Array CGH and Computational Genome Annotation in Constitutional Cytogenetics: Suggesting Candidate Genes for Novel Submicroscopic Chromosomal Imbalance Syndromes

STEVEN VAN VOOREN[1,*], BERT COESSENS[1], BART DE MOOR[1], YVES MOREAU[1], JORIS R. VERMEESCH[2]

[1]Department of Electrotechnical Engineering, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium,
[2]Center for Human Genetics, Leuven University Hospital, Herestraat 49, B-3000 Leuven
[*] to whom correspondence should be addressed. Steven.VanVooren@esat.kuleuven.be; +3216328654

Keywords: clinical informatics; cytogenetics; molecular diagnostics; array CGH; data mining; genotype-phenotype correlation

Abstract

Genome wide array CGH screening is uncovering pathogenic submicroscopic chromosomal imbalances in patients with developmental disorders. In those patients, imbalances appear now to be scattered across the whole genome and most patients carry different chromosomal anomalies. Screening patients with developmental disorders can be considered a forward functional genome screen. The imbalances pinpoint the location of genes that are involved in human development. Since most imbalances encompass regions harboring multiple genes, the challenge is (1) to identify those genes responsible for the specific phenotype and (2) to disentangle the role of the different genes located in an imbalanced region. In this review, we discuss novel tools and relevant databases that have recently been developed to aid this gene discovery process. Identification of the functional relevance of genes will not only deepen our understanding of human development but will, in addition, aid in the data interpretation and improve genetic counseling.

INTRODUCTION

Array CGH is used increasingly often as a primary genetic screening method in diagnosis and research [39, 12, 36, 34]. The technique is uncovering pathogenic submicroscopic chromosomal imbalances in patients with developmental disorders. Most patients carry different chromosomal anomalies, and anomalies occur across the whole genome [30, 24, 17, 8, 38]. These imbalances pinpoint the location of genes that are involved in human development [49]. Since most imbalances encompass regions harboring multiple genes, the challenge is (1) to identify those genes responsible for the specific phenotype and (2) to disentangle the role of the different genes located in an imbalanced region.

The high resolution at which array CGH has been used to define candidate regions for putative genes responsible for human genetic diseases is instrumental in defining and refining the critical region for a disease or phenotype and reducing the number of candidate genes for (an aspect of) the phenotype [39]. This higher resolution has led to a dramatic increase in gene identification through molecular karyotyping, and it is likely that the function of many more genes will be identified in this way [7].

However, some specific challenges apply to correlating genotype and phenotype in the context of human disease. Firstly, it is clear that etiology of rare chromosomal imbalances greatly benefits from large scale efforts in collection and organization of case reports from different genetic testing and research centers around the world. Especially for rare diseases, the need for large and well annotated case report resources is obvious.

Secondly, the identification of critical genes and pathways involved in a disease or biological process is helped by interpreting aberrations within the context of broader knowledge [34]. In understanding the functional basis of genetic conditions, it is therefore instrumental to incorporate information from different sources, other than mere genotype and phenotype information present in case reports. Integration of publicly available data sources pertaining to genome and gene function permits the development of bioinformatics methods for candidate gene selection. Such information sources range from the large corpus of biomedical literature to protein-protein interaction, pathway, and genome annotation databases in general.

In this review, we discuss relevant databases that have recently been developed to elucidate the role of genes in different aspects of the phenotype. We go on by giving an overview of published methods and tools that can help in the gene discovery process. Finally, we identify relevant issues in management and use of genotype-phenotype databases, and elaborate on issues encountered when annotating phenotypic characteristics to patient case reports ('phenotyping').

PUBLIC DATABASES

A group of phenotypically related cases can be used to delineate a minimal genomic region that segregates with a clearly defined common part of the phenotype. Through such correlation of components of a phenotype with the loci or genes within the affected chromosomal region, novel clinical entities can be defined. In order for this to be possible on a large scale, tools and databases are needed. Databases need to be extensive and publicly accessible, and computational approaches need to be compatible with these databases. Both are necessary tools in large scale studies for association of phenotypic information with genomic data.

Collaborative databases of case reports have been set up in support of discovering new clinical entities such as deletion and duplication syndromes, and correlating aberrant genotypes with phenotypes. Both global and local case repositories exist; some initiatives are closed or consortium-based while others are public. Although these initiatives differ in approach and setup, they share the common goal of supporting association studies and efforts in delineating novel syndromes by aggregating patient case reports, and in most cases, encouraging data exchange. DECIPHER and ECARUCA are considered the two most important databases for constitutional cytogenetics. An overview of chromosomal aberration databases is given in Table 1. Usually, the data mining facilities of these databases are limited to search and retrieval. Features such as clustering and gene prioritization are planned in future releases of at least some of these tools.

DECIPHER (DatabasE of Chromosomal Imbalance and Phenotype in Humans using ENSEMBL Resources, `www.sanger.ac.uk/PostGenomics/decipher`) has been inspired by the need to distinguish clinically significant imbalances from transmitted imbalances or polymorphisms detected using microarrays. One of the aims of this project is facilitating research on genetics in human development and health. The database collects information about clinical cases of submicroscopic chromosomal imbalances. Submitted clinical and genetic information is mapped onto the human genome through the ENSEMBL Genome Browser. DECIPHER has already supported the identification of new syndromes.

ECARUCA (European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations, `www.ecaruca.net`) is a European database that covers both common and rare chromosome aberrations. It contains details of thousands of published cytogenetic imbalances and is prospectively gathering rare cytogenetic and molecular cytogenetic aberrations, bringing together cytogenetic, molecular and clinical data [11].

COMPUTATIONAL APPROACHES IN CORRELATING GENOTYPE AND PHENOTYPE

A primary goal in the context of constitutional cytogenetics is elucidating the role of genes in different aspects of a phenotype without falsely associating normal variations to disease. While the above mentioned databases enable associations between phenotype and genotype to be stored, queried, annotated and exchanged, they fall short in identifying the genes underlying the phenotypic anomalies. Many tools and methods have been set up to filter high probability candidates. In what follows, we provide a list of the most important resources and published computational methodologies to generate genotype-phenotype leads and select gene candidates for further investigation.

Querying genotype-phenotype correlations in literature databases

With a focus on human as organism of interest, a primary resource for connecting genes to disease related phenotypes in a general rather than case based manner is the Online Mendelian Inheritance in Men database (OMIM). It contains curated records of genetically inherited human disorders with references to causative genes or genetic loci. Despite the highly reliable information it contains, its usefulness in computational analyses remains limited due to the unstructured way in which the phenotypes are described. Therefore, several approaches have arisen to transform the information in OMIM to be amenable for computational analysis. van Driel *et al.* [47], for example, created a human phenotype similarity map by text mining OMIM. Starting from an OMIM record or disease name, their MimMiner application retrieves the most phenotypically similar disorders. Based on the observation that similar phenotypes are often caused by functionally related or interacting genes, a researcher can easily go through the list of genes associated with similar phenotypes to select leads for further investigation. Most of the methods for computational prioritization of disease genes described in this review (see Table 2) make use of the information in OMIM, either as a starting point to further investigate genotype-to-phenotype associations [16, 18, 1, 46], or as a reference to benchmark their proposed method [3, 33, 13, 44, 25, 32].

Akin to OMIM, the MEDLINE database of biomedical literature contains a large amount of useful information on genotype-phenotype relations in free-text format. Here too, several published approaches exist to quickly guide researchers to information relevant to their research interest. The iHOP resource created by Hoffmann *et al.* [23] provides an intuitive access to the published literature by hyperlinking abstracts and sentences via the gene or protein symbols and names they contain. The approach taken by Van Vooren *et al.* [48] uses overrepresentation statistics to correlate specific biomedical terms from various targeted biomedical vocabularies with cytogenetic bands cited in MEDLINE abstracts. Through a web application named aBandApart, researchers can easily fetch the most relevant concepts for a genetic region of interest, look up chromosomal bands associated with a query term, and retrieve related literature. Another rich source of phenotypical information about genes is provided by Entrez Gene's GeneRIFs (Gene References into Function) [31]. These direct associations between genes and published literature allow construction of accurate textual representations of a gene using standard text mining techniques [19, 20]. Yet only Aerts *et al.* [3] and Lage *et al.* [25] make use of this valuable information source for candidate gene prioritization.

Other noteworthy resources include PhenomicDB [21] and PhenoGO [27]. PhenomicDB's value lies in how it integrates phenotype information from multiple speciesoriented databases into one repository. Phenotype-genotype associations are grouped based on gene orthology to allow exploration across different species. PhenoGO is mentioned here because it is a good example of how more advanced text mining techniques can help to bridge the gap between functional annotations of genes and phenotype descriptions [27]. The PhenoGO system uses Natural Language Processing of MEDLINE abstracts to connect phenotypic contextual information with GO annotations (and hence genes) and other biomedical ontologies.

Finding phenotype-rich genotypic features

Apart from retrieval of phenotype information from databases, also certain aspects (or features) of the gene or protein sequence can be used to infer or predict associated phenotypes. For instance, it is known that disease genes tend to code for longer proteins and are in general evolutionary more highly conserved. Both López-Bigas and Ouzounis [26] and Adie *et al.* [2] take this approach to calculate the correlation between genes and disease. Both started from the list of genes known to be involved in hereditary disease in the morbid map table of OMIM to define discriminating features, and subsequently classified all known human genes using a decision tree-based model. López-Bigas and Ouzounis used information about length, phylogenetic extent, degree of conservation, and paralogy of proteins in their Disease Gene Prediction (DGP) method. For their Prospectr method, Adie at al. used a more elaborate feature set that reflects the structure, content, and evolutionary conservation of both the DNA and protein sequence. The outcome of both methods is a score that indicates the probability of a gene to be disease causing.

While these kinds of methods provide valuable information, their applicability in connecting genes to specific phenotypes or diseases remains limited. This is mainly because they do not rely on the existing knowledge of a particular disease, contrary to the methods described in the following paragraph.

Pinpointing phenotype-related genes: guilt-by-association

The published methods described here can be broadly divided in three categories. An overview of methods is presented in Table 2. The first category covers *ab initio* methods. These try to identify genes by defining whether their characteristics or features are related to a specific disease. Examples of such features are genomic location (e.g., within a linkage region), sequence features, sequence phylogeny, functional annotation, gene expression, etc. Most methods take into account a combination of features to prioritize the candidates. The method by Turner *et al.* [45], POCUS, calculates statistical overrepresentation of Gene Ontology annotations and InterPro protein domains for genes in a given set of genomic loci to identify successful leads. GeneSeeker is a web application implemented by van Driel *et al.* [46] that filters the genes in a certain region based on user-specified characteristics of interest (tissues, phenotypic features of a syndrome, etc.). The Genes2Diseases (G2D) application presented by Perez-Iratxeta *et al.* [33] calculates the association between a gene and a disease based on the co-occurrence in a set of MEDLINE abstracts of MeSH terms in the ’Diseases’ and ’Chemical and Drugs’ categories with the gene’s Gene Ontology annotations. Other methods in this category include TEAM [14], the method by Tiffin *et al.* [44], and the genomic convergence approach described by Hauser *et al.* [22].

A second category of methods to link genes and phenotypes are network methods. Here, the emphasis is on the creation of an interaction network of genes or proteins. The rationale behind these methods is that similar phenotypes are often caused by functionally related genes, (i.e., genes that belong to the same functional process, take part in related pathways, or code for proteins that are part of the same protein complex). They differ mainly in the way the protein network is constructed and how interactions partners of known disease proteins are associated with known disease phenotypes. Franke *et al.* [13] created a Bayesian classifier to first predict protein-protein interactions not present in a gold-standard data set, using GO annotation, gene coexpression, and protein-protein interaction data. Then, their Prioritizer application establishes whether candidate genes in known disease loci are closer together in the network than expected. Oti *et al.* [32] used a hybrid protein-protein interaction network to find interaction partners of known disease proteins. They went on by checking whether the genes coding for these partners were in a disease associated locus for which no genes were previously identified. Lage *et al.* [25] follow a similar approach by constructing a quality-controlled human protein interaction network and deriving candidate protein complexes that contains the product of each of the candidate genes from it. The input phenotype is then compared to phenotypes of disease causing proteins present in these complexes and the protein coding candidate genes are scored accordingly using a Bayesian predictor.

We identify *similarity* methods as a third category, because here, prioritization of candidate genes is based on similarity between candidate and known disease genes, rather than on putatively involved features or on their presumed disease causing interaction partners. The Endeavour application by Aerts *et al.* [3] uses a sound statistical framework based on order statistics to reconcile a large number of different data sources. Used data includes both existing knowledge (literature, functional annotations, pathway information, etc.) and experimentally derived data (gene expression, protein interaction, etc.) to balance out a bias towards known genes. Candidate genes are compared to a user-selected or automatically retrieved list of training genes that represents the disease or phenotype under study, and prioritised according to their similarity with the training set thus obtained. It is worth noting that Endeavour is one of the only computational methods with which an in vivo validated new disease gene was revealed. Adie *et al.* [1] devised a similar method named SUSPECTS. Here, a more generic candidate gene scoring approach is used. Contrary to Endeavour, the set of training genes can not be customised, only four data sources can be included in the analysis, and the method shows no flexibility towards filtering the candidate genes at a user-defined locus. A new method, CAESAR, is described by Gaulton *et al.* [18] and is the most recent addition to similarity based methods. It also takes a more generic approach in scoring candidates. CAESAR uses a myriad of different data sources and integrates similarity measures from these different information spaces through the use of four different arithmetic operations. Although already an older method, the approach of Freudenberg and Propping [16] is also worth a mention in this category.

It must be noted that some approaches can be classified under more than one category. This is, for instance, the case with the CGI method of Ma *et al.* [29] in which *ab initio* derived gene-condition coexpression biclusters are combined with data from a protein interaction network. Endeavour also takes into account data from known protein interaction networks (BIND and Kegg) and, like SUSPECTS, from disease probabilities described earlier (DGP and Prospectr).

An upcoming trend in computational gene identification is the use of a concert of prioritization methods based on a combination of prioritization results. This approach was presented by Tiffin *et al.* [43] and Elbers *et al.* [9] who both conducted a study to find genes commonly associated with obesity and type II diabetes.

CHALLENGES FOR AUTOMATED GENOTYPE-PHENOTYPE CORRELATIONS

As more and more biological data are stored on computers, the problem of efficient retrieval and analysis of these data becomes the most important scientific bottleneck. This problem is particularly acute in biology because biological data are notorious for their complex form and semantics [41]. Case report databases can only provide value when the data is of sufficient quality and is rigorously evaluated, annotated, and interpreted within the richest possible context [34]. This is not a straightforward task. In a review paper on novel computational tools that allow researchers to amass, access, integrate, organize and manage phenotypic databases, Lussier and Liu [28] state that the development of phenotypic databases lags behind the advance in genomic databases, and creates the need for novel computational methods to unlock gene-disease relationships. In the next section, we will discuss database quality issues with regard to phenotyping in the context of chromosomal aberration and phenotype databases. We also discuss some specific challenges in gene prioritization for constitutional cytogenetics.

To avoid an inconsistent evaluation of a phenotype by multiple clinical geneticists, a single observer can be designated so that classification criteria can be uniformly applied (for example, Zhang *et al.* [50]). In their genotype-phenotype mapping efforts for Cri du Chat Syndrome, Zhang *et al.* attributed inconsistent results in previous mapping efforts partly to issues regarding inconsistent evaluation of the phenotype by multiple observers and lack of consideration for age dependence of prominence of phenotypic characteristics.

Appointing a single observer is clearly impossible for studies on rare disorders where case reports are gathered from a multitude of genetic research and testing centers. However, for case reports to be amenable to large scale data integration, exchange, and mining, phenotypic annotations need to be uniform and unbiased. Formalizations are crucial for organizing and executing experiments, as well as storing and sharing the experiment results [41].

In a context that spans multiple diagnostic or research entities (cross-departmental, cross-clinic, or in international collaborations), the proper use of dedicated ontologies can partly address this issue, allowing clinical geneticists to use a uniform vocabulary of clearly defined phenotype features to annotate case reports. Sound ontologies are instrumental to mapping function to gene products in the genome [42, 10]. However, even if a detailed and highly descriptive standardized vocabulary of phenotype characteristics is available, some important issues remain. Firstly, phenotypic traits can be age dependent or linked to a certain developmental stage, and can evolve over time. Secondly, phenotypes can vary in penetrance or severity, leading to the need for qualifiers and not just concepts. Thirdly, across databases, phenotype annotations often happen at different levels of granularity, in different formats, and with different aims [28].

Representation of phenotypic information is more complicated than biological data, and consequently there are few data standards and models for managing phenotypes within human repositories [28]. With OMIM as an example, Lussier *et al.* state that while OMIM has the largest collection of human diseases [40], the unstructured narrative content of its phenotypes makes it unsuitable for computational analysis, data mining and fusion, and integration between databases, as was mentioned before. It is clear that, in addition to proper interpretation of clinical features, unambiguous and complete identification and annotation of developmental anomalies, dysmorphic features, and any phenotype aspect in general is crucial for databases to be useful and interoperable. Several common terminologies to describe phenotypic aspects of a patient are presently available. Some of them can be licensed or obtained under certain conditions, others are freely available. Some well known ontologies and vocabularies are listed in Table 3.

DECIPHER and CGHGate (a database tool for storage, reporting and mining Array CGH Case reports, `www.esat.kuleuven.be/cghgate`) make use of a structured vocabulary present in LNDB (London Neurology Database), a hierarchical list of human dysmorphology concepts. Although LNDB is adequate for describing human dysmorphology in the context of constitutional developmental disorders, this vocabulary has some issues with regard to disambiguation between concepts and uniqueness and consistency of identifiers. It was not designed to be used as a standard for phenotype annotation amenable to mining, database integration and automated annotation.

OMD LNDB is not the only vocabulary that suffers from such issues. Soldatova [41] states that ontologies are often primarily designed to provide biologists with a common vocabulary for standard annotation purposes, and are not always structured with standard practice in mind. This approach is not compatible with the increasing use of computational reasoning in biology and its dependence on ontological data. Soldatova further states that although expert biologists may be able to deal with poorly designed and inconsistent ontologies, this is not currently possible for computer programs that do machine learning or text mining. As such programs are set to dominate the analysis and retrieval of biological data, Soldatova argues that biological ontologies should be designed with these needs in mind as well.

Some challenges are specific to gene prioritization for human development and constitutional cytogenetics. For one, it is important to note that phenotype characteristics are often complex traits that are not a function of state, but rather an end or even intermediate point that can be reached trough different and very unrelated developmental processes. In short, variations or mutations in different genes may yield identical or related phenotypes. This contributes to the complexity of gene prioritization for phenotype traits. Secondly, environment interactions during human development are likely to be an important cause of heterogeneity in phenotype. Attribution of phenotype traits not only to the genotype but also to the environment (nature vs. nurture) increases the order of complexity of the task at hand.

While parts of a phenotype can be explained by the action of a single gene, other characteristics are caused by multiple genes. For this reason, it is important that tools for phenotype based candidate gene prioritization are conceived with complex disorders in mind.

Positional or epigenetic effects may play a role in developmental disorders, so that genes responsible for the phenotype may actually lie outside the aberrant region.

Redon *et al.* recently showed the large extent to which non-pathogenic copy number variations are present throughout the human genome through analysis of Array CGH and SNP genotyping data [35]. Hurles *et al.* [35] have shown that this affects 12 per cent of the human genome, around the same level as SNP variation. Understanding benign copy number polymorphism is further complicated by the fact that some so-called normal variation may underlie a phenotypic characteristic such as disease susceptibility [34] or involvement in a late onset phenotype.

CONCLUSIONS

Array CGH is increasingly being used to define candidate regions for putative genes responsible for human genetic diseases. The increase in gene identification through molecular karyotyping will be driven by building, operating, extending, and disclosing genotype-phenotype databases, by integration of these databases and by making them interoperable, searchable, and amenable to large scale data mining initiatives. Ontologies and standardization of data can support these efforts.

Currently, there is a gap between existing candidate gene prioritization tools and existing case report and genotype-phenotype correlation databases. It can be expected that future prioritization tools will increasingly make use of publicly available case report repositories, and that database efforts in turn will move towards offering tools for intelligent search, clustering, candidate gene prioritization and data mining in general.

Standardization of ontologies, conventions on storage and annotation of raw experiment data to make them available to the community in a useful way (such as the MGED (`www.mged.org`) initiative MIAME (Minimum information about a microarray experiment) [6, 4]) and the use of novel data mining algorithms for data integration will improve the automated gene annotation processes of chromosomal aberrations and the delineation of novel and complex clinical entities. The tools and databases being developed to identify the functional relevance of genes will not only deepen our understanding of human development but will, in addition, aid in the data interpretation and improve genetic counseling.

ACKNOWLEDGEMENTS

Our research is supported by the following grants. Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, several PhD/postdoc and fellow grants. Flemish Government: FWO: PhD/postdoc grants, projects G.0407.02 (support vector machines), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (ontologies in bioi), G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, GBOU-McKnow-E (Knowledge management algorithms), GBOUSQUAD (quorum sensing), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos; SBOBioFrame, Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011); EU-RTD: ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain.

REFERENCES

[1] E A Adie, R R Adams, K L Evans, D J Porteous, and B S Pickard. Suspects: enabling fast and effective prioritization of positional candidates. *Bioinformatics (Oxford, England)*, 22(6).

[2] Euan A Adie, Richard R Adams, Kathryn L Evans, David J Porteous, and Ben S Pickard. Speeding disease gene discovery by sequence based candidate prioritization. *BMC bioinformatics*, 6.

[3] Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, Bert Coessens, Frederik De Smet, Leon-Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem Hassan, Peter Carmeliet, and Yves Moreau. Gene prioritization through genomic data fusion. *Nature biotechnology*, 24(5).

[4] Catherine A Ball and Alvis Brazma. Mged standards: work in progress. *Omics : a journal of integrative biology*, 10(2).

[5] DS Borgaonkar. Chromosomal variation in man; a catalogue of chromosomal variants and anomalies. *Wiley*, 1997.

[6] A Brazma, P Hingamp, J Quackenbush, G Sherlock, P Spellman, C Stoeckert, J Aach, WAnsorge, C A Ball, H C Causton, T Gaasterland, P Glenisson, F C Holstege, I F Kim, V Markowitz, J C Matese, H Parkinson, A Robinson, U Sarkans, S Schulze-Kremer, J Stewart, R Taylor, J Vilo, and M Vingron. Minimum information about a microarray experiment (miame)-toward standards for microarray data. *Nature genetics*, 29(4).

[7] de Ravel, Devriendt, Fryns, and Vermeesch. What's new in karyotyping? The move towards array comparative genomic hybridisation (CGH*). Eur J Pediatr*, 2007.

[8] Bert B A de Vries, Rolph Pfundt, Martijn Leisink, David A Koolen, Lisenka E L M Vissers, Irene M Janssen, Simon van Reijmersdal, Willy M Nillesen, Erik H L P G Huys, Nicole de Leeuw, Dominique Smeets, Erik A Sistermans, Ton Feuth, Conny M A van Ravenswaaij-Arts, Ad Geurts van Kessel, Eric F P M Schoenmakers, Han G Brunner, and Joris A Veltman. Diagnostic genome profiling in mental retardation. *American journal of human genetics*, 77(4).

[9] Clara C Elbers, N Charlotte Onland-Moret, Lude Franke, Anne G Niehoff, Yvonne T van der Schouw, and Cisca Wijmenga. A strategy to search for common obesity and type 2 diabetes genes. *Trends in endocrinology and metabolism: TEM*, 18(1).

[10] I Feenstra, H G Brunner, and C M A van Ravenswaaij. Cytogenetic genotypephenotype studies: improving genotyping, phenotyping and data storage. *Cytogenetic and genome research*, 115(3-4).

[11] I Feenstra, J Fang, D A Koolen, A Siezen, C Evans, R M Winter, M M Lees, M Riegel, B B A de Vries, C M A Van Ravenswaaij, and A Schinzel. European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations (ECARUCA); an online database for rare chromosome abnormalities. *European journal of medical genetics*, 49(4), 2006.

[12] Lars Feuk, Andrew R Carson, and Stephen W Scherer. Structural variation in the human genome. *Nature reviews*. Genetics, 7(2).

[13] Lude Franke, Harm van Bakel, Like Fokkens, Edwin D de Jong, Michael Egmont- Petersen, and Cisca Wijmenga. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *American journal of human genetics*, 78(6).

[14] Lude Franke, Harm van Bakel, Begoña Diosdado, Martine van Belzen, Martin Wapenaar, and Cisca Wijmenga. Team: a tool for the integration of expression, and linkage and association maps. *European journal of human genetics : EJHG*, 12(8).

[15] Nelson Freimer and Chiara Sabatti.

[16] J Freudenberg and P Propping. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics (Oxford, England)*, 18 Suppl 2.

[17] J M Friedman, Agnes Baross, Allen D Delaney, Adrian Ally, Laura Arbour, Linlea Armstrong, Jennifer Asano, Dione K Bailey, Sarah Barber, Patricia Birch, Mabel Brown-John, Manqiu Cao, Susanna Chan, David L Charest, Noushin Farnoud, Nicole Fernandes, Stephane Flibotte, Anne Go, William T Gibson, Robert A Holt, Steven J M Jones, Giulia C Kennedy, Martin Krzywinski, Sylvie Langlois, Haiyan I Li, Barbara C McGillivray, Tarun Nayar, Trevor J Pugh, Evica Rajcan-Separovic, Jacqueline E Schein, Angelique Schnerch, Asim Siddiqui, Margot I Van Allen, GaryWilson, Siu-Li Yong, Farah Zahir, Patrice Eydoux, and Marco A Marra. Oligonucleotide microarray analysis of genomic imbalance in children with mental retardation. *American journal of human genetics*, 79(3).

[18] Gaulton, Mohlke, and Vision. A computational system to select candidate genes for complex human traits.

[19] P Glenisson, P Antal, J Mathys, Y Moreau, and B De Moor. Evaluation of the vector space representation in text-based gene clustering. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*.

[20] Patrick Glenisson, Bert Coessens, Steven Van Vooren, Janick Mathys, Yves Moreau, and Bart De Moor. Txtgate: profiling gene groups with text-based information. *Genome biology*, 5(6).

[21] Philip Groth, Nadia Pavlova, Ivan Kalev, Spas Tonov, Georgi Georgiev, Hans- Dieter Pohlenz, and Bertram Weiss. Phenomicdb: a new cross-species genotype / phenotype resource. *Nucleic acids research*, 35(Database issue).

[22] Michael A Hauser, Yi-Ju Li, Satoshi Takeuchi, Robert Walters, Maher Noureddine, Melinda Maready, Tiffany Darden, Christine Hulette, Eden Martin, Elizabeth Hauser, Hong Xu, Don Schmechel, Judith E Stenger, Fred Dietrich, and Jeffery Vance. Genomic convergence: identifying candidate genes for parkinson's disease by combining serial analysis of gene expression and genetic linkage. *Human molecular genetics*, 12(6).

[23] Robert Hoffmann and Alfonso Valencia. A gene network for navigating the literature. *Nature genetics*, 36(7).

[24] Adrian S Ishkanian, Chad A Malloff, Spencer K Watson, Ronald J DeLeeuw, Bryan Chi, Bradley P Coe, Antoine Snijders, Donna G Albertson, Daniel Pinkel, Marco A Marra, Victor Ling, Calum MacAulay, andWan L Lam. A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet*, 36(3):299–303, 2004.

[25] Kasper Lage, E Olof Karlberg, Zenia M Størling, Pall I Olason, Anders G Pedersen, Olga Rigina, Anders M Hinsby, Zeynep Tumer, Flemming Pociot, Niels Tommerup, Yves Moreau, and Søren Brunak. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology*, 25(3).

[26] Nuria Lopez-Bigas and Christos A Ouzounis. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic acids research*, 32(10).

[27] Yves Lussier, Tara Borlawsky, Daniel Rappaport, Yang Liu, and Carol Friedman. Phenogo: assigning phenotypic context to gene ontology annotations with natural language processing. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*.

[28] Yves A Lussier and Yang Liu. Computational approaches to phenotyping: highthroughput phenomics. *Proc Am Thorac Soc*, 4(1):18–25, 2007.

[29] Xiaotu Ma, Hyunju Lee, Li Wang, and Fengzhu Sun. Cgi: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics (Oxford, England)*, 23(2).

[30] B Menten, N Maas, B Thienpont, K Buysse, J Vandesompele, C Melotte, T de Ravel, S Van Vooren, I Balikova, L Backx, S Janssens, A De Paepe, B De Moor, Y Moreau, P Marynen, J-P Fryns, G Mortier, K Devriendt, F Speleman, and J R Vermeesch. Emerging patterns of cryptic chromosomal imbalance in patients with idiopathic mental retardation and multiple congenital anomalies: a new series of 140 patients and review of published reports. *Journal of medical genetics*, 43(8).

[31] Joyce A Mitchell, Alan R Aronson, James G Mork, Lillian C Folk, Susanne M Humphrey, and Janice M Ward. Gene indexing: characterization and analysis of NLM's generifs. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium.*

[32] M Oti, B Snel, M A Huynen, and H G Brunner. Predicting disease genes using protein-protein interactions. *Journal of medical genetics*, 43(8).

[33] Carolina Perez-Iratxeta, Matthias Wjst, Peer Bork, and Miguel A Andrade. G2d: a tool for mining genes associated with disease. *BMC genetics*, 6.

[34] Daniel Pinkel and Donna G Albertson. Comparative genomic hybridization. *Annu Rev Genomics Hum Genet*, 6:331–54, 2005.

[35] Richard Redon, Shumpei Ishikawa, Karen R Fitch, Lars Feuk, George H Perry, T Daniel Andrews, Heike Fiegler, Michael H Shapero, Andrew R Carson, Wenwei Chen, Eun Kyung Cho, Stephanie Dallaire, Jennifer L Freeman, Juan R Gonzalez, Monica Gratacos, Jing Huang, Dimitrios Kalaitzopoulos, Daisuke Komura, Jeffrey R MacDonald, Christian R Marshall, Rui Mei, Lyndal Montgomery, Kunihiro Nishimura, Kohji Okamura, Fan Shen, Martin J Somerville, Joelle Tchinda, Armand Valsesia, Cara Woodwark, Fengtang Yang, Junjun Zhang, Tatiana Zerjal, Jane Zhang, Lluis Armengol, Donald F Conrad, Xavier Estivill, Chris Tyler-Smith, Nigel P Carter, Hiroyuki Aburatani, Charles Lee, Keith W Jones, StephenWScherer, and Matthew E Hurles. Global variation in copy number in the human genome. *Nature*, (7118):444–54, 2006.

[36] Damien Sanlaville, Jean-Michel Lapierre, Catherine Turleau, Aurélie Coquin, Guntram Borck, Laurence Colleaux, Michel Vekemans, and Serge Pierrick Romana. Molecular karyotyping in human constitutional cytogenetics. *European journal of medical genetics*, 48(3).

[37] A. Schinzel. Catalogue of unbalanced chromosome aberration in man. *de Gruyter,* 2001.

[38] Jonathan Sebat, B Lakshmi, Dheeraj Malhotra, Jennifer Troge, Christa Lese- Martin, TomWalsh, Boris Yamrom, Seungtai Yoon, Alex Krasnitz, Jude Kendall, Anthony Leotta, Deepa Pai, Ray Zhang, Yoon-Ha Lee, James Hicks, Sarah J Spence, Annette T Lee, Kaija Puura, Terho Lehtimaki, David Ledbetter, Peter K Gregersen, Joel Bregman, James S Sutcliffe, Vaidehi Jobanputra, Wendy Chung, Dorothy Warburton, Mary-Claire King, David Skuse, Daniel H Geschwind, T Conrad Gilliam, Kenny Ye, and Michael Wigler. Strong association of *de novo* copy number mutations with autism. *Science (New York, N.Y.),* 316(5823).

[39] L G Shaffer and B A Bejjani. Medical applications of array CGH and the transformation of clinical cytogenetics. Cytogenet *Genome Res*, 115(3-4):303–9, 2006.

[40] Leah C Solberg, William Valdar, Dominique Gauguier, Graciela Nunez, Amy Taylor, Stephanie Burnett, Carmen Arboledas-Hita, Polinka Hernandez-Pliego, Stuart Davidson, Peter Burns, Shoumo Bhattacharya, Tertius Hough, Douglas Higgs, Paul Klenerman,William O Cookson, Youming Zhang, RobertMDeacon, J Nicholas P Rawlins, Richard Mott, and Jonathan Flint. A protocol for highthroughput phenotyping, suitable for quantitative trait analysis in mice. *Mammalian genome : official journal of the International Mammalian Genome Society*, 17(2).

[41] Larisa N Soldatova and Ross D King. Are the current ontologies in biology good ontologies? *Nat Biotechnol*, 23(9):1095–8, 2005.

[42] Paul D Thomas, Huaiyu Mi, and Suzanna Lewis. Ontology annotation: mapping genomic regions to biological function. *Curr Opin Chem Biol*, 11(1):4–11, 2007.

[43] Nicki Tiffin, Euan Adie, Frances Turner, Han G Brunner, Marc A van Driel, Martin Oti, Nuria Lopez-Bigas, Christos Ouzounis, Carolina Perez-Iratxeta, Miguel A Andrade-Navarro, Adebowale Adeyemo, Mary Elizabeth Patti, Colin A M Semple, and Winston Hide. Computational disease gene

identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic acids research,* 34(10).

[44] Nicki Tiffin, Janet F Kelso, Alan R Powell, Hong Pan, Vladimir B Bajic, andWinston A Hide. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic acids research, 33(5).

[45] Frances S Turner, Daniel R Clutterbuck, and Colin A M Semple. Pocus: mining genomic sequence annotation to predict disease genes. *Genome biology*, 4(11).

[46] M A van Driel, K Cuelenaere, P P C W Kemmeren, J A M Leunissen, H G Brunner, and Gert Vriend. Geneseeker: extraction and integration of human diseaserelated information from web-based genetic databases. *Nucleic acids research*, 33(Web Server issue).

[47] Marc A van Driel, Jorn Bruggeman, Gert Vriend, Han G Brunner, and Jack A M Leunissen. A text-mining analysis of the human phenome. *European journal of human genetics* : EJHG, 14(5).

[48] Van Vooren, Thienpont, Menten, Speleman, De Moor, Vermeesch, and Moreau. Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations. *Nucleic Acids Research*, vol. Advance Access, no. 10.1093/nar/gkm054, Apr. 2007, pp. 1-11.

[49] Lisenka E L M Vissers, Joris A Veltman, Ad Geurts van Kessel, and Han G Brunner. Identification of disease genes by whole genome cgh arrays. *Human molecular genetics*, 14 Spec No. 2.

[50] Xiaoxiao Zhang, Antoine Snijders, Richard Segraves, Xiuqing Zhang, Anita Niebuhr, Donna Albertson, Huanming Yang, Joe Gray, Erik Niebuhr, Lars Bolund, and Dan Pinkel. High-resolution mapping of genotype-phenotype relationships in cri du chat syndrome using array comparative genomic hybridization. *American journal of human genetics*, 76(2):312–26, 2005.

TABLES (3 in total)

Catalogue of Unbalanced Chromosome Aberration in Man - Albert Schinzel's comprehensive catalogue of chromosomal aberrations in man in book form. The catalogue is a standard reference for clinicians treating patients with autosomal chromosome aberrations and for physicians and biologists working in cytogenic laboratories and human genetic institutes [37]
Chromosome Abnormality Database (`www.ukcad.org.uk/cocoon/ukcad/`) - The UK Association of Clinical Cytogeneticists (ACC) Chromosome Abnormality Database (CAD) is a collection of both constitutional and acquired abnormal karyotypes reported by UK Regional Cytogenetics Centres. It is open to all Genetics Professionals, and available for searches on different abnormalities and karyotypes in both a clinical context as for medical research
Chromosome Anomaly Collection (`www.ngrl.org.uk/Wessex/collection. htm`) - a catalogue of unbalanced structural chromosome abnormalities (USCA) without phenotypic effect. The Collection also includes the cytogenetically visible euchromatic variants as part of the continuum of copy number variation in the human genome
DECIPHER (`www.sanger.ac.uk/PostGenomics/decipher`) - DatabasE of Chromosomal Imbalance and Phenotype in Humans using ENSEMBL Resources, see text
Database of Chromosome Aberrations in Cancer (`cgap.nci.nih.gov/ Chromosomes/Mitelman`) – the Mitelman catalog for cancer cytogeneticists is a standard reference database that compiles information on chromosome changes identified in human neoplasms. The electronic version supports searches by karyotype, reference, tumor type, and location
ECARUCA (`www.ecaruca.net`) - European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations, see text
The Human Phenome Project - a proposed international effort to create comprehensive phenomic databases of systematically collected phenotypic information, and to develop approaches for analyzing such phenotypic data [15]
Mendelian Cytogenetics Network DataBase (`www.mcndb.org`) - an online database on disease associated balanced chromosomal rearrangements, containing information on breakpoints and clinical features and disease potential. Aims at initiating collaborative studies of specific disorders
Online Database of Chromosomal variation in man (`www.wiley.com/legacy/products/subject/life/borgaonkar/access.html`) - a systematic collection of important citations from the world's literature reporting on all common and rare chromosomal alterations, phenotypes, and abnormalities in humans. The database is organized by variations and anomalies, numerical anomalies, and chromosomal breakage syndromes [5]
Progenetix (`www.progenetix.de`) - a database of published cytogenetic abnormalities in human malignancies, mostly from CGH experiments

Table 1 – an overview of chromosomal aberration databases. These efforts aim at aggregating chromosomal aberration information at various levels of resolution and annotate the genome with case reports, congenital anomalies or phenotypes.

Computational methods	Available online	Used resources	Application to complex traits	Supported species	In-vivo validation	Data integration method	Website and reference
POCUS (2003)	No	InterPro, GO, UniGene	-	Human	No	Overrepresentation statistics	Turner [45]
TEAM (2004)	Yes (download)	GO, gene expression data	Yes	Human	No	Filtering functionality	`humgen.med.uu.nl/~lude/team` Francke *et al*. [14]
(Tiffin *et al.* 2005)	No	MEDLINE, eVOC	-	Human	No	Term co-occurrence statistics	`www.sanbi.ac.za/tiffin_et_al` Tiffin *et al.* [44]
GeneSeeker (2005)	Yes	MGD, GDB, MEDLINE, OMIM, UniProt, GXD	-	Human, mouse	No	Boolean logic	`www.cmbi.ru.nl/GeneSeeker/` van Driel *et al.* [46]
G2D (2005)	Yes	MeSH, MEDLINE, GO	Yes	Human	No	Fuzzy set theory	`www.ogic.ca/projects/g2d_2/` Perez-Iratxeta *et al.* [33]
(Freudenberg & Propping 2002)	No	OMIM, GO	-	Human	No	Generic scores	Freudenberg *et al.* [16]
SUSPECTS (2006)	Yes	OMIM, HGMD, GAD, Prospectr, InterPro, GO, gene expression data,	-	Human	No	Generic scores	`www.genetics.med.ed.ac.uk/suspects/` Adie *et al.* [1]
Endeavour (2006)	Yes (download)	MEDLINE, EST, KEGG, GO, TRANSFAC, Jaspar, InterPro, BIND, DGP, Prospectr, gene expression data	Yes	Human, mouse, fly	Yes	Order statistics	`www.esat.kuleuven.be/endeavour/` Aerts *et al.* [3]
Caesar (2007)	No	MP, eVOC, GO, OMIM, Entrez Gene, EnsEMBL, UniProt, InterPro, BIND, HPRD, KEGG, MGD, GAD	Yes	Human	No	Generic scores	`visionlab.bio.unc.edu/caesar/` Gaulton *et al.* [18]
(Oti *et al.* 2006)	No	HPRD, DIP, interactions from high-throughput experiments	-	Human	No	Generic approach	Oti *et al.* [32]
Prioritizer (2006)	Yes (download)	BIND, HPRD, Reactome, KEGG, GO, SMD, GEO, GeneNetwork	Yes	Human	No	Bayesian classifier	`www.prioritizer.nl` Franke *et al.* [13]
CGI (2007)	No	Yeast gene expression compendia (knock-out, stress response, and cell cycle), MPPI, GO	Yes	Yeast, human	No	Markov Random Field theory	Ma *et al.* [29]
(Lage *et al.* 2007)	No	MINT, BIND, IntAct, KEGG PPrel, KEGG ECrel, Reactome	Yes	Human	No	Bayesian classifier	Lage *et al.* [25]

Table 2. Overview of published methodologies for gene prioritisation. Remarks: the *Used resources* column only contains the data sources used in the prioritisation methodology, not the data sources used to validate or benchmark the approach; the column *Application to complex traits* contains *Yes* if the paper describing the method explicitly mentioned it's applicability to or application on complex traits, a dash otherwise; the *Supported species* column contains the species the method was applied on previsously, not necessarily all species the method could be applied on.

GO	Gene Ontology, a systematic terminology for functional features of genes and proteins
ICD-9	International Classication of Diseases Clinical Modification
LDDB, LNDB	Oxford Medical Dictionary London Dysmorphology and Neurology Databases
MPO	Mammalian Phenotype Ontology
SNOMED	Systematized Nomenclature of Medicine
UMLS	Unified Medical Language System, groups and links a host of ontologies

Table 3. Well known and widely used ontologies and vocabularies relating to phenotypic traits and human disease.