin of a disease. The function of each protein and the complexities of protein–protein interactions are critical for understanding physiological processes. In addition, recent studies show that non-coding parts of the genome produce small conserved ribonucleic acids (non-coding RNA, ncRNA) that control molecular and cellular processes (Alexander et al., 2010; Tutar, 2012). Thus, in order to develop effective diagnostic techniques and disease treatments, genomics, transcriptomics, and proteomics should be studied integrally and systematically as a whole system.

Through a system-based approach, Lusi et al. integrated genomic, molecular, physiological data with traditional genetic and biochemical methods to study complex disease including diabetes and cardiovascular disease. He pointed out that analyzing the individual components of the whole system is far from sufficient, since in reality, these components interact with each other and these interactions play crucial roles in development of diseases (Lusi et al., 2008).

A number of recent studies have successfully applied network models in describing and simplifying such complex systems (Akutekwe & Seker, 2014; Barabási et al., 2011; Gilman et al., 2011; O’Roak et al., 2012; Vandin et al., 2011, Vidal et al., 2011). In these studies, network topology is used to investigate biological networks including metabolic networks, protein–protein interaction networks, gene regulatory networks, transcriptional profiling networks, etc., and their interactions. For example, in the gene network clusters created by Gilman et al. (2011) using NETBAG (network-based analysis of genetic associations), many proteins are found to participate in the formation of autism. These proteins may become new biomarkers for the diagnosis of autism. In another study conducted by Akutekwe & Seker (2014), a biomarker identification method used a dynamic Bayesian network to model the temporal relationship among stratified features for early diagnosis of ovarian cancer. Gstaiger et al. tried to bridge the gap between genotype and phenotype by studying the inference of genetically perturbed molecular networks based on a combination of genomics, proteomics, and phenomics data (Gstaiger & Aebersold, 2009). All these innovative strategies may provide a deeper understanding of disease development and help us discover new indicators for early-stage diseases.

Early diagnosis of osteoarthritis

Osteoarthritis (OA), one of the leading causes of chronic disability worldwide, is a form of arthritic disease characterized by the progressive destruction of articular cartilage. The pathogenesis of OA is multifactorial: aging, injury, and genetic predisposition may all be contributing factors that cause joint cartilage degeneration. Currently, clinical diagnosis of OA relies on radiographic assessment, pain symptoms, and mobility of the joint. Unfortunately, OA develops asymptotically its early stages and when it becomes detectable, extensive and irreversible deterioration of joint has already occurred. Therefore, there is a need for new diagnostic methods, such as new specific biological markers, to detect OA at before such deterioration happens. However, without understanding the biological mechanisms of OA, the search for effective early biomarkers among billions of molecules is like finding a “needle in a haystack”. In the past few years,

development “omics” and bioinformatics techniques have impacted the etiology and diagnostics of complex diseases like OA. High-throughput fast screening of biomarkers at the whole “omic” level becomes a reality. As an example, we describe recent progress and challenges in early-stage OA diagnosis using these high-throughput techniques.

Genomics in OA diagnosis

Genome-wide association studies have examined thousands of SNPs in the whole genome and OA. So far, approximately 15 OA susceptibility loci have been identified by GWAS, although some of them are gender or racial specific (Tsezou, 2014). Elliott et al. (2013) found significant overlap between OA and height and OA and body mass index (BMI) by comparing OA and BMI GWAS data, suggesting that OA and obesity may share genetic background. In a more comprehensive meta-analysis study, Rodriguez-Fontenla et al. (2014) summarized nine GWAS of OA, they identified two genes (*COL11A1* and *VEGF*) that are significantly associated with hip OA development.

In order to find the rare variants that are missed in common GWAS studies, Boer et al. conducted a whole exome-sequencing study of 1524 participants, of whom 199 had hip OA. Besides three genes already identified in previous GWAS studies, they found that gene fibroblast growth factor 3 (*FGF3*) may contribute to hip OA by suppressing endochondral bone formation (Boer et al., 2014). Unfortunately, to our knowledge, this is the only OA-related whole genome/exome sequencing study published to date. To obtain a better understanding of the genomic architecture of OA, additional whole genome large-scale NGS studies on various cohorts should be undertaken.

Several recent genome-wide DNA epigenetic studies using high-throughput arrays have revealed new potential OA biomarkers. DNA methylation (one of the common DNA epigenetic modifications in promoter regions of genomic DNA) may influence DNA stability, chromatin structure, and regulate gene expression. Several studies have examined the genome-wide DNA methylation profile of human articular chondrocytes in cartilage and trabecular bone samples from OA patients and healthy controls to identify profiles of DNA methylation in OA disease (Delgado-Calle et al., 2013; Fernandez-Tajes et al., 2014). All these studies found significant differential methylation levels in certain genes between the patient and normal groups, and it is possible that these methylation sites and the genes in which they are contained could be used as new diagnostic markers for OA.

Transcriptomics in OA diagnosis

Many microarray-based gene expression studies on various tissue types from OA patients have identified differentially expressed genes and profiles that could contribute to the development of new biomarkers. For example, Blom et al. (2014) identified approximately 200 differentially expressed genes (fold change $\geq \pm 2$) in synovium, whereas in peripheral blood, 86 genes were expressed with at least 1.5-fold difference (Ramos et al., 2014). As increased evidence indicates that the subchondral bone plays a major role in the initiation and progression of OA, Chou et al. (2013) performed a

whole-genome gene expression study of subchondral bone. They found a total of 972 genes that were differentially expressed (fold change $\geq \pm 2$) between normal and OA bone samples. Interestingly, these studies identified only very few of the same differentially expressed genes, suggesting that in OA, disease-related gene expression changes with time, or may be highly tissue and/or patient specific. Although a few molecular models can explain a small portion of tissue-dependent gene expression regulation, the full regulation mechanisms in different tissues are not clear (Fu et al., 2012). Nevertheless, it is essential that we consider the complex (and in some cases, non-canonical) roles of genes and their pathways in diverse tissue and cell types. Hence, it is important that different studies use expression data from the same tissues to maintain comparability and assess the association between genes and disease.

Proteomics and metabolomics in OA diagnosis

Although proteomics and metabolomics approaches in OA diagnostic studies are relatively new, they have already identified a great number of potential disease biomarkers. A broad range investigation of proteomic profiles in different tissues has been conducted, including femoral head, humeral head, meniscus, explants, etc. (Hsueh et al., 2014). Additional studies are more focused on human body fluid as the harvest is comparatively non-invasive and consequently easier to translate to clinical practice. Serum and urine are the most commonly used body fluids for proteomic analysis of OA (Takinami et al., 2013). However, since they are spatially removed from the affected tissues it is possible that some key proteins may be diluted. Synovial fluid (SF), although sometimes difficult to obtain, can be studied as a compromise between non-invasiveness and sensitivity (Balakrishnan et al., 2014). A metabolomics analysis of synovial fluid has successfully classified OA phenotypes into two metabolically distinct subgroups using the concentration of acylcarnitine, which may be related to the carnitine metabolism pathway (Zhang et al., 2014). These types of studies will help to unravel the complex pathogenesis of OA and simplify new biomarker discovery by dividing OA into several subtypes.

A problem with proteomic and metabolomic studies of early OA is that abnormal protein or metabolite expression is relatively dynamic compared with gene mutation. Usually samples are obtained from patients who are already clinically diagnosed with OA; therefore, the proteomic and metabolomics profiles can only represent the status of the patients at the advanced or even end stage of the disease. Without knowing the biomarker profile changes during OA progression, we should be careful in assuming that differentially expressed proteins or metabolites in late OA are also potential biomarkers for early OA diagnosis. Takinami et al. (2013) conducted a study which followed knee OA patients for 2 years to overcome this problem. However, OA is known to have a much longer pathogenic in some patients (even up to decades), and some evidence shows that cartilage degeneration which could ultimately lead to OA can start in youth. Therefore, it is essential to develop long term follow-up studies now, so that the next generation will be able to benefit from these types of diagnostic studies in OA.

System biology in OA diagnosis

Extensive “omics” data have been screened so far and many biomarkers have been proposed, but their sensitivity or specificity is not high enough for clinical use and the reliability varies among studies (Table 1). One possible explanation for this is the multifactorial pathogenesis of OA: aging, injury, and genetic predisposition may all act as contributing factors, and consequently, single biomarker diagnostics are not efficient enough to comprehensively classify all early-stage OA patients of various etiologies. Although system biology is an effective technique for complex disease research, very few studies have been conducted on OA. Olex et al. (2014) integrated time-course microarray gene expression data from a mouse model into a PPI network. However, mice are known to have a much different genetic response than humans following an injury, and mouse models might be poor representatives of the human inflammatory response (Seok et al., 2013). Nacher et al. (2014) applied a PageRank-based diffusion algorithm to recognize OA-related proteins in a chondrocyte protein network and found that protein Q6EEV6 could play a key role in OA development. In another similar study, some of the top hub genes in the PPI network are also differentially expressed, indicating that these genes may be potential targets for OA diagnosis and treatment (Wang et al., 2014).

All these studies share some common limitations. First, further genetic and experimental studies are needed to eliminate the possibility of false positive results from computational analysis. Second, all these studies are trying to find one or several biomarkers, which departs from the original purpose of system biology study in complex disease: to study complex intracellular and intercellular networks as a whole. Lack of effective methods to interpret biologic network results might be one of the reasons. Pilot works are needed to put computational analysis into perspective in the future.

Importance of patient cohort characterization

Although high-throughput “omics” platforms coupled with the application of complex bioinformatics approaches have had a number of successes in identifying potential biomarkers in complex diseases such as cancer (Wang et al., 2009; Zhang et al., 2013), sepsis (Lukaszewski et al., 2008), arthritis (Heard et al., 2014; Swan et al., 2013) and others, it is important to realize that some, if not all complex diseases have numerous associated co-morbidities and risk factors. Therefore, it is essential to have extremely well-characterized patient cohorts to be sure we are not identifying biomarkers associated with those co-morbidities and/or risk factors. This is particularly important in diseases where no early diagnostic tests exist to assist in the confirmation/validation of the novel biomarkers.

Conclusion

The high-throughput “omics” techniques bring new energy to diagnostics, offering a comprehensive data resource from micro (e.g. genomics) to macro (e.g. phenomics). Facing the “big data” generated by such techniques, more powerful computational resources and efficient models or algorithms are needed for data storage, transfer, and mining. Systems biology is one of the most successful methods for studying

Table 1. Biomarker approaches examined for the diagnosis of osteoarthritis.

Publication name	Publish date	Biochemical biomarkers	Body fluid	Sample size	Age	Gender	Comparison group	Statistical method	Sensitivity/ specificity
Role of hyaluronic acid in early diagnosis of knee osteoarthritis (Singh et al., 2014a)	2014.12	Hyaluronic acid	Serum	100/50 (cases/controls)	Case 51.28 ± 7.93 Control 46.08 ± 4.81 <i>p</i> < 0.001	Case: 34 males, 66 females; Control: 16 males 34 females	Normal/mild OA	ROC curve	87.6%/86.0%
Identification and characterization of osteoarthritis patients with inflammation derived tissue turnover (Siebuhr et al., 2014)	2014.1	hsCRP, CRPM	Serum	12/202/57/60 (mild/moderate/ severe/TKR)	Mean: 65.3	155 Males; 176 females	Mild/moderate/ severe/TKR	ROC curve	33.3%/87.18%
TSG-6 activity as a novel biomarker of progression in knee osteoarthritis (Wisniewski et al., 2014)	2014.2	TSG-6	Synovial fluid, serum	91 (OA)			High/low risk of rapid progression to end-stage OA	ROC curve	91%/82%
A computational method to differentiate normal individuals, osteoarthritis and rheumatoid arthritis patients using serum biomarkers (Heard et al., 2014)	2014.6	38 cytokines	Serum	100/100 (cases/controls)	Normal: 40.0 ± 9.5 OA: 60.4 ± 10		Normal/OA	Artificial neural network	100%/100%
Serum Cartilage Oligomeric Matrix Protein: tool for early diagnosis and grading of severity of primary knee osteoarthritis (Singh et al., 2014b)	2014.8	COMP	Serum	100/50 (cases/controls)	Cases: 58.93 ± 9.20 Control: 54.68 ± 9.68	50 Males; 100 females	Normal/OA (KL = 0/KL ≥ 1)	ROC curve	98%/98%
Usefulness of urinary CTX-II and NTX-I in evaluating radiological knee osteoarthritis: the Matsudai knee osteoarthritis survey (Tamishi et al., 2014)	2014.2	uCTX-II	Urine	794/616 (OA/normal)	Males: KL0: 1:65.3 ± 9.7 KL2: 72.3 ± 5.3 KL3: 72.5 ± 4.9 KL4: 74.7 ± 3.2 Females KL0: 1: 60.7 ± 9.5 KL2: 70.3 ± 6.0 KL3: 71.6 ± 6.1 KL4: 72.9 ± 5.1	435 Males; 605 females	KL0, 1/KL ≥ 2	ROC curve	≈ 60%/40% men ≈ 70%/40 women
Cartilage Oligomeric Matrix Protein – inflammation biomarker in knee osteoarthritis (Zivanovic et al., 2011)	2011.2	COMP	Serum	66/22 (effusion/control)	69.97 ± 9.37	20 Males; 68 females	With/without effusion (inflammation)	ROC Curve	59%/50%
Identification of Osteoarthritis Biomarkers by	2012.12	S100A12 and other two	Synovial fluid	36/24 (OA/RA)	OA: 70.2 ± 5.4 RA: 68.4 ± 4.9	OA: 9 males, 27 females	OA/RA	Artificial neural network	89.4%/91.2%

(continued)

Table 1. Continued

Publication name	Publish date	Biochemical biomarkers	Body fluid	Sample size	Age	Gender	Comparison group	Statistical method	Sensitivity/specificity
Proteomic Analysis of Synovial Fluid (Han et al., 2012) Fibulin 3 peptides Fib3-1 and Fib3-2 are potential biomarkers of osteoarthritis (Henrotin et al., 2012)	2012.07	unknown proteins Fib3-1, Fib3-2	Urine	10/5 (cases/controls)	Cases: 76.0 ± 5.0 Control: 25.6 ± 2.6	RA: 4 males, 20 females 15 Females	Normal/late OA	ROC curve	Fib3-1: 68.4%/ 77.1%, Fib3-2: 74.6%/ 85.7

biologic processes and integrating multiple data resources. Many studies have applied network models in describing etiopathogenesis and immune responses that may help the discovery of novel biomarkers for early diagnosis. However, we should be careful when applying such models, especially when there is uncertainty regarding the bias of clinical data and no other diagnostic tests are available for validation.

Acknowledgements

The authors would like to thank Catherine Leonard for her assistance in copyediting the document.

Declaration of interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

References

- Akutekwe A, Seker H. (2014). Two-stage computational bio-network discovery approach for metabolites: ovarian cancer as a case study. 2014 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), 2014. Valencia: IEEE, 97–100.
- Alexander RP, Fang G, Rozowsky J, et al. (2010). Annotating non-coding regions of the genome. *Nat Rev Genet* 11:559–71.
- Balakrishnan L, Nirujogi RS, Ahmad S, et al. (2014). Proteomic analysis of human osteoarthritis synovial fluid. *Clin Proteomics* 11:6.
- Barabasi AL, Gulbahce N, Loscalzo J. (2011). Network medicine: a network-based approach to human disease. *Nat Rev Genet* 12:56–68.
- Berg JS, Khoury MJ, Evans JP. (2011). Deploying whole genome sequencing in clinical practice and public health: meeting the challenge one bin at a time. *Genet Med* 13:499–504.
- Blom A, van Lent P, van den Bosch M, et al. (2014). THU0455 transcriptomics to identify synovial genes and pathways associated with disease progression in a cohort of early osteoarthritis patients (CHECK). *Ann Rheum Dis* 73:340–1.
- Boer CG, Rooij JV, Peters M, et al. (2014). Discovery and analysis of rare coding variants for hipOA by exome-sequencing. *Osteoarthritis Cartilage* 22:S226–7.
- Chen EZ, Chiu RW, Sun H, et al. (2011). Noninvasive prenatal diagnosis of fetal trisomy 18 and trisomy 13 by maternal plasma DNA sequencing. *PLoS One* 6:e21791.
- Chen L, Liu R, Liu ZP, et al. (2012). Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Sci Rep* 2:342.
- Chou CH, Wu CC, Song IW, et al. (2013). Genome-wide expression profiles of subchondral bone in osteoarthritis. *Arthritis Res Ther* 15:R190.
- Clinge J, Bennett KL. (2007). Introduction to computational proteomics. *PLoS Comput Biol* 3:e114.
- Craig-Schapiro R, Kuhn M, Xiong C, et al. (2011). Multiplexed immunoassay panel identifies novel CSF biomarkers for Alzheimer's disease diagnosis and prognosis. *PLoS One* 6:e18850.
- Delgado-Calle J, Fernandez AF, Sainz J, et al. (2013). Genome-wide profiling of bone reveals differentially methylated regions in osteoporosis and osteoarthritis. *Arthritis Rheum* 65:197–205.
- Delles C, Schiffer E, von Zur Muhlen C, et al. (2010). Urinary proteomic diagnosis of coronary artery disease: identification and clinical validation in 623 individuals. *J Hypertension* 28:2316–22.
- Ding L, Wendl MC, McMichael JF, Raphael BJ. (2014). Expanding the computational toolbox for mining cancer genomes. *Nat Rev Genet* 15:556–70.
- Elliott KS, Chapman K, Day-Williams A, et al. (2013). Evaluation of the genetic overlap between osteoarthritis with body mass index and height using genome-wide association scan data. *Ann Rheum Dis* 72:935–41.
- Fernandez-Tajes J, Soto-Hermida A, Vazquez-Mosquera ME, et al. (2014). Genome-wide DNA methylation analysis of articular chondrocytes reveals a cluster of osteoarthritic patients. *Ann Rheum Dis* 73:668–77.

- Fu J, Wolfs MG, Deelen P, et al. (2012). Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet* 8:e1002431.
- Gerszten RE, Asnani A, Carr SA. (2011). Status and prospects for discovery and verification of new biomarkers of cardiovascular disease by proteomics. *Circ Res* 109:463–74.
- Gilman SR, Iossifov I, Levy D, et al. (2011). Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* 70:898–907.
- Good DM, Zurbig P, Argiles A, et al. (2010). Naturally occurring human urinary peptides for use in diagnosis of chronic kidney disease. *Mol Cell Proteomics* 9:2424–37.
- Gstaiger M, Aebersold R. (2009). Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat Rev Genet* 10:617–27.
- Hampel H, Prvulovic D, Teipel S, et al. (2011). The future of Alzheimer's disease: the next 10 years. *Prog Neurobiol* 95:718–28.
- Han MY, Dai JJ, Zhang Y, et al. (2012). Identification of osteoarthritis biomarkers by proteomic analysis of synovial fluid. *J Int Med Res* 40:2243–50.
- Heard BJ, Fritzler MJ, Wiley JP, et al. (2013). Intraarticular and systemic inflammatory profiles may identify patients with osteoarthritis. *J Rheumatol* 40:1379–87.
- Heard BJ, Rosvold JM, Fritzler MJ, et al. (2014). A computational method to differentiate normal individuals, osteoarthritis and rheumatoid arthritis patients using serum biomarkers. *J R Soc Interface* 11:20140428.
- Helmy A, Antoniadou CA, Guilfoyle MR, et al. (2012). Principal component analysis of the cytokine and chemokine response to human traumatic brain injury. *PLoS One* 7:e39677.
- Henrotin Y, Gharbi M, Mazzucchelli G, et al. (2012). Fibulin 3 peptides Fib3-1 and Fib3-2 are potential biomarkers of osteoarthritis. *Arthritis Rheum* 64:2260–7.
- Hoyert DL, Xu J. (2012). Deaths: preliminary data for 2011. *Natl Vital Stat Rep* 61:1–51.
- Hsueh MF, Onnerfjord P, Kraus VB. (2014). Biomarkers and proteomic analysis of osteoarthritis. *Matrix Biol* 39:56–66.
- Kaplan RM, Chambers DA, Glasgow RE. (2014). Big data and large sample size: a cautionary note on the potential for bias. *Clin Transl Sci* 7:342–6.
- Liu J, Page D, Peissig P, et al. (2014). New genetic variants improve personalized breast cancer diagnosis. *AMIA Jt Summits Transl Sci Proc* 2014:83–9.
- Lukaszewski RA, Yates AM, Jackson MC, et al. (2008). Presymptomatic prediction of sepsis in intensive care unit patients. *Clin Vaccine Immunol* 15:1089–94.
- Lusis AJ, Attie AD, Reue K. (2008). Metabolic syndrome: from epidemiology to systems biology. *Nat Rev Genet* 9:819–30.
- Madsen R, Lundstedt T, Trygg J. (2010). Chemometrics in metabolomics – a review in human disease diagnosis. *Anal Chim Acta* 659:23–33.
- Martin JA, Wang Z. (2011). Next-generation transcriptome assembly. *Nat Rev Genet* 12:671–82.
- Marx V. (2013). Biology: the big challenges of big data. *Nature* 498:255–60.
- Mehrotra R, Gupta DK. (2011). Exciting new advances in oral cancer diagnosis: avenues to early detection. *Head Neck Oncol* 3:1–9.
- Nacher JC, Keith B, Schwartz J-M. (2014). Network medicine analysis of chondrocyte proteins towards new treatments of osteoarthritis. *Proc R Soc B: Biol Sci* 281:20132907.
- O'Roak BJ, Vives L, Girirajan S, et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 485:246–50.
- Olex AL, Turkett WH, Fetrow JS, Loeser RF. (2014). Integration of gene expression data with network-based analysis to identify signaling and metabolic pathways regulated during the development of osteoarthritis. *Gene* 542:38–45.
- Rahman SJ, Gonzalez AL, Li M, et al. (2011). Lung cancer diagnosis from proteomic analysis of preinvasive lesions. *Cancer Res* 71:3009–17.
- Ramos YF, Bos SD, Lakenberg N, et al. (2014). Genes expressed in blood link osteoarthritis with apoptotic pathways. *Ann Rheum Dis* 73:1844–53.
- Rodriguez-Fontenla C, Calaza M, Evangelou E, et al. (2014). Assessment of osteoarthritis candidate genes in a meta-analysis of nine genome-wide association studies. *Arthritis Rheumatol* 66:940–9.
- Schadt EE, Linderman MD, Sorenson J, et al. (2010). Computational solutions to large-scale data management and analysis. *Nat Rev Genet* 11:647–57.
- Seok J, Warren HS, Cuenca AG, et al. (2013). Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc Natl Acad Sci USA* 110:3507–12.
- Shaffer LG, Dabell MP, Fisher AJ, et al. (2012). Experience with microarray-based comparative genomic hybridization for prenatal diagnosis in over 5000 pregnancies. *Prenat Diagn* 32:976–85.
- Siebuhr AS, Petersen KK, Arendt-Nielsen L, et al. (2014). Identification and characterisation of osteoarthritis patients with inflammation derived tissue turnover. *Osteoarthritis Cartilage* 22:44–50.
- Singh S, Kumar D, Sharma NR. (2014a). Role of hyaluronic acid in early diagnosis of knee osteoarthritis. *J Clin Diagn Res* 8:LC04–7.
- Singh S, Shahi U, Kumar D, Shahi NT. (2014b). Serum Cartilage Oligomeric Matrix Protein: tool for early diagnosis and grading of severity of primary knee osteoarthritis. *Int J Osteol Orthop* 1:1–7.
- Stein LD. (2010). The case for cloud computing in genome informatics. *Genome Biol* 11:207.
- Swan AL, Hillier KL, Smith JR, et al. (2013). Analysis of mass spectrometry data from the secretome of an explant model of articular cartilage exposed to pro-inflammatory and anti-inflammatory stimuli using machine learning. *BMC Musculoskelet Disord* 14:349.
- Takinami Y, Yoshimatsu S, Uchiumi T, et al. (2013). Identification of potential prognostic markers for knee osteoarthritis by serum proteomic analysis. *Biomark Insights* 8:85–95.
- Tanishi N, Yamagiwa H, Hayami T, et al. (2014). Usefulness of urinary CTX-II and NTX-I in evaluating radiological knee osteoarthritis: the Matsudai knee osteoarthritis survey. *J Orthop Sci* 19:429–36.
- Tsezou A. (2014). Osteoarthritis year in review 2014: genetics and genomics. *Osteoarthritis Cartilage* 22:2017–24.
- Tutar Y. (2012). Pseudogenes. *Comp Funct Genomics* 2012:424526.
- Vandin F, Upfal E, Raphael BJ. (2011). Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol* 18:507–22.
- Vidal M, Cusick ME, Barabasi A-L. (2011). Interactome networks and human disease. *Cell* 144:986–98.
- Visscher PM, Brown MA, McCarthy MI, Yang J. (2012). Five years of GWAS discovery. *Am J Hum Genet* 90:7–24.
- Wang HQ, Wong HS, Zhu H, Yip TT. (2009). A neural network-based biomarker association information extraction approach for cancer classification. *J Biomed Inform* 42:654–66.
- Wang Q, Li Y, Zhang Z, et al. (2014). Bioinformatics analysis of gene expression profiles of osteoarthritis. *Acta Histochem* 117:40–6.
- Wapner RJ, Martin CL, Levy B, et al. (2012). Chromosomal microarray versus karyotyping for prenatal diagnosis. *N Engl J Med* 367:2175–84.
- Wellcome Trust Case Control, C. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447:661–78.
- Wen T, Stucke EM, Grotjan TM, et al. (2013). Molecular diagnosis of eosinophilic esophagitis by gene expression profiling. *Gastroenterology* 145:1289–99.
- Wiseman SM, Haddad Z, Walker B, et al. (2013). Whole-transcriptome profiling of thyroid nodules identifies expression-based signatures for accurate thyroid cancer diagnosis. *J Clin Endocrinol Metab* 98:4072–9.
- Wisniewski HG, Colon E, Liublinska V, et al. (2014). TSG-6 activity as a novel biomarker of progression in knee osteoarthritis. *Osteoarthritis Cartilage* 22:235–41.
- Zurbig P, Jerums G, Hovind P, et al. (2012). Urinary proteomics for early diagnosis in diabetic nephropathy. *Diabetes* 61:3304–13.
- Zhang A, Sun H, Wang X. (2012). Saliva metabolomics opens door to biomarker discovery, disease diagnosis, and treatment. *Appl Biochem Biotechnol* 168:1718–27.
- Zhang F, Chen J, Wang M, Drabier R. (2013). A neural network approach to multi-biomarker panel discovery by high-throughput plasma proteomics profiling of breast cancer. *BMC Proc* 7:S10.
- Zhang W, Likhodii S, Zhang Y, et al. (2014). Classification of osteoarthritis phenotypes by metabolomics analysis. *BMJ Open* 4:e006286.
- Zivanovic S, Rackov LP, Zivanovic A, et al. (2011). Cartilage oligomeric matrix protein-inflammation biomarker in knee osteoarthritis. *Bosn J Basic Med Sci* 11:27–32.