Array comparative genomic hybridization and computational genome annotation in constitutional cytogenetics: suggesting candidate genes for novel submicroscopic chromosomal imbalance syndromes

Steven Van Vooren, MSc,[1] Bert Coessens, PhD,[1] Bart De Moor, PhD,[1] Yves Moreau, PhD,[1] and Joris R. Vermeesch, PhD[2]

Genome-wide array comparative genomic hybridization 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. Because most imbalances encompass regions harboring multiple genes, the challenge is to (1) identify those genes responsible for the specific phenotype and (2) 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. ***Genet Med*** **2007:9(9):642–649.**

Key Words: *clinical informatics, cytogenetics, molecular diagnostics, array comparative genomic hybridization, data mining, genotype-phenotype correlation*

Array comparative genomic hybridization (CGH) is used increasingly, often as a primary genetic screening method in diagnosis and research.[1–4] 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.[5–9] These imbalances pinpoint the location of genes that are involved in human development.[10] Because most imbalances encompass regions harboring multiple genes, the challenge is to (1) identify those genes responsible for the specific phenotype and (2) 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.[1] 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.[11]

However, some specific challenges apply to correlating genotype and phenotype in the context of human disease. First, 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.

Second, 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.[4] 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 permit the development of bioinformatics methods for candidate gene selection. Such information sources range from the large body 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 continue by giving an overview of published methods and tools that can help in the gene discov-

From the [1]Department of Electrotechnical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium; [2]Center for Human Genetics, Leuven University Hospital, Leuven, Belgium.

Steven Van Vooren, MSc, Department of Electrotechnical Engineering, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium; E-mail: Steven.VanVooren@esat.kuleuven.be

Disclosure: The authors declare no conflict of interest.

Submitted for publication May 8, 2007.

Accepted for publication June 11, 2007.

DOI: 10.1097/GIM.0b013e318145b27b

ery 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. 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, and 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 being implemented in at least some of these tools.

DECIPHER [DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (http://decipher.sanger.ac.uk/)] 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, such as a 17q21.3 microdeletion syndrome associated with developmental delay and learning disability[12] and a 14q11.2 microdeletion syndrome associated with mental retardation and other minor anomalies.[13]

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.[14]

Table 1
An overview of chromosomal aberration databases

Catalogue of Unbalanced Chromosome Aberration in Man[54]: Albert Schinzel's comprehensive catalog of chromosomal aberrations in humans in book form. The catalog 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.

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 Centers. 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 catalog of unbalanced structural chromosome abnormalities 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.[53]

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. It aims to initiate collaborative studies of specific disorders.

The online database of Chromosomal Variation in Man: A Catalog of Chromosomal Variants and Anomalies[52] (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.

Progenetix (www.progenetix.de): A database of published cytogenetic abnormalities in human malignancies, mostly from comparative cytogenetic hybridization experiments.

These efforts aim to aggregate chromosomal aberration information at various levels of resolution and annotate the genome with case reports, congenital anomalies, or phenotypes.

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. Although the above-mentioned databases enable associations between phenotype and genotype to be stored, queried, annotated, and exchanged, most 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 the human as the organism of interest, a primary resource for connecting genes to disease-related phenotypes in a general rather than in a case-based manner is the Online Mendelian Inheritance in Man 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.,[15] for example, created a human phenotype similarity map by text mining OMIM. Starting from an OMIM record or disease name, their MimMiner application (www.cmbi.ru.nl/MimMiner) 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 (Table 2) make use of the information in OMIM, either as a starting point to further investigate genotype-to-phenotype associations[16–19] or as a benchmark.[20–25]

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.[26] 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.[27] 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 find 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 phenotypic information about genes is provided by Entrez Gene's GeneRIFs (Gene References into Function).[28] These direct associations between genes and published literature allow construction of accurate textual representations of a gene using standard text mining techniques.[29,30] Yet only Aerts et al.[20] and Lage et al.[24] make use of this valuable information source for candidate gene prioritization.

Other noteworthy resources include PhenomicDB[31] and PhenoGO.[32] PhenomicDB's value lies in how it integrates phenotype information from multiple species-oriented 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.[32] The PhenoGO system uses Natural Language Processing of MEDLINE abstracts to connect phenotypic contextual information with gene ontology annotations (and hence genes) and other biomedical ontologies.

Finding phenotype-rich genotypic features

Apart from retrieval of phenotype information from databases, certain aspects (or features) of the gene or protein sequence can also 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 evolutionarily more highly conserved. Both López-Bigas and Ouzounis[33] and Adie et al.[34] 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 method. For their Prospectr method, Adie et 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.

Although 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.,[35] 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.[19] that filters the genes in a certain region based on user-specified characteristics of interest (e.g., tissues, phenotypic features of a syndrome). The Genes2Diseases application presented by Perez-Iratxeta et al.[21] calculates the association between a gene and a disease based on the co-occurrence

Table 2
Overview of published methodologies for gene prioritization

Computational methods	Available online	Resources used	Application to complex traits	Supported species	In vivo validation	Data integration method	Web site, ref.
POCUS, 2003	No	InterPro, GO, UniGene	—	Human	No	Overrepresentation statistics	Turner et al.[35]
TEAM, 2004	Yes, download	GO, gene expression data	Yes	Human	No	Filtering functionality	humgen.med.uu.nl/~lude/team, Francke et al.[36]
Tiffin et al., 2005	No	MEDLINE, eVOC	—	Human	No	Term co-occurrence statistics	www.sanbi.ac.za/tiffin_et_al, Tiffin et al.[23]
GeneSeeker, 2005	Yes	MGD, GDB, MEDLINE, OMIM, UniProt, GXD	—	Human, mouse	No	Boolean logic	www.cmbi.ru.nl/GeneSeeker/. van Driel et al.[19]
G2D, 2005	Yes	MeSH, MEDLINE, GO	Yes	Human	No	Fuzzy set theory	www.ogic.ca/projects/g2d_2/, Perez-Iratxeta et al.[21]
Freudenberg and 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.[18]
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.[20]
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.[17]
Oti et al., 2006	No	HPRD, DIP, interactions from high-throughput experiments	—	Human	No	Generic approach	Oti et al.[25]
Prioritizer, 2006	Yes, download	BIND, HPRD, Reactome, KEGG, GO, SMD, GEO, GeneNetwork	Yes	Human	No	Bayesian classifier	www.prioritizer.nl, Franke et al.[22]
CGI, 2007	No	Yeast gene expression compendia (knockout, stress response, and cell cycle), MPPI, GO	Yes	Yeast, human	No	Markov random field theory	Ma et al.[38]
Lage et al., 2007	No	MINT, BIND, IntAct, KEGG PPrel, KEGG ECrel, Reactome	Yes	Human	No	Bayesian classifier	Lage et al.[24]

The 'Resources used' column contains data sources used in the prioritization methodology, not the data sources used for validation or as a benchmark. The column 'Application to complex traits' contains 'Yes' if the paper describing the method explicitly mentions application or applicability in a complex traits context. The column 'Supported species' only contains those species the method was effectively applied to. See appendix for definitions of abbreviations.

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[36], the method by Tiffin et al.,[23] and the genomic convergence approach described by Hauser et al.[37]

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.[22] created a bayesian classifier to first predict protein-protein interactions not present in a gold-standard data set, using gene ontology 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.[25] 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.[24] 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 with 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.[20] uses a sound statistical framework based on order statistics to reconcile a large number of different data sources. The data used include both existing knowledge (e.g., literature, functional annotations, pathway information) and experimentally derived data (e.g., gene expression, protein interaction) to balance out a bias toward known genes. Candidate genes are compared with a user-selected or automatically retrieved list of training genes that represent the disease or phenotype under study and prioritized 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.[20] Adie et al.[18] devised a similar method named SUSPECTS. Here, a more generic candidate gene scoring approach is used. Contrary to Endeavour, the set of training genes cannot be customized, only four data sources can be included in the analysis, and the method shows no flexibility toward filtering the candidate genes at a user-defined locus. A new method, CAESAR, is described by Gaulton et al.[17] 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.[38] 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 (disease gene prediction method 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.[39] and Elbers et al.[40] who both conducted a study to find genes commonly associated with obesity and type 2 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 are of sufficient quality and is rigorously evaluated, annotated, and interpreted within the richest possible context.[4] 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[42] 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 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.[43]). 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]

Standard nomenclature

In a context that spans multiple diagnostic or research entities (cross-departmental, cross-clinic, 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.[44,45] However, even if a detailed and highly descriptive standardized vocabulary of phenotype characteristics is available, some important issues remain. First, phenotypic traits can be age dependent or linked to a certain developmental stage and can evolve over time. Second, phenotypes can vary in penetrance or severity, leading to the need for qualifiers and not just concepts. Third, across databases, phenotype annotations often happen at different levels of granularity, in different formats, and with different aims.[42]

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.[42] With OMIM as an example, Lussier et al.[42] state that although OMIM has the largest collection of human diseases, 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 currently available. Some of them can be licensed or obtained under certain conditions, and others are freely available. Some well-known ontologies and vocabularies are listed in Table 3.

DECIPHER (see above) and CGHGate [a database tool for storage, reporting, visualization, and mining array CGH case reports (www.esat.kuleuven.be/cghgate)] allow case report phenotypes to be described using the structured vocabulary present in the Oxford Medical Dictionary (OMD) London Neurology Database (LNDB). Both LNDB and LDDB (London Dysmorphology Database) contain a hierarchy of human dysmorphology concepts. Although these are adequate for describing human dysmorphology in the context of constitutional developmental disorders, these vocabularies have some issues with regard to concept uniqueness and disambiguation, and consistency of the hierarchical structure and of concept identifiers. They were 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 and King[41] state 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 and King further state 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 and King argue that biological ontologies should be designed with these needs in mind as well.

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

GO	Gene ontology: systematic terminology for functional features of genes and proteins
ICD-9	International Classification 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

It is clear that standardization of clinical terminology is crucial when accurate phenotype subgroups need to be defined for genetic analysis and are collected from several genetic centers or databases. In 2005, an international nomenclature group was set up to begin the task of defining clinical terms to construct a standardized nomenclature[47,48] aiming at consensus definitions for an unambiguous set of clinical terms. This work is currently in progress. The global adoption of such a standard for the description of abnormal phenotypes will not only support information exchange between clinical geneticists and allow database interoperability, but, if certain technical specifications are met, will aid a deeper integration of databases and will support data mining through computational biology methods.

Quality standards

Case report genotype information stems from wet lab experimentation. Array CGH data contain noise that can obfuscate actual genetic aberrations. The role of raw array CGH data normalization methods and aberration-calling algorithms is key in delineating higher level genomic aberrations from raw experimental output. As statistical methods and algorithms for data normalization and analysis are subjects of ongoing research, it is important that information on which data analysis approaches are used be added to reports stored in case repositories. It is also advisable that all raw data be stored with the patient report to allow future analysis with novel algorithms to be conducted. Storage of unmodified numerical output of scanner software is a minimal requirement. The advent of array platforms that feature a huge number of reporters makes this a nontrivial issue.

Conventions on storage and annotation of raw microarray experiment data, such as the MGED initiative MIAME (minimum information about a microarray experiment), have been developed.[49,50] A draft version of an extension on the MIAME initiative to array CGH is present as a checklist from the MGED Web site (www.mged.org).

Additionally, it is relevant to define within the context of a case report database initiative when a case should be entered to the repository. Duplicate entries should be avoided. Inclusion of experiment reports that do not show an aberrant phenotype is debatable, but it is important to keep in mind that the aggregation of genotypic data on healthy individuals and storage of parent and sibling genetic profiles is useful to ease identification of disease causing alterations, to chart normal human copy number variation, and to elucidate disease susceptibility loci.

Gene prioritization

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 through 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. Second, environmental 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 versus nurture) increases the order of complexity of the task at hand.

Although 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.[51] recently showed the large extent to which nonpathogenic copy number variations are present throughout the human genome through analysis of array CGH and single nucleotide polymorphism genotyping data. The authors show that this affects 12% of the human genome, around the same level as single nucleotide polymorphism 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[4] 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 toward 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 initiative MIAME[49,50]) 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 data interpretation and improve genetic counseling.

ACKNOWLEDGMENTS

Our research is supported by the following grants: GOA AMBioRICS, CoE EF/05/007 SymBioSys; FWO projects 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 projects GBOU-McKnow-E (knowledge management algorithms), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos, SBOBioFrame; Belgian Federal Science Policy Office projects IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011); EU-RTD projects ERNSI (European Research Network on System Identification), FP6-NoE Biopattern, FP6-IP e-Tumours, FP6-MC-EST Bioptrain, FP6-STREP Strokemap.

APPENDIX

BIND: Biomolecular Interaction Network Database, bond.unleashedinformatics.com
DGP: Disease Gene Prediction method, cgg.ebi.ac.uk/services/dgp
DIP: Database of Interacting Proteins, dip.doe-mbi.ucla.edu
Ensembl: www.ensembl.org
Entrez Gene: www.ncbi.nlm.nih.gov/sites/entrez?db=gene
EST: Expressed Sequence Tags
eVOC: www.evocontology.org
G2D: Candidate Genes to Inherited Diseases
GAD: Genetic Association Database, geneticassociationdb.nih.gov
GDB: Human Genome Database, www.gdb.org
GeneNetwork: www.genenetwork.nl
GEO: Gene Expression Omnibus, www.ncbi.nlm.nih.gov/geo
GO: Gene Ontology, www.geneontology.org
GXD: Gene Expression Database, www.informatics.jax.org
HGMD: Human Gene Mutation Database, www.hgmd.cf.ac.uk
HPRD: Human Protein Reference Database, www.hprd.org
IntAct: www.ebi.ac.uk/intact/site/index.jsf
InterPro: www.ebi.ac.uk/interpro
Jaspar: The high-quality transcription factor binding profile database, jaspar.cgb.ki.se
KEGG: Kyoto Encyclopedia of Genes and Genome, www.genome.jp/kegg
MeSH: Medical Subject Headings, www.nlm.nih.gov/mesh
MGD: Mouse Genome Database, www.informatics.jax.org
MINT: Molecular Interactions Database, mint.bio.uniroma2.it/mint
MP: Mouse Phenotypes, www.informatics.jax.org

MPPI: Mammalian Protein-Protein Interaction Database, mips.gsf.de/proj/ppi

OMIM: Online Mendelian Inheritance in Man database, www.ncbi.nlm.nih.gov/omim

Prospectr: PRiOrization by Sequence and Phylogenetic Extent of CandidaTe Regions, www.genetics.med.ed.ac.uk/prospectr

Reactome: Curated database of biological processes in humans, www.reactome.org

SMD: Stanford MicroArray Database, genome-www5.stanford.edu

TRANSFAC: database on eukaryotic transcription factors, www.gene-regulation.com

Unigene: www.ncbi.nlm.nih.gov/sites/entrez?db=unigene

UniProt: Universal Protein Resource, www.uniprot.org

References

1. Shaffer LG, Bejjani BA. Medical applications of array CGH and the transformation of clinical cytogenetics. *Cytogenet Genome Res* 2006;115:303–309.
2. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet* 2006;7:85–97.
3. Sanlaville D, Lapierre J-M, Turleau C, Coquin A, et al. Molecular karyotyping in human constitutional cytogenetics. *Eur J Med Genet* 2005;48:214–231.
4. Pinkel D, Albertson DG. Comparative genomic hybridization. *Annu Rev Genomics Hum Genet* 2005;6:331–354.
5. Menten B, Maas N, Thienpont B, Buysse K, et al. 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. *J Med Genet* 2006;43:625–633.
6. Ishkanian AS, Maloff CA, Watson SK, et al. A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet* 2004;36:299–303.
7. Friedman JM, Baross A, Delaney AD, Ally A, et al. Oligonucleotide microarray analysis of genomic imbalance in children with mental retardation. *Am J Hum Genet* 2006;79:500–513.
8. Bert B, de Vries A, Pfundt R, Leisink M, et al. Diagnostic genome profiling in mental retardation. *Am J Hum Genet* 2005;77:606–616.
9. Sebat J, Lakshmi B, Malhotra D, Troge J, et al. Strong association of de novo copy number mutations with autism. *Science* 2007;316:445–449.
10. Lisenka EL, Vissers M, Veltman JA, van Kessel AG, et al. Identification of disease genes by whole genome CGH arrays. *Hum Mol Genet* 2005;14 Spec No. 2:R215–R223.
11. de Ravel TJ, Devriendt K, Fryns JP, Vermeesch JR. What's new in karyotyping? The move towards array comparative genomic hybridisation (CGH). *Eur J Pediatr* 2007;166:637–643.
12. Zahir F, Firth H, Baross A, Delaney A, et al. Novel deletions of 14q11.2 associated with mental retardation and similar minor anomalies in three children. *J Med Genet.* In press.
13. Shaw-Smith C, Pittman AM, Willatt L, Martin H, et al. Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nat Genet* 2006;38:1032–1037.
14. Feenstra I, Fang J, Koolen DA, Siezen A, et al. European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations (ECARUCA); an online database for rare chromosome abnormalities. *Eur J Med Genet* 2006;49:279–291.
15. van Driel MA, Bruggeman J, Vriend G, Brunner HG. A text-mining analysis of the human phenome. *Eur J Hum Genet* 2006;14:535–542.
16. Freudenberg J, Propping P. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics* 2002;18(Suppl 2):S110–S115.
17. Gaulton KJ, Mohlke KL, Vision TJ. A computational system to select candidate genes for complex human traits. *Bioinformatics* 2007;23:1132–1140.
18. Adie EA, Adams RR, Evans KL, Porteous DJ, et al. Suspects: enabling fast and effective prioritization of positional candidates. *Bioinformatics* 2006;22:773–774.
19. van Driel MA, Cuelenaere K, Kemmeren PPCW, Leunissen JAM, et al. Geneseeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res* 2005;33(Web Server issue):W758–761.
20. Aerts S, Lambrechts D, Maity S, Van Loo P, et al. Gene prioritization through genomic data fusion. *Nat Biotechnol* 2006;24:537–544.
21. Perez-Iratxeta C, Wjst M, Bork B, Andrade MA. G2D: a tool for mining genes associated with disease. *BMC Genet* 2005;6:45.
22. Franke L, van Bakel H, Fokkens L, de Jong ED, et al. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* 2006;78:1011–1025.
23. Tiffin N, Kelso JF, Powell AR, Pan H, et al. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res* 2005;33:1544–1552.
24. Lage K, Karlberg EO, Størling ZM, Olason PI, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* 2007;25:309–316.
25. Oti M, Snel B, Huynen MA, Brunner HG. Predicting disease genes using protein-protein interactions. *J Med Genet* 2006;43:691–698.
26. Hoffmann H, Valencia A. A gene network for navigating the literature. *Nat Genet* 2004;36:664.
27. Van Vooren S, Thienpont B, Menten B, Speleman F, et al. Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations. *Nucleic Acids Res* 2007;35:2533–2543.
28. Mitchell JA, Aronson AR, Mork JG, Folk LC, et al. Gene indexing: characterization and analysis of NLM's generifs. *AMIA Annu Symp Proc* 2003;460–464.
29. Glenisson P, Antal P, Mathys J, Moreau Y, De Moor B. Evaluation of the vector space representation in text-based gene clustering. *PAC Symp Biocomput.* 2003:391–402.
30. Glenisson P, Coessens B, Van Vooren S, Mathys J, et al. Txtgate: profiling gene groups with text-based information. *Genome Biol* 2004;5:R43.
31. Groth P, Pavlova N, Kalev I, Tonov S, et al. Phenomicdb: a new cross-species genotype/phenotype resource. *Nucleic Acids Res* 2007;35(Database issue):D696–699.
32. Lussier Y, Borlawsky T, Rappaport D, Liu Y, et al. PhenoGo: assigning phenotypic context to gene ontology annotations with natural language processing. *PAC Symp Biocomput* 2006:64–75.
33. Lopez-Bigas N, Ouzounis CA. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res* 2004;32:3108–3114.
34. Adie EA, Adams RR, Evans KL, Porteous DJ, et al. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics* 2005;6:55.
35. Turner FS, Clutterbuck DR, Semple CAM. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol* 2003;4:R75.
36. Franke L, van Bakel H, Diosdado B, van Belzen M, et al. Team: a tool for the integration of expression, and linkage and association maps. *Eur J Hum Genet* 2004;12:633–638.
37. Hauser MA, Li Y-J, Takeuchi S, Walters R, et al. Genomic convergence: identifying candidate genes for Parkinson's disease by combining serial analysis of gene expression and genetic linkage. *Hum Mol Genet* 2003;12:671–677.
38. Ma X, Lee H, Wang L, Sun F. CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics* 2007;23:215–221.
39. Tiffin N, Adie E, Turner F, Brunner HG, et al. Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Res* 2006;34:3067–3081.
40. Elbers CC, Onland-Moret NC, Franke L, Niehoff AG, et al. A strategy to search for common obesity and type 2 diabetes genes. *Trends Endocrinol Metab* 2007;18:19–26.
41. Soldatova LN, King RD. Are the current ontologies in biology good ontologies? *Nat Biotechnol* 2005;23:1095–1098.
42. Lussier YA, Liu Y. Computational approaches to phenotyping: high-throughput phenomics. *Proc Am Thorac Soc* 2007;4:18–25.
43. Zhang X, Snijders A, Segraves R, Zhang X, et al. High-resolution mapping of genotype-phenotype relationships in cri du chat syndrome using array comparative genomic hybridization. *Am J Hum Genet* 2005;76:312–326.
44. Thomas PD, Mi H, Lewis S. Ontology annotation: mapping genomic regions to biological function. *Curr Opin Chem Biol* 2007;11:4–11.
45. Feenstra I, Brunner HG, van Ravenswaaij CMA. Cytogenetic genotype-phenotype studies: improving genotyping, phenotyping and data storage. *Cytogenet Genome Res* 2006;115:231–239.
46. Solberg LC, Valdar W, Gauguier D, Nunez G, et al. A protocol for high throughput phenotyping, suitable for quantitative trait analysis in mice. *Mamm Genome* 2006;17:129–146.
47. Biesecker LG. Mapping phenotypes to language: a proposal to organize and standardize the clinical descriptions of malformations *Clin Genet* 2005;68:320–326.
48. Merks JHM, van Kamebeek CDM, Caron HN, Hennekam RC. Phenotypic abnormalities: terminology and classification *Am J Med Genet A* 2003;123:211–230.
49. Brazma A, Hingamp P, Quackenbush J, Sherlock G, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 2001;29:365–371.
50. Ball CA, Brazma A. Mged standards: work in progress. *Omics* 2006;10:138–144.
51. Redon R, Ishikawa S, Fitch KR, Feuks L, et al. Global variation in copy number in the human genome. *Nature* 2006;444:444–454.
52. Borgaonkar DS. Chromosomal variation in man; a catalogue of chromosomal variants and anomalies. New York: John Wiley & Sons, 1997.
53. Freimer N, Sabatti C. The Human Phenome project. *Nat Genet* 2003;34:15–21.
54. Schinzel A. Catalogue of unbalanced chromosome aberration in man. Berlin:de Gruyter, 2001.