

GVF: a computable standard for personal genomes data

Mark Yandell
Dept. of Human Genetics
Eccles Institute of Human Genetics
University of Utah & School of Medicine

Genomics enabled EHRs will need to*:

(1) integrate structured genotype and phenotypic information (for accurate clinical interpretation),

(2) insert genetic/genomic information into the clinical workflow (for streamlined processes)

(3) enable secondary use of the data

* *Emerging Landscape of Genomics in the Electronic Health Record for Personalized Medicine* Mollie H. Ullman-Cullere and Jomol P. Mathew. Human mutation 2011

This means that:

- EHR systems need to be adapted for the personalization of medicine enabled through genetics/genomics data.
- As a first step, structured genetic/genomic data must be available within the EHR in a **computable** and **consistent** format.

Outline

- GVF – a computable, consistent & clinically orientated data standard for personal genome sequences
- GVF can be used to describe a single variant, a gene's variants, or an entire genome's
- Proof-of-principle whole-genome comparative analyses enable by GVF
- GVF enabled software tools for clinical decision support

Its all in the variant file

- Variant files have become the *de facto* 'standard' for personal genome sequences
- Each variant file contains about 3 million SNVs compared to the reference human genome

Variant files are not standardized

This is a problem.

Soap SNP

```
chr1 SoapSNP SNP 4793 4793 25 + . ID=YHSNP0128643; status=novel; ref=A; allele=A/G; support1=48; support2=26;
chr1 SoapSNP SNP 6434 6434 48 + . ID=YHSNP0128644; status=novel; ref=G; allele=A/G; support1=10; support2=11;
chr1 SoapSNP SNP 93896 93896 51 + . ID=rs4287120; status=dbSNP; ref=T; allele=C/T; support1=5; support2=4;
location=MSTB1:LTR/MaLR;
```

Watson Genome SNP

```
BJW-1117373 chr1 41921 G C . novel . 2 0 4 het
BJW-1117523 chr1 42101 T G Y rs2691277.1 . 1 0 1 ?
BJW-1119675 chr1 45408 C T Y rs28396308 . 3 0 3 .
```

Venter Genome SNP

```
1 1103675000013 heterozygous_SNP 556001 556002 . + A/C;RMR=0;TR=0 Method1
1 1103675000017 homozygous_SNP 652719 652720 . + G/A;RMR=0;TR=1 Method1
1 1103675000019 homozygous_SNP 694229 694230 . + T/C;RMR=1;TR=0 Method1
```

Korean Genome SNP

```
chr10 56397 C CT rs12262442 28 C/T 17 11
chr10 61776 T CT rs61838967 15 T/C 7 8
chr10 65803 T CT KOREFSNP1 27 T/C 19 8
```

Complete Genomics SNP

```
6,chr1,31843,31844,snp,snp,A,G,G,235
21,chr1,36532,36533,snp,snp,A,G,G,36
23,chr1,36970,36971,snp,snp,G,C,C,109
```

Badly needed : a standard format

- ◆ Currently every sequence provider uses their own idiomatic data format for variant files.
- ◆ The first step towards enabling analyses of multiple personal genomes is a standardized data format to facilitate comparisons.
- ◆ This is not the first time the genomics community has faced this problem.

This isn't a new problem for genomics

- 1998 onward the model organism community worked together to facilitate genome annotation

The Generic Model Organism Database Community

- Genomic browsers like Apollo and GBrowse
 - Data storage such as Chado relational Schema
 - Data exchange such as GFF3
 - Annotation pipelines such as MAKER
- One of the main take homes was that an ontology was needed to type fields in both files and databases; hence the Sequence Ontology

SO

The Sequence Ontology Project

[Home](#) [Browser](#) [Wiki](#) [GFF3](#) [Resources](#) [About](#) [Request A Term](#) [Site Map](#)

[Home](#) > [Resources](#) > [GVF](#)

SO: tool for the unification of genome Annotations.

Eilbeck K, Lewis SE, Mungall CJ, Yandell MD, Stein L, Durbin R, Ashburner M.
Genome Biology 2005, 6:R44

GVF Pragmas

GFF3 allows for pragmas that define file-wide directives to processing software. All pragmas from GFF3 are included and GVF adds the following pragmas and defines a set of tag-value pairs for use with these pragmas.

News

- ▶ **2009 December** Release 2.4.1 available.
- ▶ **2009 October** Release 2.4 available.
- ▶ **2009 July** SO presented at the International Conference of Biomedical Ontology.
- ▶ **2009 June** Chris Conley -Undergrad from BYU joins SO for the summer
- ▶ **2009 May** Graduate student, John Naylor joins SO.

This is what a gene annotation looks like naked

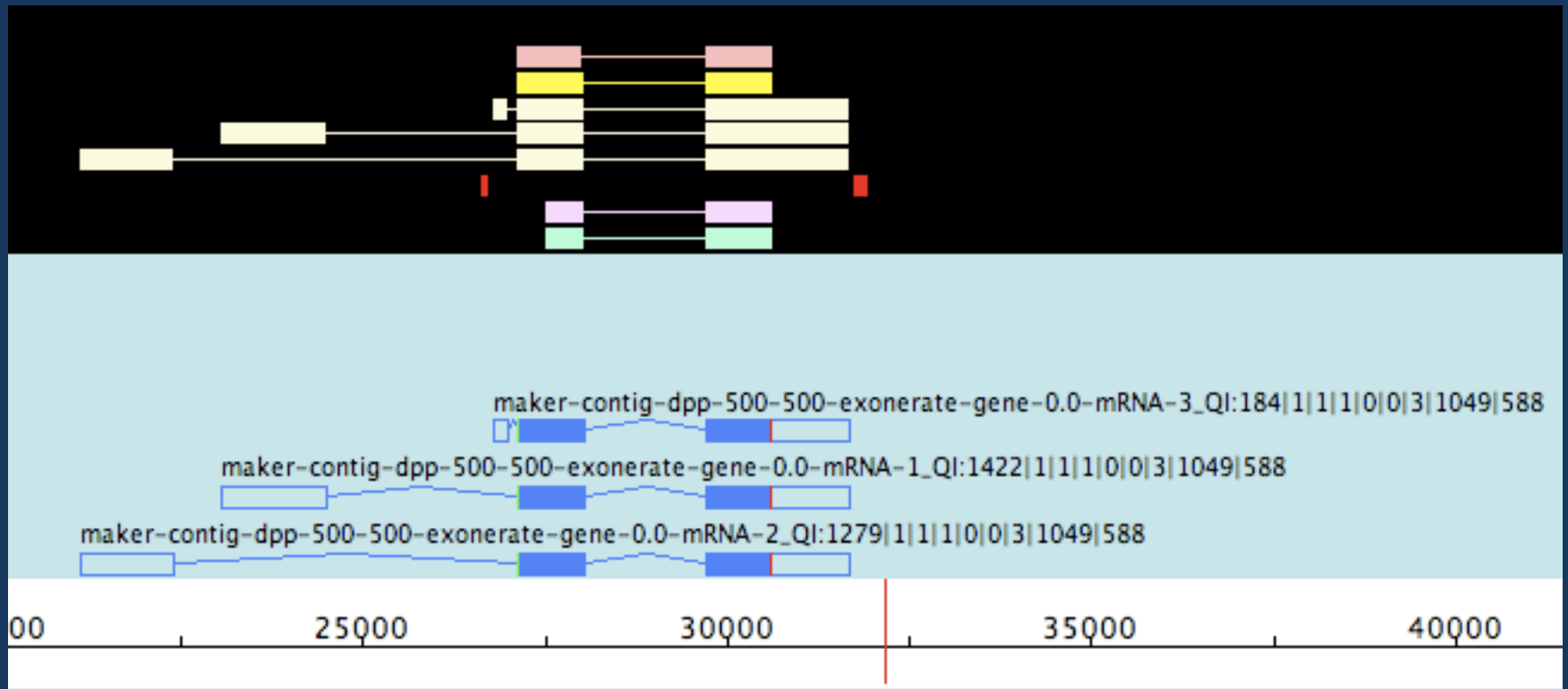
GFF3 representation of a gene annotation

```
##gff-version3
ID      source  feature  start    stop      score    strand  phase  attributes
chr17  UCSC    mRNA     62467934 62469545  .        -        .      ID=A00469;...
chr17  UCSC    three_prime_UTR` 62467934 62468038  .        -        .      Parent=A00469
chr17  UCSC    CDS      62468039 62468236  .        -        1      Parent=A00469
Chr17  UCSC    CDS      62468490 62468654  .        -        2      Parent=A00469
Chr17  UCSC    CDS      62468747 62468866  .        -        1      Parent=A00469
chr17  UCSC    CDS      62469076 62469236  .        -        1      Parent=A00469
chr17  UCSC    CDS      62469497 62469506  .        -        0      Parent=A00469
chr17  UCSC    five_prime_UTR 62469507 62469545  .        -        .      Parent=A00469
chr9   UCSC    mRNA     90517946 90527968  .        -        .      ID=AB000114;Ontology_term=GO:0007155,GO:0005194,GO:0005578;Dbxref=AFFX-U95:41031_at,Genbank-protein:BAA19055;;
chr9   UCSC    three_prime_UTR 90517946 90518841  .        -        .      Parent=AB000114
chr9   UCSC    CDS      90518842 90519167  .        -        1      Parent=AB000114
chr9   UCSC    CDS      90520309 90521248  .        -        0      Parent=AB000114
chr9   UCSC    five_prime_UTR 0521249 90521264  .        -        .      Parent=AB000114
chr9   UCSC    five_prime_UTR 90527892 90527968  .        -        .      Parent=AB000114
```

SO
terms

SO
relations

Standardized file formats enable downstream software applications for analysis



A GFF3 genome annotation visualized in the Apollo Genome Browser

Data standards have great power to unify and enable the community

Main Page - GMOD

http://gmod.org/wiki/Main_Page

page discussion view source history

Welcome to GMOD

GMOD is the **Generic Model Organism Database** project, a collection of open source software tools for creating and managing genome-scale biological databases. You can use it to create a small laboratory database of genome annotations, or a large web-accessible community database. GMOD tools are in use at [many large and small community databases](#).

How do I Get Started?

See [Overview](#) for the big picture. For an introduction to specific GMOD components see the list of the most popular tools at the right, or visit [GMOD Components](#) for a comprehensive list of GMOD tools. If GMOD looks promising for your needs, consider attending the next [GMOD community meeting](#).

How do I Get Support?

GMOD support is available from several different sources. [Support](#) introduces each support option (this web site, [GMOD Mailing Lists](#), [Training and Outreach](#) activities (including [GMOD Schools](#)), and the [GMOD Help Desk](#)) and offers guidance on which one is the most appropriate for your question.

How do I Get Involved?

As an open source project GMOD relies on the [donation of time and software](#) by groups and individuals. Contribution of new tools, adoption of existing ones, and [improving the documentation](#) are all welcome. [Existing](#) and potential users are encouraged to provide feedback via [mailing lists](#) or the [help desk](#). The [GMOD Project Page](#) lists projects in need of ideas and developers. You can also attend [project meetings](#). The next project meeting will be held in [October 2011](#) at the [Ontario Institute for Cancer Research \(OICR\)](#) in Toronto, Canada.

Contributing Organizations

FlyBase WormBase wFleaBase CSH GRAMENE NESCent MGI dictyBase RGD BioCyc SGD EcoliWiki Berkeley Lab Penn State iPlant Collaborative CUIGI Emory University

GMOD HELP DESK

Now Hiring: GMOD Help Desk

Galaxy 2011

Community Conference

Register by April 24 and save 20%

GMOD News Add

FlyBase jobs

Entagen Bioinformatican Wanted Positions @ Bayer CropScience

JBrowse is Hiring

Galaxy Events: April 2011

Account Creation Temporarily Disabled

GMOD Americas 2011 Report

march 2011 meeting twitter hashtag

Planned downtime for gmod.org

InterMine 0.96 Release

New & Revised Pages

- MAKER Tutorial
- News/FlyBase jobs
- Sandbox
- News/Entagen
- Bioinformatican Wanted
- GBrowse
- PopUp Balloons
- GBrowse Configuration HOWTO
- News/Positions @ Bayer CropScience
- Galaxy
- Talk:Bio::Chado::Schema
- Bio::Chado::Schema

Popular GMOD Tools

Genome Browsing and Editing

- GBrowse: Genome annotation viewer
- Apollo: Genome annotation editor

Comparative Genomics

- CMap: Comparative map viewer
- GBrowse_syn: Synteny viewer

Database Tools

- Chado: Biological database schema

GVF: a standardized file format for variation files



[Home](#) [Browser](#) [Wiki](#) [GFF3](#) [Resources](#) [About](#) [Request A Term](#) [Site Map](#)

[Home](#) > [Resources](#) > [GVF](#)

GENOME VARIATION FORMAT

A standard variation file format for human genome sequences

Martin G. Reese¹⁺⁺, Barry Moore², Colin Batchelor³, Fidel Salas¹, Fiona Cunningham⁶, Gabor Marth⁵, Lincoln Stein⁵, Paul Flicek⁶, Mark Yandell², and Karen Eilbeck²⁺⁺.
Genome Biology 2010.

GVF Pragmas

GFF3 allows for pragmas that define file-wide directives to processing software. All pragmas from GFF3 are included and GVF adds the following pragmas and defines a set of tag-value pairs for use with these pragmas.

Omicia Inc.
U of Utah
Royal Society of Chemistry
Boston College
1000 genomes project
Ontario Institute for Cancer Research
EBI/Ensembl
Sequence Ontology project

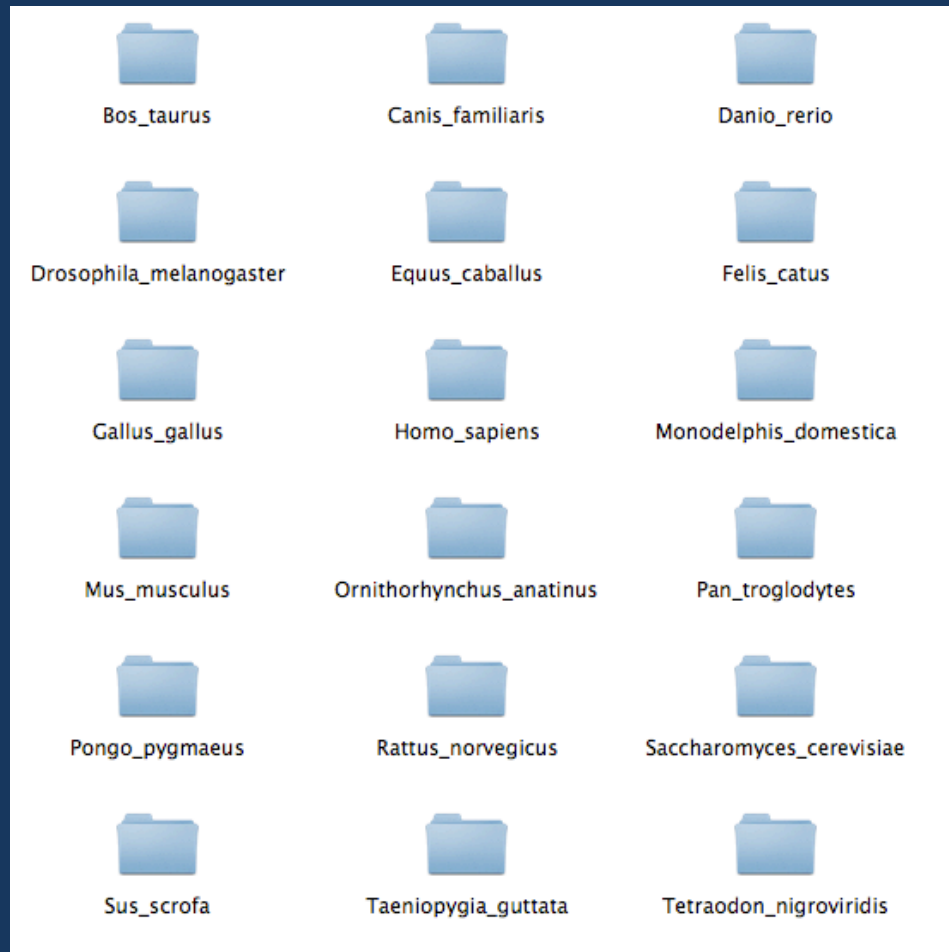
Advantages of GVF

- Describe a personal genome, a population or collection of variants.
- Descriptive terms are Ontology typed. Therefore file can be reasoned over computationally
- Library of existing software for analysis and visualization
- Multiple technologies can be represented
- Permits detailed annotation of physical manifestation of variant

How is GVF different from VCF?

- Descriptive terms typed via the Sequence Ontology.
- Library of existing software for analysis and visualization (GMOD tools)
- Focus is on clinical annotation and functional consequence of variant; e.g. splice junction variant causing exon skipping in DMD.

GVF is now the output of EBI variant annotation pipeline



<ftp://ftp.ensembl.org/pub/current/variation/>

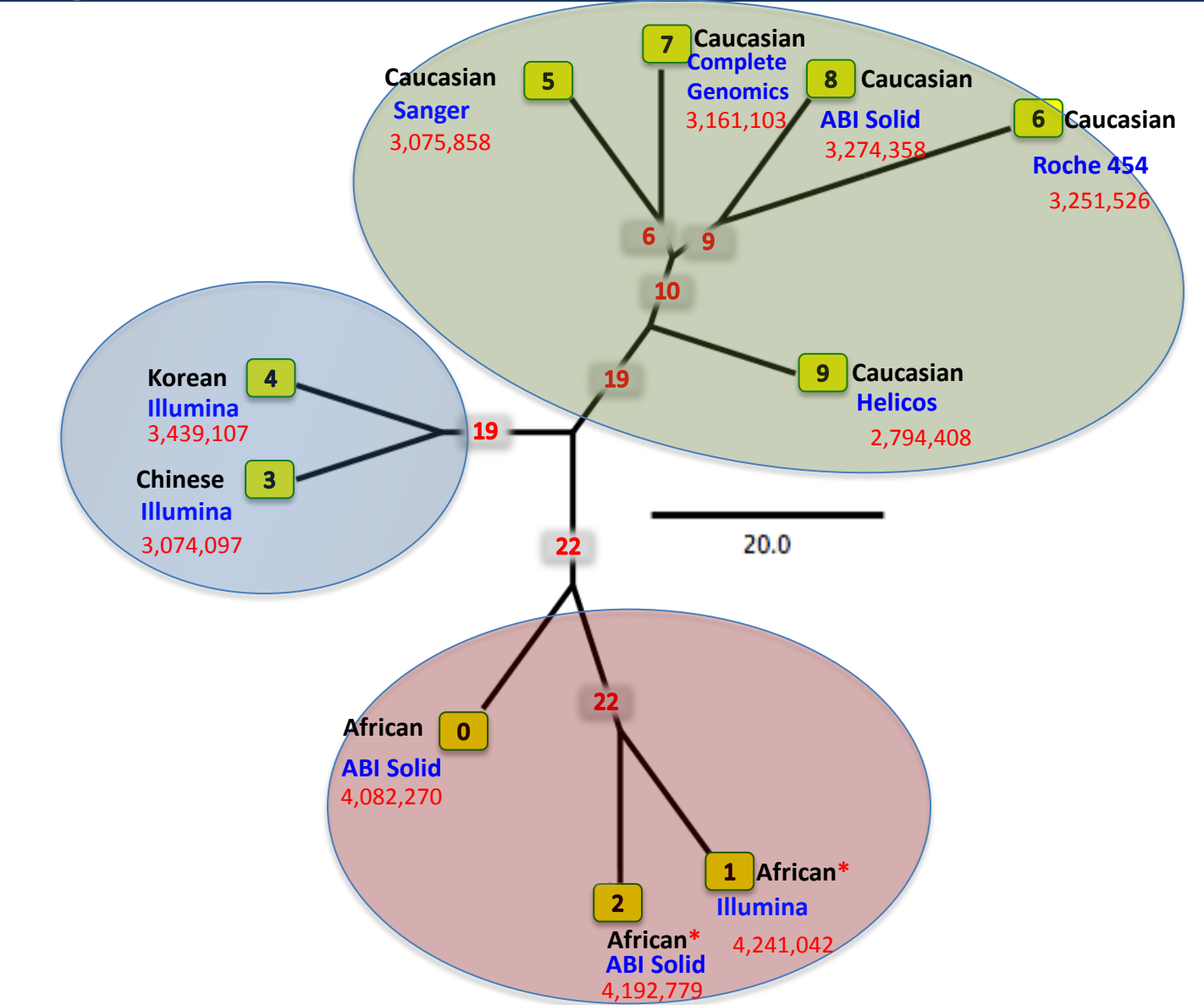
10Gen Dataset

<http://www.sequenceontology.org/resources/10Gen.html>

Genome	Individual	Ethnicity	Platform	Reference
0	NA19240	African	ABI SOLiD	De la Vega, et al. 2009
1	NA18507	African	Illumina	Bentley et al. 2008
2	NA18507	African	ABI SOLiD	McKernan, et al. 2009
3	Chinese	Asian	Illumina	Wang et al. 2008
4	Korean	Asian	Illumina	Ahn et al. 2009
5	Venter	Caucasian	Sanger	Levy et al. 2007
6	Watson	Caucasian	Roche 454	Wheeler et al. 2007
7	NA07022	Caucasian	CGenomics	Drmanac, et al. 2009
8	NA12878	Caucasian	ABI SOLiD	De la Vega, et al. 2009
9	Quake	Caucasian	Helicos	Pushkarev et al. 2009

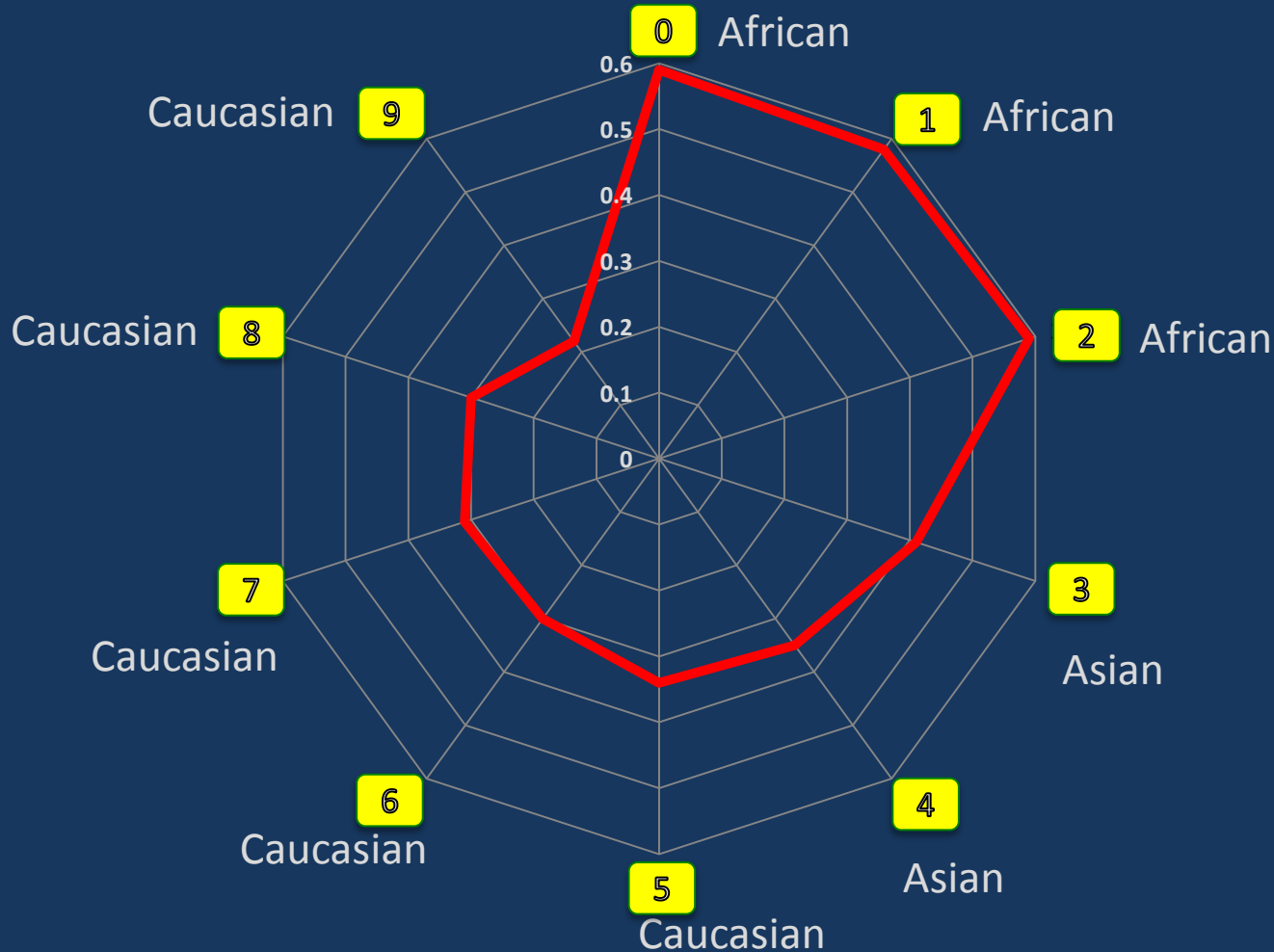
A standard variation file format for human genome sequences Martin G. Reese¹⁺⁺, Barry Moore², Colin Batchelor³, Fidel Salas¹, Fiona Cunningham⁶, Gabor Marth⁵, Lincoln Stein⁵, Paul Flicek⁶, Mark Yandell², and Karen Eilbeck²⁺⁺.
Genome Biology 2010.

Standards enable comparative genomics of people, platforms and variant calling methods



Global analysis of disease-related DNA sequence variation in 10 healthy individuals: Implications for whole-genome-based clinical diagnostics. Barry Moore, Hao Hu, Marc Singleton, Martin G. Reese, and Mark Yandell. Genetics in Medicine 2011

African Genomes Are Generally Homozygous For More OMIM Alleles*



*Ratio homozygous/heterozygous positions

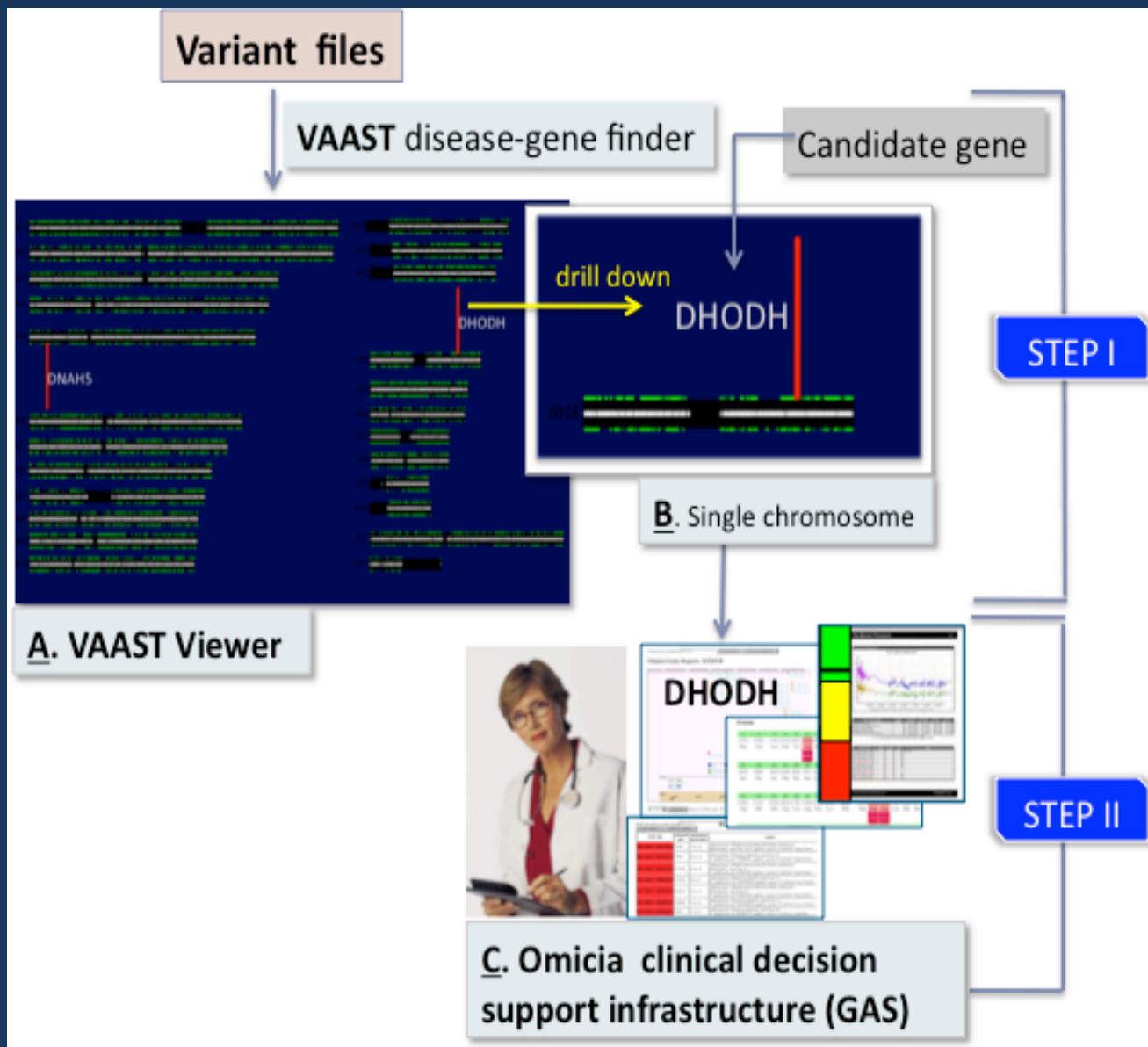
Global analysis of disease-related DNA sequence variation in 10 healthy individuals: Implications for whole-genome-based clinical diagnostics

Barry Moore, Hao Hu, Marc Singleton, Martin G. Reese, and Mark Yandell. *Genetics in Medicine* 2011

Standards enable tools for clinical decision support

VAAST

Omicia Tools



GVF is input to the Omicia Inc. Genome Annotation Station

Omicia Annotation Station - Variant Mining Report - 10Gen Korean with indels

http://sun.omicia.com:8080/test/index.php?id=11&genomelid=200018 UGT1A1

Edward ▾ News ▾ Blogs ▾ Travel ▾ Shop ▾ Home ▾ FootyTube ▾ Twitter / Home ▾ Omicia Mail ▾ Annotation Station

Omicia Mail - Priorit... Omicia Annotation St... Omicia Gene Report: ... Variant Annotation R... Variant Annotation R... UDP glucuronosyltra...

Omicia Annotation Station

ekiruluta@omicia.com | Home | Settings | Report a bug | Help | Sign out

Gene Symbol

Disease Category

- All
- Aging
- Cardiovascular
- Dental
- 17 more ▾

Omicia Gene Sets

- Cardiology
- Cancer
- Neurological - Parkinsons
- Neurological - Alzheimers
- Neurological - Epilepsies
- Respiratory Systems 2
- Psychiatric
- Aging
- Aging (b)

My Gene Sets

Drug Target

Pathway

Filter by

Variant Quality

Frequency

Disease Evidence

Show Only

- Homozygous
- Nonsynonymous
- Stop Gained/Lost
- Insertion/Deletion

Exclude

Export

Variant ID Rs #	Chrom Position	Change	Gene	Zygoty	Consequence	Phred Score Reads	Frequency	Disease Evidence
18 (rs1061170)	chr1 194925860	C→T,T	CFH	homozygous	non-synon	NA 19(0:19)	44.04%	<i>pgkb</i> : Macular Degeneration
19 (rs1065489)	chr1 194976397	G→T,T	CFH	homozygous	non-synon	NA 24(0:24)	19.72%	<i>hgmd</i> : Haemolytic uraemic syndrome, association with (pubmed , omim)
20	chr1 205010882	C→C,T	IL10	heterozygous	non-synon	NA 22(12:10)	0.23%	
21 (rs1051740)	chr1 224086256	T→C,T	EPHX1	heterozygous	non-synon	NA 37(17:20)	27.75%	<i>omim</i> : Lymphoproliferative Disorders, Susceptibility To Preeclampsia, Susceptibility To, Included., Emphysema, Susceptibility To, Included., Pulmonary Disease, Chronic Obstructive, Susceptibility To, Include <i>hgmd</i> : Epoxide hydrolase deficiency, association with (pubmed , omim) <i>pgkb</i> : Craniofacial Abnormalities
22 (rs2234922)	chr1 224093029	A→A,G	EPHX1	heterozygous	non-synon	NA 15(7:8)	19.27%	<i>omim</i> : Epoxide Hydrolase Polymorphism <i>hgmd</i> : Preeclampsia, association with (pubmed , omim)
23 (rs1937)	chr10 59815348	G→C,G	TFAM	heterozygous	non-synon	NA 10(2:8)	8.72%	<i>hgmd</i> : Alzheimer disease, late-onset, reduced risk, association with (pubmed , omim)

100 Page 1 of 2 Displaying 1 to 100 of 138 items

Disease categories

Known disease genes

Support

Kind of alteration

GVF is the input for VAAST*

- A Probabilistic disease-gene finder for personal genomes
- Rapidly search personal genome sequences for genes having significant differences in variant frequencies vs. controls
- Identify novel disease-causing genes & their variants
- Can be used to hunt for both rare and common disease genes and their causative alleles
- Determine the statistical significance of candidate genes

*A probabilistic disease-gene finder for personal genomes Mark Yandell, Chad Huff, Hao Hu, Marc Singleton, Barry Moore, Jinchuan Xing, Lynn Jorde and Martin G. Reese.
Manuscript under review.

VAAST rhymed with BLAST

BLAST

query

database

hits

Expect

Fast

VAAST

target genomes

background genomes

hits

P-value

Fast

BLAST searches for statistically significant *similarity* between sequences.

VAAST searches for statistically significant *dissimilarity* between sequences.

A Test Case: MILLER SYNDROME

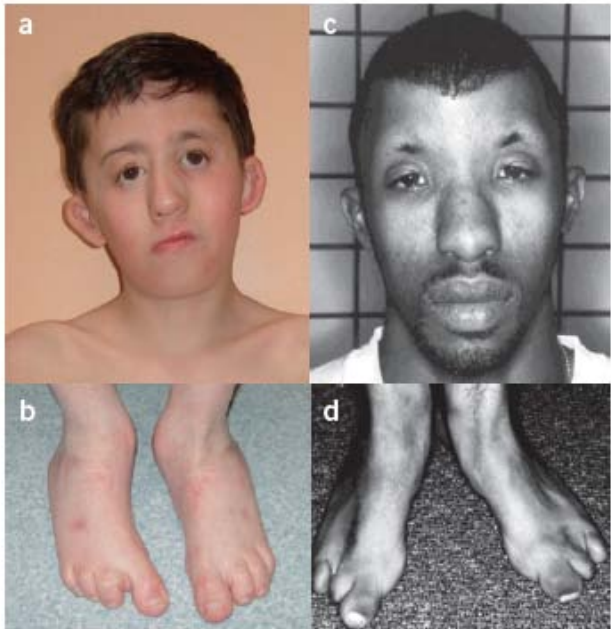


Figure 1 Clinical characteristics of an individual with Miller syndrome and an individual with methotrexate embryopathy. (a,b) A 9-year-old boy with Miller syndrome caused by mutations in *DHODH*. Facial anomalies (a) include cupped ears, coloboma of the lower eyelids, prominent nose, micrognathia and absence of the fifth digits of the feet (b). (c,d) A 26-year-old man with methotrexate embryopathy. Note the cupped ears, hypertelorism, sparse eyebrows and prominent nose (c) accompanied by absence of the fourth and fifth digits of the feet (d). c and d are reprinted with permission from ref. 30.

ARTICLES

nature
genetics

Exome sequencing identifies the cause of a mendelian disorder

Sarah B Ng^{1,10}, Kati J Buckingham^{2,10}, Choli Lee¹, Abigail W Bigham², Holly K Tabor^{2,3}, Karin M Dent⁴, Chad D Huff⁵, Paul T Shannon⁶, Ethylin Wang Jabs^{7,8}, Deborah A Nickerson¹, Jay Shendure¹ & Michael J Bamshad^{1,2,9}

Science

AAAS

Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing

Jared C. Roach, *et al.*

Science **328**, 636 (2010);

DOI: 10.1126/science.1186802

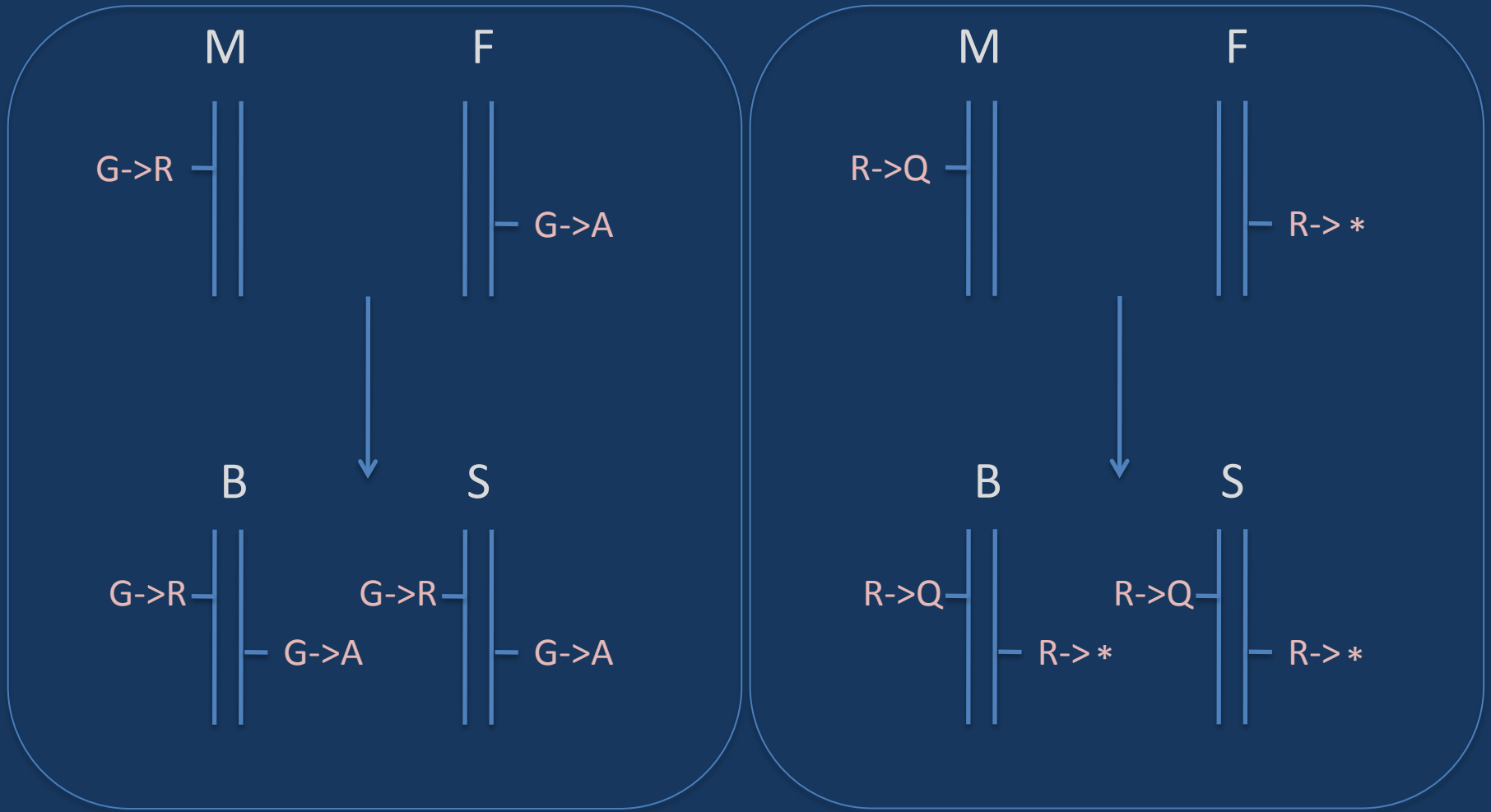
Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing

Jared C. Roach,^{1*} Gustavo Glusman,^{1*} Arian F. A. Smit,^{1*} Chad D. Huff,^{1,2*} Robert Hubley,¹ Paul T. Shannon,¹ Lee Rowen,¹ Krishna P. Pant,³ Nathan Goodman,¹ Michael Bamshad,⁴ Jay Shendure,⁵ Radoje Drmanac,³ Lynn B. Jorde,² Leroy Hood,^{1†} David J. Galas^{1†}

Alleles Responsible for MILLER SYNDROME in Utah Kindred*

CHR 16: DHODH

CHR 5: DNAH5



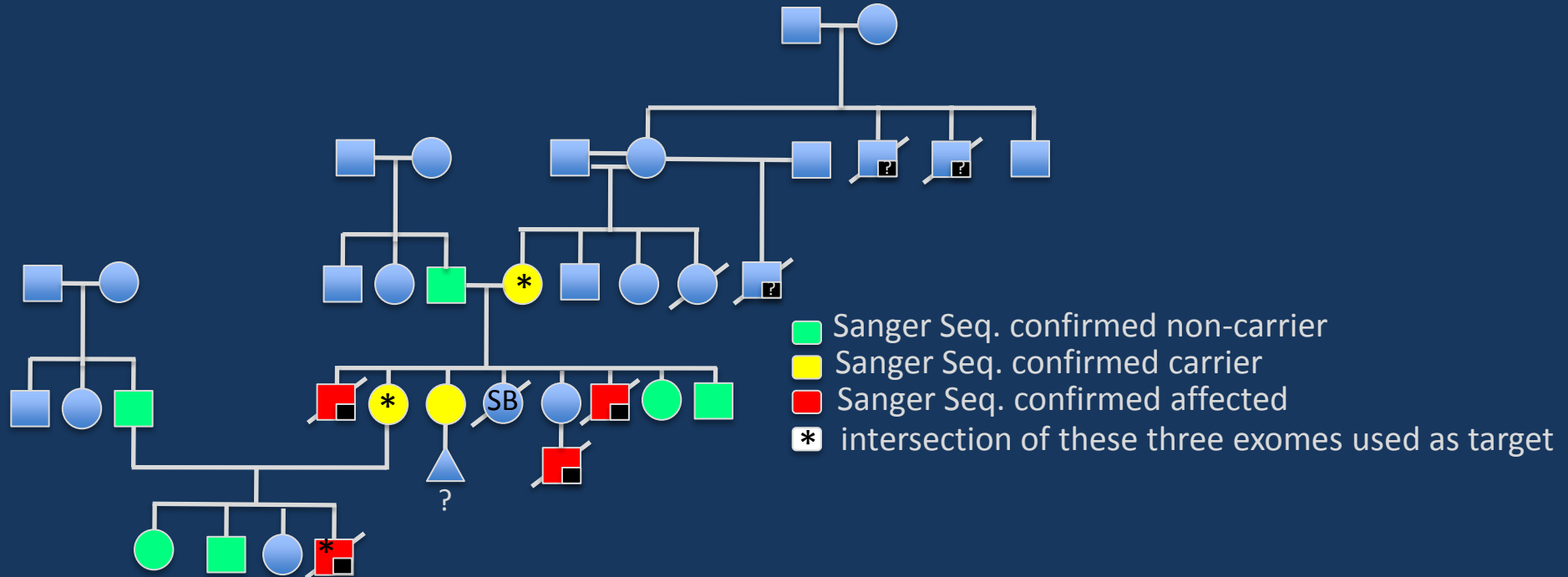
•Ng *et al*, Nature Genetics 42, 30–35 (2010) doi:10.1038/ng.499

•Roach, *et al*, Science, 328 636, 2101

Schematic of VAAST Analysis of MILLER Kindred 1 using a single quartet : *only two candidate genes*



10/20/2010: VAAST identifies its first new human genetic disease



High-throughput sequencing identifies an infantile lethal disorder caused by protein N-terminal acetyltransferase deficiency

N- α -acetyltransferase 10

$P < 1 \times 10^{-8}$

Alan F. Rope, Kai Wang, Rune Evjenth, Jinchuan Xing, Jennifer J. Johnston, Jeffrey J. Swensen, W. Evan Johnson, Barry Moore, Chad D. Huff, Lynne M. Bird, John C. Carey, John M. Opitz, Catherine A. Stevens, Christa Schank, Heidi Deborah Fain, Reid Robison, Brian Dalley, Steven Chin, Sarah T. South, Theodore J. Pysher, Lynn Jorde, Hakon Hakonarson, Johan R. Lillehaug, Leslie G. Biesecker, Mark Yandell, Thomas Arnesen, Gholson J. Lyon; submitted

Next Steps for GVF

- Expand support for phenotype and clinical annotation
- Flesh out representation of single gene diagnostic descriptions (Collaboration with K. Voelkerding, ARUP)
- Develop HL7 compliant XML DTD for embedding GVF in EHRs

Acknowledgements (SO)

- **P41** – NHGRI (PI Blake)
 - Supports the development of The Sequence Ontology & The Gene Ontology
- **R01** – NHGRI (PI Eilbeck)
 - Supports the development of ontology enabled software
- **SBIR** – NLM (multi PI Reese & Eilbeck)
 - Supports variant/disease annotation

Acknowledgements (VAAST project)

Leppert Lab
Jorde Lab



Marth Lab
Gabor Marth



Fidel Salas
Edward S. Kiruluta
Steve Chervitz
Archie Russell
George Miklos
Paul Billings
Erwin Frise
Martin Reese



Francisco de la Vega
Kevin McKernan

This work was supported by NIH SBIR grants 1R4HG003667 to Omicia/Yandell, SBIR 1R44HG002991 to Omicia and an NIH ARRA GO grant 1RC2HG005619-01 to Yandell/Omicia all administered by the National Human Genome Research Institute (NHGRI).

OMIM Alleles Are Distributed Along Ethnic Lines

