Integrating Exome Variants with Other Genomic Data and Functional Annotations

David Adams MD PhD
Pediatrics/Biochemical Genetics
William Gahl Laboratory/MGB/NHGRI
Undiagnosed Diseases Program/OD/DIR

Introduction

- Practicing pediatrician/medical geneticist
- Research interests
  - Diagnostic dilemmas
  - Biochemical genetics
  - Inherited pigmentation disorders
- Next generation sequencing
  - Undiagnosed Diseases program
    - Families/individuals with mystery syndromes
    - Often requires an "agnostic" approach
      - No preexisting clues, similarity to prior projects, etc
- Will present examples and ideas from multiple UDP projects and work with collaborators

Terminology

- *Next Generation or NextGen*
  - Any of the new technologies that attempt to sequence an entire cell's worth of DNA or genes or transcripts
  - e.g. the "-omes" exome, genome, transcriptome
- *Variant*
  - A difference from a defined reference sequence
- *Pathogenic variant*
  - A variant that is wholly or partially responsible for a phenotype of interest (≈ mutation)
- *Candidate variant or candidate*
  - A variant with characteristics suggesting that it may be a pathogenic variant

Outline and Scope

1. Next Generation Project Design Considerations
2. Integration of Next Generation Techniques with Other Genetic Analyses and Data
   1. SNP arrays
   2. Phenotype and family history data
3. Validation and Reanalysis
   1. Functional validation
   2. Strategies to reanalyze uninformative datasets

Outline and Scope

- Included
  - Mostly exome sequencing
  - Rare variants
    - High penetrance, high effect, small number of genes
  - Humans
- Not-included
  - Cancer/somatic comparison projects
  - Common variants
    - Low penetrance, small effect, possibly many genes
  - Non-Humans
- Nonetheless, some overlap

Project Design: The Biggest Question

How will I know if a candidate variant is the pathogenic variant I am looking for?

Project Design: How Can I Improve My Chances of Success?

- Careful project selection
- Parallel analyses (a few examples)
  - SNP chip array
  - Extensive phenotyping
  - Expression Analysis
- Consider variables in experimental design
  - Number of pedigree members to sequence
  - Spectrum of collaboration
    - Sequencing → Analysis → Validation
  - Involvement in Analysis
    - Alignment, genotype assignment, quality measurement, annotation, candidate variant identification, filtrating, other analyses

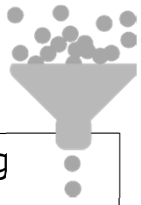
Project Design: Project Selection Example Tool

Criterion	**Less Interesting (1)**	**…**	**Intermediate (3)**	**…**	**More Interesting (5)**
Phenotype	Multifactorial	…		…	Genetic (early onset, severe, developmental pattern)
Material	Single Individual	…	Trio	…	Better than quartet or equivalent (one unaffected sib allowed)
Interest	Mild phenotype, overlaps with common conditions	…	…	…	Severe/compelling phenotype, unique presentation, treatments imaginable
Family	One affected individual	…	>2 affected individuals who are not sibs	…	>2 affected children of same parents (AR) or transmitted new dominant pattern (AD)

Data Integration

- Criteria for applying external data
- An extended example: combining exome and SNP array data
  - Explore various types of information obtainable from SNP chips
  - Integration
- Other examples:
  - Clinical phenotyping and pedigrees
  - Using biological clues
  - Using accumulated data from multiple exome projects

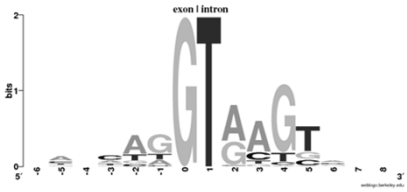
Data Integration: Criteria



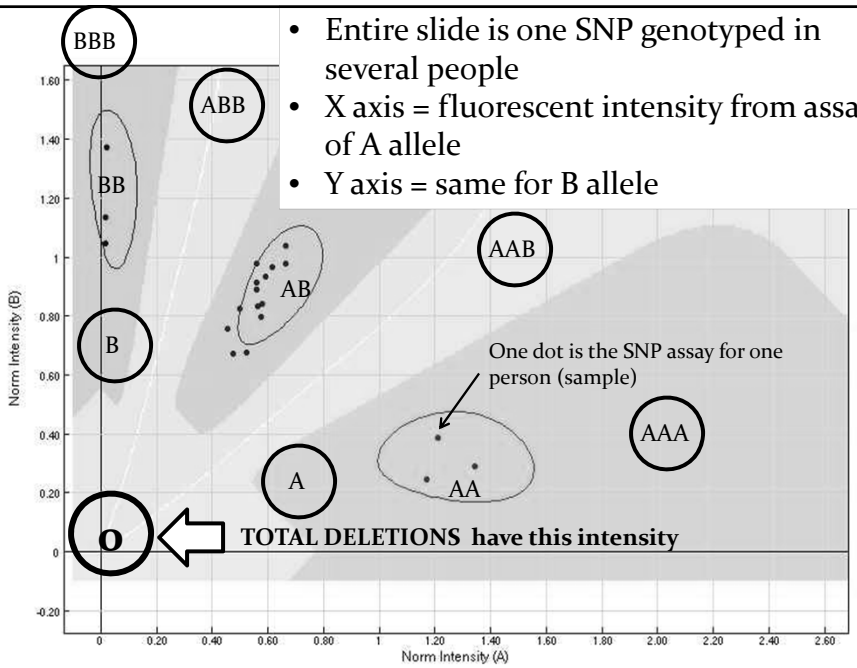
- Applies to both "filtering" and integrating external data
  - How much is the candidate variant list reduced?
    - (Is it worth the trouble?)
  - How error prone is it?
    - (Did it throw out the true variant or include many false variants it was designed to exclude?)
- Examples
    - dbSNP (especially pre 130): frequently used, can remove many variants, can exclude true pathogenic variants, can fail to exclude common variants
    - Segregation filtering: IF high quality data and correct genetic model, has favorable characteristics

Data Integration: What is a SNP?

- Single Nucleotide Polymorphism
  - A single base at a defined genomic position
    - Exact nucleotide varies in population
    - Location is defined by conserved oligo nearby



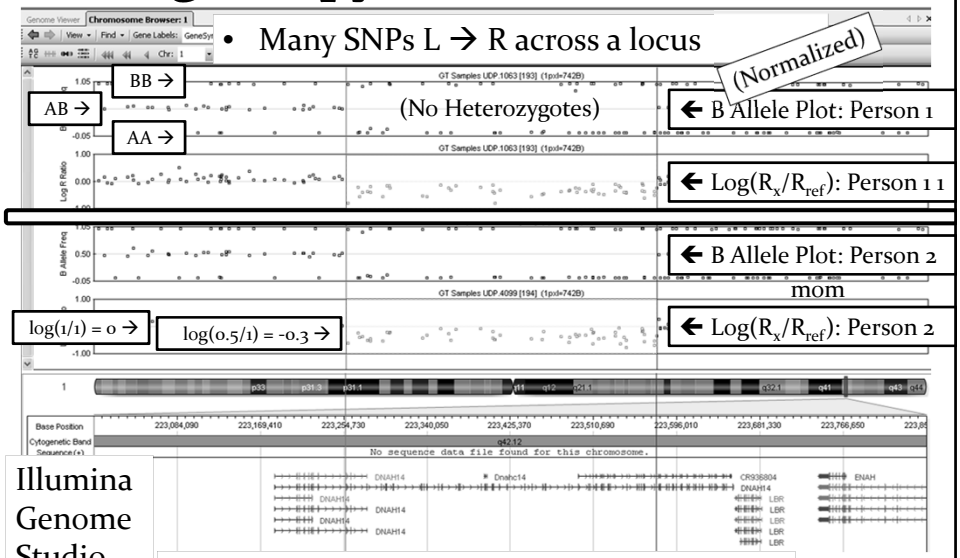
- Most common allele is called "A" by convention
- Less common "minor" allele is called "B" by convention



- Entire slide is one SNP genotyped in several people
- X axis = fluorescent intensity from assay of A allele
- Y axis = same for B allele

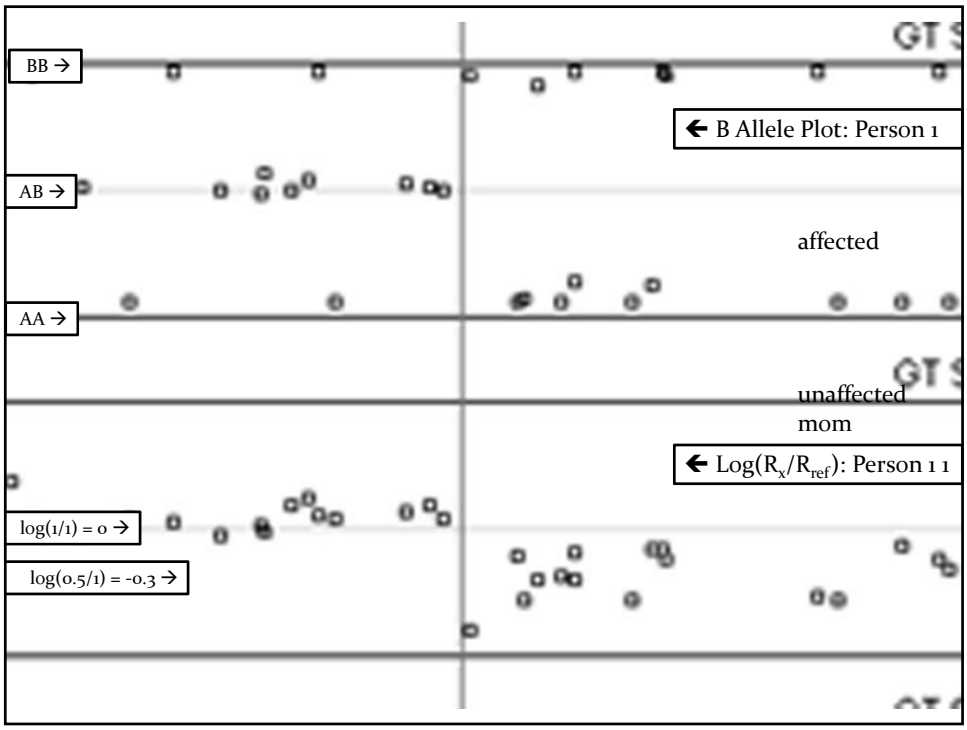
Data Integration: Two People with a Single Copy DNA Deletion

- Many SNPs L → R across a locus

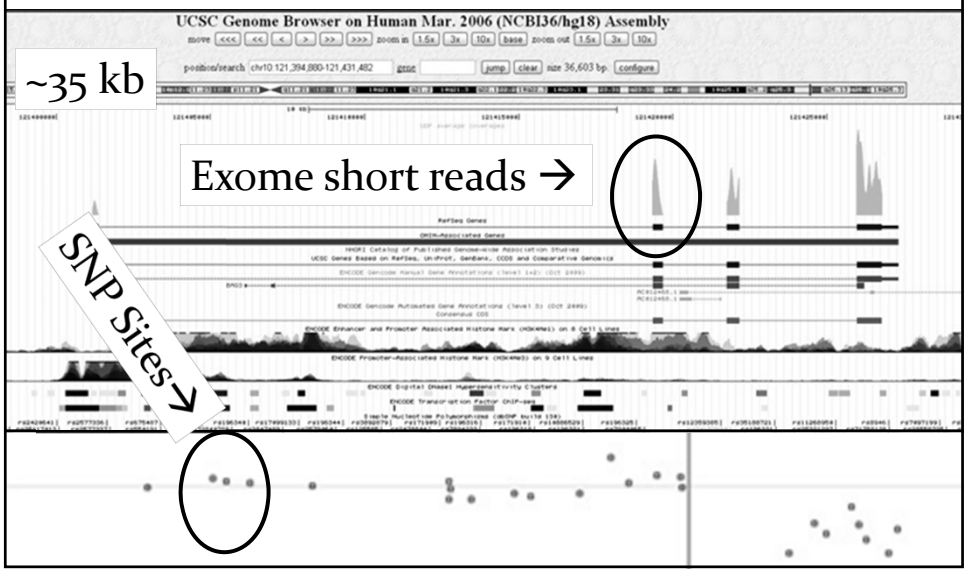


Illumina Genome Studio

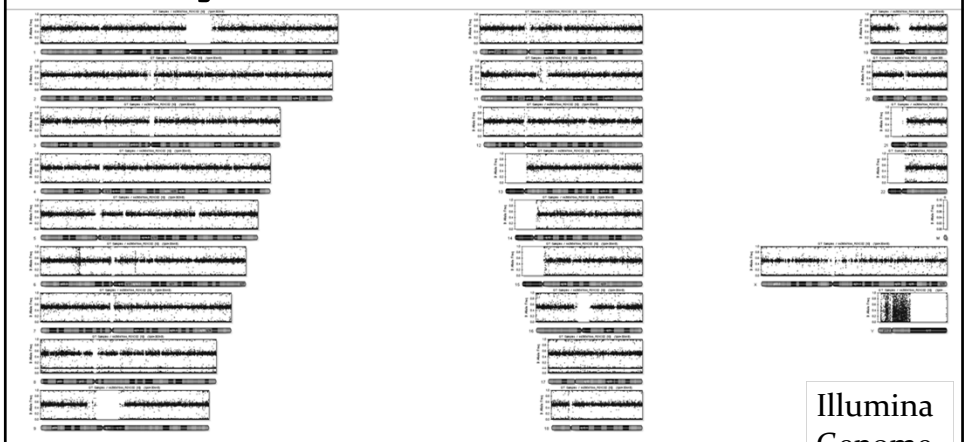
Duplications and double-copy deletions follow similar rules



Data Integration: Why Combine Exomes and SNP Arrays?



Data Integration: SNPs Provide A Survey of Genomic Structure



Illumina Genome Studio

- Normal B allele plot of whole chromosome genome 1-22 plus X and Y
- Illumina SNP 1M Duo: ~ 3 kb average between SNP
- Surveys genomic structure, Cheaper than genomes
- The same data will likely be available from genome sequencing, but is not from exomes

Data Integration: What Types Information Can SNP Chips Add?

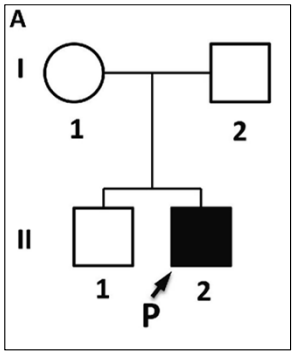
- Dosage changes (reliably above 10 – 50 kb)
  - Single and double copy deletions, duplications
- Chromosomal mosaicism
- Consanguinity
- Uniparental Disomy
- Regions of "anomalous continuous homozygosity"
  - Contiguous homozygous regions that are markedly longer than expected for a given genomic region
- Recombination mapping (with pedigrees)

Data Integration: Detecting Dosage Changes

- Manufacturer software/visual inspection
  - Illumina, Affymetrix
- PennCNV
  - A open source program to automatically detect dosage abnormalities (deletions/duplications) in SNP chip data
  - http://www.openbioinformatics.org/penncnv/
  - Generates a list of genomic spans with potential copy number changes

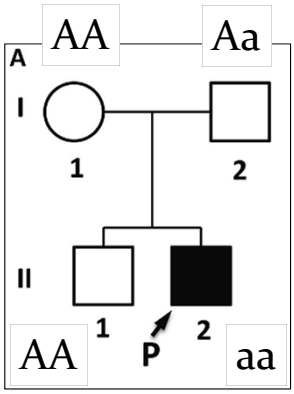
Data Integration: Using Dosage Abnormalities

- 10 y/o male
  - Complex neurological phenotype (balance problems, sensory deficits, weakness, intellectual disability)
  - Guessed autosomal recessive, applied multiple filters as discussed
  - Didn't find anything



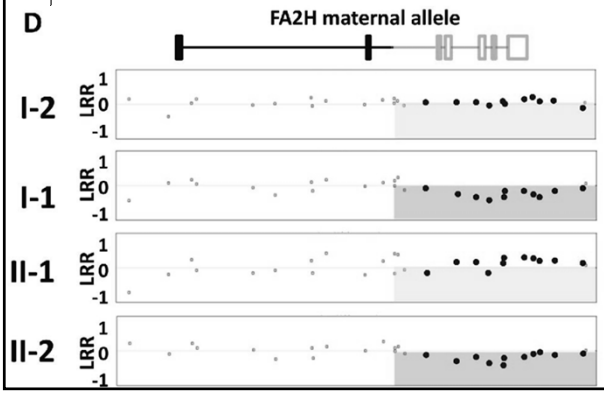
Data Integration: Using Dosage Abnormalities

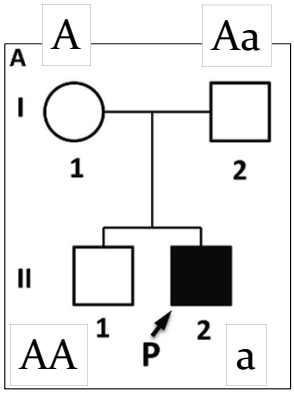
- Reanalyzed data with new, automated filtering tool (VAR-MD) → relaxed filtering constraints → found a candidate
- The candidate had been filtered out initially because the pattern of variants in the pedigree did not follow segregation rules



Data Integration: Using Dosage Abnormalities

- In fact, the mother was not homozygous but hemizygous
- SNP Chip data confirmed a small deletion

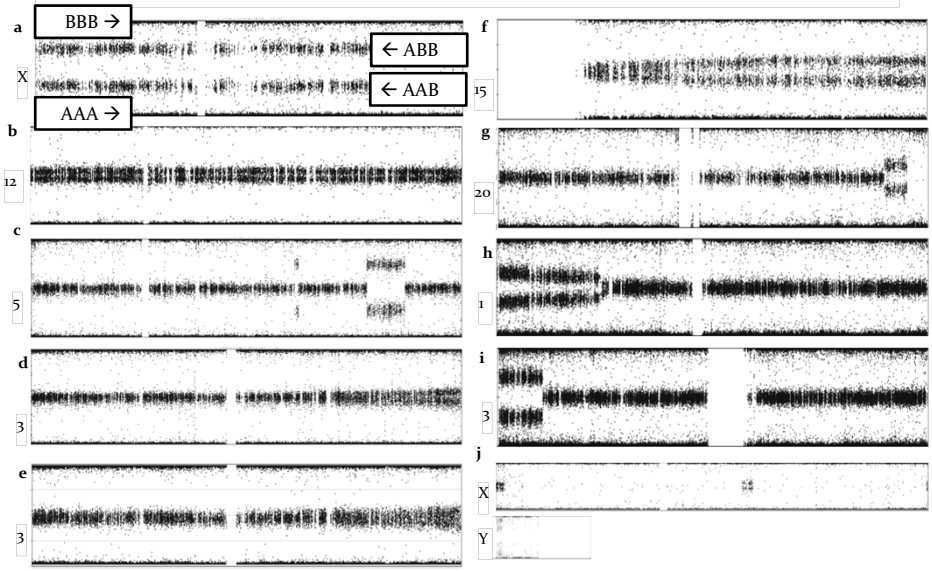




Data Integration: Using Dosage Abnormalities

- Dosage abnormalities should be correlated with sequence variants
  - Single copy deletions may pair with deleterious sequence variants
  - Duplications may result in subtle/important changes in dosage (50% to 33% may matter, especially with multi-meric proteins)
  - Can create a BED file of PennCNV output and filter with VarSifter or other tool

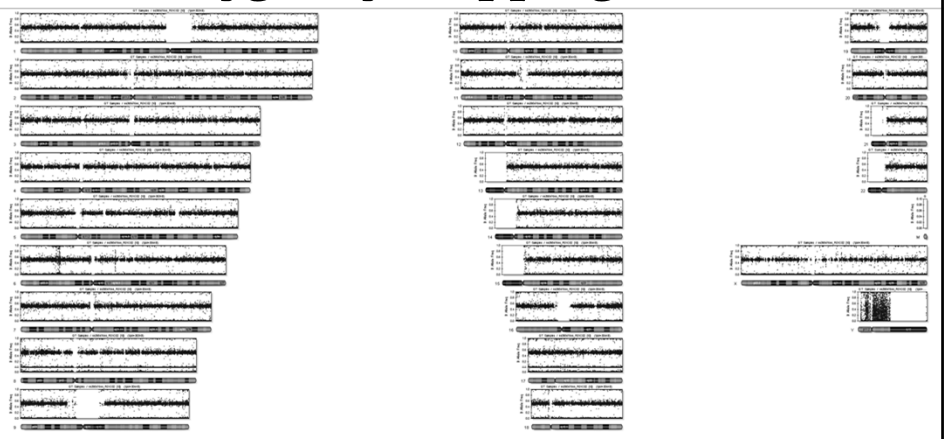
Data Integration: Chromosomal Mosaicism



Data Integration: Chromosomal Mosaicism

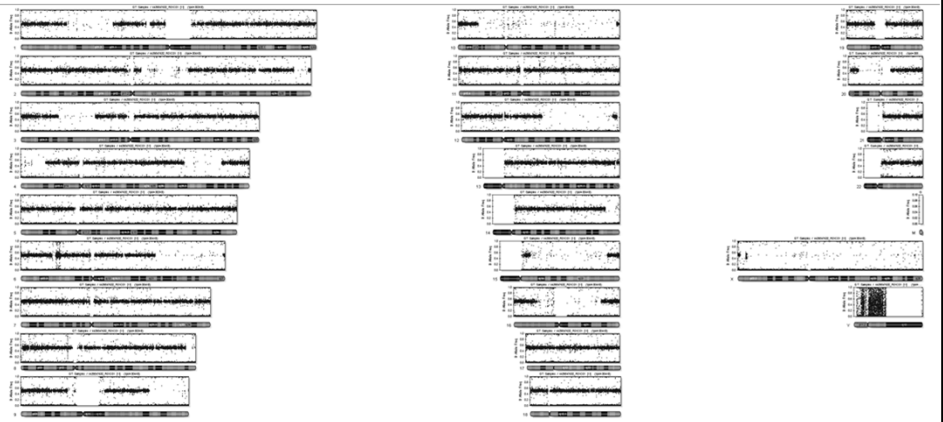
- Consider the effect of mosaicism on sequencing quality
  - Homozygous and heterozygous base calling uses the relative proportions of short sequence reads with different genotypes
  - Mosaicism directly affects the quality of such base calling
- May indicate regions of interest in the genome
- Important in somatically evolving cells, e.g. cancer

Data Integration: Consanguinity/ Homozygosity Mapping—Normal

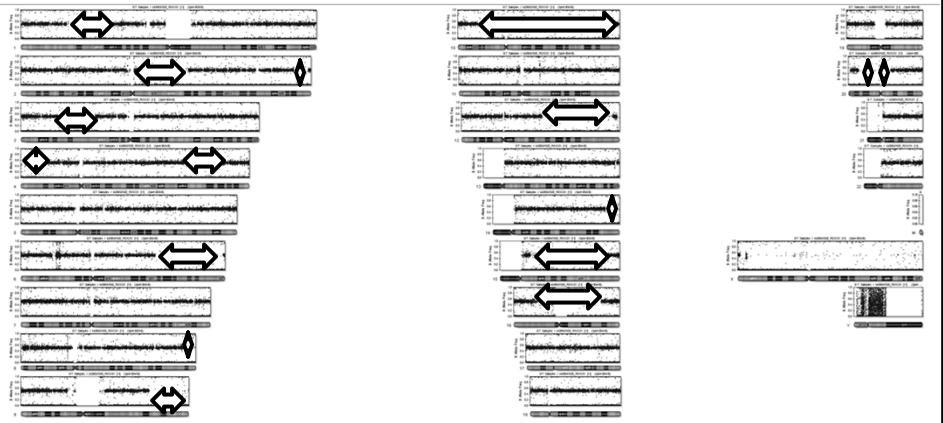


•Normal B allele plot of whole chromosome genome 1-22 plus X and Y

Data Integration: Consanguinity/ Homozygosity Mapping— Increased Homozygosity



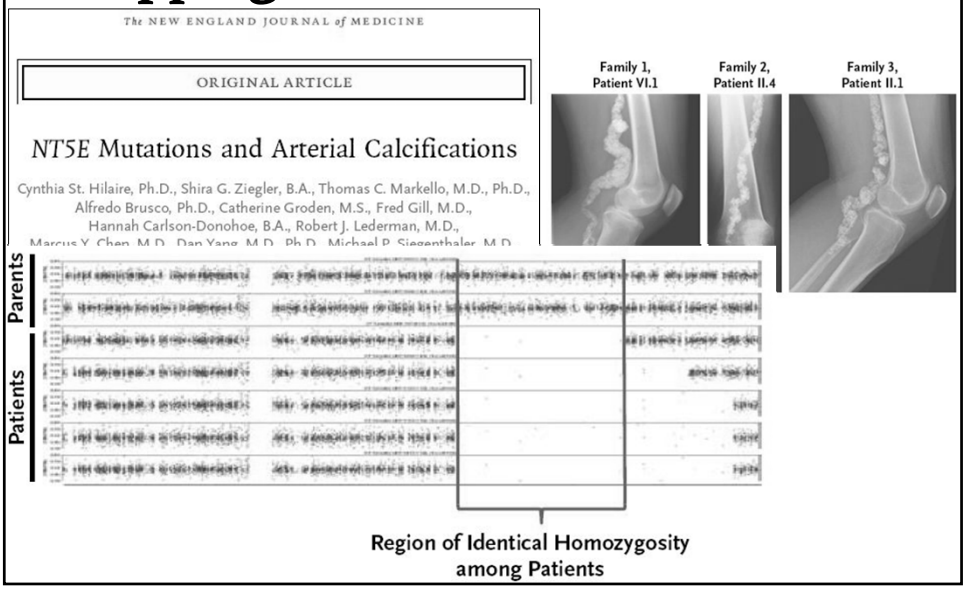
Data Integration: Consanguinity/ Homozygosity Mapping— Increased Homozygosity



Data Integration: Detection of Homozygosity/Consanguinity

- Manufacturer software/visual inspection
  - Illumina, Affymetrix
- PLINK
  - http://pngu.mgh.harvard.edu/~purcell/plink/
  - "PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner."
  - Can auto-detect regions of homozygosity

Data Integration: Homozygosity Mapping



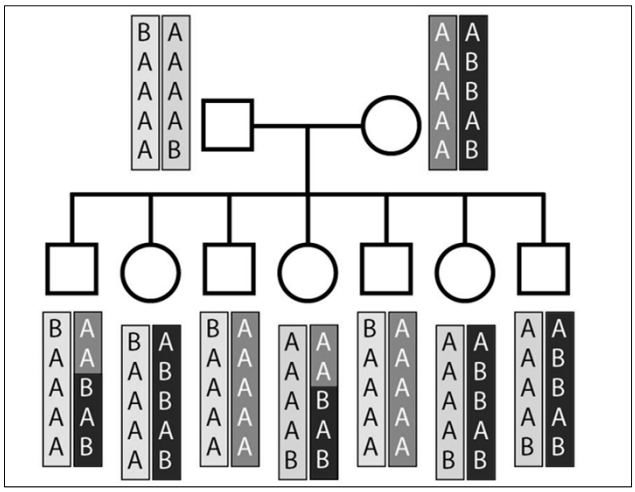
Data Integration: Homozygosity Mapping

- Can identify regions of homozygosity using "B allele" plots
- Can look at the subset of homozygous variants
- May alter planning of NextGen experiments
  - Custom capture instead of exome capture, esp. if standard kits don't cover region well
  - Specific genes can be investigated with Sanger sequencing
- Optimal consanguinity level is probably ~2nd (3%) to 3rd cousins (0.8%).

Data Integration: Intensity Measurements → Boolean Queries

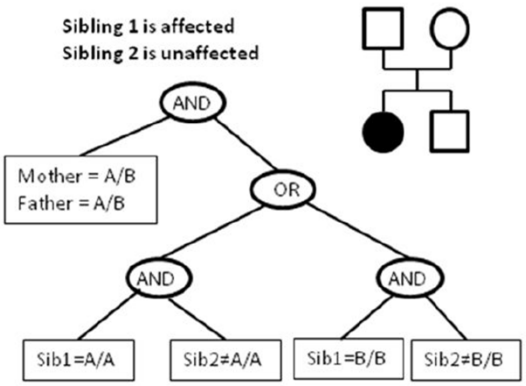
- Examples
  - Uniparental Disomy (not explored today)
  - Mapping recombination events onto chromosomes
- Based on Boolean logic that filters SNPs based on Mendelian segregation
  - Examples (straightforward genetics)
    - If a mother is AB and a father is AA, then a child who is AB had to get the B allele from the mother
    - At the next locus (SNP), the same is true
    - If some children are AB_1/AB_2 and some are AB_1/AA_2, a recombination is suggested

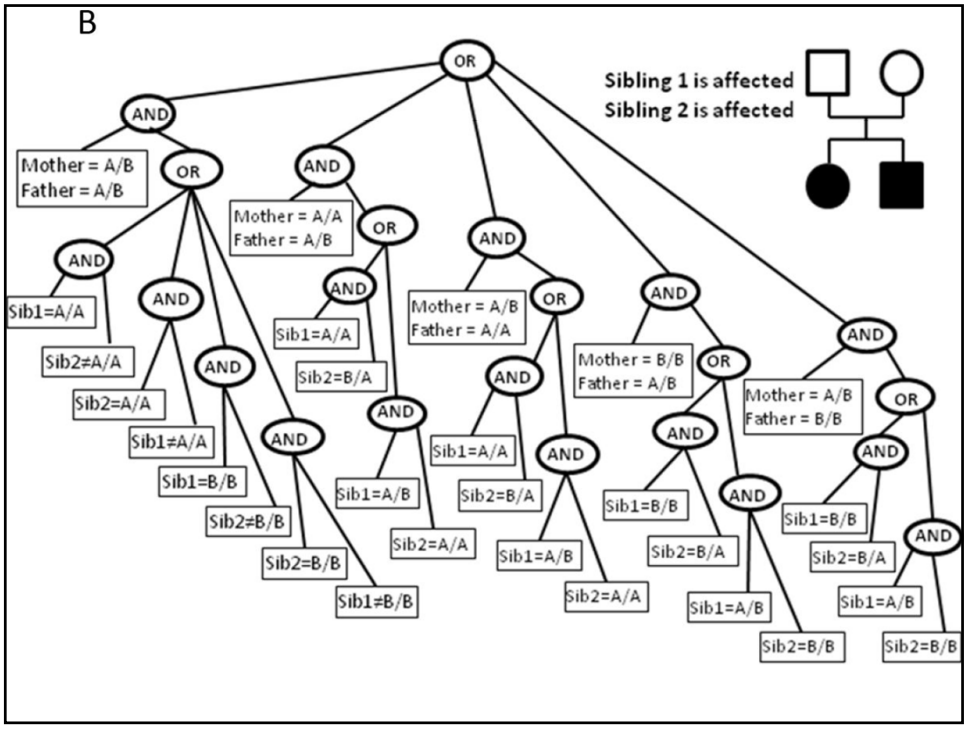
Data Integration: Recombination Mapping



Data Integration: Boolean Logic to Map Simple Pedigree

A

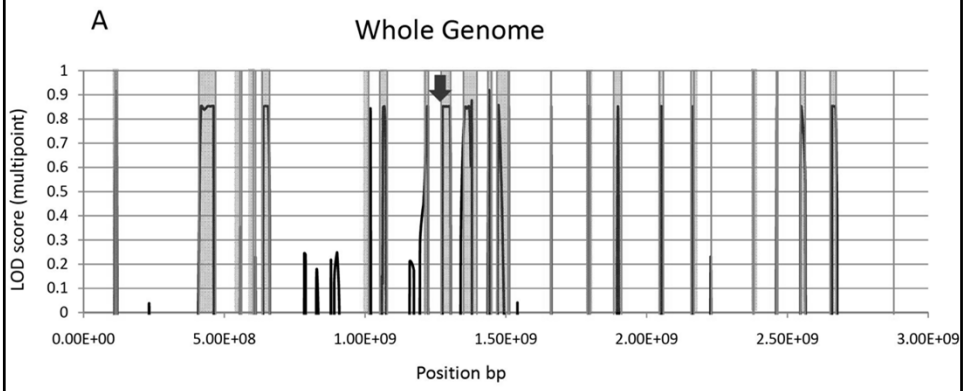




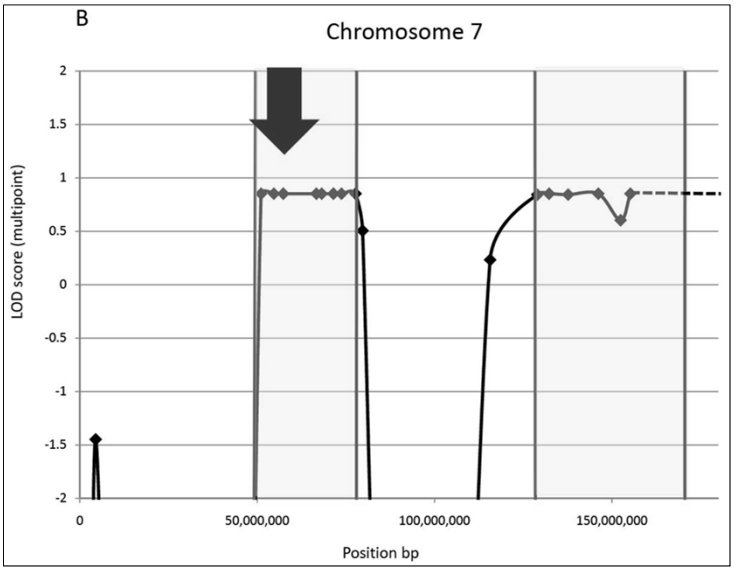
Data Integration: Recombination Mapping Versus Linkage Analysis

- Classic linkage analysis
  - Robust markers (tandem repeats, etc)
  - Fewer/more widely spaced (440 in ABI set)
  - Analysis (LOD score) must take into account the chance of double recombinations between markers
- SNP-based linkage mapping
  - Less robust markers (SNP genotype more likely to be wrong or uninformative)
  - Much higher density of markers (30,000 on average)
  - Many "assays" to test for recombinations
  - Double-recombination errors unlikely

Data Integration: Mapped Discrete Intervals Versus LOD Score

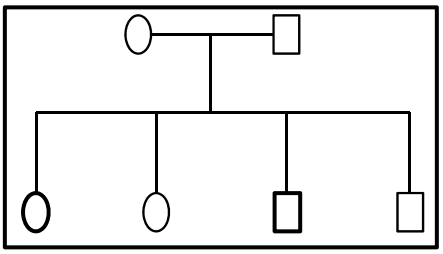


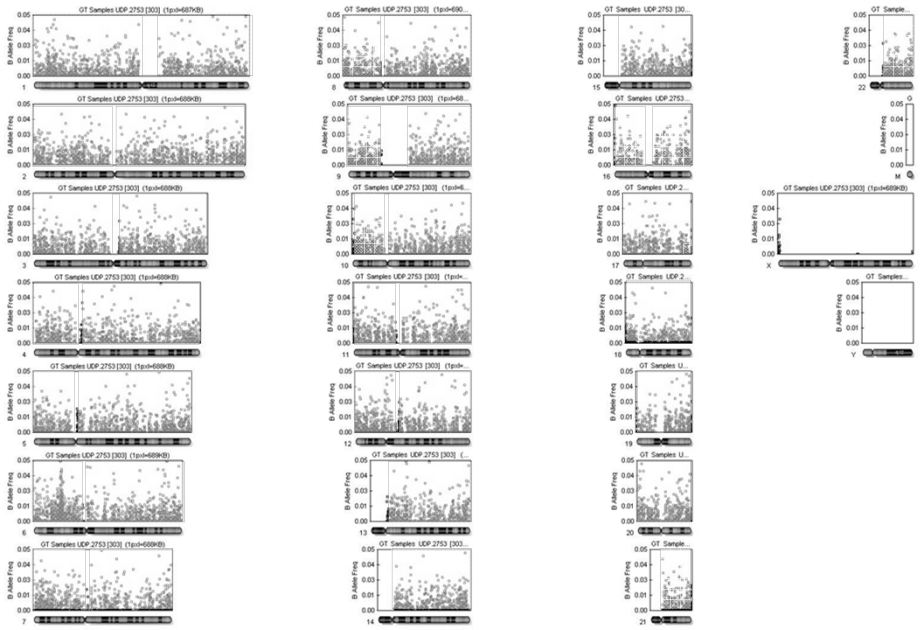
Data Integration: Mapped Discrete Intervals Versus LOD Score

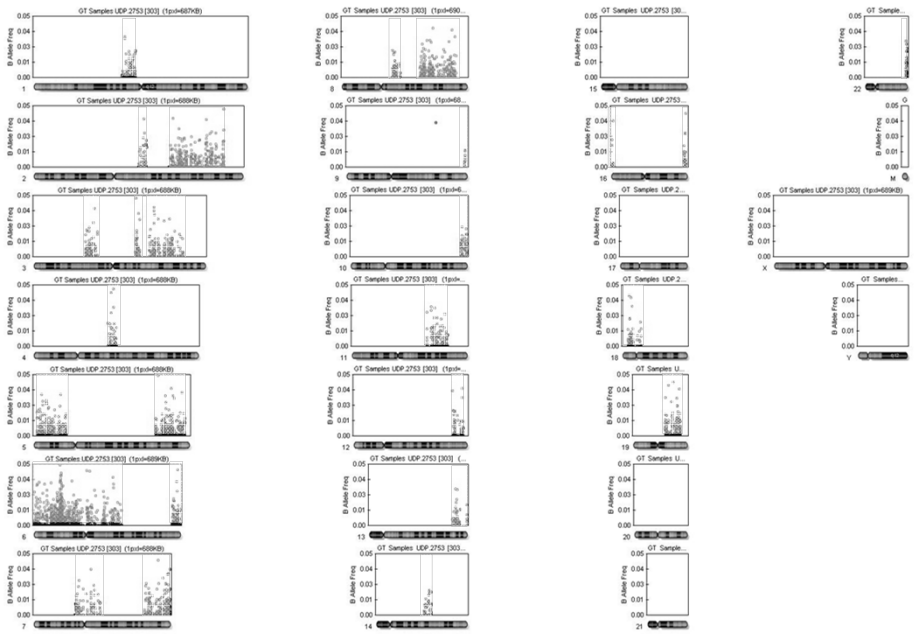


Data Integration: Recombination Mapping Example

- 2 children out of 4 are affected with a neurodegenerative disorder
- 6 family members sent for exome sequencing
- ~112,000 variants
- Recomination mapping applied







Data Integration: Recombination Mapping Example

Total variants	=	112936
<1% frequency (1Kgenome)	=	51025
Gene name kill list (pseudogenes, etc.)	=	51008
Chromosome segregation(SNP linkage)	=	4638
Mendelian segregation (locus by locus)	=	198
Stop/frameshift/splice/Nonsynonymous	=	43
Deleterious prediction (CDpred)	=	13
Genes with 2 variants(passing all above)	=	2

VarSifter - P:\Varsifter\Varsifter_Dec_13\UDP2753\UDP2753_3173unaff.vs.1Kfilter

File View Help

Index	Chr	LeftFlank	RightFlank	refseq	type	ref_allele	var_allele	ref_aa	var_aa	aa_pos	CDPred...	seen_in_more_than_2_samples	
73851	chr3	156380910	156380912	MME	Non-syn...	T	C	M	T	741	-11	0	73849,
92624	chr7	65195280	65195282	ASL	Non-syn...	G	A	V	I	448	-4	0	92619,9
35403	chr16	1334980	1334982	BAIAP3	Non-syn...	A	C	E	A	642	-3	0	35404,
83380	chr5	150905154	150905156	FAT2	Non-syn...	G	A	T	I	1909	-3	0	83375,8
92648	chr7	65741468	65741470	KCTD7	Non-syn...	G	T	D	Y	229	-3	0	92649,
92649	chr7	65741595	65741597	KCTD7	Non-syn...	T	A	L	H	271	-3	0	92648,
20774	chr11	101824757	101824759	TMEM123	Non-syn...	G	A	S	L	51	-2	0	20772,
51726	chr19	43681720	43681722	RYR1	Non-syn...	A	G	N	S	2342	-2	0	51730,
61449	chr2	215993643	215993645	FN1	Non-syn...	C	T	V	I	558	-2	0	61431,6
95467	chr7	157056922	157056924	PTPRN2	Non-syn...	C	G	G	A	898	-2	0	95465,9
15834	chr10	134849596	134849598	KNDC1	Non-syn...	G	A	R	Q	252	1	0	15840,
98186	chr8	106883117	106883119	ZFPM2	Non-syn...	G	A	M	I	544	1	0	98188,9
98219	chr8	110516687	110516689	PKHD1L1	Non-syn...	G	C	G	A	1145	2	0	98215,9
52240	chr19	48276950	48276952	PSG2	Non-syn...	C	G	G	R	118	-9	1	52231,5
52230	chr19	48267744	48267746	PSG2	Non-syn...	C	T	R	H	304	-4	1	52231,5
51467	chr19	40541861	40541863	FFAR3	Non-syn...	A	G	N	S	77	-2	1	51465,5
51472	chr19	40554330	40554332	GPR42	Non-syn...	A	G	N	S	77	-2	1	51471,5
51540	chr19	41028276	41028278	NPHS1	Non-syn...	T	C	E	G	588	-2	1	51542,
51962	chr19	45927006	45927008	ITPKC	Non-syn...	G	A	R	H	439	-2	1	51961,
51465	chr19	40541652	40541654	FFAR3	Non-syn...	G	C	Q	H	7	0	1	51467,5
51471	chr19	40554121	40554123	GPR42	Non-syn...	G	C	Q	H	7	0	1	51472,
52753	chr19	53492141	53492143	CCDC114	Non-syn...	C	T	S	N	432	1	1	52754,

Sample	Genotype	MPG score	coverage
UDP2753.NA	AT	192	120
UDP2755.NA	AT	93	117
UDP3165.NA	AT	192	123
UDP3168.NA	TT	66	100
UDP3171.NA	TT	64	89
UDP3173.NA	TT	135	200

The second change from the Mother is passed down only to both affected children

Number of Variant Positions: 22

Data Integration: Recombination Mapping

- Requires
  - A defensible genetic model
  - Multiple family members, but fewer than for a linkage study
- Can be used to
  - Define segments of the genome that segregate according to a given genetic model
  - Exclude segregation-inconsistent regions and their associated variants

Data Integration: Phenotype and Pedigree Incorporation

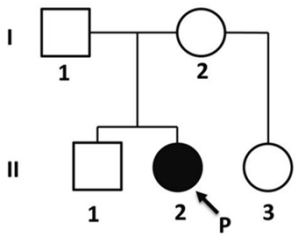
- Phenotyping:
  - May implicate pathways
  - May provide clues for candidate validation
    - Model organism rescue experiments, etc
  - Clues as to an appropriate genetic model
- Pedigrees/Family History
  - A powerful resource for variant filtering
  - Phenotyping critical, just as with linkage projects
    - Affected/unaffected status
    - Penetrance estimation

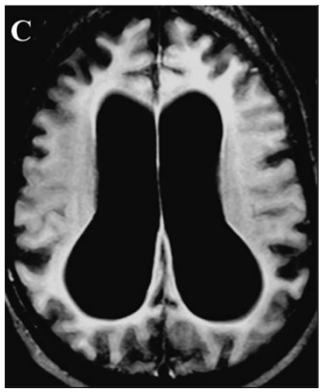
Data Integration: Phenotyping and Gene Lists

- Phenotyping may allow for the construction of gene lists:
  - Functional
    - Mitochondrial genes
    - Metabolic genes interacting with a given metabolite
  - Pathways
    - Developmental
  - Clinical syndromes
    - Multiple diagnostic hypotheses
    - Genetic hetrogeneity
      - Hereditary spastic paraparesis
      - Spinocerebellar ataxia
- VarSifter can incorporate gene include lists

Data Integration: Phenotyping

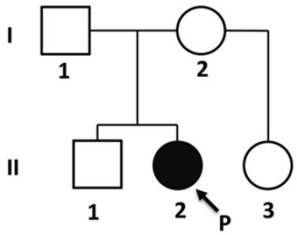
- 19 y/o female with slowly progressive neurological disease
- Course suggestive of several known neurological disorders including G_{M1} gangliosidosis
- However, that diagnosis had been excluded by the "gold standard" of enzymatic testing

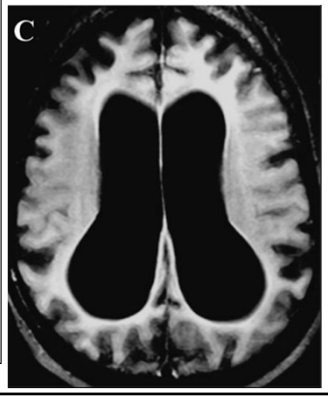




Data Integration: Phenotyping

- Exome sequencing detected candidate variants in the beta-galactosidase gene, the gene associated with G_{M1} gangliosidosis
- Molecular results plus strong clinical suspicion prompted retesting of enzyme activity
- Retesting showed enzymatic deficiency consistent with G_{M1} gangliosidosis



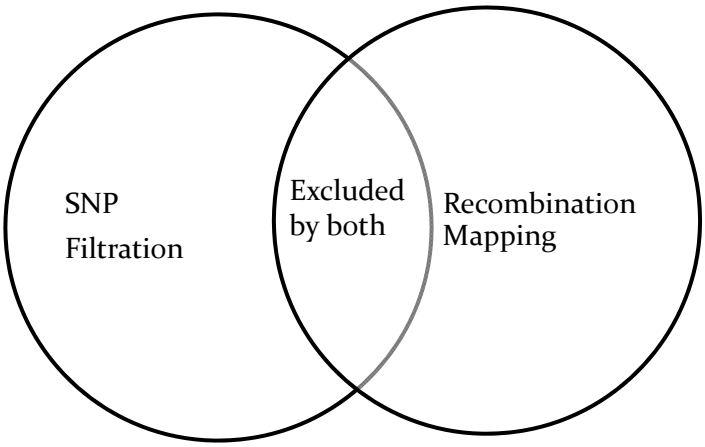


Data Integration: Single Exome vs Small Pedigree

- Single exome
  - Less expensive
  - Analysis more straightforward (fewer tools required)
  - Generates more candidate variants
- Small pedigree
  - More expensive
  - Analysis requires additional tools
  - Fewer candidate variants
  - Filtration using this data can have low error rates with correct model and high quality data

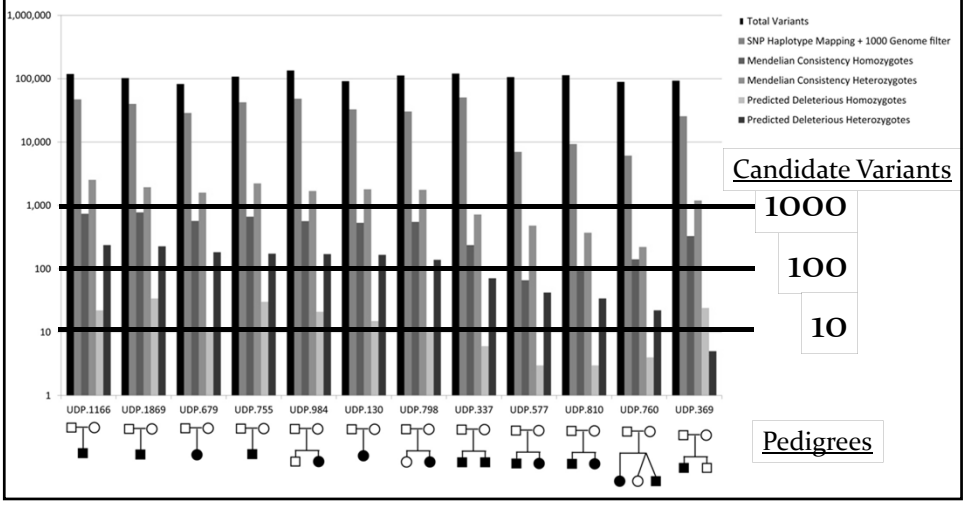
Data Integration: Single Exome vs Small Pedigree

$$\cup \neq \cap$$



Data Integration: Single Exome vs Small Pedigree

Incorporating Segregation: Pedigree Composition



Data Integration: Single Exome vs Small Pedigree

- Single Exome
  - Use when other clues available
    - Likely pathway or cellular process implicated
    - Homozygosity mapping/region of anamalous homozygosity
    - Genetic heterogeneity/Gene list
- More family members
  - Few or no clues → "Agnostic" approach
  - Good phenotyping is available → much less helpful without this information
  - For mapping, should have both parents and at least one sibling of the proband (trios much less useful, esp for recessive models)

Data Integration Summary

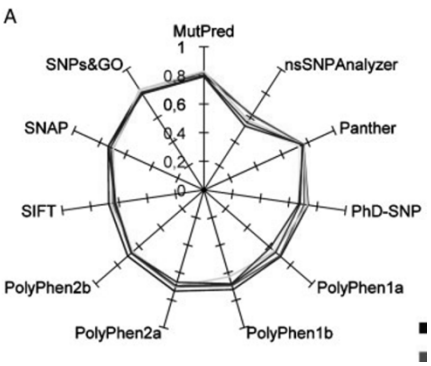
- Use all available resources when planning an next generation sequencing project
- For exome sequencing, consider using SNP arrays to evaluate genomic structure
- Study design should include information gleaned from careful phenotyping and family history
- New approaches are being published on a regular basis

Validation and Reanalysis: Evaluation of Candidate Variants

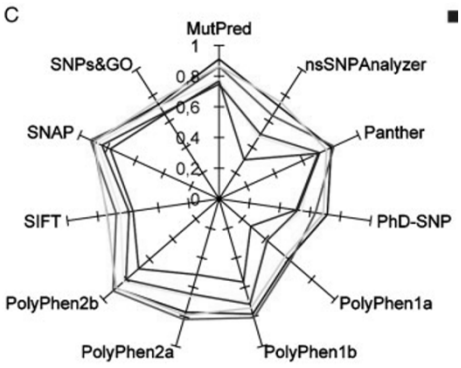
- Sequence validation
  - Research Sanger sequencing
    (CLIA sequencing for clinical reporting)
  - Likelihood of verification is based on filtering techniques
    - AR model, passed all filters: can be 90+%
    - AD model, passed all filters: can be 30% or less, (especially with new dominants)
- Functional validation
  - Determining the biological effect of the variant
  - **No *in silico* methods can replace functional analysis for previously uncharacterized variants**

Validation and Reanalysis: *In Silico* Pathogenicity Prediction

Accuracy



Sensitivity



Thusberg, et al, 2011

Validation and Reanalysis: Evaluation of Candidate Variants

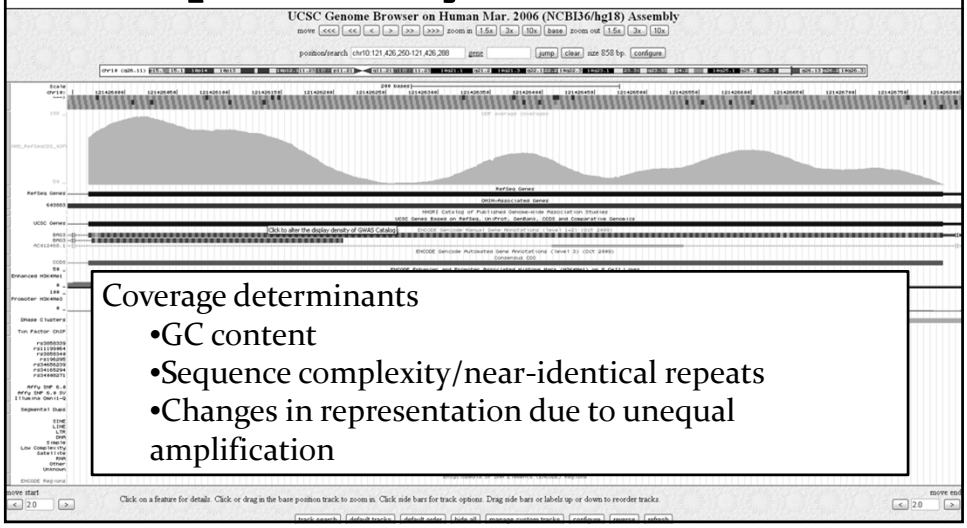
- Editors will ask for evidence of functional consequences:
  - Protein and/or RNA measurements
  - Enzyme activity
  - Rescue experiments
  - Model organisms
  - Etc.
- Exceptions
  - Previously well characterized variants
  - Severe variants in well characterized genes

Validation and Reanalysis: Coming Up Empty Handed

http://www.officialpsds.com

- Revisit Assumptions
  - Heritability
  - Genetic models
  - Variables/parameters used in filters
  - Phenotype assignments
- Know what the technique measures and doesn't
  - Targeting, capturing, sequencing, base calling
- Explore sources of false negative results
  - Study data quality and actual coverage

Functional Validation: Sequencing Success Varies in Expected and Unexpected Ways



Coverage determinants
- GC content
- Sequence complexity/near-identical repeats
- Changes in representation due to unequal amplification

Functional Validation: Methods to Evaluate Coverage

- Genotyping quality and completeness in exome sequencing is complex and can fail differently than Sanger sequencing
  - Targeting → BED file showing "baits"
  - Capture/Complexity → involved topic, but historical data can be used
  - Sequencing/Alignment → coverage and other metrics, historical data
  - Base Calling → MPG and other metrics, historical data
- An accumulated set of data using the same techniques is an invaluable resource

Revisiting Unrevealing Data: Prior Data Usage

- Using previously collected data
  - Used exome sequencing data from UDP and ClinSeq comprising several hundred exomes
  - Looked for genotypes out of Hardy-Weinberg Equilibrium
    - Fischer's Exact Test
    - Bonferroni Correction for 10^6 positions
- Two Error types
  - All homozygous non-reference: ref has minor allele
  - All heterozygous genotypes: likely two similar regions aligned together to form "compression"
- Data used to make site exclusion list

Validation and Reanalysis: Example —Clinical Sequencing

- Given a set of genes associated with a known disorder, how well are they covered?
  - 114 exomes from 27 families
  - Gene lists (Dias et al sumbitted/unpublished)
    - 64 genes associated with various muscle disorders
    - 24 genes associated with hereditary spastic paraparesis
- Assumed standard for clinical sequencing

  "If a clinical sequencing test comes back negative, then all of the sequenced gene regions were sequenced with sufficient quality to detect all variants in those regions."

Validation and Reanalysis: Example —Clinical Sequencing

- Observations
  - Targeted capture kits (SureSelect 38 Mb and 50 Mb) included from 47% to 73% of nucleotides within the gene list (this is probably lower than average)
  - While average coverage was high (~40x to >100x), 2 – 3% of nucleotides had < 4 fold coverage
  - Overall:
    - Most sequenced nucleotides could be genotyped
    - For these particular lists, not all regions were sequenced adequately to rule out all pathogenic variants
  - In other words: know your assay characteristics

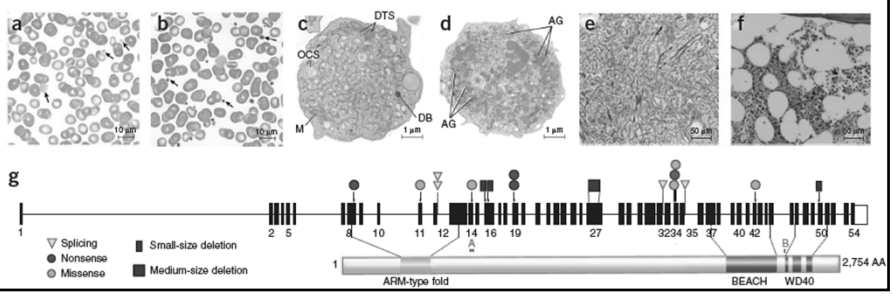
Validation and Reanalysis: Example —The Missing Gene

NBEAL2 is mutated in gray platelet syndrome and is required for biogenesis of platelet α-granules

Meral Gunay-Aygun[1,2,13], Tzipora C Falik-Zaccai[3,4,13], Thierry Vilboux[1], Yifat Zivony-Elboum[3], Fatma Gumruk[5], Mualla Cetin[5], Morad Khayat[3], Cornelius F Boerkoel[1], Nehama Kfir[3], Yan Huang[1], Dawn Maynard[1], Heidi Dorward[1], Katherine Berger[1], Robert Kleta[1], Yair Anikster[6,7], Mutlu Arat[8], Andrew S Freiberg[9], Beate E Kehrel[10], Kerstin Jurk[10], Pedro Cruz[11], Jim C Mullikin[11], James G White[12], Marjan Huizing[1] & William A Gahl[1,2]

- Large linkage region
- Many genes sequenced
- Exome sequenced
- Early kit missed exon
- Sanger sequencing revealed gene

Gray pla
bleeding
lack α-g
(neurob
domain
and plat
combina
of sucro
indicate
(endopl



Validation and Reanalysis: Summary

- Functional validation is required to prove that a candidate variant is THE pathogenic variant
- If there are no good candidates at the end of the analysis
  - Revisit assumptions and analysis parameters
  - Study quality/coverage issues of project
  - Use historical data if available
- Data quality is constantly improving, but
  - Failure modes need to be studied for each set of techniques/conditions

Conclusions

- Give time to experimental design
- Consider using adjunct technologies to compliment exome analysis
- Phenotyping is critical
- Consider using additional family members in certain cases
- Functional proof of pathogenicity is *de rigueur*
- Analyze data in an integrative manner, altering assumptions and filtering constraints as needed

Acknowledgements

William Gahl

Tom Markello
Cyndi Tifft
Neal Boerkoel
Murat Sincan
Karin Fuentes Fajardo
Taylor Davis
Charles Markello
Roxanne Fischer
Richard Hess
Thierry Vilboux

Camilo Toro
Tyler Pierson
Gretchen Golas
Lynne Wolfe
Michele Nehrebecky
Rena Godfrey
Catherine Grodin
Val Robinson
Joy Bryant
Hanna Carlson-Donohue

Jim Mullikin
Jamie Teer
Nancy Hanson
Pedro Cruz
Les Biesecker
Praveen Cherukuri
Meral Gunay-Aygun
Dennis Landis