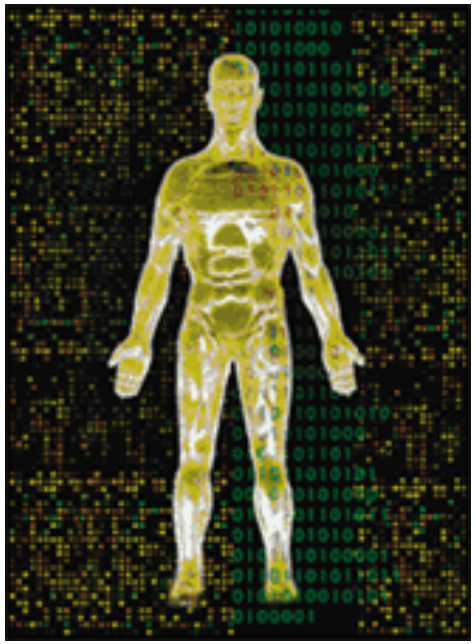
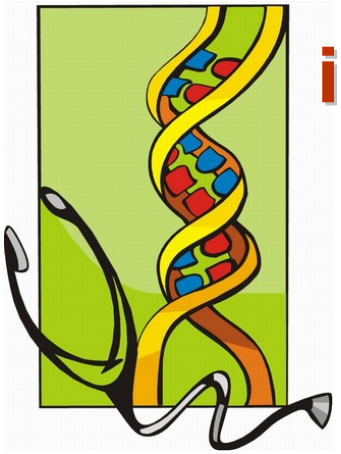
UMC St Radboud



Optimized data interpretation through reliable database integration



Array in daily practice: promises and pitfalls

Amsterdam, May 27, 2011

Nicole de Leeuw

Department of Human Genetics

Radboud University Nijmegen Medical Centre

The Netherlands

Human Genetics Nijmegen

Quality Criteria for array analysis

Objective Classification of CNVs

Available sources for array data interpretation

- In-house sources
- Publicly available internet sources
- Proprietary sources?

In-house sources for interpretation of array data

- I.e., set of healthy individuals used in Nijmegen:
  - ECI study (Excel list) (McMullan et al., 2009, Hum Mutat 30:1082-92)
  - BIG study (Excel list)
  - Healthy parents
  - Access database with all (potentially) pathogenic CNVs

gryNotulen_totaal

Id	A_uitslag array	not_ACTIE	achternaam	aanvrager	not_loss/ga	chromosom	not_aberrat	not_groep	pat/mat
5075	PN10-4868: 460 kb gain in 4q27	array ouders?			gain	4	4q27	Potential	
5077	PN10-4906: 300 kb gain in Xp21.2	array ouders? Ook ~5,4 % homozygoot			gain	23	Xp21.2	Potential	
5080	PN10-4915: 500 kb gain in 15q26.1	array ouders			gain	15	15q26.1	Potential	
5082	PN10-4916: 1,43 Mb loss in 17q12	array ouders? Deletie HNF1B met MLPA aangetoond			loss	17	17q12	De Novo	niet bepaald
5081	PN10-4916: 160 kb gain in 17q12	array ouders?			gain	17	17q12	Potential	
5076	PN10-4937, 2,14 Mb gain 11q	gain op 11 patiënt uit Rotterdam			gain	11	11q12.3q13.1	Potential	
5074	PN10-4942	zeer kleine deletie op 8q			loss	8	8q24.13	Potential	
5083	PN10-5010, 1 Mb gain 15q25.3	1 Mb gain op 15q			gain	15	15q25.3	Potential	
5088	PN10-5032	del proximaal van SMS			loss	17	17p11.2	Potential	
5085	PN10-5064: 9,22 Mb loss in 8p12	array / CO ouders			loss	8	8p12	Potential	
5089	PN10-5122, deletie 2q	6 Mb deletie 2q			loss	2	2q12.2q13	Potential	
5090	PN10-5123, deletie 11q	5,4 Mb deletie 11q22			loss	11	11q22.1q22.3	Potential	
5093	PN10-5144	deletie 4q (net als in PN10-2087)			loss	4	4q21.21q21.22	Potential	
5094	PN10-5145	deletie 16pter (ATR-16 syndroom)			loss	16	16p13.3	Potential	
5095	PN10-5146	22q11 duplicatie			gain	22	22q11.21	Potential	
5106	PN10-5147, dupje 17q en deletie [	DMD deletie exon 45			gain	17	17q11	Potential	
5096	PN10-5154	19q13.11q13.12 duplicatie			gain	19	19q13.11q13.1	Potential	
5097	PN10-5156	zeer kleine deletie 17q24			loss	17	17q24.2	Potential	
4857	PN10-7,44 Mb loss in 22q13.2q13.3	moeder draagster van t(16;22)			loss	22	11q13.2q13.33	De Novo	mat
5057	PN11-0024	grote deletie op 14 ook bij moeder			loss	14	14q24.3q31.3	Inherited	mat
5099	PN11-0026: 1,5 Mb gain in Xp22.33	array ouders			gain	23	Xp22.33	Potential	
5101	PN11-0028: 7,65 Mb gain in 12q21.	array / CO? Ouders			gain	12	12q21.2q21.31	Potential	
5102	PN11-0034: 360 kb loss in 21q22.11	array ouders			loss	21	21q22.11q22.1	Potential	

Publicly available sources for data interpretation

- UCSC Genome Browser: genome.ucsc.edu/
- Database of Genomic Variants: projects.tcag.ca/variation/
- ECARUCA: www.ecaruca.net
- www.ncbi.nlm.nih.gov/
- Ensembl Genome Browser: www.ensembl.org/Homo_sapiens/index.html
- DECIPHER: decipher.sanger.ac.uk/
- Geneimprint: www.geneimprint.com/site/genes-by-species
- Etc.

Proprietary web-based data visualization software

PerkinElmer For the Better

Genoglyphix®

SIGNATURE GENOMICS

Genoglyphix Genome Browser (Build hg18 Mar. 2006)

STATIC BROWSER VIEW

Search Result Genomic Coordinates	chr18:51040559-51406858
Image Display Genomic Coordinates	chr18:50940559-51506858

Another Search | Interactive Browser | Printable Version | Help

SYNDROMES IN REGION

Pitt-Hopkins | More Info

OMIM GENES IN REGION (1 TOTAL)

TCF4 | OMIM

Various options in array diagnostics

Multi-platform comparison for array-CGH in diagnostic constitutional applications and new high throughput workflow

Shuwen Huang

National Genetics Reference Laboratory (Wessex)

New and Developing Technologies for Genetics Diagnostics 2010

DH Department of Health

Aknowledgement

NGRL (W) Microarray Laboratory (NWML)

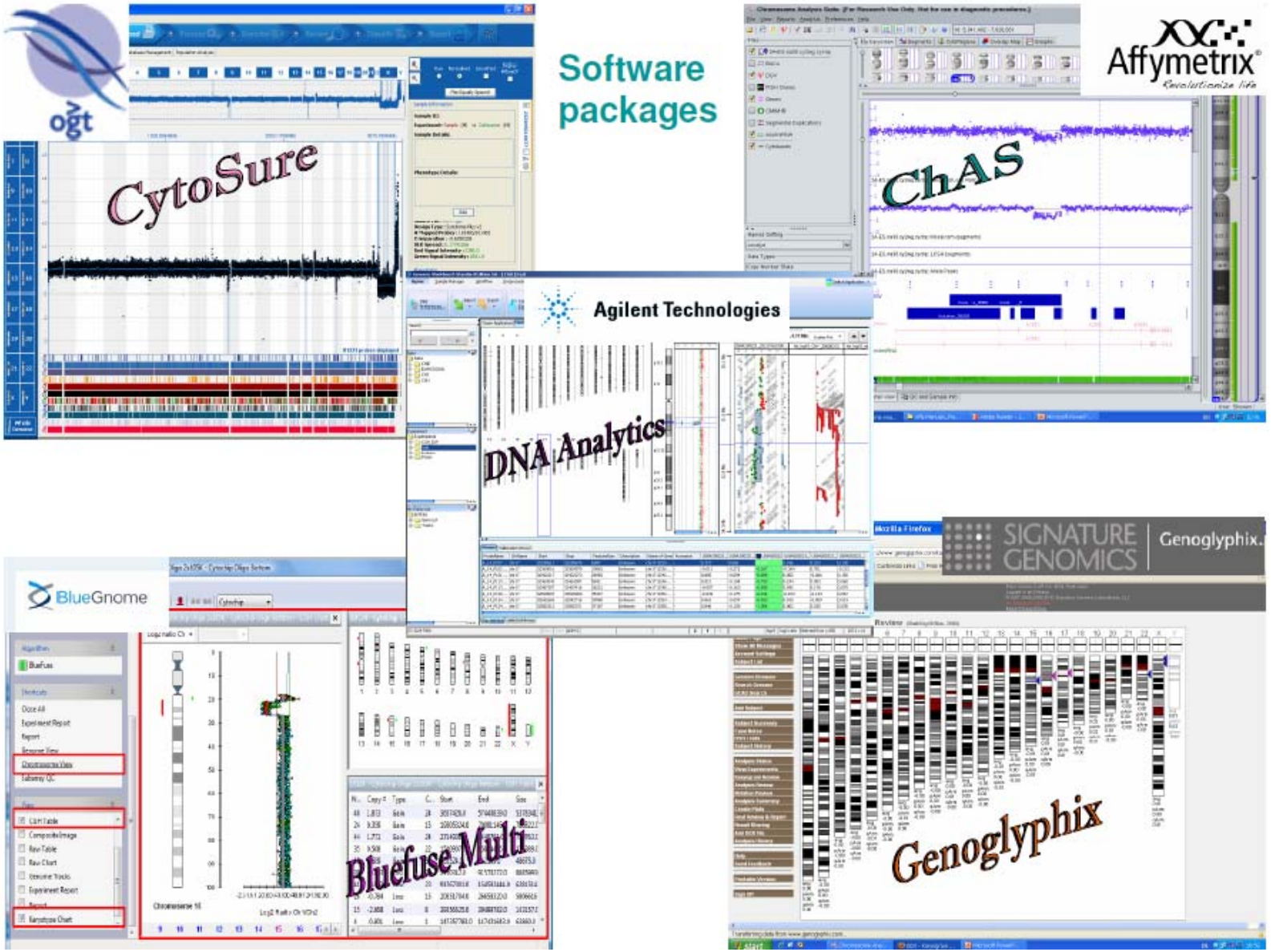
- Agilent Technology
- Affymetrix
- NimbleGen
- Signature Genomics
- Oxford Gene Technology
- BlueGnome
- Wessex Regional Genetics Laboratory (WRGL) high throughput lab (Dan Ward)

Various options in array diagnostics

The basic comparison elements:

- **Core sample set**
  - 12 samples with known and varied chromosomal abnormalities
  - Enough genomic DNA to be run on different platforms
- **Platforms**
  - **Agilent**
    - NGRL 4x44k
    - ISCA 8x60k (OGT)
    - ISCA 4x180k
  - **Affymetrix** — Affymetrix 2.7M Cytogenetics Research Solution
  - **NimbleGen** — NimbleGen CGX-12
- **Software**
  - **For Agilent array**
    - Agilent DNA Analytics
    - BlueGnome Bluefuse Multi
    - OGT CytoSure Interpret
  - **For Affymetrix array** — ChAS
  - **For NimbleGen array** — NimbleScan v2.5 & Genoglyphix

Various options in array diagnostics



Various options in array diagnostics

Software and analysis processing

Array Company	Affymetrix	NimbleGen	Agilent				
Software	**ChAS v1.0.1**	**Genoglyphix v2.4 (Signature Genomics)**	**DNA Analytics v4.0**			**BlueFuse Multi v2.1 (BlueGnome)**	**CytoSure Interpret v3.0.6 (OGT)**
Array platform	**2.7M**	**CGX-12**	**4x180k**	**8x60k**	**4x44k**	**4x44k**	**4x44k**
Total analytical time per case* (Average) in minutes	40	30	45	35	30	45	30
Aberration filters	100kb with 85 markers for amp + del, 50kb with 20 markers for cyto relevant regions 5Mb for LCSH$^{\Psi}$ with 3 markers and mosaicism with 500 markers, 85% Confidence	5 contiguous probes	4 contiguous probes	4 contiguous probes	3 contiguous probes	3 contiguous probes	3 contiguous probes
Algorithms		NG packager segmentation algorithm	ADM-2			Multi v1.0	CBS
Tracks included	BACs, DGV, FISH Clones, Genes, OMIM, Segmental Duplications, sno/miRNA	FISH probes, Sequence Gaps, Segmental Duplications, GC Content, SignatureSelect Clones, SignatureSelect OS 105K Probes, NimbleGen CGX Probes, SignatureSelect OS 44K Probes, Abnormal Region(s), MyGCAD, Community GCAD, GCAD, Benign CNVs, Genes, RefSeq Genes, SGL GPS, SGL CNVs, DGV	Genes, DGV, CpGIsland, miRNA, PAR			Disease, Genes, BlueFISH, BAC Gain/ BAC Loss, Oligo Gain/ Oligo Loss, DGV Gain/ DGV Loss, BG Gain/ BG Loss	Syndrome, Gene, Exon, CHOP CNV, ECARUCA, Recombination hotspot, DGV, Confirmation (FISH and MLPA probes), DECIPHER, Redon CNV
Possibility of adding custom tracks	Yes	Yes	Yes	Yes	Yes	No	Yes

* Total analytical time including analysing, checking and authorising

$^{\Psi}$ LCSH= Long contiguous stretches of homozygosity

How to achieve a reliable interpretation (fast)?

How to achieve a reliable interpretation (fast)?

In-house data set(s)

ECARUCA

ISCA

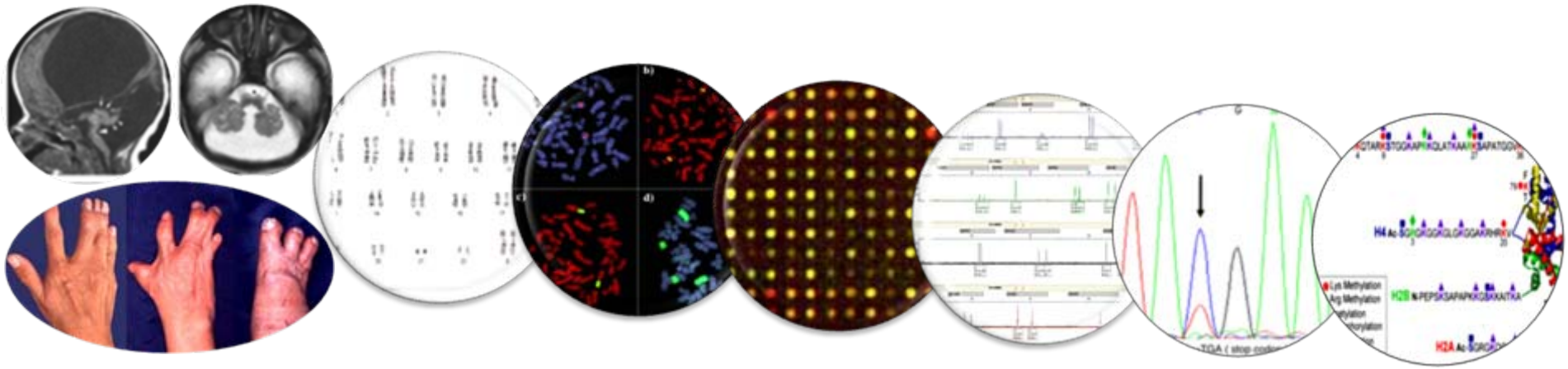
DECIPHER v5.0

U C S C

Database of Genomic Variants

OMIM
Online Mendelian Inheritance in Man

Johns Hopkins University



How to achieve a reliable interpretation (fast)?

GeCCO by hehir-kwa

Hehir-Kwa et al. PLoS Comput Biol. 2010 Apr 22;6(4):e1000752.

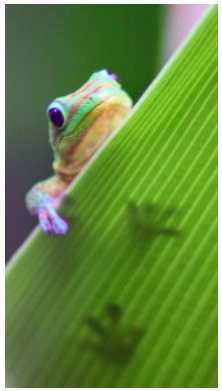
Summary | Files | Reviews | Support | Develop | Tracker | Mailing Lists | Forums | Code

GeCCO (Genomic CNV Classification Objectively) is a bioinformatics tool for classifying copy number variants as either benign or pathogenic.

Project Home
genomegecco.sf.net

Recommend Me
Show some love

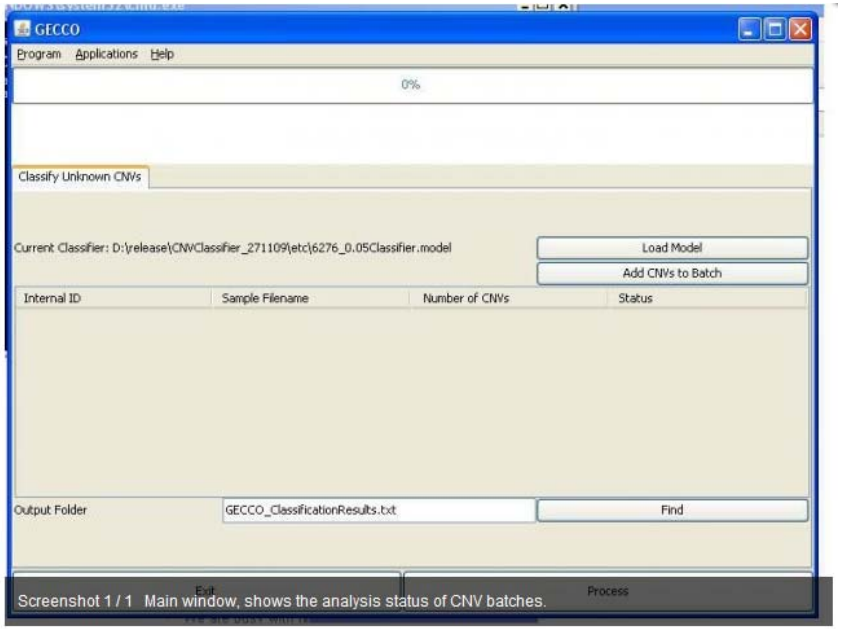
sf Download
test script to check...



Develop
sf.net/projects/genomegecco/develop

Last Update
2010-06-11

Support
sf.net/project/memberlist.php?group_id=281552



UMC St Radboud

How to achieve a reliable interpretation (fast)?

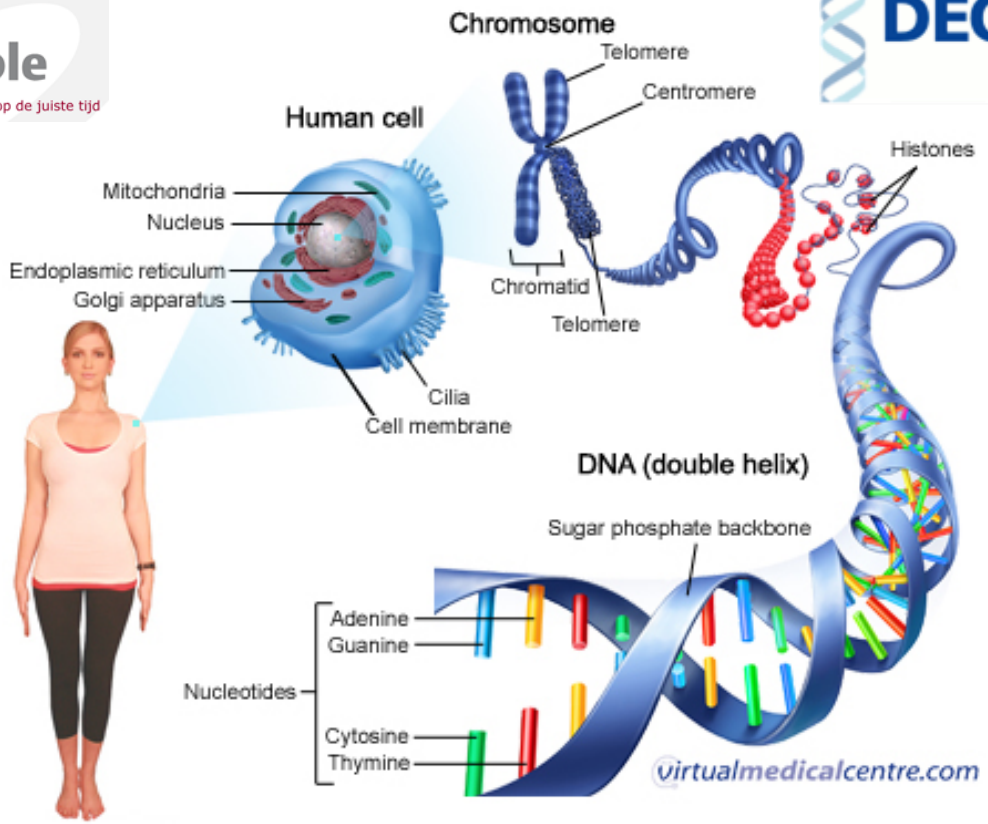
ECARUCA

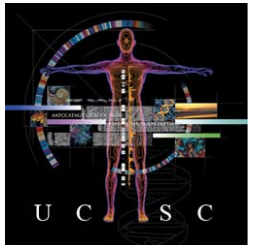
OMIM Online Mendelian Inheritance in Man — Johns Hopkins University

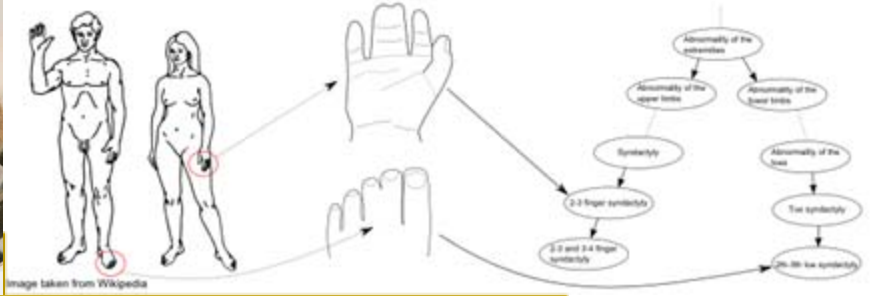
DECIPHER v5.0

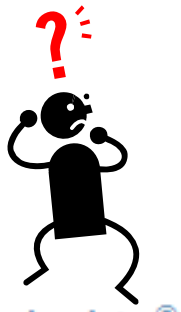
Rimarcable
Kennis op de juiste plek op de juiste tijd

POEMA 3.0

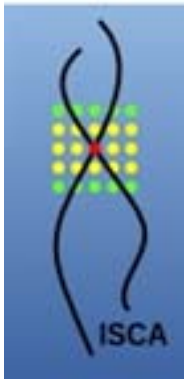








Genoglyphix®



A growing set of sources....

Genomic research generates data on a massive scale and methods to manage and interpret large-scale data are essential to translation of research into biological understanding.

The Wellcome Trust Sanger Institute has developed - often with colleagues outside the institute - a suite of databases to help researchers. These can be accessed from this page.

[Wellcome Library, London]

ARNIE
AVEXIS Receptor Network with Integrated Expression
more

BLAST
The sequencing projects' Blast Search Services
more

COSMIC
Database of somatic mutation information in human cancers
more

DECIPHER
Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources
more

ENCODE
Aims to identify all functional elements in the human genome sequence
more

GeneDB
Annotation from a growing number of organisms is available from the GeneDB web database
more

Ensembl Genome Browser
Ensembl produces and maintains automatic annotation on selected eukaryotic genomes
more

MEROPS
Provides the internationally recognised classification of peptidases and their inhibitors
more

GLIDERS
Genome-wide linkage disequilibrium repository and search engine
more

Rfam
Provides a classification of RNA families using covariance models
more

iPfam
Describes Pfam domain interactions that are observed in PDB entries
more

TreeFam
Provides curated phylogenetic trees of animal genes, that give reliable ortholog and paralog assignments
more

Pfam
Provides a classification of proteins into families and domains using hidden Markov models
more

Wormbase
The primary database on the biology and genome of the model organism Caenorhabditis elegans
more

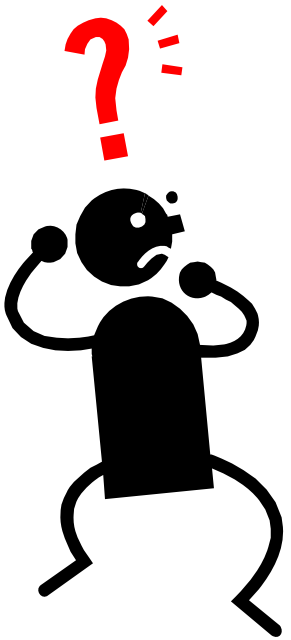
Tiffin
A database of predicted regulatory motifs, a subset with predicted functional annotation
more

ZF-Models
Use of zebrafish to make models for human disease and development
more

VEGA Genome Browser
Central database repository for high quality manual annotation of vertebrate finished genome sequence
more

Related Links

- Projects
- Academic Faculty
- Mouse resources
- Zebrafish resources
- Cancer genome project
- The Genome Campus
- What we do



How to further improve reliable data interpretation?

How to improve the interpretation of genetic data?

→ Search and Submit by using the **DICE**!

DECIPHER – **I**SCA database – **C**ARTAGENIA – **E**CARUCA

To what extent is integration of the available sources possible and achievable?

cartagenia

All tomorrow's genomes

How to further improve reliable data interpretation?

14.10 – 16.30 **Workshops**

(including one coffeebreak at 15.15)

1. Quality criteria and platforms to be used in routine diagnostics (J. Vermeesch)
Aim: define minimum criteria for the quality of platforms and how the platform used is appropriately reflected in the patient report.

2. The interpretation of CNVs and the use of databases (N. de Leeuw)
Aim: how to reach an efficient interpretation strategy? Is web-based integration of the existing databases needed / achievable? How is information in the databases curated?

3. How to deal with unexpected findings / informed consent (G. de Wert)
Aim: guideline on what should or should not be reported, also depending on method of informed consent used. Proposition for information that can be used for pre-test counselling and consent forms with different levels of consent.

4. Array in prenatal diagnostics (D. Ledbetter)
Aim: guideline how to implement, what indications, what to report, informed consent and how to collaborate in order to gain a high level of experience asap

16.30 hrs Plenary session: Feedback from the workshops
17.30 hrs Closure

How to further improve reliable data interpretation?

1. Who can submit to and search the database?
2. To what extent is data **curated**? By who?
3. What are the **objectives** of the database? Research and / or diagnostic purposes?
4. Is the data **publicly available** or are costs involved?
5. Should all array data be made publicly available, regardless whether it comes from diagnostics or research?
6. Differentiation between causative CNVs and 'normal genome variants'. Define the **categories** and indicate whether a database collects all or only certain categories.
7. Is there a differentiation between **prenatal** and postnatal findings?
8. How to ensure reliable details on **clinical information**? Should there be minimum requirements? If so, what kind of minimum requirements?
9. How to ensure **single case registration** per database and between databases? We don't want the same data shown more than once.
10. Is **standardization** of genetic and clinical information achievable?
11. How to maximally **simplify submission** of genetic and clinical data (ideally: submission is only a mouse click away; from local database to international database OR to make use of 'viewing' instead of copying data. Could that be realized? What about national and international security measures?
12. **Informed consent**: the submitter's reliability?
13. Data on **control individuals**: gender and age are important. Do we prefer data from i.e. controls older than 70 years of age to rule out late-onset-diseases?
14. Delicate topic, but may be crucial: information on **ethnicity** of both control and study individuals.

How to further improve reliable data interpretation?

Do we continue with multiple, similar databases or do we want one universal database?

a. Is it an option to combine (parts of) the available databases such as DECIPHER, ISCA and ECARUCA or is it sufficient to have one search engine to query multiple databases?

b. What are the reasons NOT to have one universal database?

c. If we prefer one universal database, how can we achieve this? What do we need to it?

Reliable data interpretation in the near future

Array diagnostics is daily practice
BUT
Next generation sequencing in diagnostics is next

Research:

Re-sequencing of disease loci

Exome sequencing

Whole genome sequencing

Diagnostics:

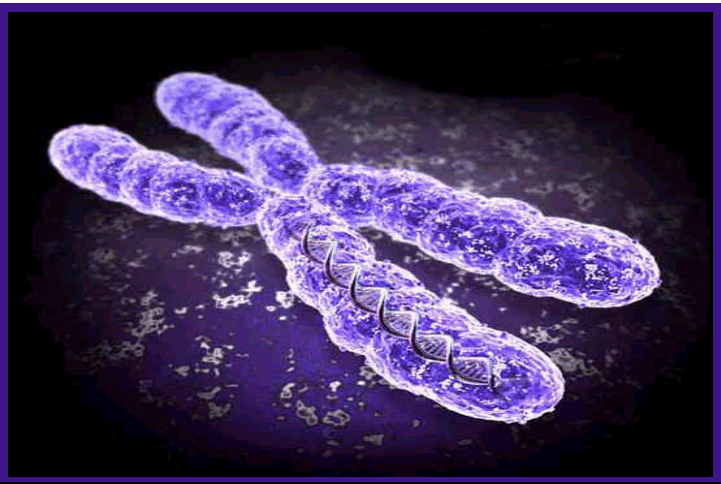
Targeted re-sequencing in complex monogenic diseases

Exome sequencing

Whole genome sequencing

Human Genetics Nijmegen

UMC St Radboud



Sunday, May 29, 2011 - 15.15 - 16.45 hrs

WS03. UCSC Genome Browser - Introductory: Basic functionality and navigation, sharing sessions
Chair: R. Kuhn
Attendees are requested to bring your own laptop for this workshop.
The UCSC Genome Browser (http://genome.ucsc.edu) integrates information from a wide variety of genomic resources, including, but not limited to gene predictions; disease associates, including OMIM and locus-specific databases; gene-expression data; copy-number
variation; comparative genomics; SNPs; HapMap data; gene-chip mappings; and histone- and DNA- modification data.

The large number of genome-wide datasets available allows users to pursue inquiries requiring multiple lines of evidence all in one location. By allowing users the flexibility to select only those data relevant to making a particular scientific argument, the Browser allows the pursuit of inquiry-driven data analysis by displaying all pertinent information in one graphical view.

This session will introduce participants to the basic functions of the Genome Browser (http://genome.ucsc.edu), including interface conventions, navigation and search strategies for effective use of the Browser. The Session tool for sharing the results of inquiries with colleagues will be presented, as will export of camera-ready pdf files for publication.

The Table Browser, a powerful tool for direct mining of the underlying data tables will be demonstrated. Brief mention of the Galaxy toolkit will set the stage for its introduction in the Intermediate session.

Monday, May 30, 2011 - 15.15 - 16.45 hrs

WS10. UCSC Genome Browser 2 / Galaxy
Chairs: R. Kuhn, D. Clements

Gene variant databases, or Locus-Specific Databases (LSDBs), are used to collect and display information on sequence variants on a gene-by-gene basis. They are used most frequently in relation to DNA-based diagnostics, facilitating clinicians', scientists' and patients' and their families' easy access to an up-to-date overview of all gene variants identified world-wide. The databases are taken care of by curators, experts in the field who guard the database, ensure regularly updates and check submitted data.

The purpose of this course is to teach the tasks involved in database curation. Using lectures and practicals, we will go through the entire process of database curation, from starting a database, to entering the first data, curating submissions and advertising the resource which has been built. With the experience gained, participants should feel confident to curate a gene variant database.

The course will be organised with the help of the Human Variome Project (HVP) / Human Genome Variation Society (HGVS) and the EU FP7 GEN2PHEN project. During the course we will use the Leiden Open Variation Database (LOVD) platform.