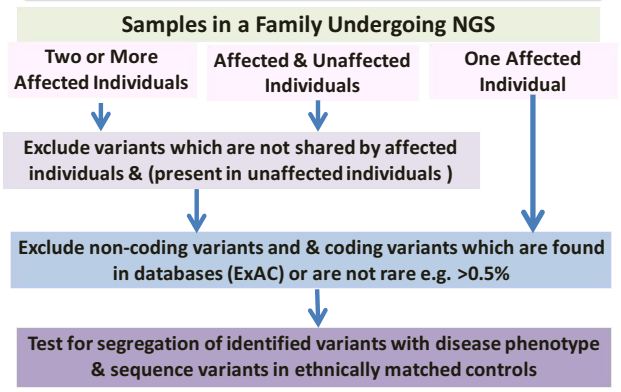
Table of Contents

Introduction to Finding Pathogenic Variants Using NGS Data: Linkage Analysis and Filter Approaches

Suzanne M. Leal
sleal@bcm.edu
Center for Statistical Genetic
Baylor College of Medicine
https://www.bcm.edu/research/labs/center-for-statistical-genetics

Analysis of Family Data Via Filtering Strategies



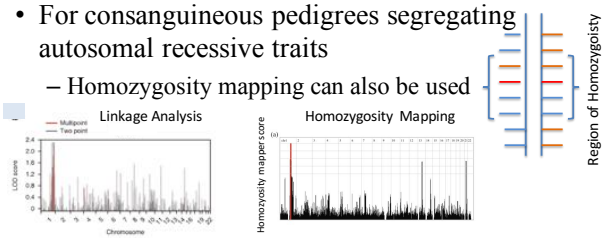
This Strategy can Fail!

None of the variants completely segregate with disease status

- Affected individuals are phenocopies or incorrectly diagnosed
- Unaffected individuals are disease variant carriers (reduced penetrance)
- Sample swaps have occurred
- Locus heterogeneity within the pedigree

Performing Linkage Analysis

- DNA samples from all informative pedigree members are genotyped using arrays
- Parametric two-point and multipoint linkage analysis performed
- For consanguineous pedigrees segregating autosomal recessive traits
  - Homozygosity mapping can also be used



Trouble Shooting Using Linkage Analysis

- Linkage analysis can be performed using genotyping arrays or sequence data
- Observed LOD scores compared to
  - Expected maximum LOD (EMLOD)
  - Maximum LOD (MLOD)
- Deflated LOD scores can be due to
  - Incorrect phenotype information
  - Locus heterogeneity within the pedigree
- Genotypes can also aid in detection of incorrect familial relationships

Benefits of Performing Linkage Analysis Using Genotyping Arrays

- Aids in selection of individuals for sequencing
- Maps the disease locus to specific genomic region(s)
- Filtering can be performed within several Mb, i.e. linkage region, instead of the entire genome
  - Reducing the number of variants which need to be followed-up
    - Testing for segregation in pedigrees
    - Evaluating frequencies in ethnically matched controls

Non-syndromic Hearing Impairment (NSHI)

- 893 NSHI families ascertained
  - Pakistan, USA, Switzerland, Turkey, Jordan, Hungry (Roma), Poland & Germany
- Intra-familial heterogeneity in the collection
  - 15.3% (95% CI 11.9 - 19.9%) Santos-Cortez et al. 2015 EJHG
- Linkage analysis followed by exome sequencing led to the identification of a number of NSHI genes
  - *KARS* (Santos-Cortez et al. 2013 AJHG)
  - *ADCY1* (Santos-Cortez et al. 2014 Hum Mol Genet)
  - *TBC1D24* (Rehman et al. 2014 AJHG)

REPORT American Journal of Human Genetics

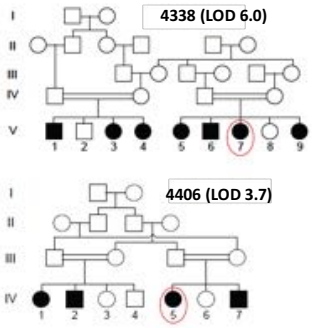
Mutations in *KARS*, Encoding Lysyl-tRNA Synthetase, Cause Autosomal-Recessive Nonsyndromic Hearing Impairment DFNB89

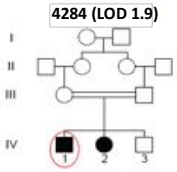
Regie Lyn P. Santos-Cortez,[1,8] Kwanghyuk Lee,[1,8] Zahid Azeem,[2,3] Patrick J. Antonellis,[4,5] Lana M. Pollock,[4,6] Saadullah Khan,[2] Irfanullah,[2] Paula B. Andrade-Elizondo,[1] Ilene Chiu,[1] Mark D. Adams,[6] Sulman Basit,[2] Joshua D. Smith,[7] University of Washington Center for Mendelian Genomics, Deborah A. Nickerson,[7] Brian M. McDermott, Jr.,[4,5,6] Wasim Ahmad,[2] and Suzanne M. Leal[1,*]

DFNB89 Locus (16q21-q23.2)

Bilateral symmetric moderate-to-profound hearing impairment across all frequencies

SNP genotyping (Illumina linkage panel) Region of homozygosity 16.1 Mb: Containing ~180 genes





Basit et al., Hum. Genet. 2010

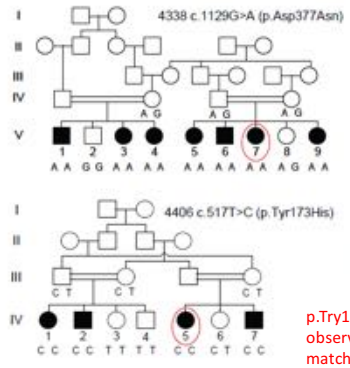
One individual from each family selected for exome sequencing based upon region of homozygosity

Rare Homozygous Variants in the DFNB89 Region

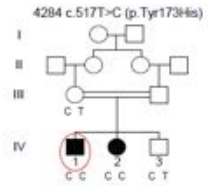
Family	Gene	Variant	Frequency ExAC	Damaging*
4406	COG4	p.Ile271Val	0.0005	MT, LRT
4406	ZFHX3	p.Pro1929Ser	0.0005	None
4406,4284	KARS	p.Tyr173His	0.00002	All
4338	KARS	p.Asp377Asn	0	All
4338	CNTNAP 4	p.Ala1235Thr	0.00002	MT, LRT

*Bioinformatics Tools: CADD, LRT, MutationAssessor, MutationTaster (MT), PolyPhen-2, SIFT
All variant sites were deemed to be conserved (PhyloP & GERP)

KARS Variants Segregate with HI in DFNB89 Families



KARS encodes lysyl-tRNA synthetase



p.Try173His & pAsp377Asn not observed in 750 ethnically matched Pakistani chromosomes

Analysis of Family Based Data (Mendelian)

- Genotype Informative family Pedigree Members
- Perform Linkage Analysis
- Select Pedigree Member(s) for Sequencing
- Remove Variants Which are not Rare in ExAC, e.g. MAF> 0.5%
- Investigate Functionality using Bioinformatic Tools
- Determine if Variant Segregates with Phenotype
- Population Specific Frequencies for Variant
- Acquire Additional Families with Variants with the Same Gene

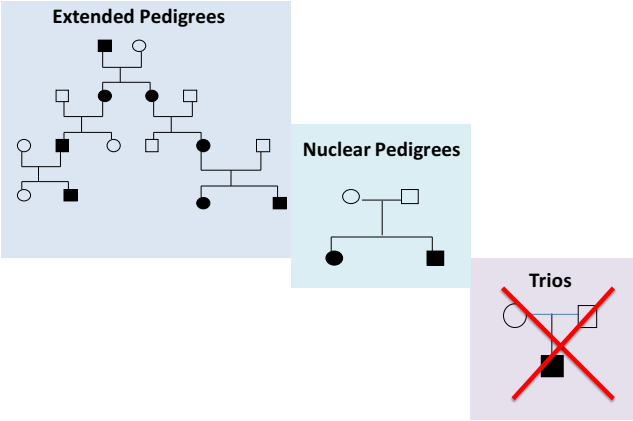
Introduction to the Basics

Suzanne M. Leal
sleal@bcm.edu
Center for Statistical Genetic, Baylor College of Medicine
https://www.bcm.edu/research/labs/center-for-statistical-genetics
www.statgen.us

Goal of Linkage Studies

- To localize disease/trait/susceptibility loci to a unique position on the genome
  - Only family data can be used to carry out linkage studies
    - Extend families
      - Pedigrees with multiple branches and/or multigenerational
    - Nuclear Families
      - Parents and offspring
  - Trios (parents and proband) cannot be used for linkage studies
    - Suitable for association studies

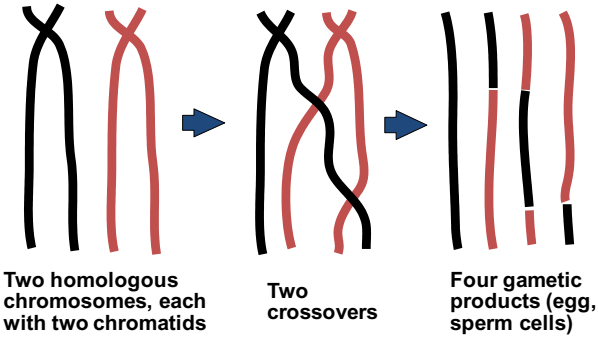
Types of families for Linkage Analysis



Linkage Analysis & Homozygosity Mapping

- Can be used to reduce the region to be followed up with sequencing
  - Thus greatly reducing the number of variants
  - May lead to identification of the causal variant where other approaches have failed
- Genotype all available informative families member to perform linkage analysis/homozygosity mapping

Chromosome in meiosis with two crossovers



Parametric Linkage Analysis

- For Mendelian traits
  - Mode of inheritance must be known
    - Autosomal Recessive
    - Autosomal Dominant
    - X-linked
  - Trait can have reduced penetrance or phenocopies

Linkage Analysis – Allele Sharing Methods

- Also known as nonparametric or model free method
  - Neither nonparametric or model free
  - Mode of Inheritance does not need to be known
    - Complex traits
  - Underlying genetic model is not specified in the analysis

Parametric Linkage - Analysis

- Goal
  - To test whether there is linkage between a disease locus and a marker or set of marker loci
  - Null hypothesis
    - No linkage - recombination fraction (θ=0.5)
      - Recombination rate 50%
        - » Disease locus and marker locus/loci far apart
      - Loci on two different chromosomes

Parametric Linkage - Analysis

- Alternative hypothesis
  - linkage θ<0.5
  - Wish to reject the null hypothesis of no linkage
    - Use a LOD score criterion of 3.3 ($p \leq 0.05$)
  - Estimate the recombination fraction (genetic distance) between the disease and the marker loci

Linkage Analysis - Allele Sharing Methods

- Compare the amount of allele sharing between
  - Affected Sibling
  - Other affected relative pairs
    - Avuncular
      - e.g. uncle-Niece
    - Cousins

Linkage Analysis - Allele Sharing Methods

- Variety of tests to elucidate if there is an excess of allele sharing
  - Mean test
    - Null hypothesis
      - Under no linkage
        - Affected siblings share 50% of their alleles
    - Alternative Hypothesis
      - Under linkage
        - Affected siblings share > 50% of their alleles

Polymorphisms & Variants

- Polymorphism
  - A region of the genome that varies between individual members of a population
  - Usually with a frequency of at least 1 or 5%
- Variant or mutation
  - An alteration in a genome compared to some reference state
    - Does not have to be causal or functional
- Types of Variants
  - Pathogenic
  - Of unknown significance
  - Benign

Loci & Alleles

- Locus: A specific position on the genome
  - For the autosomes 2 alleles are observed at each locus
- Alleles: Are alternative forms of DNA sequence that occur at a locus
  - e.g. the A, B, 0 alleles of the AB0 gene

Loci & Alleles

- Codominant
  - Both alleles are expressed in the heterozygous state
- Dominant
  - Expression is the same in heterozygous as in the homozygous state
    - The homozygous state can sometimes produce a more severe phenotype than the heterozygous state
      - Homozygous lethal
- Recessive
  - Homozygous state is necessary for expression

Hardy Weinberg Equilibrium (HWE)

- For the autosomes the proportion of each genotype follows the laws of HWE
  - p^2, 2pq & q^2
- Which is based upon the observed allele frequencies

HWE

- The organism is diploid
- Reproduction is sexual
- Generations are non-overlapping
- Mating is random
- Population size is very large
- Migration is negligible
- Mutation can be ignored
- Natural selection does not affect the alleles under consideration

HWE

Example 2 allelic system (e.g. SNP marker)

Allele 1 frequency $p = (2N_{11} + N_{12})/2N$
Allele 2 frequency $q = (2N_{22} + N_{12})/2N$

Expected proportions of heterozygotes and homozygotes under HWE

1 1= p^2
1 2 = 2pq
2 2 = q^2

HWE

The following genotype counts are observed

	Observed	Expected
1 1	300	?
1 2	500	?
2 2	200	?

Allele frequencies

1 allele: p=(600+500)/2000=0.55
2 allele: q=(500+400)/2000=0.45

Note q=(1-p)

HWE

Expected genotype frequencies under HWE

11	p^2=0.3025
12	2pq=0.495
22	q^2= 0.2025

Expected number of genotypes under HWE*

11	302.5
12	495
22	202.5

*For a sample size of 1,000 individuals

HWE

$$X^2=\Sigma\frac{(\text{observed-Expected})^2}{\text{Expected}}$$

	Observed	Expected
1 1	300	302.5
1 2	500	495.0
2 2	200	202.5

X^2= $(300-302.5)^2/302.5+(500-495)^2/495+(200-202.5)^2/202.5=0.102$

X^2 = 0.102 p=0.75 1 df

Testing for deviations in HWE

- Chi-square tests
- Exact tests
- Likelihood ratio tests

Reasons for Deviation from HWE

- Population Admixture
- Heterozygous Advantage
- Copy number variants
- Genotyping Error
- Chance

Loci, Genotypes & Haplotypes

- Multiple marker on a chromosome
  - Microsatellites
  - Single nucleotide polymorphisms (SNPs)
  - Single nucleotide variants (SNVs)
- Genotype
  - The two alleles at a locus comprise a genotype
- Haplotype
  - The alleles on each chromosome

Locus, Genotype & Haplotype

Genotypes are known

Genotype for

Locus A: 1 1

Locus B: 2 2

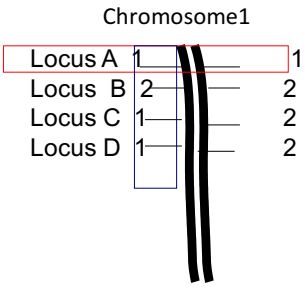
The haplotype for each chromosome of a pair usually needs to be reconstructed

Haplotypes for

Locus A & B: 1 2, 1 2

Locus C & D: 1 1, 2 2

or 1 2, 2 1



Linkage Studies-Genetic Maps

- A map provides the position and order of marker loci
  - Physical position
  - Genetic position
    - Based upon interpolation for SNV (single nucleotide variant) and SNP (single nucleotide polymorphism
- Genetic position necessary to perform multipoint linkage analysis

Genetic Maps

- Map distance given in Centimorgans (cM)
- Recombination (Θ) fractions cannot be added
  - Except in the case of complete interference $x=\Theta$
    - Under complete interference multiple crossovers between two loci can be excluded
      - It can also be assumed there is complete interference when two loci are closely linked ($\theta<0.05$)
        - » Then recombination fractions can be added

Genetic Maps

- Can convert Θ to map distances using
  - Map functions
    - Haldane
    - Kosombi
    - Sturt
  - The distances can then be summed
- No one-to-one correspondence between map distance and number of base pairs
  - Recombination events variable across the genome

Haldane Map Function (Haldane 1919)

- Assumption that crossovers in different intervals occur according to a Poisson probability law
  - Note x is given in Morgans
- $X = \begin{cases} -1/2 \ln(1-2\theta) & \text{if } 0 \leq \theta < ½ \\ \text{infinity} & \text{otherwise,} \end{cases}$
-
- The Inverse is

$$\theta = ½[1-\exp(-2|x|)]$$

Genetic Maps

- Most SNP and SNVs are not on genetic maps
- Physical position and order known
  - Unknown genetic map distance
- Using Genetic Maps such as
- Rutgers Combined Linkage-Physical Map
  - http://compgen.rutgers.edu/mapinterpolator
- Interpolation can be used to estimate the genetic distance of markers to perform linkage analysis

Genome Scan Data (Marker Loci) for Linkage Analysis

- Microsatellite Marker loci
  - Not currently usually used
- Genotyping Arrays
  - SNP and SNV marker Loci
- Exome and whole genome sequence data
  - SNV and SNP marker loci

Microsatellite Markers

- For the most part have been replaced by SNP marker loci
- Microsatellite markers have many alleles
  - Heterozygosity >0.71
- Usually denoted by a D#
- Linkage whole genome scans
  - 10 cM scan
    - ~400 marker loci
  - 5 cM whole genome scan
    - ~800 marker loci

Heterozygosity (H)

- Provides information on what proportion of individuals that will be heterozygous for a particular marker locus
  - Assumption Hardy Weinberg Equilibrium

$$H=1-\Sigma p_i^2$$

SNP and SNV Marker loci

- Most commonly used markers for linkage analysis are SNP loci
  - Base change at a single nucleotide
    - Most have only two alleles (diallelic)
      - But can have up to four alleles
    - Those which have more than two alleles are not used
    - Heterozygosity $\leq$0.5

SNP and SNV Marker loci

- Denoted by an rs#
- SNP have a minor allele frequency (MAF) of $\geq$5%
  - Can also be defined as having a MAF $\geq$1%
- SNVs have a MAF $\leq$ 1%
  - Usually diallelic but can have up to four alleles

Genotyping Arrays SNP/SNV Marker Loci*

- Illumina HumanCore-24 Bead Chip
  - ~300,000 SNP marker loci
    - Up to a additional 300,000 custom markers
- Illumina HumCoreExome-24 Bead Chip
  - ~300,000 SNP marker loci
  - ~240,000 Exome marker loci
    - Up to an additional 100,000 custom markers
- Illumina HumanOmni5-Quad
  - ~4.2 Million SNP/SNV marker loci
    - Up to an additional 500,000 custom markers
- Illumina HumanOmni5Exome
  - ~4.5 Million SNP/SNV marker (including exome content)
  - Up to an additional additional 200,000 custom markers

*These arrays were all developed for association studies- the Illumina linkage array has been discontinued

Genotyping Arrays

- Higher density arrays are overkill for linkage analysis
  - A subset of informative markers can be used
    - e.g. ~0.20cM
    - Once linkage has been established denser maps of markers can be analyzed within the linkage region
- Using entire set of markers extremely slow to analyze
  - May not be able to complete linkage analysis within a reasonable amount of time

Features of Mendelian Traits

- Non-Allelic/Locus/Linkage Heterogeneity
- Allelic heterogeneity
- Phenocopies
- Reduced penetrance
- Age specific reduced penetrance

Heterogeneity

- Allelic Heterogeneity
  - Multiple separate alleles at the same locus are responsible for the disease phenotype
    - Cystic fibrosis
- Non-allelic/Locus/Linkage Heterogeneity
  - Different individual genes are responsible for disease etiology
    - Charcot-Marie-tooth disease
    - Adult polycystic kidney disease (APKD)
    - Non-syndromic hearing loss

Phenocopies

- Traditional definition
  - An environmentally induced phenotype that resembles the phenotype produced by a mutation
- Examples
  - Individuals taking meperidien which is tainted with its by product MPTP
    - Causes the destruction of dopaminergic neurons and produces a Parkinson disease phenotype
  - Epilepsy due to traumatic brain injury

Phenocopies

- The term phenocopy (although used incorrectly) is also used to describe
  - Genetic heterogeneity
    - An individual(s) within a pedigree which is affected due to a different gene than the other pedigree members
      - E.g. BRCA1 families with breast cancer patients with out a BRCA1 variant
  - Misdiagnosed cases within a pedigree
    - Alzheimer's disease pedigrees with cases of dementia which are not Alzheimer's disease

Reduced Penetrance

- Age specific
- Sex specific/Sex limited
- Exposure specific
- Incomplete penetrance
  - A proportion of disease gene carriers never develop the phenotype
- Can reduce the power of detecting linkage
- Unaffected individuals below the age of onset provide no linkage information

Familial & Founder Effect

- Familial
  - Any trait which is more common in relatives of an affected individual than in the general population
  - Can be genetic or environmental or both
    - Prion disease Kuru
- Founder Effect
  - A high frequency of a disease allele in a population founded by a small ancestral group due to one or more founders being carriers of this allele

Assortative Mating

- Selection of mate with preference to a certain phenotype/genotype (that is non-random mating)
  - Positive
    - preference for a mate with the same phenotype
  - Negative
    - Preference for a mate with a different phenotype

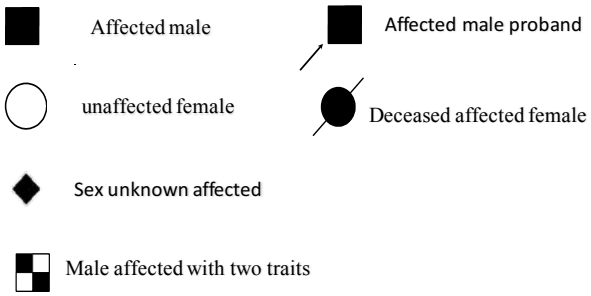
Epistasis & Pleiotropy

- Epistasis
  - Interaction between alleles at two different loci
- Pleiotropy
  - Multiple phenotype effects of a single gene
    - Example Marfans Syndrome

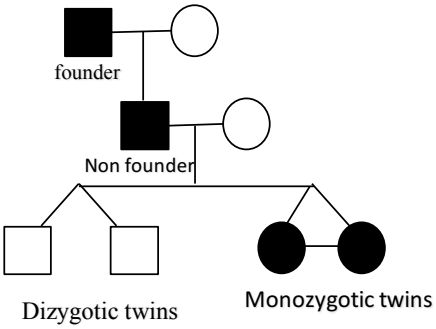
Mendelian Traits

- Modes of Inheritance
  - Autosomal Dominant Inheritance
  - Autosomal Recessive Inheritance
  - X-linked Inheritance
    - Dominant
    - Recessive

Pedigree Drawings -Symbols



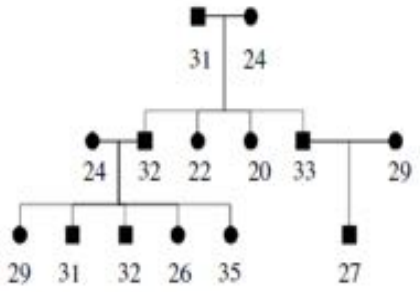
Pedigree Drawings



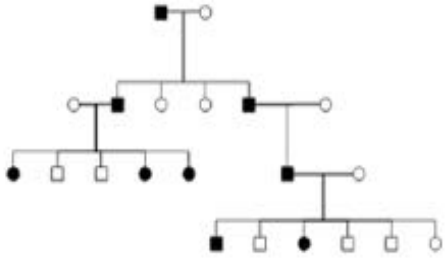
Phenotype Quantitative & Qualitative Traits

- Quantitative trait
  - Continuous
  - Dichotomizing based upon an arbitrary or clinical cut-off
    - Can lead to loss of power (due to misclassification)
- Qualitative – binary disease trait
  - Affected or unaffected

Quantitative Trait - Example BMI



Autosomal Dominant Pedigree



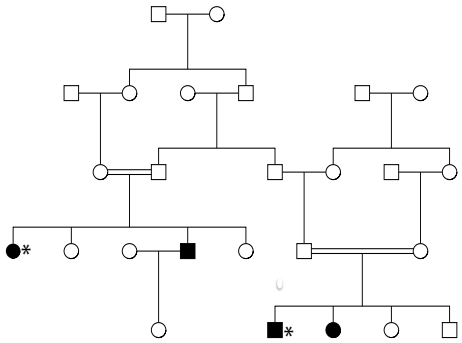
Autosomal Dominant Mode of Inheritance

- If trait is fully penetrant with no phenocopies the following is true:
- Each affected individual carries at least one copy of the disease/trait allele
- Each unaffected individual must be homozygous wild type
- If an affected individual has an unaffected parent they must by heterozygous for the disease/trait allele

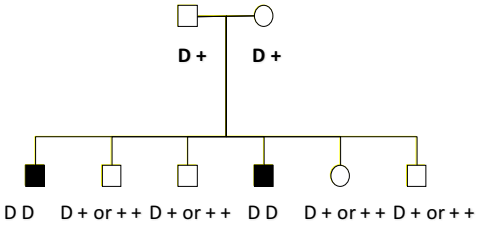
Autosomal Dominant Mode of Inheritance

- On average 50% of the children from an heterozygous affected individual will also be heterozygous for the disease allele and affected
  - 100% of all children of an affected homozygous individual will be affected
- Equal number of males and females affected:
  - There are exceptions
    - e.g. sex limited traits
- Affected (heterozygous) and unaffected individuals provide equal linkage information

Consanguineous Autosomal Recessive Pedigree



Autosomal Recessive Pedigree with Unrelated Parents



Autosomal Recessive Mode of Inheritance

- The following hold true for fully penetrant diseases with with no phenocopies
- Each affected individual must be either homozygous or a compound heterozygous for the pathogenic variant(s)

Autosomal Recessive Mode of Inheritance

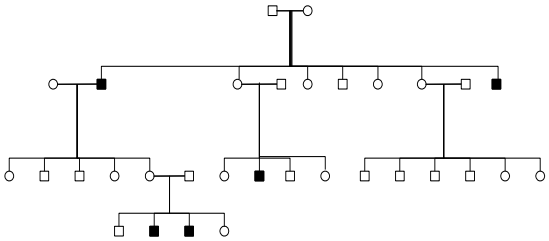
- Unaffected individuals can either by homozygous wild type or carry one copy (heterozygous) of the pathogenic variant
  - 1/3 homozygous wild type
  - 2/3 carriers, heterozygous for the pathogenic variant

- Approximately 25% of all children whose parents are carriers will be affected

Autosomal Recessive Mode of Inheritance

- Offspring of two affected individuals will all be affected
  - If both parents have the same pathogenic variant or pathogenic variants within the same gene
- If pathogenic variant(s) are rare
  - Usually only nuclear families are observed, with both parents unaffected.
    - Exceptions are
      - For consanguineous kindreds where multiple affected sibships can be observed in the pedigree.
      - Quasidominant/ Pseudodominant Inheritance

X-Linked Recessive Pedigree



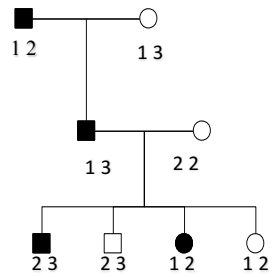
X-linked Recessive Mode of Inheritance

- No male to male transmission
- For fully penetrant traits disease with no phenocopies the following is true
  - 50% female children of female carriers will also be carriers
  - 50% of male children of carriers females will be affected
  - All female children of affected males will be carriers
- In some circumstances carrier females are also affected
  - But have a milder phenotype than affected males

Calculation of LOD scores

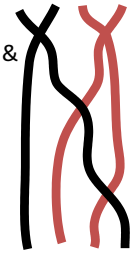
Suzanne M. Leal
sleal@bcm.edu
Center for Statistical Genetic
Baylor College of Medicine
https://www.bcm.edu/research/labs/center-for-statistical-genetics
www.statgen.us

Pedigree Drawing with Marker Loci

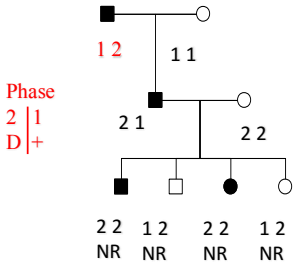


Informative Individuals for Linkage

- An individual to be informative
  - Most be heterozgyous at the marker locus &
  - And a second locus
    - Disease locus
    - Marker locus
- In order to observe whether a recombination event occurred or not between the two loci



Autosomal Dominant Pedigree Phase Known

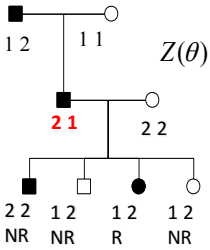


Autosomal Dominant Pedigree Phase Known

$H_0: \Theta = ½$
$H_1: \Theta < ½$

$$\hat{\Theta} = \frac{R}{NR + R}$$

$\theta = 0.25$

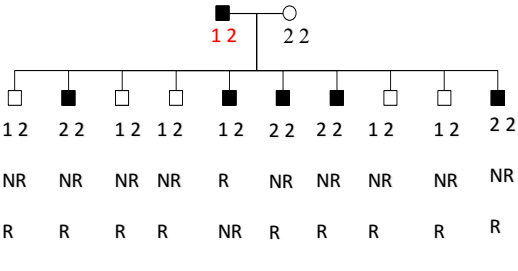


$$Z(\theta) = \log_{10} \frac{\theta^1 (1-\theta)^3}{½^1 (1-½)^3}$$

$Z(\theta) = 0.227$

Autosomal Dominant - Phase Unknown

Phase I 2 1 D +
Phase II 2 1 + D



Autosomal Dominant Phase Unknown

H_0: Θ = ½

H_1: Θ <½

$$Z(\theta) = \log_{10} \frac{\theta^1(1-\theta)^9 + \theta^9(1-\theta)^1}{\frac{1}{2}^{10} \quad + \quad \frac{1}{2}^{10}}$$

Maximum LOD Score occurs at 1.3 at Θ=0.1

LOD Scores

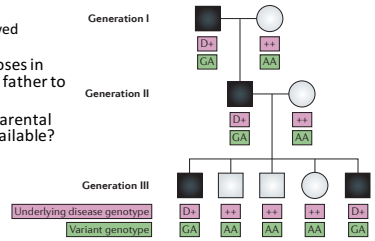
- LOD Scores can be added across families
  - Must be summed at the same theta value or map distance
- If all members of the family are genotyped
  - The genotype frequencies are not used in the LOD score calculation
    - Misspecification of allele frequencies will not bias the LOD score
- When genotype data is not available for all family members
  - Misspecification of allele frequencies can increase type I error

Linkage Analysis

- For traits which are fully penetrant with no phenocopies
  - When θ=0 and there are not recombination events
  - When the marker is fully informative
- Autosomal dominant traits (phase-known pedigrees)
  - Each affected and unaffected individuals adds 0.3 to the LOD score

Linkage Information Obtained from an Autosomal Dominant Pedigree

- Each offspring both affected and unaffected adds 0.3 to the Lod score
  - LOD Score 1.5 for displayed pedigree
- The only informative meioses in this example are from the father to his offspring
- What is the LOD score if parental genotype data was not available?

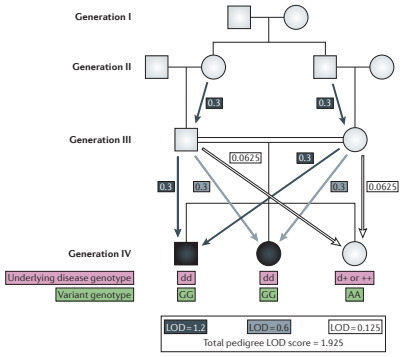


$$Z(\theta) = \log_{10}[((1-\theta)^5 + \theta^5)/((\tfrac{1}{2})^5 + (\tfrac{1}{2})^5)].$$

Linkage Analysis

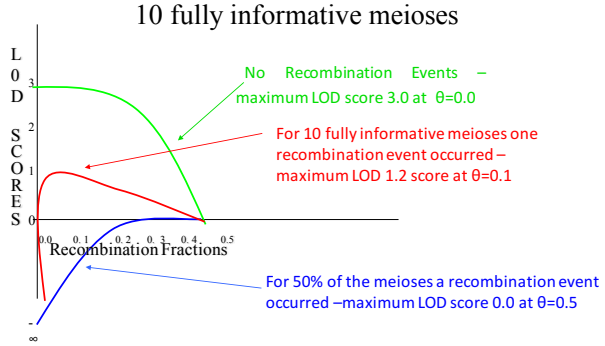
- Autosomal recessive traits
  - First affected individual is not informative for linkage
  - Except if parental mating is consanguineous
    - How much information the first affect individual provides depends on the frequency of the haplotype/marker
    - How distantly related are the parents
      - The more distantly related the parents and the lower the frequency of the haplotype/variant the higher the LOD score
        - » Maximum LOD score first cousin mating one affected LOD=1.2
        - » Maximum LOD score second cousin matting one affected LOD=1.8
  - Each additional affected individual adds
    - Adds 0.6 to the LOD score
  - Each additional unaffected individual
    - Adds 0.125 to the LOD score

Linkage Information obtained from a Consanguineous Autosomal Recessive Pedigree



Lod Score Curve for Autosomal Dominant Pedigree -Phase Known

10 fully informative meioses



Size of Mapped Regions

- For Mendelian disease/traits
  - Where a large sample (many informative meiosis ~200) are available
    - Highly unusual that a disease locus can be mapped to a region which is < 1cM (~ < 1 Mb)
      - However the genetic/physical region is usually much larger
  - For consanguineous kindreds
    - Even when linkage can be established
      - A limited number of informative meiosis
        - » Large genetic region of homozygosity with many genes

The Effect of Using Incorrect Marker Allele Frequencies on LOD Scores

- If there are pedigree members with missing genotype data
- Using incorrect marker allele frequencies
  - Can increase type I error
- Important to obtain accurate <u>population specific</u> estimates of allele frequencies
- If there is missing genotype data it is advisable not to use equal allele frequencies for marker loci

Obtaining Allele Frequencies

- Estimate from pedigree founders
  - Must have a sufficient number of founders
- Obtain from the manufacture of genotype array
  - Usually allele frequencies provided for Europeans, African Americans and Asians
    - For e.g. Illumina HumanOmni5-Quad
      - http://support.illumina.com/array/array_kits/humanomni5-4-beadchip-kit/downloads.html

Obtaining Allele Frequencies

- Alohomora
  - Provides frequencies for Europeans, African Americans and Asians for popular SNP arrays
    - Creates datafile with allele frquencies
    - http://gmc.mdc-berlin.de/alohomora/maps/

Obtaining Allele Frequencies

- UCSC Genome Binoinformatics
  - For customized SNP arrays & population specific allele frequencies
  - Use Table browser
    - http://genome.ucsc.edu/cgi-bin/hgTables
  - Populations specific allele frequencies can be downloaded using
    - HapMap project or
    - HGDP (Human Genome Diversity Project).
  - Select 'Variation' in group menu
  - Select 'HapMap SNPs' or 'HGDP Allele Freq' in track menu
  - Then SNP list can be pasted or uploaded to appropriate file

Two-point Linkage Analysis

- Can be performed between the disease and marker loci for parametric linkage analysis
- For SNP data two-point linkage analysis is not very informative
- Can be used to elucidate linked regions
  - Which can be followed-up with multipoint analysis

Multipoint-point Linkage Analysis

- Can increase the informativeness of markers within the region
  - Extremely important when SNP marker loci are analyzed
- Helps to fine map a locus to a smaller region
  - Compared to two-point linkage analysis

Multipoint-point Linkage Analysis

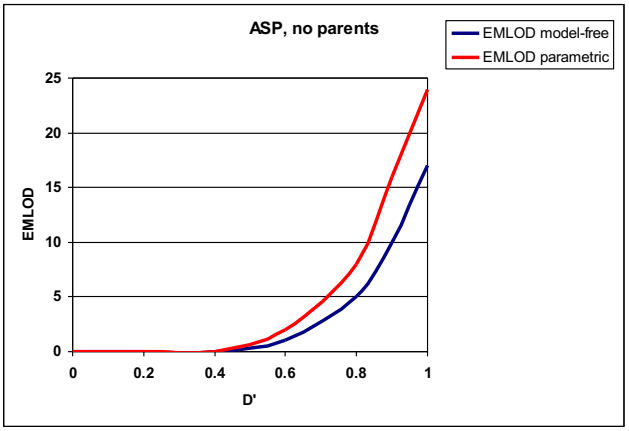
- Incorrect specified genetic map
  - Can bias LOD scores
  - Bias the position of the genetic locus
- For parametric linkage analysis when the genetic model is mis-specified and a susceptibility locus is placed between two flanking markers:
  - Can result in false negative results
  - Pushes the disease locus outside of the map of markers

Multipoint-point Linkage Analysis

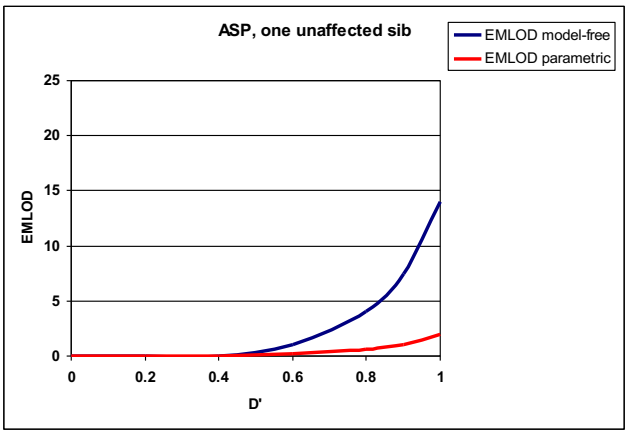
- Intermarker linkage disequilibrium (LD)
  - Can increase type I error
- When parental genotypes are missing
- For consanguineous pedigrees
  - when parental, grandparent, etc. genotypes are missing

Examples

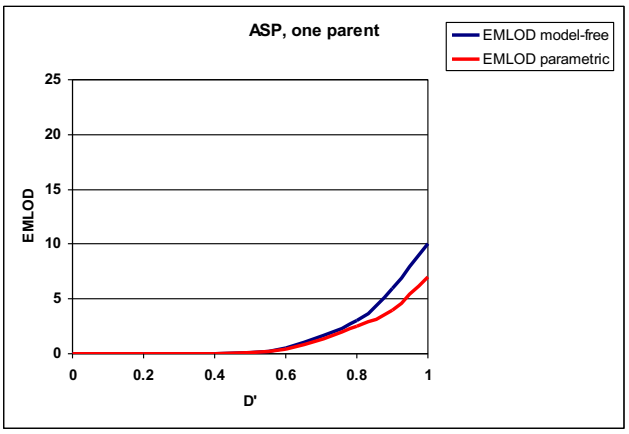
- Affected sibpairs
  - Without parental genotype data
  - Without parental genotype data and genotype data from one unaffected sibling
  - Missing genotype data from one parent
- Consanguineous pedigree – first cousin mating
  - Parents genotypes
    - Various relative missing genotype data



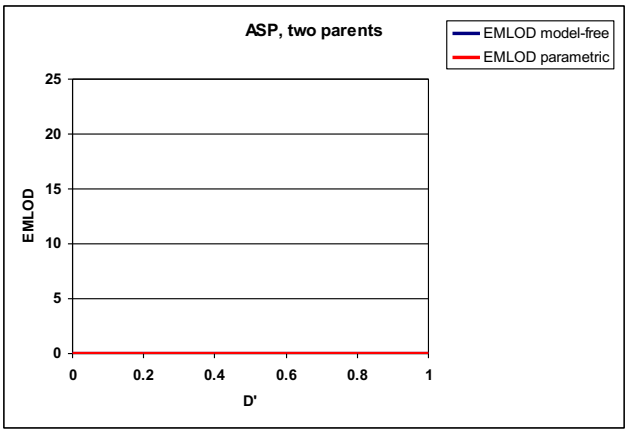
Huang et al. 2004



Huang et al. 2004

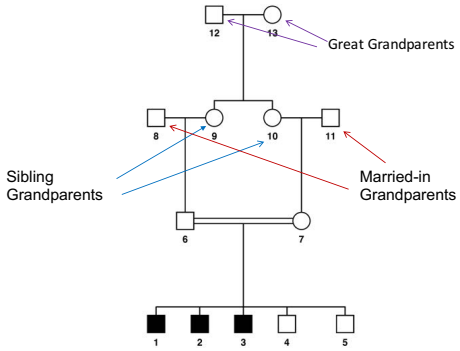


Huang et al. 2004

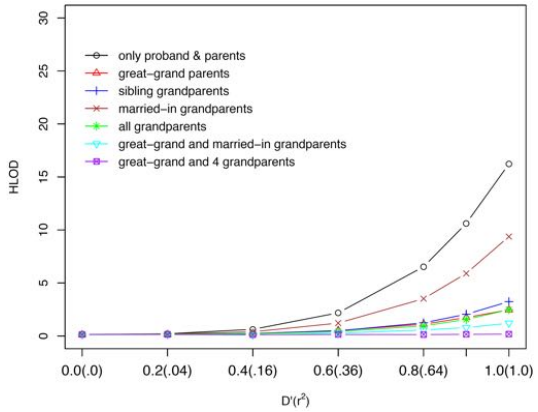


Huang et al. 2004

Pedigree Structure- Indicating Individuals with Missing Genotype data



Parents of the Proband Genotyped



Avoiding Inflation of LOD Scores due to Inter-marker LD

- Trim marker loci so that LD is weak between marker loci e.g. $r^2 < 0.5$
  - Can lead to a loss of power
- Analyze data using programs that incorporate haplotype frequencies
  - e.g. Merlin

Advantages of Two-point Linkage Analysis

- Not influenced by intermaker-LD
  - Therefore no inflation of the LOD score
- Not influenced by incorrect genetic maps
  - Which can cause incorrect map position and deflation of the LOD score

Error Detection in Pedigree data

- First need to remove markers which are missing a large number of genotypes
  - e.g. ≥5%
- A more stringent criterion can be used for SNPs with MAF<5%
  - e.g. >1%
- These markers can have higher genotyping error rates for the non-missing genotypes

Error Detection in Pedigree data

- Check for Mendelian errors
  - Marker should be removed for the entire pedigree
    - Do not just remove individuals involved in the Mendelian inconsistency
  - PedCheck
    - Useful program to detect Mendelian inconsistencies
      - https://watson.hgen.pitt.edu/register/docs/pedcheck.html

Error Detection in Pedigree data

- SNP markers are not very informative and therefore often not possible to detect errors through Mendelian inconsistencies
  - Those markers which are most informative (H=0.5) produce the least number of Mendelian inconsistencies
- Can check for double recombination events over short genetic distances
  - This is an indication that a genotyping error has occurred
    - Merlin (Abecasis et al. 2002 Nat Genet)
      - Can be used to detect double recombination events
      - http://csg.sph.umich.edu//abecasis/Merlin/

Type I error

- Reject the null hypothesis even when it is true
  - e.g. reject the null hypothesis of no linkage even when it is true
    - The null hypothesis of no linkage should have not been rejected

Type I error

- If a nominal criterion of p=0.05 is use as the criterion to reject the null hypothesis
  - One test performed 1 out of 20 chance null hypothesis rejected when it should not have been
    - False positive
  - If 1,000 tests are performed
    - By chance for ~50 tests the null will be rejected
      - Even though the null hypothesis is true

Type I error-Parametric Linkage Analysis

- If many tests are performed must adjust for multiple testing
  - Family wise error rate
- LOD score criterion takes into consideration
  - Multiple testing
  - Size of the genome
  - Number of chromosomes

Type I error-Parametric Linkage Analysis

- A LOD Score of 0.59
  - Nominal p-value 0.05 [one sided])
  - Is not used to reject the null hypothesis of no linkage
- For parametric linkage analysis a LOD score of 3.3* is used to reject the null hypothesis
  - Nominal one sided p-value 0.000049
  - Genome wide p-value 0.05

*Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241-247

Type II error (Power)

- Type II error
  - When the null hypothesis is false and it is not rejected
  - Represented by β
- Power
  - The ability to reject the null hypothesis when it is false
  - Most studies require a power (1-β) of at least 0.8

False positive

true positive

Statistical decision	True state of null hypothesis	
	Ho True	Ho False
Reject Ho	Type I error	Correct
Do not reject HO	Correct	Type II error

True negative

False negative

Cloud Computing

Michael Nothnagel, michael.nothnagel@uni-koeln.de, 2015

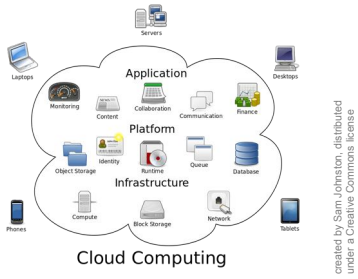
Outline

- Motivation
- Basic idea
- Providers
- Costs
- Advantages vs. disadvantages
- Concerns

Motivation

- Own IT infrastructure is expensive to set up, maintain and update
  - Projects may be one-time endeavors
  - Not cost-efficient for single, short-time projects
- Urgent need for immediate access to computing power
  - IT resources may not be present at location
  - Promised/Planned setup is delayed
- Setup of an IT infrastructure serving high demand in computing, storage and archiving requires substantial expertise
  - Limited pool of personnel

Basic idea

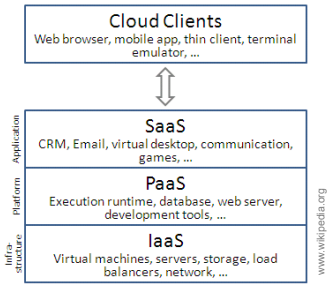


Instant, demand-driven access to a network of pooled configurable resources for computing and storage without or with minimal management effort by the service provider

Types of cloud services

IaaS	PaaS	SaaS
Infrastructure as a service ("cloud foundation")	Platform as a service	Software as a service
Own administration of virtual servers, complete access to instance	Running own application within the cloud, no administration	Use of an exisiting application offered by the cloud
Amazon's Web Services (Elastic Compute Cloud, Simple Storage Service S3)	Microsoft's Windows Azure, Goolge's App Engine	Apple's iCloud, Google's Drive, Microsoft's oneDrive, ownCloud, DropBox

Levels of cloud computing



Location of clouds

- Public cloud
  - Access to virtual IT infrastructure via the internet
  - Commercial service providers
- Private cloud
  - Access to virtual IT infrastructure within an organization
  - Usually located within the same country as the users
  - Protected against outside access
  - Increasingly used at high-performance computing (HPC) centers, e.g. at universities, by provision of virtual computers to users instead of real ones
- Hybrid & Community clouds

Providers

- Amazon
  - EC2 cor computing
  - S3 for storage (Web services)
- Google
  - Compute Engine
- IBM
  - Focus on businesses
- T Systems (Deutsche Telekom)
  - Focus in businesses
- There are many more providers.

Amazon

- Amazon's EC2 for elastic web-based computing
- Virtual servers ("instances") with
  - look & feel of a real server, with own IP address
  - root privileges
  - choice of operating system
  - flexible configuration of working memory, cores (CPUs), hard disk space
- Storage of data using Amazon's S3
- Booking as:
  - On-demand instance: payment by the hour, extremely flexible
  - Reserved instance: reservation of computing capacity for one or three years, less expensive than on-demand
  - Spot instance: bidding for unused EC2 capacity, execution of instance as long as bid is above actual spot price

Costs for Amazon: storage

AWS S3 Frankfurt (Germany): prices per GB per month

Region: EU (Frankfurt)

	Standardspeicher	Reduced Redundancy Storage	Glacier-Speicherung
Erstes TB pro Monat	$0.0324 pro GB	$0.0260 pro GB	$0.0120 pro GB
Nächste 49 TB pro Monat	$0.0319 pro GB	$0.0255 pro GB	$0.0120 pro GB
Nächste 450 TB pro Monat	$0.0314 pro GB	$0.0251 pro GB	$0.0120 pro GB
Nächste 500 TB pro Monat	$0.0308 pro GB	$0.0247 pro GB	$0.0120 pro GB
Nächste 4 000 TB pro Monat	$0.0303 pro GB	$0.0242 pro GB	$0.0120 pro GB
Über 5 000 TB pro Monat	$0.0297 pro GB	$0.0238 pro GB	$0.0120 pro GB

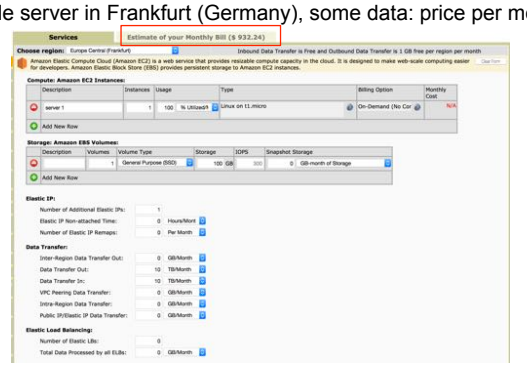
e.g. 3 TB per month: ~ $ 35-100

Additional fees apply for access and data transfer.

(August 2015)

Costs for Amazon: computing

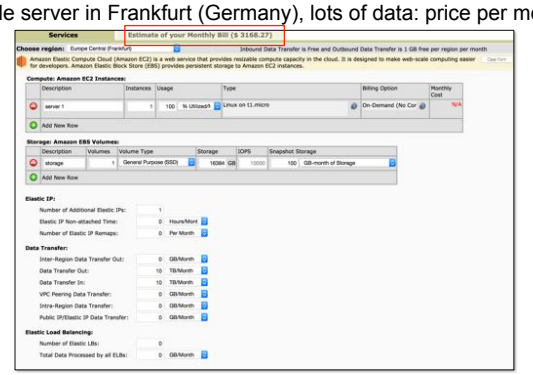
Single server in Frankfurt (Germany), some data: price per month



(August 2015)

Costs for Amazon: computing

Single server in Frankfurt (Germany), lots of data: price per month



(August 2015)

Costs for Google: storage

AWS S3 Frankfurt (Germany): prices per GB per month

Persistent Disk Pricing	
Standard Persistent Disk Provisioned Space	$0.04 GB / month
SSD Persistent Disk Provisioned Space	$0.17 GB / month
Snapshot storage*	$0.026 GB / month
IO operations	No additional charge

* Addition of egress fees for restore will be delayed as a promotional offer until September 1, 2015.

Local SSD Pricing	
Local SSD Provisioned Space	$0.218 GB / month

Image Storage	
Image storage	$0.085 GB / month

e.g. 3 TB per month: ~ $ 120-510

Additional fees apply.

(August 2015)

Costs for Google: computing

Single server in Europe/APAC: price per hour

High Memory

Machines for tasks that require more memory relative to virtual CPUs.

Machine type	Virtual CPUs	Memory	GCEU[1]	Lowest price[2] (USD) per hour with full sustained usage	Typical price[3] (USD) per hour	Full price[4] (USD) per hour without sustained use	Preemptible price[5] (USD) per hour
n1-highmem-2	2	13GB	5.50	$0.097	$0.106	$0.139	$0.0385
n1-highmem-4	4	26GB	11	$0.194	$0.212	$0.278	$0.077
n1-highmem-8	8	52GB	22	$0.388	$0.424	$0.556	$0.154
n1-highmem-16	16	104GB	44	$0.776	$0.848	$1.112	$0.308
n1-highmem-32[6] (Beta)	32	208GB	88	$1.552	$1.696	$2.224	$0.616

High CPU

Machines for tasks that require more virtual CPUs relative to memory.

Machine type	Virtual CPUs	Memory	GCEU[1]	Lowest price[2] (USD) per hour with full sustained usage	Typical price[3] (USD) per hour	Full price[4] (USD) per hour without sustained use	Preemptible price[5] (USD) per hour
n1-highcpu-2	2	1.80GB	5.50	$0.059	$0.064	$0.084	$0.022
n1-highcpu-4	4	3.60GB	11	$0.118	$0.128	$0.168	$0.044
n1-highcpu-8	8	7.20GB	22	$0.236	$0.256	$0.336	$0.088
n1-highcpu-16	16	14.40GB	44	$0.472	$0.512	$0.672	$0.176
n1-highcpu-32[6] (Beta)	32	28.80GB	88	$0.944	$1.024	$1.344	$0.352

e.g. per month: ~ $185

(August 2015)

Costs for Google: computing

OS for single server in Europe/APAC: price per hour

Premium OS Pricing

Pricing for premium operating systems differ based on the machine type where the premium operating system image is used. For example, an f1-micro instance will be charged $0.02 per hour for a SUSE image, while an n1-standard-8 instance will be charged $0.11 per hour. All prices for premium operating systems are in addition to charges for using a machine type.

Pricing for premium operating systems are the same worldwide and do not differ based on zones or regions, as machine type prices do.

More details

Red Hat Enterprise Linux (RHEL) images	$0.06/hour for machine types with less than 8 virtual CPUs $0.13/hour for machine types with 8 virtual CPUs or more
SUSE images	$0.02/hour for f1-micro and g1-small machine types $0.11/hour for all other machine types
Windows server images	$0.02/hour for f1-micro and g1-small machine types $0.04 per core/hour for all other machine types

e.g. per month: ~ $95

(August 2015)

Advantages vs. disadvantages

- Scalability both
  - by computational demand
  - by storage demand
- Costs
  - proportional to usage
  - occur at time of usage
- Immediate availability
  - No need to wait for that new HPC cluster for another six months
- Data backup
  - Usually automatically provided
- Limited bandwidth:
  - NGS datasets can be huge (up to several TB), depending on the type of transmitted file (fastq, bam, vcf)
  - Transfer can last several days or even weeks, with risks of interruption
- Costs:
  - No “flatrate” for usage
- Potential lock-in effect with provider when using more cloud-specific elements

Concerns: data privacy

- Security during transfer between client and server
  - Known and unknown but potentially exploited bugs in encryption software (e.g. SSL/TSL)
- Security at server
  - encryption of databases and file systems
- Profiling based on user data (Google)
- NSA and other secret services
- Account hijacking

Concerns: data protection

- >90% of all cloud infrastructure is located in the USA.
- National laws may prohibit transfer to another country or outside the European Union.
- Germany: Regular checks for compliance of used cloud with standards set by law (“Bundesdatenschutzgesetz”) at physical location are mandatory for personalized data (genetic identifiably!).
- Some US providers have set up computing centers within the European Union.
  - Ireland & Germany (Amazon), Denmark (Google)
  - They are still required under the US Patriot Act to transfer data to the US government if requested.
  - Big Brother Award 2012 for cloud computing as a technology awarded by digitalcourage, Chaos Computer Club e.V., and others

Literature on cloud computing

- Cloud computing security risk assessment by the European Union: http://www.enisa.europa.eu/activities/risk-management/files/deliverables/cloud-computing-risk-assessment
- Cloud Security Alliance: Top Threats https://cloudsecurityalliance.org/topthreats/csathreats.v1.0.pdf

Penetrance

Suzanne M. Leal
sleal@bcm.edu
Center for Statistical Genetic
Baylor College of Medicine
https://www.bcm.edu/research/labs/center-for-statistical-genetics
www.statgen.us

Elementary Usage of the Affection Status Locus Type

- Penetrance: Probability(Phenotype |genotype)
- Autosomal dominant - Complete penetrance no phenocopies

	1/1	1/2	2/2
P(affected|genotype)	0	1	1
P(unaffected|genotype)	1	0	0

All unaffected individuals must be 1/1
Affected individuals can be either 1/2 or 2/2

Note: 2 denotes the disease allele

Elementary Usage of the Affection Status Locus Type

- Penetrance: Probability(Phenotype |genotype)
- Autosomal recessive - Complete penetrance no phenocopies

	1/1	1/2	2/2
P(affected|genotype)	0	0	1
P(unaffected|genotype)	1	1	0

All affected individuals must be 2/2
Unaffected individuals can be either 1/1 or 1/2

Note: 2 denotes the disease allele

Elementary Usage of the Affection Status Locus Type

- X-linked recessive - Complete penetrance no phenocopies
  - Specify separate penetrances for males & females

	1/1	1/2	2/2
P(affected female|genotype)	0	0	1
P(unaffected female|genotype)	1	1	0

	1	2
P(affected male|genotype	0	1
P(unaffected male|genotype	1	0

All affected females must be 2/2 and affected males hemizygous 2

Note: 2 denotes the disease allele

Elementary Usage of the Affection Status Locus Type

- X-linked dominant or X-linked recessive with milder expression in females - Complete penetrance no phenocopies
  - Specify separate penetrances for males & females

	1/1	1/2	2/2
P(affected female|genotype)	0	0	1
P(unaffected female|genotype)	1	1	0

	1	2
P(affected male|genotype	0	1
P(unaffected male|genotype	1	0

Affected females are either 1/2 or 2/2 & affected males are hemizygous 2

Note: 2 denotes the disease allele

Use of Affection Status Locus

- Do not get confused by thinking the following is always true
  - 2 for affected individuals
  - 1 for unaffected individuals
- Code 2
  - Use penetrances as denoted in the datafile
- Code 1
  - 1-penetrances are used

Use of Affection Status Locus - Example

Autosomal Dominant

	1/1	1/2	2/2
Coded in datafile	0	1	1

Penetrance for individuals coded with a 2

1/1	1/2	2/2
0	1	1

Penetrance for individuals coded with a 1.

	1/1	1/2	2/2
	1-0	1-1	1-1
→	1	0	0

Multiple Liability Classes

- Used when penetrance is not the same for all individuals due to reduced penetrance and phenocopies
  - Age of onset
  - Different penetrance for males and females
- In pedigree file
  - First column "affection status"
  - Second column liability class

"Affection" status ↖ ↗ liability class

1	0	0	1	2 1	5 6
2	0	0	2	2 2	4 4
3	1	2	2	2 3	5 4
4	1	2	1	2 3	5 4
5	1	2	1	2 4	6 4

Pedfile.pre

Examples

- The ratios between genotypes are the most important factor in any penetrance model.
  - A risk ratio of 1:1 gives no linkage information at the disease locus.
  - Can be used for an individual whose phenotype is unknown

Example	1/1	1/2	2/2
	0.5	0.5	0.5

  - In the cases where the risk ratio is ∞:1 the individual either does or does not carry a copy of the disease allele. In the example below all individuals assigned to this penetrance class must be 2/2 at the disease locus.

Example	1/1	1/2	2/2
	0.0	0.0	1.0

Example

- Autosomal dominant disease with reduced penetrance and no phenocopies.

1/1	1/2	2/2	Codes in Pedigree File: Aff. Ind.	Unaff Ind.
0.0	0.6	0.6	2 1	1 1
0.0	0.8	0.8	2 2	1 2
0.0	1.0	1.0	2 3	1 3

1.) Does it matter which liability class an affected individual is assigned to? Why?

Example

- Autosomal dominant disease with reduced penetrance and phenocopies

	1/1	1/2	2/2	Codes in Pedigree File: Aff. Ind.	Unaff Ind.
0-10 yrs	0.01	0.6	0.6	2 1	1 1
11-25 yrs	0.02	0.8	0.8	2 2	1 2
>25 yrs	0.05	1.0	1.0	2 3	1 3

How would the following be coded?

- An unaffected 15 year old
- An unaffected 9 year old
- An affected 25 year old
- An affected 10 year old

Do the Penetrances make Sense on a Population Level?

- If the population prevalence of a disease Φ is known then the disease gene frequency p and penetrances f for an autosomal trait should satisfy:

$$\Phi = f_{DD}p^2 + 2f_{Dd}p(1-p) + f_{dd}(1-p)^2$$

- For sex-linked recessive traits

$$\Phi = pf_d + (1-p)f_+$$

Do the Penetrances make Sense on a Population Level?

- If the population prevalence of a disease Φ is known then the disease gene frequency p and penetrances f for an autosomal trait should satisfy:

$$\Phi = f_{DD}p^2 + 2f_{Dd}p(1-p) + f_{dd}(1-p)^2$$

- The population frequency for genetic cases for an autosomal dominant trait:
  - $A = f_{DD}p^2 + 2f_{Dd}p(1-p)$

Do the Penetrances make Sense on a Population Level?

- The frequency of phenocopies is given by:
  - $C = f_{dd}(1-p)^2$
- If the disease is rare then
  - $A \approx 2pf_{Dd}$
  - $C \approx (1-2p)$ fdd
- The phenocopy rate
  - The proportion of phenocopies amongst all affected individuals is
  - equal to C/(A+C).

Example

- Calculate the population prevalence and phenocopy rate for an autosomal dominant trait where:
  - $f_{DD} = f_{Dd} = 0.8$
  - $f_{dd} = 0.02$
  - $p = 0.001$
  - $\Phi = f_{DD}p^2$ (0.0000008) $+ 2f_{Dd}p(1-p)$ (0.00160) $+ f_{dd}(1-p)^2$ (0.01996)
- $\Phi = 0.0216$
- The phenocopy rate equals 0.926

How can Penetrance Data be Obtained

- From the literature
  - All necessary information not always available
- Estimate it from the data
  - Usually biased due to way data was ascertained
  - May not have enough data for reliable estimates
    - Linkage programs
    - Ageon (SAGE 3.1)
    - Approximate methods

Software to Perform Linkage Analysis Genotype Array Data

Suzanne M. Leal
sleal@bcm.edu
Center for Statistical Genetic
Baylor College of Medicine
https://www.bcm.edu/research/labs/center-for-statistical-genetics
www.statgen.us

Frequent Questions

- Why are you telling me about so many programs?
  - Please just let me know the best one
- Unfortunately not so easy
  - No single program works equally well in all situations
- Some programs can handle
  - Large pedigrees but not many markers
  - Many markers but not large pedigrees
  - Both large pedigrees and many markers but does not provide exact LOD scores
- Here is an abbreviated list of linkage programs

LINKAGE(Lathrop et al. 1984)/ FASTLINK (Cottingham et al. 1993)

- Parametric analysis only
- Suitable for relatively large pedigrees
- Limited in the number of loci for multipoint analysis
  - LINKAGE can allow for slightly more alleles/markers but slower than FASTLINK
- Elston-Stewart Algorithm scales exponentially with the number of loci and linearly with the number of non-founders

LINKAGE(Lathrop et al. 1984)/ FASTLINK (Cottingham et al. 1993)

- Quantitative and Qualitative analysis
- Allows for estimation of various parameters: theta, penetrance, allele frequencies
  - ILINK
- Two-point linkage can be performed using
  - MLINK
  - ILINK
- Can estimate haplotype frequencies and incorporate them in the analysis
  - MLINK
  - LINKMAP

LINKMAP LINKAGE/FASTLINK

- LINKMAP can be used to calculate multipoint LOD scores
- Due to Elston-Stewart algorithm can calculate LOD scores for large pedigrees but very limited in the number marker loci
  - Number of marker loci dependent on number of alleles
    - Maxhap
      - Product of the allele frequencies including disease locus
    - For SNP marker loci can only perform multipoint analysis using ~7 marker loci
  - Can use sliding window

Sliding the Disease Locus Across a Map of Marker Loci

1.) 1___2___3___4
2.) 2___1___3___4
3.) 1___3___4___5 (only one step)
4.) 3___1___4___5
5.) 1___4___5___6 (only one step)
6.) 4___1___5___6
7.) 4___5___1___6
8.) 4___5___6___1

1-Disease Locus
2, 3, 4, 5 and 6-Marker Loci

Superlink (Silberstein et al. 2006)

- Can analyze complex pedigrees quickly
  - Parametric linkage analysis
    - Computes exact LOD scores
  - Ideal for pedigrees with many loops (marriage or consanguinity)
    - In particular animal pedigrees
      - Dogs
      - Cattle
- Can perform multipoint linkage analysis
  - Limited in the number of marker loci 2-4
  - Implements Bayesian networks

Superlink (Silberstein et al. 2006)

- Can quickly calculate genome wide two-point LOD scores
- Multipoint linkage analysis
  - Use sliding window to calculate LOD scores for >~3 marker loci
  - Not suitable for genome-wide multipoint linkage analysis
- Efficient use of parallelization of the algorithm
- No need to install program
  - Use of Superlink is available free online

Genehunter (Kruglyak et al. 1996)

- Parametric and Non-parametric linkage analysis
- Provides exact rapid calculation of multipoint LOD scores through the implementation of hidden Markov Models
  - This approach scales linearly with the number of loci, but exponentially with the number of non-founders
  - Implements the Lander & Green Algorithm

Genehunter (Kruglyak et al. 1996)

- Handles a large number of marker loci
  - But only pedigrees of small to moderate to moderate size
    - Maxbit $(2n-f) \leq 21$
- Qualitative Traits
- NPL counts the numbers of alleles shared IBD amongst 2 or more affected relatives
  - Calculates p-values using either exact distribution or normal approximation

Genehunter 2.0 (Daly et al. 1998)

- Performs variance component analysis for mapping quantitative traits
- Performs all sib-pair analysis contained in the Mapmaker/sib software
- Constructs Haplotypes
- Implements a large pedigree approximation for the computation of a non-parametric allele sharing statistic on extended pedigrees of arbitrary size and complexity
- Computes traditional and multilocus Transmission Disequilibrium Test (TDT)

Allegro (Gudbjartsson et al. 2000)

- Allegro has the same basic functionality as Genehunter
  - Includes the features of Genehunter plus
- Supported features
  - Parametric and nonparametric LOD scores
  - Nonparametric NPL scores,
  - Information
  - Exact p-values
  - Expected crossover rate
  - Constructs Haplotypes
  - Simulation

Allegro (Gudbjartsson et al. 2000)

- Typical speedup compared to Genehunter is 30-fold.
- On a computer with four Gb of memory the program can handle pedigrees with up to about 28 bits
- Same data format as Genehunter
- ALLEGRO2 can handle even larger pedigrees

MERLIN (Abecasis et al. 2002)

- Handles small to medium sized pedigrees
  - Implements Lander & Green Algorithm
- Parametric analysis
- Non-parametric analysis
- Variance Components Analysis
- Regression based linkage analysis (quantitative traits)
- Incorporates LD in analysis
- Error checking – double recombination events over small genetic distances

SIMWALK2 (Sobel and Lange 1996)

- A Markov Chain Monte Carlo (MCMC) algorithm is implemented in order to transverse the space of inheritance vectors for each pedigree
- The initial legal descent state is found for using an iterative genotype elimination technique.
  - Simulated annealing is then performed to search for find the single most likely descent graph.

Simwalk2

- The MCMC random walk proceeds to sample the possible underlying configurations in proportion to their likelihood
  - A sample average is then used to give estimated results for the pedigree
- Can analyze large families with complex structures
  - >1000 individuals
- Handles a large number of markers
  - >30 markers
- Performs
  - Constructs Haplotypes
  - Parametric Analysis
  - Nonparametric analysis

Integrated Suites for Linkage Analysis - Alohmora

- Facilitates Analysis of a large number of markers
  - Incorporating genetic mappings
  - Allows for Analysis of a subset of markers
- Error Checking
  - Pedcheck
  - Merlin
- Linkage Analysis
  - Allegro
  - Merlin
  - Genehunter
  - Simwalk2

Integrated Suites-Easy Linkage

- Runs on windows
- Data preparation
  - Allows for analysis of a subset of markers
- Calls
  - Genehunter
  - Allegro, etc
- Graphical representation of results

Haplotypes/3-Unit Support Interval

- After completion of linkage analysis
  - Haplotypes should be constructed
    - e.g Allegro, SimWalk2
- Additionally a 3-unit support interval should be obtained
- If linkage was established the causal variant should lie within the haplotype and/or 3-unit support interval

In Summary

Error Detection in Pedigree data

- PedCheck
  - Mendelian errors
- Merlin
  - Double recombination events over short genetic distances

Analysis Programs

- Elston-Stewart Algorithm
  - Large Pedigrees
  - limited number of markers
    - Linkage
    - Fastlink
    - Vitesse
    - Superlink

Analysis Programs

- Lander-Green Algorithm
  - Small-medium sized pedigrees
  - Large number of Marker loci
    - Genehunter
    - Allegro
    - Merlin

Analysis Programs

- Other methods – Bayesian networks
  - Superlink
    - Suited for pedigrees with many inbreeding or marriage loops
- Approximate methods - MCMC
  - Simwalk2
  - LOKI

Pedigree Drawing Programs

- Haplopainter
  - Can draw pedigrees and haplotypes
- Pelican

Integrated Suites

- Easy Linkage
- Alohomora

Genotyping Error Detection Using MERLIN

– Introduction –

Michael Nothnagel, michael.nothnagel@uni-koeln.de, 2015

Sources of error in genetic data

- Genotyping errors
- False paternities
- Probes mix-up
- Data mix-up
- Wrong model (e.g. marker distances)
- Wrong phenotype definition
- ...

Appear as:

- Mendelian errors
- Unlikely genotypes
- Double recombinants over short genetic distances

The MERLIN software

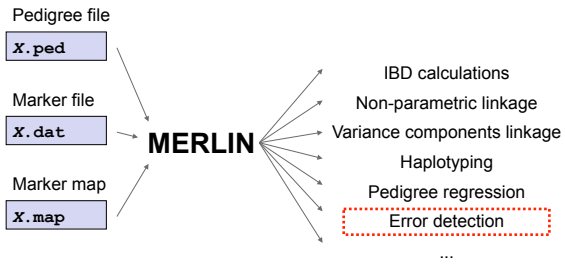
Merlin—rapid analysis of dense genetic maps using sparse gene flow trees

Gonçalo R. Abecasis[1,2], Stacey S. Cherny[1], William O. Cookson[1] & Lon R. Cardon[1]

Published online: 3 December 2001, DOI: 10.1038/ng786

- „Multipoint Engine for Rapid Likelihood INference“
- „Works like magic!“ (G. Abecasis)
- Available for Linux, Solaris , MacOS X, Windows
- Efficient data storage through sparse binary trees for modelling gene flow in pedigrees and corresponding algorithms
- Used via command line mode
- Reference:
  - Abecasis et al. (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97-101.
- Download & Docs:
  http://www.sph.umich.edu/csg/abecasis/Merlin/
- Current version: 1.1.2

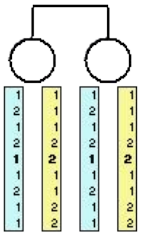
What can MERLIN be used for?



(linkage / QTDT / [PLINK] format)

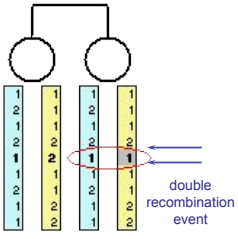
Unlikely genotypes (I)

Sib-ship 1



Sibs are identical at all markers
→ sibs very likely share this stretch of the chromosome
(identity-by-descent [IBD] = 2)

Sib-ship 2



One marker contradicts sharing information from all other markers
→ very unlikely case → check!

(from MERLIN online tutorial)

Unlikely genotypes (II)

Question: Do a particular marker with genotype g and its neighboring markers (G\g) provide consistent information?

Ratio of likelihood ratios (LR) for two marker map models:

$$r = \frac{LR_{linked}}{LR_{unlinked}} = \frac{\frac{L(G\backslash g|\theta)}{L(G|\theta)}}{\frac{L(G\backslash g|\theta=0.5)}{L(G|\theta=0.5)}}$$

Likelihood for:

- g set to unknown, all markers — using model map distances
- g set to unknown, all markers — assuming unlinked markers

g consistent with $G\backslash g$ under θ → $r << 1$

g inconsistent with $G\backslash g$ under θ → $r >> 1$

MERLIN reports $1/r$ as mistyping score!!

Running MERLIN

At the command line:

```
merlin -p X.ped -d X.dat [-m X.map] <options>
```

Pedigree file **Marker file** **Marker map file** (optional)

Analysis options (listed whenever MERLIN starts)

- General: --error [ON], --information, …
- IBD States: --ibd, --kinship, --matrices, …
- NPL Linkage: --npl, --pairs, --qtl, …

…

Other programs in bundle:

- `pedstats`
- `pedwipe`

…

Error detection using MERLIN

1. **Are there genotype errors in the data?**
   → Detection of Mendelian errors / unlikely genotypes:
   `merlin -p … -d … --error`

2. **Are reported errors simply due to chance?**
   → Estimation of the false-positive rate for error detection
   `merlin -p … -d … --error --simulate -r <seed> --reruns <reps>`

3. **„Wipe“ errors from data!**
   → Erase problematic genotypes (requires error file `merlin.err`)
   `pedwipe -p … -d …`

Other programs

- Other multipoint error detection methods implemented in:
  - SimWalk2
    (Sobel & Lange. Am J Hum Genet 1996;58:1323–1337)
    https://www.genetics.ucla.edu/software/simwalk
  - Mendel v14.4.2
    (Sobel E, et al. Am J Hum Genet 2002;70:496–508,
    Lange K, et al. Bioinformatics 2013;29:1568-1570)
    https://www.genetics.ucla.edu/software/mendel
  - Sibmed
    (Douglas JA, et al. Am J Hum Genet 2000; 66:1287–1297)
    http://csg.sph.umich.edu/boehnke/sibmed.php
- Performance comparison in:
  - Mukhopadhyay, et al. Comparative study of multipoint methods for genotype error detection. Hum Hered 2004;58:175-189.

Homozygosity Mapping

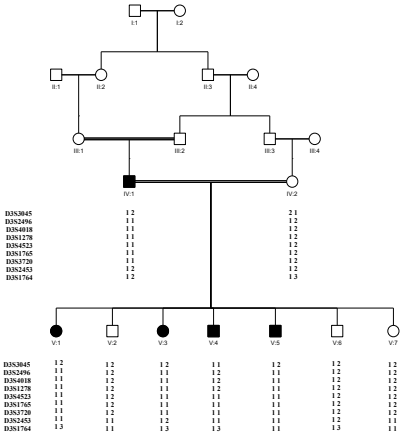
Suzanne M. Leal
sleal@bcm.edu
Center for Statistical Genetic, Baylor College of Medicine
https://www.bcm.edu/research/labs/center-for-statistical-genetics
www.statgen.us

Homozygosity Mapping - Concept

- Useful tool to map autosomal recessive traits
  - Particularly for consanguineous pedigrees
- Surrounding the pathogenic variant, multiple markers will be homozygous
  - Pointing to one or several regions of the genome where the pathogenic variant occurs

Homozygosity Mapping - Concept

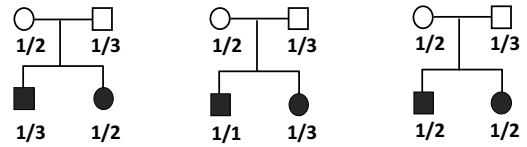
- Can look at homozygosity within a single individuals
- However information from several affected individuals
  - Usually but not necessarily from the same pedigree
    - Can help to reduce the number of regions
      - And the size of the region containing the putative causal variant



Concept

- Two segments of the chromosome are inherited from a common ancestor
  - Sharing is identical-by-descent (IBD)
- Often occurs in consanguineous pedigrees
- Two different haplotypes, surrounding the pathogenic variant, in affected individuals who are offspring of a consanguineous mating
  - Disease variant(s) entered the pedigree more than once in order to observe this phenomenon
    - Highly unusual

Identify by Descent (IBD)/Identify by State (IBS)



IBD=0 IBS=1 IBD=1 IBS=1 IBD=2 IBS=2

Concept

- Can also observe regions of homozygosity in "outbreed" populations
  - Small breeding pool
- Individuals, although not known to be, related (1st or 2nd cousins)
  - Are in reality quite closely related
  - Often inbreeding coefficients can be high due to generations of intermarriage

Concept

- Can occur in small populations
  - Geographically isolated
    - Mountains, Island populations
  - Socially isolated
- An individual can also inherit two copies of the same variant by "chance"
  - Usually parents are distantly related but this relationship is unknown

Performing Homozygosity Mapping

- In a single individual
- Often more than one run of homozygosity
- Difficult to determine which run of homozygosity contains the causal variant

Performing Homozygosity Mapping

- Information can be used from multiple individuals affected individuals (same phenotype) who may or may not related
  - If multiple individuals are homozygous for an overlapping interval on the chromosomes
  - Can lead to identifying the correct regions of homozygosity
    - Also aids in reducing the size of the interval

Performing Homozygosity Mapping

- Can determine that two individuals are distantly related because they are homozygous for the same haplotype
- Examine the region of homozygosity across individuals in order to obtain the smallest region in common
  - Likely to contain the pathogenic variant

Performing Homozygosity Mapping

- Even if two or more individuals are not homozygous for the same haplotypes
- Can still examine the haplotypes to determine the smallest interval containing the causal gene
- Caveat it can be possible that not all individuals have the same phenotype due to the same gene
  - Unusual but can occur when there are multiple genes responsible for the same phenotype within a small genetic region
    - Nonsyndromic hearing impairment
      - 13q11-13q12
        - » GJB2 and GJB6

Performing Homozygosity Mapping

- In this situation by examining the region of overlap between individuals
- Can accidently exclude region containing the causal variant
- This can also occur when examining smallest region of homozygosity between families
  - i.e. when analyzing families which do not have the same haplotype within the region of homozygosity

Performing Homozygosity Mapping

- Most beneficial in consanguineous pedigrees
- If pedigree is sufficiently large
  - Can usually map the causal variant to one region
- Caution should be used when trying to refine interval using unaffected individuals
  - May <u>not</u> have disease phenotype due to reduced penetrance
    - Carrying two copies of causal variant and thus are homozygous where the disease variant lies

Performing Homozygosity Mapping

- May be advantageous to only use affected individuals
- Dependent on disease etiology
- Likewise phenocopies can cause rejection of true region
- Phenotyping is extremely important

Performing Homozygosity Mapping

- Can help to quickly zoom in on the region containing the causal variant
- For homozygosity mapping analyzing thousands of marker loci takes seconds
  - Can use a wide variety of genotyping arrays
    - Illumina HumCoreExome-24 Bead Chip
  - Also can use exome or whole genome data
- Multipoint linkage analysis can be time consuming
  - Homozygosity mapping can be used to elucidate the region where initial linkage analysis should be carried out
  - And most likely contains the pathogenic variant

Performing Homozygosity Mapping

- Region of homozygosity and 3-unit linkage support interval usually perfectly overlap
- Performing multipoint linkage analysis not correcting for intermarker linkage disequilibrium can inflate LOD scores
  - This can occur if family members are missing genotype data
    - e.g. parental genotypes
      - For consanguineous pedigrees missing grandparental data can also cause an increase in false positive LOD scores

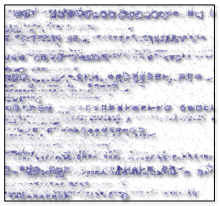
Performing Homozygosity Mapping

- Regions of homozygosity can give additional support to a linkage finding for autosomal recessive traits when analysis is performed in consanguineous pedigrees
  - Robust to intermarker linkage disequilibrium

Programs

- HomozygosityMapper (Seelow et al. 2009)
  - http://www.homozygositymapper.org/
- IBDfinder (Carr et al. 2009)
  - http://dna.leeds.ac.uk/ibdfinder/
- AutoSNPa (Carr et al. 2006)
  - http://dna.leeds.ac.uk/autosnpa/
- PLINK (Purcell et al. 2008)
  - http://pngu.mgh.harvard.edu/~purcell/plink/

File formats for sequence data

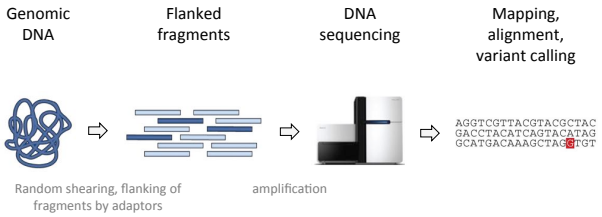


Michael Nothnagel, michael.nothnagel@uni-koeln.de, 2015

Outline

- NGS technologies
- Workflow and corresponding data files
- FASTQ files: reads fresh from the sequencer
- SAM/BAM files: read mapping
- VCF files: variants, genotypes and more

Whole-genome sequencing (WGS)

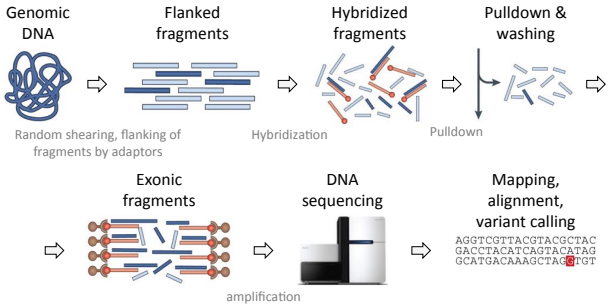


Whole-genome sequencing (WGS) does not cover the whole genome.
WGS is cheap on a per-base basis, but still expensive in total costs.
Sequencing allows assessment of genetic variation that is unknown in advance (must be known for genotyping).

modified from Bamshad et al. (2011) Nat Rev Genet

Whole-exome sequencing (WES)

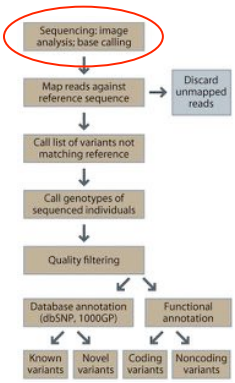
Exons as prime functional candidates represent 1-2% of total sequence



Vendors for exome capture kits: Agilent, Illimuna, Nimblegen, and others.
WES does not cover the whole exome. Covered regions depend on the used library.

modified from Bamshad et al. (2011) Nat Rev Genet

Bioinformatic workflow



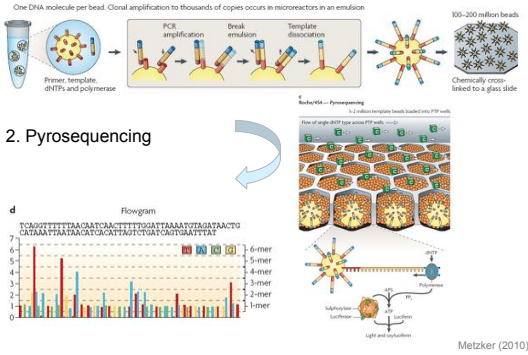
Jobling, et al. (2014)

Next-Generation Sequencing (NGS)

- Aim:
  - Full sequences, rare variants
  - Direct assessment of genetic variation directly
- 'Generations':
  - First: Sanger sequencing
  - Second: 'next-generation sequencing'
  - Third and Fourth are coming
- Platforms:
  - Roche 454
  - Illumina / Solexa / Sequenom
  - Applied Biosystems (ABI) SOLiD
  - Helicos BioSciences
  - Pacific Biosciences
  - Ion Torrent, …

NGS: Roche 454

1. Emulsion PCR

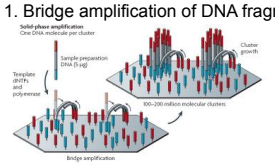


2. Pyrosequencing

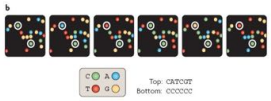
Metzker (2010) Nat Rev Genet

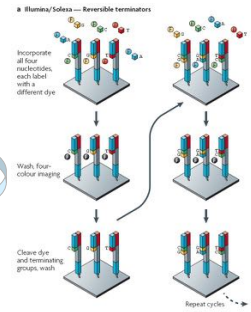
NGS: Illumina/Solexa

1. Bridge amplification of DNA fragments



2. Sequencing by labeled reversible terminators

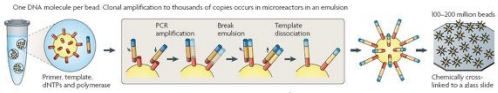




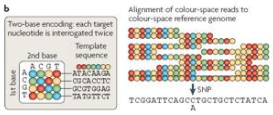
Metzker (2010) Nat Rev Genet

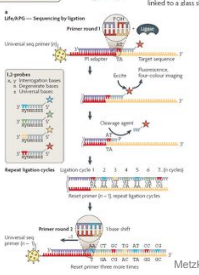
NGS: Applied Biosystems (ABI) SOLiD

1. Emulsion PCR



2. Sequencing by ligation





Metzker (2010) Nat Rev Genet

NGS: read length

Table 1 | Comparison of next-generation sequencing platforms

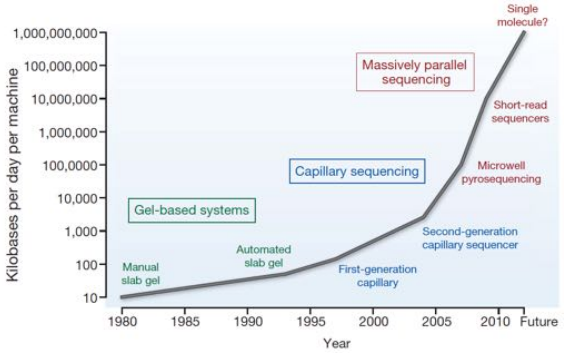
Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330[a]	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo-polymer repeats	Bacterial and insect genome de novo assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA_x	Frag, MP/ solid-phase	RTs	75 or 100	4[a], 9[b]	18[a], 35[b]	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7[a], 14[b]	30[a], 50[b]	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Polonator G.007	MP only/ emPCR	Non-cleavable probe SBL	26	5[c]	12[c]	170,000	Least expensive platform; open source to adapt alternative NGS chemistries	Users are required to maintain and quality control reagents; shortest NGS read lengths	Bacterial genome resequencing for variant discovery	J. Edwards, pers. comm.
Helicos BioSciences HeliScope	Frag, MP/ single molecule	RTs	32[a]	8[a]	37[a]	999,000	Non-bias representation of templates for genome- and seq-based applications	High error rates compared with other reversible terminator chemistries	Seq-based methods	91
Pacific Biosciences (target release: 2010)	Frag only/ single molecule	Real-time	964[a]	N/A	N/A	N/A	Has the greatest potential for reads exceeding 1 kb	Highest error rates compared with other NGS chemistries	Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks	S. Turner, pers. comm.

Metzker (2010) Nat Rev Genet

Capacity of sequencing instruments

Mardis (2011) Nature

Data generation throughput



Stratton, et al. (2009) Nature

NGS: sequenced individuals back then

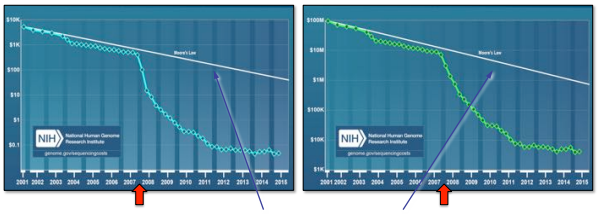
Table 2 | **Sequencing statistics on personal genome projects**

Personal Genome	Platform	Genomic template libraries	No. of reads (millions)	Read length (bases)	Base coverage (fold)	Assembly	Genome coverage (%)*	SNVs in millions (alignment tool)	No. of runs	Estimated cost (US$)
J. Craig Venter	Automated Sanger	MP from BACs, fosmids & plasmids	31.9	800	7.5	*De novo*	N/A	3.21	>140,000	70,000,000
James D. Watson	Roche/454	Frag: 500 bp	93.2[a]	250[b]	7.4	Aligned[c]	95[d]	3.32 (BLAT)	234	1,000,000[e]
Yoruban male (NA18507)	Illumina/ Solexa	93% MP: 200 bp	3.410[g]	35	40.6	Aligned[c]	99.9	3.83 (MAQ) 4.14 (ELAND)	40	250,000[f]
		7% MP: 1.8 kb	271	35						
Han Chinese male	Illumina/ Solexa	66% Frag: 150–250 bp	1,921[g]	35	36	Aligned[c]	99.9	3.07 (SOAP)	35	500,000[h]
		34% MP: 135 bp & 440 bp	1,029	35						
Korean male (AK1)	Illumina/ Solexa	21% Frag: 130 bp & 440 bp	393[g]	36	27.6	Aligned[c]	99.8	3.45 (GSNAP)	30	200,000[h]
		79% MP: 130 bp, 390 bp & 2.7 kb	1,156	36, 88, 106						
Korean male (SJK)	Illumina/ Solexa	MP: 100 bp, 200 bp & 300 bp	1,647[g]	35, 74	29.0	Aligned[c]	99.9	3.44 (MAQ)	15	250,000[h,i]
Yoruban male (NA18507)	Life/APG	9% Frag: 100–500 bp	213[g]	50	17.9	Aligned[c]	98.6	3.87 (Corona-lite)	9.5	60,000[h,j,k]
		91% MP: 600–3,500 bp	2,075[g]	25, 50						
Stephen R. Quake	Helicos BioSciences	Frag: 100–500 bp	2,725[g]	32[l]	28	Aligned[c]	90	2.81 (Index-DP)	4	48,000[h]
AML female	Illumina/ Solexa	Frag: 350–200 bp[m]	2,730[g,n]	32	32.7	Aligned[c]	91	3.81[o] (MAQ)	91	1,600,000[p]
		Frag: 150–200 bp[m]	1,081[g,n]	35	13.9		83	2.92[o] (MAQ)	34	
AML male	Illumina/ Solexa	MP: 200–250 bp[m]	1,620[g,n]	35	23.3	Aligned[c]	90.5	3.46[o] (MAQ)	16.5	500,000[p]
		MP: 200–250 bp[m]	1,351[g,n]	50	21.3		97.4	3.45[o] (MAQ)	13.1	
James R. Lupski CMT male	Life/APG	16% Frag: 100–500 bp	230[g]	35	29.6	Aligned[c]	99.8	3.42 (Corona-lite)	3	75,000[h,q]
		84% MP: 600–3,500 bp	1,211[g]	25, 50						

Metzker (2010) Nat Rev Genet

Sequencing costs

per Megabase (Mb) per genome

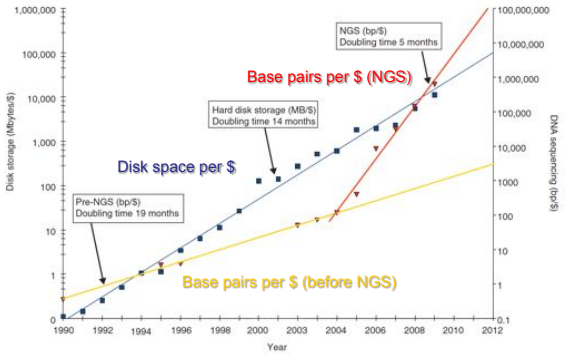


Moore's law (1965, co-founder of Intel)

"The number of transistors in a dense integrated circuit doubles approximately every two years"

http://www.genome.gov/SequencingCosts/

In-silico storage of NGS data



Stein (2010) Genome Biology

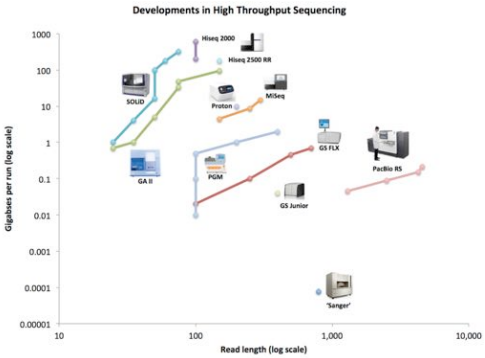
Output of sequencers

TECHNOLOGY CLASS	SEQUENCING MACHINE	READ LENGTH (NUCLEOTIDES)	READS PER RUN	RUN TIME
Conventional (chain termination sequencing)	ABI prism 3730 Sanger dideoxy sequencing	400–900	96	20 minutes to 3 hours
Massively parallel sequencing of PCR-amplified DNAs	Roche/454 pyrosequencer	400–600	1 million	7 hours
	Illumina/Solexa HiSeq 2000	150 × 2	many hundreds of millions	2 days to 10 days
	ABI SOLiD 4	35–75	hundreds of millions	7 days
	Life Technologies Ion Torrent	200	5 million	4 hours
Massively parallel sequencing of unamplified (single-molecule) DNAs	Pacific Biosciences SMRT (single-molecule real time) sequencing	~3000	up to 75,000	1 hour

Table 3.3 Genetics and Genomics in Medicine (© Garland Science 2015)

Strachan, et al. (2015)

Read length vs. throughput



Lex Nederbragt (2013): developments in NGS. figshare. http://dx.doi.org/10.6084/m9.figshare.100940.

Illumina HiSeq X for WGS

Table 1: HiSeq X System Sequencing Capacity

	HiSeq X Ten System	HiSeq X Five System
Minimum Number of Instruments	10	5
Annual Genome Capacity	> 18,000	> 9000
Price per 30× Genome	< $1000	< $1500

Table 2: HiSeq X System Performance Parameters[a]

Parameter	Specification
Output per Run	Dual flow cell: 1.6-1.8 Tb
Single Reads Passing Filter	Dual flow cell: 5.3-6 billion
Supported Read Length	2 × 150 bp
Run Time	< 3 days
Quality	≥ 75% of bases above Q30 at 2 × 150 bp
Supported Library Preparation	TruSeq DNA PCR-Free Library Prep Kit TruSeq Nano DNA Library Prep Kit

All information has been obtained from the Illumina web site.

http://www.illumina.com/systems/hiseq-x-sequencing-system/

FASTQ file format

Contains all unaligned reads obtained from the sequencer

Plain text file, four lines per sequence

Sequence identifier
(optionally some more description)

Sequence letters

```
…
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
…
```

"+"
(optionally some more description)

Quality code for respective letter

```
!"#$%&'()*+,-./0123456789:;<=>@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

low → high

Base quality: Phred score Q (I)

Sanger sequencing:

$$Q = -10 \log_{10} p$$

Ewing & Green (1998)
Genome Res

Illumina/Solexa have used a slightly different quality score in the past:

$$Q = -10 \log_{10} \frac{p}{1-p}$$

p – probability of an incorrect base call

Quality: Phred score Q (II)

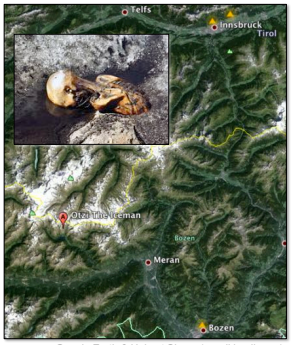
Sanger sequencing:

Quality score Q	$p=10^{-(Q/10)}$	Probability of incorrect base call	Proportion of accurate base calls
10	10^{-1} = 0.1	1 in 10	90%
20	10^{-2} = 0.01	1 in 100	99%
30	10^{-3} = 0.001	1 in 1,000	99.9%
40	10^{-4} = 0.0001	1 in 10,000	99.99%
50	10^{-5} = 0.00001	1 in 100,000	99.999%
60	10^{-6} = 0.000001	1 in 1,000,000	99.9999%

Phred score	0	1	…	20	…	50	…	92	93
ASCII coding	33	34	…	53	…	83	…	125	126
FASTQ symbol	!	"	…	5	…	S	…	}	~

Quality shows a trend of decreasing towards the end of a read.

An example: the Tyrolean Iceman ("Ötzi")



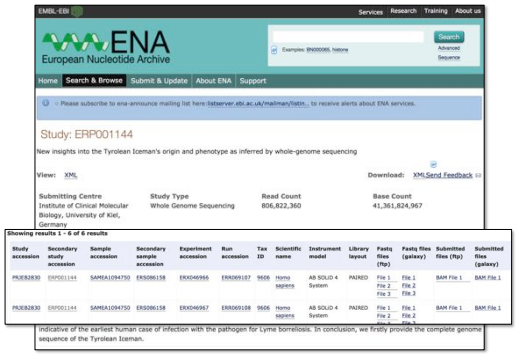
Google Earth & Helmut Simon (en.wikipedia.org)

- Discovery in Sept. 1991 off a mountain pass (3210m) near Tisenjoch (Ötztal Alps) by hiking German couple (92 m inside Italy, off Austria)
- Ötzi died ~5250 YBP during the Copper Age (Chalcolithic)
- First genetic study in 1994 (Handt et al., Science; on mtDNA variation)
- Full nuclear sequencing (WGS) in 2012 (Keller, Graefen, Ball, et al., Nat Comm)

Data deposited at European Nucleotide Archive (ENA):
http://www.ebi.ac.uk/ena; accession number: ERP001144

Ötzi's sequence data

http://www.ebi.ac.uk/ena/data/view/ERP001144



A FASTQ file from Ötzi

```
@ERR069107.83594103 61_656_767/1
GGCTGAGGCAGGAGAATTGCTTGAACCCAGGACACGGAGGTTGTG
+
IIIIIIIIIIIIIIGIHHIIIII:HI.,,(AI1(((F)&BIIIIE
@ERR069107.83594104 1361_1032_197/1
TGGAATGGAATGGAATGGAATGGAATGGAAC
+
IIGIIICEIIICCIIIIIIIICAI((D2IF;
@ERR069107.83594105 2174_789_1143/2
TGTGTGTGTGTGTGTGTGTGTGTGT
+
IIIIIIIIIIIIIIIIIIIIIIII@
@ERR069107.83594106 744_1322_1870/1
TCTCCACTTCATTCCATTCCATTTCATCCTATTCCA
+
IIIIIIIIIIIIIIIIIIIIIIIIIIB>I??IIIII
@ERR069107.83594107 533_1805_480/1
TCTCAAAAGAAGACATTTATGCAGCCAAAAAATACATGAAAAAAATGC
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@ERR069107.83594108 997_1205_1612/1
CATTCCATTGCATTCCATTCCATTCCATTAGTTTCCATTCCATTC
+
IIIIIIIIIIIIIIIIIIIIIIIIIII"""ABIIIIIIIIIIII
@ERR069107.83594109 1010_363_607/1
CGCAATGGCATTCCTAATGCTTACCGAACGAAAAAT
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHII
:
```

Sequence identifier

Sequence (31 bases)

Quality

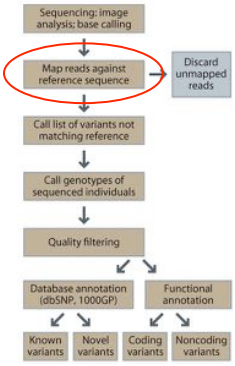
probability of incorrect base call

I → Q=73 → $p=10^{-73/10}=5.0\times10^{-8}$

(→ Q=40 → $p=10^{-40/10}=1.0\times10^{-4}$

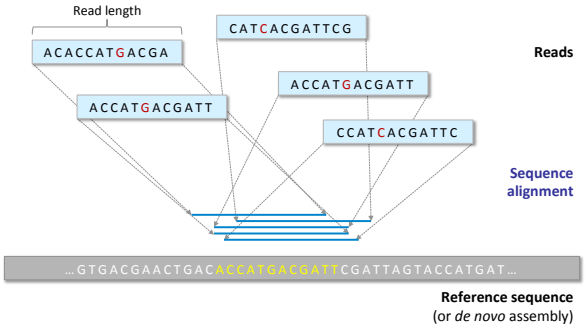
; → Q=59 → $p=10^{-59/10}=1.3\times10^{-6}$

Bioinformatic workflow



Jobling, et al. (2014)

Sequence alignment



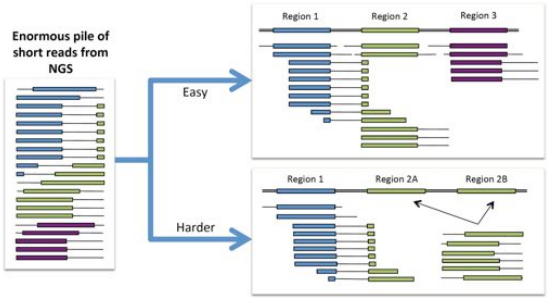
Sequence alignment

- Matching single sequences/reads (pairwise alignment) or multiple sequences/reads (multiple alignment) to a reference
- Local alignment; multitude of approaches, many NP -complete (classical algorithm: Smith-Waterman algorithm)
- Frequently used software for WGS/WES: BWA, bowtie
- Many others available, often for specialized tasks (or outdated):
  ELAND (Illumina), MAQ, Partek, VelociMapper, GEM, SOAP/SOAP2/SOAP3, …
- Memory-consuming!! (>2 GB)
- Repetitive and pseudo-autosomal regions are hard to align and therefore barely accessible to NGS

BWA & bowtie

- Aligner using Burrows-Wheeler Transform (approach for data compression, published 1994)
- Mapping low-divergent sequences against a large reference genome
- BWA: three algorithms:
  - BWA-backtrack: short reads up to 100bp (previous Illumina)
  - BWA-MEM: longer reads (70bp-1Mb)
  - BWA-SW: like BWA-MEM, but better for frequent alignment gaps
  - Li & Durbin (2009, 2010) Bioinformatics
  - Extensions: BarraCUDA, UGENE (visual interface)
- bowtie/bowtie 2:
  - Very fast, memory-efficient
  - Alignment of short reads
  - Langmead, et al. (2009) Genome Biol

Mapping of reads to a reference genome



Read mapping is complicated mismatches (errors or variants), InDels, duplications, insertions, etc.

https://www.broadinstitute.org/gatk/guide/best-practices

Mapping quality score MAPQ

Probability (on a log scale) of a read being misplaced

$$MAPQ = -10 \log_{10}(1-p)$$

Li, Ruan, Durbin (2008) Genome Res

p – probability of the read coming from the correct position

This probability is approximately modeled using a Bayesian approach, assuming that sequencing errors at different sites of read are independent of each other.

$$p_s(u|x,z) = \frac{p(z|x,u)}{\sum_{v=1}^{L-l+1} p(z|x,v)}$$

best alignment

all possible alignments

$p(z|x,..)=10^{-(\Sigma Qi)/10}$

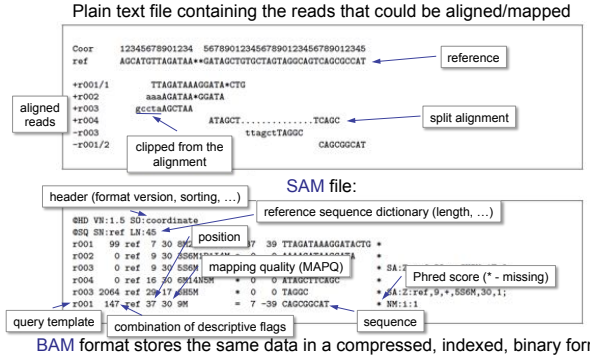
sum of single-base phred Q values at mismatch positions

e.g. two mismatches with Q=20 and Q=10: $p(z|x,..)=10^{-(20+10)/10}=0.001$

SAM/BAM file format

Sequence Alignment/Map format

Plain text file containing the reads that could be aligned/mapped

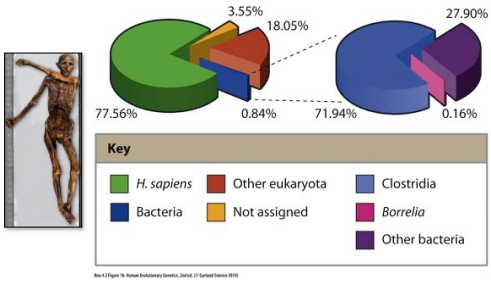


BAM format stores the same data in a compressed, indexed, binary form.

https://github.com/samtools/hts-specs

A BAM file from Ötzi

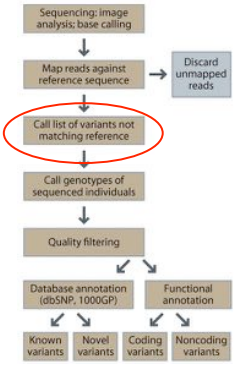
Genetic sample of Ötzi



Sequencing errors, sample contamination, and other factors can lead to unmapped reads.

Jobling, et al. (2014)

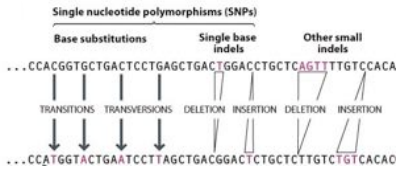
Bioinformatic workflow



Jobling, et al. (2014)

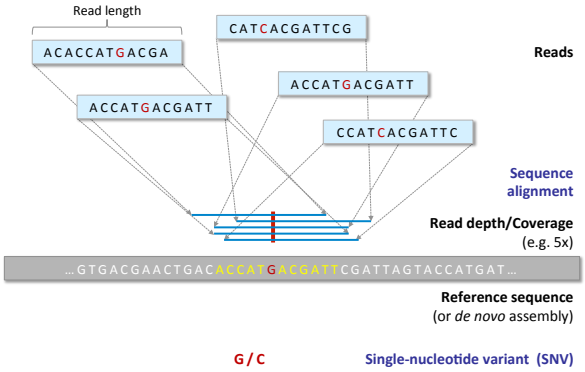
Types of variation

SNVs	InDels	SVs
• Single-nucleotide variants	• Single-base or small insertions and deletions	• Structural variants • Copy number variants (CNVs) • Inversions, translocations



Jobling, et al. (2014)

Variant calling for SNVs and InDels



Variant calling for SNVs and InDels

Unified Genotyper	Haplotype Caller
• SNVs and InDels are called separately • aka Consensus Calling (in MAP software) • Bayesian approach; the likelihood for each genotype is expressed depending on the Q and MAPQ error probabilities • The genotype with highest posterior probability is selected • First implemented in MAQ (Li, et al., 2008, Genome Biol) • Faster, any ploidy	• SNVs and InDels are called simultaneously (in local neighborhood) • Re-assembly of genomic regions with large variation; identification of possible haplotypes per region • Calculation of haplotype likelihood for given data • Bayesian approach as with consensus calling • Implemented in GATK • More accurate (for InDels)

No software is optimal for every task.
Different callers are used for different sorts of variants.

Variant calling

- SAMtools
  - Outgrew from the 1000 Genomes project
  - mpileup for calling SNVs and InDels
  - Li, et al. (2009) Bioinformatics 25, 2078-9; Li, et al. (2011) Bioinformatics 27:2987-93
  - http://samtools.sourceforge.net/
- GATK
  - Genome Analysis ToolKit; developed at the Broad Institute
  - Requires Java
  - McKanna, et al. (2010) Genome Res 20:1297-303; DePristo, et al. (2011) Nat Genet 43:491-498; Van der Auwera, et al. (2013) Curr Protocols Bioinformatics 43:11.10.1-11.10.33
  - https://www.broadinstitute.org/gatk/
- MAQ, FreeBayes, …

VCF/BCF file format

Variant Call Format

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
:
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
:
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT    QUAL  FILTER INFO                              FORMAT      NA00001        NA00002
20    14370   rs6054257 G   A      29    PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51.
20    17330   .         T   A      3     q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3
20    1230237 .         T   .      47    PASS   NS=3;DP=13;AA=T                    GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51
20    1234567 microsat1 GTC G,GTCT 50    PASS   NS=3;DP=9;AA=G                     GT:GQ:DP    0/1:35:4       0/2:17:2
:
```

meta information (field description, filter criteria, etc.)

header

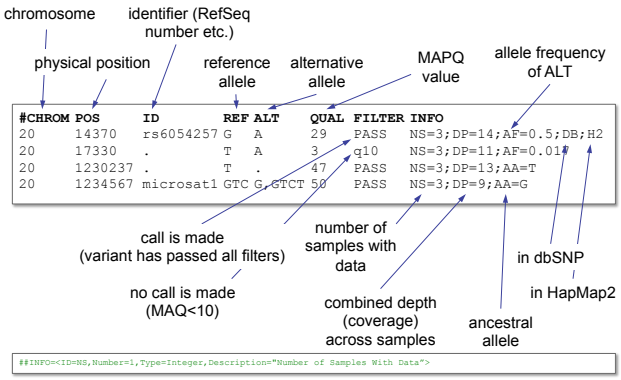
called variants

variant information | variant quality | samples

mandatory columns

BCF format stores the same data in a compressed, binary form.

VCF: variant information



See https://github.com/samtools/hts-specs for a specification of entries.

Genotypes in REF and ALT columns

…	REF	ALT	…	
…	G	A	…	substitution (SNV), 2 alleles
…	G	A,T	…	substitution (SNV), 3 alleles
…	T	.	…	monomorphic (no variant)
…	GTC	G	…	deletion, 2 bp
…	G	GTCT	…	insertion, 3 bp
…	T	<DEL>	…	large deletion (1 kb)

VCF: sample information

Data format: (described in VCF header!)

GT: inferred genotype
GQ: conditional genotype quality (phred scale)
DP: read depth (coverage)

Sample information:

```
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
```

…	FORMAT	NA00001	NA00002
…	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51
…	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3
…	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51
…	GT:GQ:DP	0/1:35:4	0/2:17:2

Individual NA00002:

1|0: genotype phased, heterozygous
48: probability of $10^{-4.8}$=0.000016 for an erroneous call
8: read depth (coverage)

0/2: genotype unphased, heterozygous for second allele in ALT
17: probability of $10^{-1.7}$=0.02 for an erroneous call
2: read depth (coverage)

VCF uses a general syntax system and flexible for coding different information.

Technical information

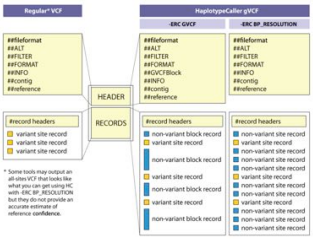
Variant information	Genotype information
AA: ancestral allele	GT: genotype: / or | un/phased 0-REF, 1,2,..-ALT allele
AC: allele count in genotypes, respectively for each ALT allele	DP: read depth at this site
BQ: base quality Q at this site	GL: $\log_{10}$ genotype likelihoods: GT:GL 0/1:-323.0,-99.1,-802.5 → L(G=0,0)=$10^{-323.0}$ L(G=0,1)=$10^{-99.1}$ L(G=1,1)=$10^{-802.5}$
DB: dbSNP membership	PL: 10*$\log_{10}$ (phred-scaled) genotype likelihoods
DP: combined read depth cross samples	GP: phred-scaled genotype posterior probabilities
H2/3: HapMap2/3 membership	PQ: phasing quality
MQ: mapping quality MAPQ	MQ: mapping quality MAPQ
MQ0: number of reads with MAPQ=0 covering this site	*and more*
1000G: 1000 genomes membership	
and more	

A VCF file from Ötzi

```
##fileformat=VCFv4.1
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order
listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad
mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
:

:
##contig=<ID=1,length=249250621,assembly=b37>
##contig=<ID=10,length=135534747,assembly=b37>
##contig=<ID=11,length=135006516,assembly=b37>
:

:
#CHROM POS    ID REF ALT QUAL   FILTER  INFO …
1      10002  .  A   C   116.55 PASS    ABHom=0.87;AC=2;AF=1.00;AN=2;BaseCounts=0,7,0,1;DP=8;Dels=0.00; …
1      10003  .  A   T   49.02  PASS    ABHet=0.13;AC=1;AF=0.50;AN=2;BaseCounts=1,0,0,7; …
1      233571 .  G   A   11.34  LowQual ABHom=1.00;AC=2;AF=1.00;AN=2;BaseCounts=2,0,0,0;DP=2;Dels=0.00; …
1      234466 .  C   T   13.41  LowQual ABHom=1.00;AC=2;AF=1.00;AN=2;BaseCounts=0,0,0,2;DP=2;Dels=0.00; …
1      546897 .  G   T   14.49  LowQual ABHom=1.00;AC=2;AF=1.00;AN=2;BaseCounts=0,0,0,2;DP=2;Dels=0.00; …
1      546900 .  G   A   14.49  LowQual ABHom=1.00;AC=2;AF=1.00;AN=2;BaseCounts=2,0,0,0;DP=2;Dels=0.00; …
1      565286 .  C   T   131.66 PASS    ABHom=1.00;AC=2;AF=1.00;AN=2;BaseCounts=0,0,0,40;DP=40;Dels=0.0 …
1      567493 .  C   T   25.63  LowQual ABHet=0.70;AC=1;AF=0.50;AN=2;BaseCounts=0,7,0,3;BaseQRankSum=- …
1      569492 .  T   C   157.27 PASS    ABHom=0.93;AC=2;AF=1.00;AN=2;BaseCounts=1,14,0,0;DP=15;DS;Dels= …
1      800383 .  C   T   14.58  LowQual ABHom=1.00;AC=2;AF=1.00;AN=2;BaseCounts=0,0,0,2;DP=2;Dels=0.00 …
:

:

:
```

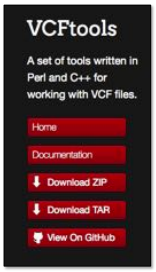
Genomic VCF (gVCF) file format

- VCF file with
  - a record for every position (also for non-called variants)
  - per-sample reference confidence estimation for invariant sites
- Produced by Haplotype Caller
- Developed at the Broad Institute



VCFtools

https://vcftools.github.io/



- Software for manipulating VCF files
- Possible tasks:
  - Filter and summarize variants, create intersections and subsets
  - Compare, validate and merge VCF files
  - Convert to different formats (e.g. PLINK, IMPUTE, BEAGLE)
  - Perform some analyses:
    - Calculation of population-genetic parameters (nucleotide diversity, FST, Tajima's D, Hardy -Weinberg proportions & test, etc.)
    - Linkage disequilibrium calculation (r^2)
    - SNP density, sample relatedness, etc.
- Danecek, et al. (2011) Bioinformatics

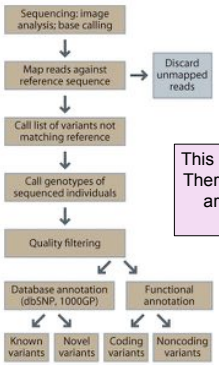
Bioinformatic workflow

FASTQ

SAM/BAM

VCF/gVCF

genotype files

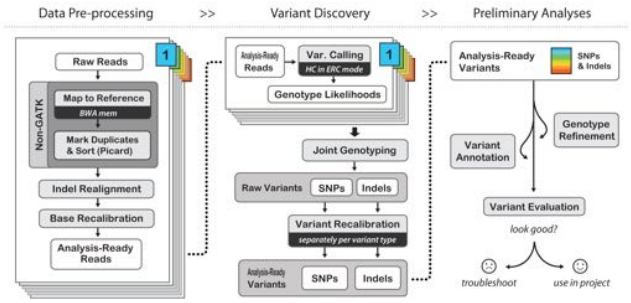


This is a simplified picture. There are more steps that are taken to produce genotype calls.

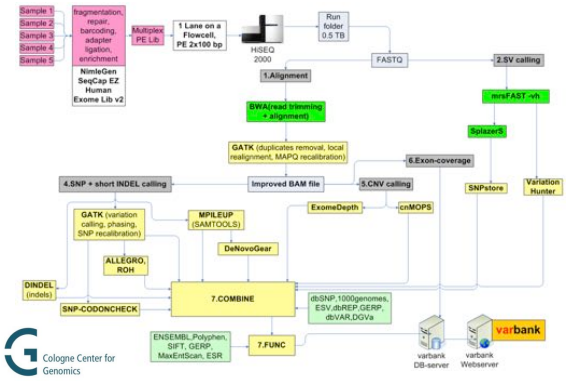
Jobling, et al. (2014)

GATK Best Practice

https://www.broadinstitute.org/gatk/guide/best-practices



Exome pipeline at the CCG



Visualization: Integrated Genome Viewer (IGV)

- Interactive visualization tool for large integrated datasets
- Many supported file formats, including SAM/BAM and VCF files
- http://www.broadinstitute.org/software/igv/
- Robinson, et al. (2011) Nat Biotech 29, 24–26; Thorvaldsdóttir, et al. (2013) Brief Bioinf 14, 178-192

IGV: view of aligned sequence reads

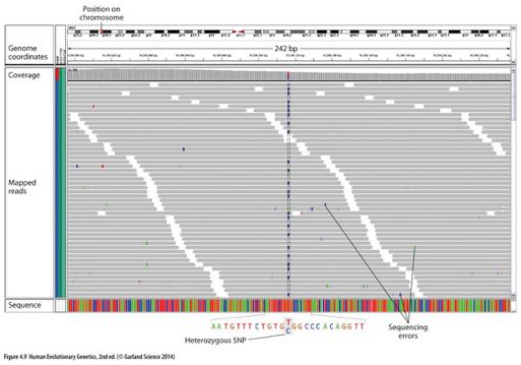
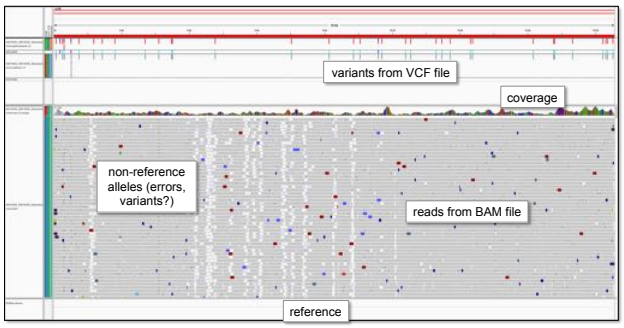


Figure 4.9 Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

Jobling, et al. (2014)

Ötzi's mtDNA in IGV



Cautionary notes on variant calling

- Models for variant calling are tuned for sensitivity
  - Project-specific trade-off between between sensitivity and specificity
- Variant calls are error-prone
  - Substantial proportions of false-positives are to be expected (!)
- Variant calling quality depends on the experiment
  - Raw DNA isolation
  - Library preparation
  - Sequencing (inter-lane differences)
- Variant calls require subsequent filtering before meaningful analyses can be conducted

Estimated proportions of false SNV detections

1000 Genomes, Pilot 2, chromosomes 1-22

[%]	NA12878 (CEU)			NA19240 (YRI)		
	All SNVs	Consensus only	P	All SNVs	Consensus only	P
454 FLX™	6.3 (6.1-6.5)	0.7 (0.5-3.6)	$<10^{-4}$	2.9 (2.7-3.2)	2.6 (1.2-4.7)	0.08
GA IIx™	8.4 (8.0-8.7)	3.5 (3.1-3.9)	$<10^{-4}$	11.1 (10.9-11.3)	3.9 (3.5-4.3)	$<10^{-4}$
SOLiD™	17.1 (16.9-17.4)	0.8 (0.1-2.6)	$<10^{-4}$	7.3 (6.8-7.8)	4.0 (3.1-4.8)	$<10^{-4}$

P values obtained from a permutation test.

Nothnagel, Herrmann, et al. (2011) Hum Genet

Literature on file formats (& Ötzi)

- Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11:31-46.
- https://github.com/samtools/hts-specs
- https://www.broadinstitute.org/gatk/guide

- http://www.ebi.ac.uk/ena
- Keller & Graefen & Ball, et al. (2012) New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat Comm 3:698.

Ötzi's museum

Hubert Berberich (de.wikipedia.org)

South Tyrolean Archeological Museum, Bozen, Italy

Filtering Approaches for the Analysis of NGS Data

Suzanne M. Leal
sleal@bcm.edu

A Few Words About Next Generation Sequencing

Generation of NGS Data

- Capture arrays can be used with sequencing to generate data on
  - Exomes
    - Aligent SureSelect 38MB
    - Aligent SureSelect 50Mb
    - Illumina TrueSeq Exome Enrichment (62Mb)
  - Targeted regions
  - Genes

Whole Exome Sequencing

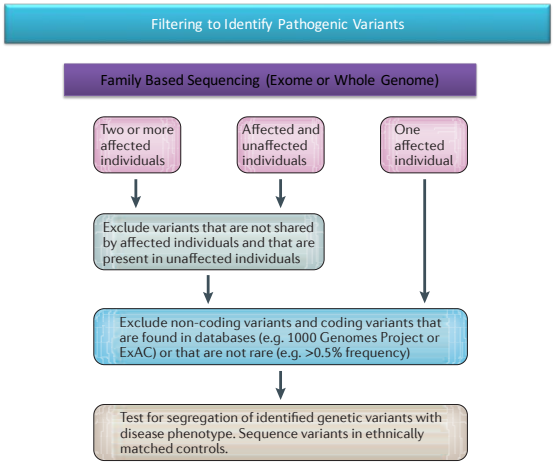
- Not really whole exome
  - Not all genes are targeted
    - Great variability between capture arrays
      - Different arrays capture different proportions of the exome
  - Not all targeted genes are captured
  - Not all targeted sequences call be aligned
  - Not all aligned sequences can be accurately called
  - Not all captured regions have sufficient depth to call variants

NGS Data

- For exome sequencing high quality data consists of of a median depth of >80X
- With >90% of the exome covered with a depth of $\geq$10X
- Whole Genome sequencing (good quality) ~30X coverage
  - Not necessary to use such high depth for whole genome as for exome sequencing
    - Reads are distributed more evenly across genome
- Sequence data for an exome is ~1/15th of the data for a genome

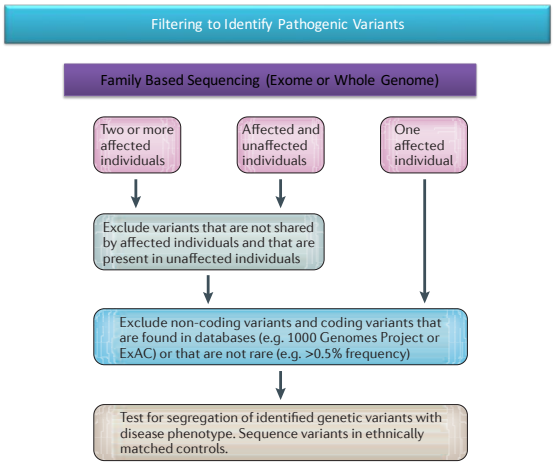
Why is Exome Sequencing Currently Used More Frequently than Whole Genome?

- Number 1 Reason - Cost!
  - An exome is ~1/3rd of the cost of a genome
- Easier interpretation of the data
  - Focuses on regions of the genome we understand best
- Ideal for the study of highly penetrant diseases
- Exome sequencing a stop-gap measure until the price of whole genome sequencing becomes more reasonable
- Already starting to see a switch
  - More studies performing whole genome sequencing



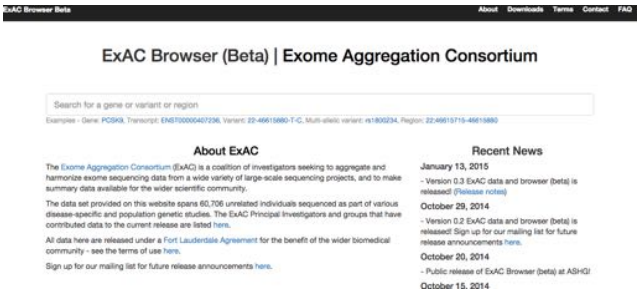
Identifying Casual Genes Using Exome/Whole Genome Sequencing

- Information on mode of inheritance may give clues to type of variants which you are looking for
  - Autosomal recessive phenotype for consanguineous pedigree
    - Homozygous variants
  - Autosomal recessive phenotype from outbreed pedigree
    - Compound heterozygous variants
    - Homozygous variants
  - Suspected *de Novo* event
    - Heterozygous variant –which is absent in parents
  - Autosomal Dominant
    - Heterozygous variant



Screening Databases

- Databases of Exome and genome Data
  - Contain individuals who have not been phenotyped
    - e.g. 1000 Genome data
  - Were ascertained because of disease phenotype
    - Coronary heart disease
  - Several databases available
    - 1000 Genomes
      - http://www.1000genomes.org/
    - Exome Variant Server
      - http://evs.gs.washington.edu/EVS/
    - ExAC
      - http://exac.broadinstitute.org/



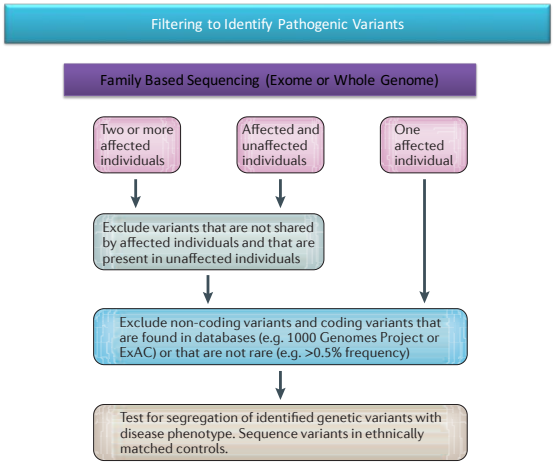
- Most extensive database with data on 60,706 individuals
- Provides break-downs by different ethnic groups
- Although contains individuals with disease, e.g. schizophrenia
  - No individuals with diagnosed early onset disease included
- Information on allele frequencies
  - Numbers of heterozyous and homozygous individuals for a variant
- Can evaluate read depth to determine if variant site of interest is covered with adequate read depth and in how many individuals

Contributing Projects

- 1000 Genomes
- Bulgarian Trios
- Finland-United States Investigation of NIDDM Genetics (FUSION)
- GoT2D
- Inflammatory Bowel Disease
- METabolic Syndrome In Men (METSIM)
- Jackson Heart Study
- Myocardial Infarction Genetics Consortium:
- Italian Atherosclerosis, Thrombosis, and Vascular Biology Working Group
- Ottawa Genomics Heart Study
- Pakistan Risk of Myocardial Infarction Study (PROMIS)
- Precocious Coronary Artery Disease Study (PROCARDIS)
- Registre Gironi del COR (REGICOR)
- NHLBI-GO Exome Sequencing Project (ESP)
- National Institute of Mental Health (NIMH) Controls
- SIGMA-T2D
- Sequencing in Suomi (SISu)
- Swedish Schizophrenia & Bipolar Studies
- T2D-GENES
- Schizophrenia Trios from Taiwan
- The Cancer Genome Atlas (TCGA)
- Tourette Syndrome Association International Consortium for Genomics (TSAICG)

Avoid 0% Cut-off When Filtering

- Databases do not consist of disease free individuals
- Mendelian variants may not be 100% penetrant
- For autosomal recessive traits
  - Carriers may be present in databases
- Frequency cut-offs should be disease specific
  - Unlikely pathogenic variants will have a frequency of >0.5%
    - There are rare exceptions
      - GJB2 35delG variant for nonsydromic hearing impairment



Test for Segregation with Disease Phenotype

- Is a variant ruled out if it does not completely segregate with the disease phenotype?
- What are reasons for incomplete segregation?
  - Variant was a false positive call
  - Not pathogenic
  - Locus heterogeneity within the pedigree
  - "Phenocopies" within the pedigree
  - Reduced penetrance
  - Incorrect pedigree structure
  - Sample swaps

Screening Control Individuals

- Is it not always necessary to screen controls given the large available databases
- Depends if the study population is well represented in the public databases
- For under represented populations variant frequencies should be examined in controls
  - Or individuals from the same populations who were ascertained for another phenotype

A Few Examples of Successful NGS Studies Using Filtering Approaches

Proof of Principal - Miller Syndrome

Nat Genet. 2010 January ; 42(1): 30–35. doi:10.1038/ng.499.

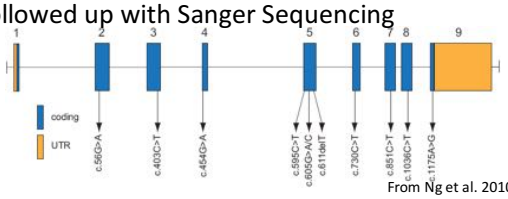
Exome sequencing identifies the cause of a Mendelian disorder

Sarah B. Ng[1,*], **Kati J. Buckingham**[2,*], **Choli Lee**[1], **Abigail W. Bigham**[2], **Holly K. Tabor**[2], **Karin M. Dent**[3], **Chad D. Huff**[4], **Paul T. Shannon**[5], **Ethylin Wang Jabs**[6,7], **Deborah A. Nickerson**[1], **Jay Shendure**[1,†], and **Michael J. Bamshad**[1,2,8,†]

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA [2]Department of Pediatrics, University of Washington, Seattle, Washington, USA [3]Department of Pediatrics, University of Utah, Salt Lake City, Utah, USA [4]Department of Human Genetics, University of Utah, Salt Lake City, Utah, USA [5]Institute of Systems Biology, Seattle WA, USA [6]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York, USA [7]Department of Pediatrics, Johns Hopkins University, Baltimore, Maryland [8]Seattle Children's Hospital, Seattle, Washington, USA

DHODH Gene Identified

- Four individuals with Miller syndrome underwent exome sequencing
  - From three families
- An additional three Miller families where followed up with Sanger Sequencing



From Ng et al. 2010

DHODH is composed of 9 exons that encode the untranslated region (orange) and protein coding region (blue). Arrows indicate the location of 11 different variants found in six Miller families

Kabuki Syndrome

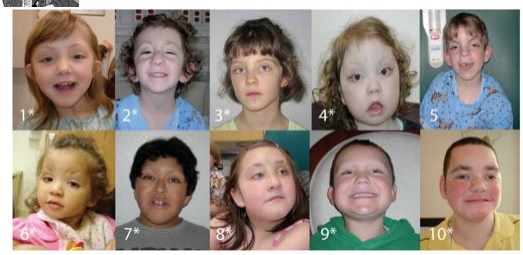
Nat Genet. 2010 September ; 42(9): 790–793. doi:10.1038/ng.646.

Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome

Sarah B. Ng[1,*], Abigail W. Bigham[2,*], Kati J. Buckingham[2], Mark C. Hannibal[2,3], Margaret McMillin[2], Heidi Gildersleeve[2], Anita E. Beck[2,3], Holly K. Tabor[2,3], Greg M. Cooper[1], Heather C. Mefford[2], Choli Lee[1], Emily H. Turner[1], Josh D. Smith[1], Mark J. Rieder[1], Koh-ichiro Yoshiura[4], Naomichi Matsumoto[5], Tohru Ohta[6], Norio Niikawa[6], Deborah A. Nickerson[1], Michael J. Bamshad[1,2,3,†], and Jay Shendure[1,†]

- Exome sequenced
  - 10 unrelated probands with Kabuki syndrome

Kabuki Syndrome



From Ng et al. 2010

- Dysmorphic, skeletal, immunologic & mild intellectual disabilities
- 1/30,000 to 1/50,0000
- Most cases simplex
  - Very few cases of parental transmission

Kabuki Syndrome

- Could have tackled problem by sequencing trios
  - Suspected to be *de Novo*
- Article describes how initial strategy failed since not all children have Kabuki syndrome due to variants in the same gene (locus heterogeneity)

Kabuki Syndrome

Table 1 Number of genes common to any subset of *x* affected individuals.

Subset analysis (any *x* of 10)	1	2	3	4	5	6	7	8	9	10
NS/SS/I	12,042	8,722	7,084	6,049	5,289	4,581	3,940	3,244	2,486	1,459
Not in dbSNP129 or 1000 Genomes	7,419	2,697	1,057	488	288	192	128	88	60	34
Not in control exomes	7,827	2,865	1,025	399	184	90	50	22	7	2
Not in either	6,935	2,227	701	242	104	44	16	6	3	1
Is loss-of-function (non-sense or frameshift indel)	753	49	7	3	2	2	1	0	0	0

The number of genes with at least one nonsynonymous variant (NS), splice-site acceptor or donor variants (SS) or coding indel (I) are listed under various filters. Variants were filtered by presence in dbSNP or 1000 Genomes (not in dbSNP129 or 1000 genomes) and control exomes (not in control exomes) or both (not in either); control exomes refer to those from 8 Hapmap[3], 4 FSS[3], 4 Miller[2] and 10 EGP samples. The number of genes found using the union of the intersection of *x* individuals is given.

Not the correct gene

From Ng et al. 2010

Kabuki Syndrome

- After failure to identify gene
- Clinicians ranked the patients from typical Kabuki syndrome to atypical
- Predicted functional assessment of variants
- Manual review of data highlighted previously unidentified nonsense variant in *MLL2* gene
  - Identified in the four highest ranked cases 1, 2, 3 & 4
  - Additional found in patients 6, 7 & 9

Kabuki Syndrome

- Additional Kabuki cases identified to have *MLL2* gene mutations
  - 26/43 cases
- 12/12 patients with available parents had *de Novo* variants

Discover of *de novo* events using Exome Sequence Data for Autism

ARTICLE

doi:10.1038/nature13908

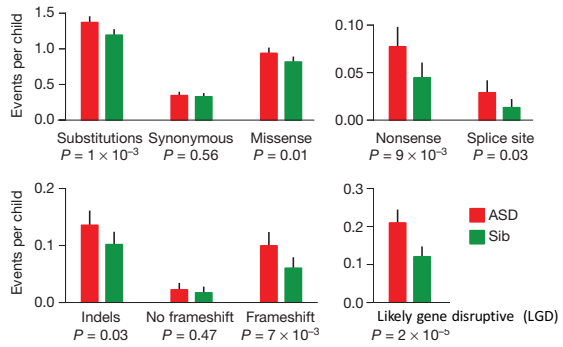
The contribution of *de novo* coding mutations to autism spectrum disorder

Ivan Iossifov[1*], Brian J. O'Roak[2,3*], Stephan J. Sanders[4,5*], Michael Ronemus[1*], Niklas Krumm[3], Dan Levy[1], Holly A. Stessman[2], Kali T. Witherspoon[6], Laura Vives[6], Karynne E. Patterson[6], Joshua D. Smith[6], Bryan Paeper[6], Deborah A. Nickerson[6], Jeanselle Dea[6], Shan Dong[7,8], Luis E. Gonzalez[7], Jeffrey D. Mandell[9], Shrikant M. Mane[10], Michael T. Murtha[7], Catherine A. Sullivan[7], Michael F. Walker[6], Zainulabedin Waqar[7], Liping Wei[7,11], A. Jeremy Willsey[9,12], Boris Yamrom[1], Yoon-ha Lee[1], Ewa Grabowska[1,10], Ertugrul Dalkic[1,11], Zihua Wang[1], Steven Marks[1], Peter Andrews[1], Anthony Leotta[1], Jude Kendall[1], Inessa Hakker[1], Julie Rosenbaum[1], Beicong Ma[1], Linda Rodgers[1], Jennifer Troge[1], Giuseppe Narzisi[1,10], Seungtai Yoon[1], Michael C. Schatz[1], Kenny Ye[12], W. Richard McCombie[1], Jay Shendure[2], Evan E. Eichler[2,13], Matthew W. State[4,5,7,14] & Michael Wigler[1]

Nature

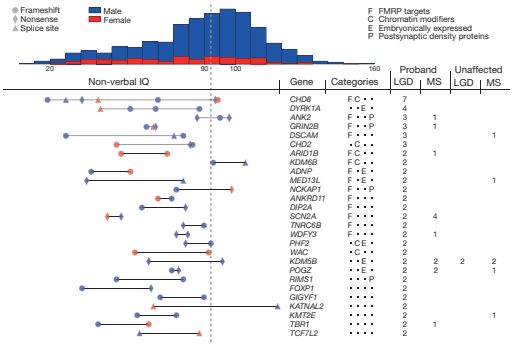
- Exome sequence data from 2,517 simplex families from the Simons Simplex Collection (SSC) was analyzed
  - Probands with autism spectrum disorder and their parents sequenced
  - 1,911 families also had sequence data on an unaffected sibling

Rates of de novo Events by Variant Type



From Iossitov et al. 2014

Genes with Recurrent Hits and Non-Verbal IQ



From Iossitov et al. 2014

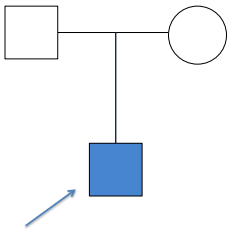
Short List of Genes Identified Using Exome Sequencing

Disease	Model	Sequencing scope	Reference
Miller syndrome	Autosomal recessive	Whole-genome, one family (two affected siblings and both parents)	18
Metachondromatosis	Autosomal dominant	Whole-genome, single proband	12
Miller syndrome	Autosomal recessive	Exome, four cases (two siblings, two other unrelated)	16
Schinzel–Giedion syndrome	Autosomal dominant	Exome, four unrelated cases	11
Fowler syndrome	Autosomal recessive	Exome, two unrelated cases	19
Kabuki syndrome	Autosomal dominant	Exome, 10 unrelated cases	13
Joubert syndrome 2	Autosomal recessive	Exomes of 2 individuals (mother and affected daughter)	15
Non-syndromic hearing loss (DFNB82)	Autosomal recessive	exome, single case	20
TARP syndrome	X-linked dominant	X chromosome exons, two unrelated carriers	21
Familial exudative vitreoretinopathy	Autosomal dominant	Linkage interval + 2 candidate genes, single proband	10
Cleriouzio-type poikiloderma with neutropenia	Autosomal recessive	Linkage interval, single case	14
Sensory/motor neuropathy with ataxia	Autosomal dominant	Linkage interval, proband and both parents	8
Non-syndromic deafness (DFNB79)	Autosomal recessive	Linkage interval, single case	17
Clinical diagnosis			
Congenital chloride-losing diarrhea	Autosomal recessive	Exome, single patient with suspected Bartter syndrome	23
Primary ciliary dyskinesia	Autosomal recessive	Exome, two siblings	
Molecular diagnosis			
Charcot–Marie–Tooth disease	Autosomal recessive	Whole-genome, single proband	22

How Many Variants will be Identified Using Filtering Approaches?

- Depends on
  - Mode of inheritance
  - Number of individuals sequenced
  - Type of sequence data
    - Exome
    - Whole genome

Pathogenic Variant Identification Using Data from a Trio or a Single individual



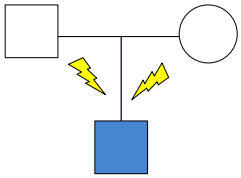
Is it possible to identify the causal gene using exome or genome sequencing?

How Many Variants will be Identified on Average for an Exome of Single Individual?

- Assume complete penetrance
  - Removing those variants with >0.5% which are in databases
    - ExAC
- Limiting analysis to protein coding mutations
  - Missense
  - Nonsense
  - Splice sites

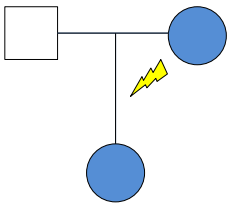
Recessive phenotypes

Rare compound heterozygous and homozygous variants



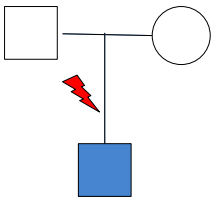
- 0-5 compound heterozygous SNVs
  - Data from parents must be available
    - To determine if variants are compound heterozygous
- 1-2 homozygous SNVs
  - A much larger number of homozygous sites will be observed when the child is an offspring of a consanguineous mating

Autosomal Dominant Mode of Inheritance



~200-300 variants

De novo mutations



0-3* coding non-synonymous mutation per individual
Exome data must be available from parents

*More may be observed due to false positive variant calls

Mode of Inheritance

- If Mode of Inheritance is unknown can try more than one model
- Filtering a single individual often leads to many variants that can reasonably be followed-up by
  - Testing for segregation
    - If family members are available
  - Functional studies
- NGS data from additional family members can be helpful in narrowing down the number of variants

Selection of Additional Family Members to Reduce the Number of Variants

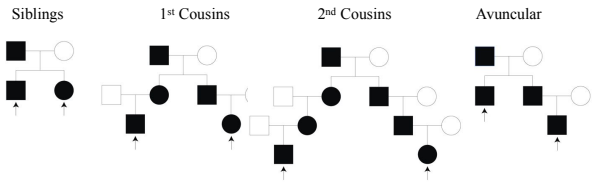
- Avoid performing NGS on unaffected family members
  - Variant frequencies can be obtained from databases
    - ExAC
  - Unaffected individuals can also be pathogenic variant carriers due to reduce penetrance
    - Can lead to exclusion of the causal variant

Selection of Additional Family Members to Reduce the Number of Variants

- Do not sequence parents of affected individuals
  - Except for the study
    - de novo events
  - Offspring will always inherit one parental allele
- More distantly related family members are the most informative
  - e.g. cousins

Selection of Additional Family Members to Reduce the Number of Variants

- Who to select can be guided by basic linkage principals
  - Those sets of individuals providing the highest "LOD" scores should be selected



Maximum LOD scores - Autosomal Dominant Pedigree

- If two affected Individuals from an a pedigree are sequenced what are the maximum LOD scores which can be obtained?

Autosomal Dominant Pedigrees	
	Maximum LOD scores
Parent-Child	0.00
Siblings	0.176
Avuncular	0.301
1st Cousins	0.602
2nd Cousins	1.201

Selecting Individuals for NGS

- Which individuals should be selected can be evaluated by simulations studies
  - SLINK/MSIM
  - Calculating maximum LOD score
- If genotype array data is available
  - GIGI- Pick (Chueng et al 2014 AJHG)
    - https://faculty.washington.edu/wijsman/progdists/gigi/software/GIGI-Pick/GIGI-Pick.html
  - ExomePick
    - http://genome.sph.umich.edu/wiki/ExomePicks

Reducing the Number of Variants For Follow-up

- Sequence multiple unrelated individuals with the same phenotype
  - Look for rare variants which are predicted to be functional that occur within the same gene
    - Due to allelic heterogeneity not all affected individuals will share the same variant

Reducing the Number of Variants For Follow-up

- If there is locus heterogeneity
  - There may be no single gene for which all affected individuals have a pathogenic variant
- For extreme locus heterogeneity
  - None of the individuals may share pathogenic variants within the same gene
    - Particularly if the sample size is small
  - Therefore not possible to narrow down results to a single gene

Selection of Individuals for NGS *de Novo*

- If it is of interest to detect *de Novo* variants
  - Child and both parents should be sequenced
- Would not expect to find *de Novo* variants in families with more than one affected individual
  - Can occur if *de novo* variant occurs in a founder that is passed to offspring

de Novo Events

- A single validated LGD *de novo* event is not sufficient to implement a gene in disease etiology
- Multiple LGD *de novo* events must be observed within a gene region
  - The number which must be observed to be significant is
    - Dependent on the sample size
    - The mutation rate within the gene region
- Significance can be evaluated
  - By comparing the *de novo* variant rate in controls
    - e.g. unaffected siblings of probands
      - Iossitov et al. 2014 Nature
  - Estimating the gene specific mutation rates
    - Neale et al. 2012 Nature

What are the Success Rates of NGS Studies?

Data Dependent

N Engl J Med. 2013 October 17; 369(16): 1502–1511. doi:10.1056/NEJMoa1306555.

Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders

Yaping Yang, Ph.D, Donna M. Muzny, M.Sc, Jeffrey G. Reid, Ph.D, Matthew N. Bainbridge, Ph.D, Alecia Willis, Ph.D, Patricia A. Ward, M.S, Alicia Braxton, M.S, Joke Beuten, Ph.D, Fan Xia, Ph.D, Zhiyv Niu, Ph.D, Matthew Hardison, Ph.D, Richard Person, Ph.D, Mir Reza Bekheirnia, M.D, Magalie S. Leduc, Ph.D, Amelia Kirby, M.D, Peter Pham, M.Sc, Jennifer Scull, Ph.D, Min Wang, Ph.D, Yan Ding, M.D, Sharon E. Plon, M.D., Ph.D, James R. Lupski, M.D., Ph.D, Arthur L. Beaudet, M.D, Richard A. Gibbs, Ph.D, and **Christine M. Eng, M.D**
Departments of Molecular and Human Genetics (Y.Y., A.W., P.A.W., A.B., J.B., F.X., Z.N., M.H., R.P., M.R.B., M.S.L., A.K., J.S., S.E.P., J.R.L., A.L.B., C.M.E.) and Pediatrics (S.E.P., J.R.L.) and the Human Genome Sequencing Center (D.M.M., J.G.R., M.N.B., P.P., M.W., Y.D., J.R.L., R.A.G.), Baylor College of Medicine, Houston.

In a clinical setting
~25% of Mendelian disoders solved

European Journal of Human Genetics (2012) 20, 490–497

www.nature.com/ejhg

REVIEW

Disease gene identification strategies for exome sequencing

Christian Gilissen*,[1], Alexander Hoischen[1], Han G Brunner[1] and Joris A Veltman[1]

Next generation sequencing can be used to search for Mendelian disease genes in an unbiased manner by sequencing the entire protein-coding sequence, known as the exome, or even the entire human genome. Identifying the pathogenic mutation amongst thousands to millions of genomic variants is a major challenge, and novel variant prioritization strategies are required. The choice of these strategies depends on the availability of well-phenotyped patients and family members, the mode of inheritance, the severity of the disease and its population frequency. In this review, we discuss the current strategies for Mendelian disease gene identification by exome resequencing. We conclude that exome strategies are successful and identify new Mendelian disease genes in approximately 60% of the projects. Improvements in bioinformatics as well as in sequencing technology will likely increase the success rate even further. Exome sequencing is likely to become the most commonly used tool for Mendelian disease gene identification for the coming years.
European Journal of Human Genetics (2012) **20,** 490–497; doi:10.1038/ejhg.2011.258; published online 18 January 2012

Keywords: Mendelian disease; gene identification; strategies; next generation sequencing; exome sequencing

- 24 families of which 14 lead to a novel gene identification
  - 58% success rate 95% CI 36%-78%
- Three families segregated known disease genes
  - Overall success rate of 71% 95 CI 51%-85%

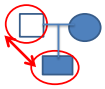
Using Genotype Array Data, Linkage Analysis and Homozygosity Mapping to Increase Success of Gene/Pathogenic Variant Identification

Benefits of Obtaining Genotype Array Data

- If multiple family members are available
  - Advantageous to perform SNP genotyping using one of the current microarrays
  - All informative individuals should be genotyped
- Can also aid in accessing the quality of DNA samples
  - Help to ensure NGS data will successfully be generated

Benefits of Obtain Genotype Array Data

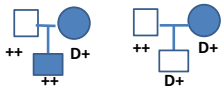
- Can be used to validate the pedigree structure
- Help to ensure that samples have not been swapped
- A variety of programs have been developed to provide probabilities on relationships within pedigrees
  - GRR
    - Abecasis et al. 2001 Bioinformatics
  - RELATIVE
    - Goring and Ott, 1997 Eur J Hum Genet
  - SIBPAIR
    - Ehm and Wagner 1998 AJHG
  - RELCHECK
    - Broman and Weber 1998 AJHG
  - RELPAIR
    - Boehnke and Cox 1997 AJHG
- Pedigree data can also be reconstructed from genotype data
  - PRIMUS
    - Staples et al. 2014 AJHG



Sample Swaps

Benefits of Linkage Analysis

- Can identify problems with pedigrees
  - Incorrect phenotype information
    - Affected individuals labeled and unaffected
    - Unaffected individuals labeled as affected
- Collaborators and Families can be re-contacted
  - To correct errors
- Errors which can not be resolved
  - Should be removed from analysis



Phenocopy Reduced Penetrance

Benefits of Linkage Analysis

European Journal of Human Genetics (2015) 23, 1207–1215

www.nature.com/ejhg

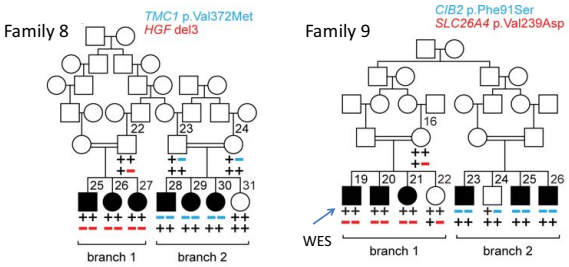
ARTICLE

Challenges and solutions for gene identification in the presence of familial locus heterogeneity

Atteeq U Rehman[1,12], Regie Lyn P Santos-Cortez[2,12], Meghan C Drummond[1], Mohsin Shahzad[3,4], Kwanghyuk Lee[5], Robert J Morell[1], Muhammad Ansar[2,5], Abid Jan[5], Xin Wang[5], Abdul Aziz[5], Saima Riazuddin[3,4], Joshua D Smith[6], Gao T Wang[2], Zubair M Ahmed[4], Khitab Gul[5], A Eliot Shearer[7], Richard J H Smith[7], Jay Shendure[6], Michael J Bamshad[8], Deborah A Nickerson[6], University of Washington Center for Mendelian Genomics[13], John Hinnant[8,14], Shaheen N Khan[3], Rachel A Fisher[9], Wasim Ahmad[5], Karen H Friderici[9,10], Sheikh Riazuddin[3,11], Thomas B Friedman[1], Ellen S Wilch[10] and Suzanne M Leal*[,2]

- Can be used to detect locus heterogeneity within pedigrees
- Linkage analysis/homozygous mapping can resolve which branches/individuals are segregating the same pathogenic variant
- Aid in selection of individuals for NGS

Inter-sibship Locus Heterogeneity



	MLOD	LOD	Region
Family 8	5.73	1.48	7q21.11-q21.3 (HGF)
Branch 1	3.59	3.59	7q21.11-q22.2 (HGF)
Branch 2	2.53	2.53	9q21.12-q21.13 (TMC1)

	MLOD	LOD	Region
Family 9	6.27	1.65	4q21.21
Branch 1	2.53	1.60	7q22.3-q31.1 (SLC26A4)
Branch 2	2.53	2.41	15q24.1-q26.1 (CIB2)

Intra-Familial Locus Heterogeneity is not Rare

- 15.3% of the families in a study of nonsydromic hearing impairment
  - 95% Confidence Interval (11.9 - 19.9%)
  - These families segregate at least one published HI gene

Classification based on variants identified and tests performed	Families with locus heterogeneity	Families without locus heterogeneity	Total
Variant identified via screening GJB2 (exon 2), CIB2 (p.Phe91Ser) or HGF (del 3)	19	98	117
Variants identified in other known HI genes			
Linkage analysis + Sanger sequencing	8	87	95
Linkage analysis + NGS	18	64	82
Total	45	249	294

Benefits of Linkage Analysis

- Unlike filtering approaches linkage analysis can incorporate reduced penetrance and phenocopies in the analysis
  - Allowing for success identification of a gene region
  - Even for pedigrees where there is phenocopies and/or reduced penetrance

Benefits of Linkage analysis

- Linkage analysis/homozygosity mapping can identify a small genomic region where the causal variant lies
  - Filtering can be applied within the linkage/homozygous region
    - Greatly reduce the number of variants which need to be followed-up
      - Test identified variant(s) for segregation within pedigree
    - This is particularly true for whole genome data where
      - Even a small genomic regions can contain hundreds of rare variants

Benefits of Linkage analysis

- Information on haplotypes can be used to select pedigree member(s) for NGS
- Selecting those individuals with the smallest possible haplotype
- Individuals which are phenocopies can be excluded from selection for NGS

Benefits of Linkage Analysis

- Examining haplotypes can also give clues if two or more families are segregating the same disease gene or variant
  - Overlapping haplotype which are not the same
    - Potentially disease phenotype due to the same disease gene
    - But unlikely due to the same pathogenic variant
  - Disease haplotype is identical – although not of the same length
    - Likely the two families are segregating the same causal variant

Benefits of Linkage Analysis

- If multiple families linked to same locus are available
  - Sequencing individual(s) from more than one family can aid in gene identification
  - When they share variants in the same gene
- Provide additional evidence of genes involvement in disease etiology
  - Compared to a single family

Selection of Family Members for NGS Autosomal Dominant Pedigrees

- Sequence ≥2 individuals from each family
  - Two individuals are often sufficient
  - Distantly related as possible
    - i.e. from two different branches
  - Choose those affected individuals within the pedigree which segregate the same haplotypes
    - Helps to exclude individuals who are potentially phenocopies
      - i.e. have the phenotype due to different causal variants
  - Select ≥2 individuals with smallest overlapping haplotypes
    - Reduces the size of the interval in which the pathogenic variant lies

Selection of Family Members for NGS Autosomal Recessive Pedigrees

- A single individual can be selected
  - With the smallest homozygous region
  - With overlapping haplotypes which span the smallest region
    - If compound heterozygous
- Sequencing additional affected family members may aid in gene identification
  - Can greatly increase cost
  - Usually not necessary

Selection of Family Members for NGS Autosomal Recessive Pedigrees

- For compound heterozygous individuals
  - Variants identified within a gene region
    - Can be sequenced in parents, e.g. Sanger
      - To determine if compound heterozygous
        - » Or lay on the same haplotype
    - Parents can also undergo NGS
      - Currently not as cost effective

Prioritize Families for Study Using NGS

- Prioritize families with multiple affected individuals
  - 1.) Significant linkage LOD ≥3.3
  - 2.) Suggestive linkage 3.3 <LOD ≤ 2.0
  - 3.) Weak linkage 2.0<LOD≥1.2
  - 4.) Small families with only 1-2 affected individuals LOD <1.2
- Single affected individuals can also be studied
  - 5.) Trios
    - Highest priority if looking for *de Novo* events
  - 6.) Single affected individuals with family history

Data Quality Control

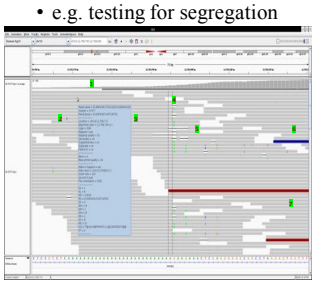
- Extremely important when testing for association for complex traits
- Also important for Mendelian traits
  - Many false positive variant sites if data is not cleaned
- Data cleaning for exome sequence data is data specific
  - e.g. remove variant sites that
    - Fail Variant Quality Score Recalibration (VQSR)
    - Fail HWE $p<5 \times 10^{-8}$
  - e.g. remove variants with
    - a read depth of <10x
    - GQ score <20

Data Quality Control

- Proceed with caution it is possible to remove true variant sites including causal variants when filtering data
  - May wish to loosen stringency of filtering/cleaning if unable to identify causal variant
- Working with "dirty data" can lead to many false positive variant sites/genotype

Integrative Genomic Viewer (IGV)

- Can be used to investigate variant calls
  - To determine if a variant is a false positive call
  - Before following-up of variants
    - e.g. testing for segregation



Robinson et al Nat Biotechnology 2011
http://www.broadinstitute.org/igv/

Software to Perform Variant Annotation and Filtering

- FamAnn
  - Yao et al. 2014 Bioinformatics
  - https://sites.google.com/site/famannotation/documentation
- Gemini
  - Paili et al. 2013 PLoS Comput Biol
  - https://gemini.readthedocs.org/en/latest/
- Jannovar
  - Jaeger et al. 2014 Hum Mutation
  - http://jannovar.readthedocs.org/en/master/install.html

Software to Perform Variant Annotation and Filtering

- VAAST
  - Hu et al. 2013 Genet Epidemiol
  - http://www.hufflab.org/software/vaast/
- Variant Mendelian Tools
  - http://varianttools.sourceforge.net/VMT/VMT
- VARank
  - Geoffroy et al. 2015 PeerJ
  - http://www.lbgi.fr/VaRank/#requirements

Data Analysis Using Filtering An Additional Note

- If multiple samples are analyzed
- Multisample calling should be used to identify variants
- A multisample VCF file should be analyzed
- If variant is calling is performed on single samples
  - No information on read depth, etc for variant sites where there is no alternative allele

Steps After Variant Identification

- Additional families with same phenotype and putatively pathogenic variants in the same gene
  - Help support involvement of the gene in disease etiology
- Form collaborations to identify additional families
- Matchmaker Exchange
  - http://www.matchmakerexchange.org/
  - Can help to identify investigators who have families with the same phenotype and variants within the same gene

Expression and Functional Studies

- Can aid in implicating a variant/gene in disease etiology
  - Particularly important if the variant/gene is found in a single family
    - Identified variant may be in LD with functional mutation
- Brings about a better understanding of disease etiology and the role the identified gene plays

Reasons for Failure of NGS

- Insufficient samples for gene identification
  - e.g. single individual with no additional family members
- Locus heterogeneity
- Phenocopies, misdiagnosed or mislabeled individuals within a pedigree
- Sample swaps
- Variant not captured
  - Can be potentially be resolved by whole genome sequencing
- Inadequate depth of coverage to call variant
- Indel/Copy number variants
  - Difficult to accurately call
    - Sensitivity can be low

Steps When NGS Does not Reveal Putatively Causal Variant

- When linkage region is known
  - Examine the region to determine which genes have not
    - been captured
    - missing data due to poor read depth coverage or
    - Variants have not been called
  - Examine regions with IGV
    - Follow-up with Sanger Sequencing if
      - Missing regions
      - Poor quality variants
- If exome sequencing was perform proceed to whole genome sequencing
  - The causal variant could lie outside of the coding region

An Example of Using Linkage Analysis and NGS to Identify Pathogenic Variants

Nat Genet. ; 44(8): 916–921. doi:10.1038/ng.2348.

TGFB2 loss of function mutations cause familial thoracic aortic aneurysms and acute aortic dissections associated with mild systemic features of the Marfan syndrome

Catherine Boileau[1,2,3,4,14,15], Dong-Chuan Guo[5,14], Nadine Hanna[1,2,3], Ellen S. Regalado[5], Delphine Detaint[1,2,6], Limin Gong[5], Mathilde Varret[1], Siddharth Prakash[5,12], Alexander H. Li[5], Hyacintha d'Indy[1,3], Alan C. Braverman[7], Bernard Grandchamp[2,8], Callie S. Kwartler[5], Laurent Gouya[2,3,4], Regie Lyn P. Santos-Cortez[9], Marianne Abifadel[1], Suzanne M. Leal[9], Christine Muti[2], Jay Shendure[10], Marie-Sylvie Gross[1], Mark J. Rieder[10], Alec Vahanian[6,8], Deborah A. Nickerson[10], Jean Baptiste Michel[1], National Heart Lung and Blood Institute (NHLBI) Go Exome Sequencing Project[11], Guillaume Jondeau[1,2,6,8,14], and **Dianna M. Milewicz[5,12,13,14,15]**

Clinical Features of Two TAA Families with *TGFB2* Variants

	Pedigree ID	
Phenotypic Information	**TAA288**	**MS239**
Number affected pedigree members	7 TAA	6 TAA
Age at diagnosis (years)	5 – 41 (median 32)	27 – 53 (median 36)
Surgical Intervention	1 TAA	1 TAA, 1 MVP*
Arterial tortuosity	No	Yes
Other cardiac disease	2 MVP	1 MVP*
Lens dislocation	No	1/6 minor
Flat cornea	Unknown	2/6
Pectus deformity	3/7 mild	2/6 definite, 1/6 mild
Scoliosis	2/7 definite, 1/7 mild	1/6 mild
Flat feet	5/7	6/7
Joint hyperflexibility	5/7	3/6
High-arched palate	6/7	3/6
Striae atrophicae	4/7	4/6

*Mitral valve prolapse

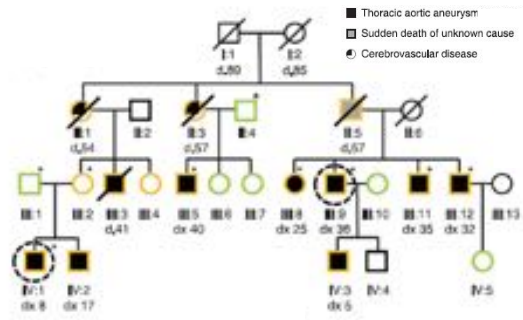
Analysis of Two TAA Pedigrees with Mild Systemic Features of Marfan Syndrome

- Whole genome linkage analysis performed
  - Pedigree TAA288
    - Affymetrix 50k SNP Array
      - Samples from 9 informative pedigree members genotyped
  - Pedigree MS239
    - 1,056 microsatellite markers (deCode array)
      - Samples from 14 informative pedigree members genotyped

Analysis of Two TAA Pedigrees with Mild Systemic Features of Marfan Syndrome

- Multipoint and two-point analysis performed
  - Autosomal dominant mode of inheritance
    - 90% Penetrance
    - Disease allele frequency 0.0001
- Computer Software
  - Pedcheck
  - Merlin, Superlink & SimWalk2
- Both pedigrees mapped to 1q41
  - TAA288
    - Multipoint LOD score 2.4
  - MS239
    - Multipoint LOD score 1.6

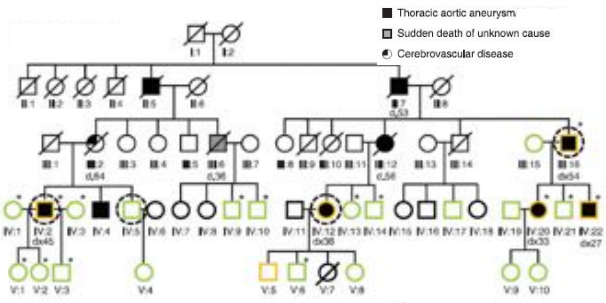
Pedigree TAA288



*Individuals underwent whole genome genotyping
Circled individuals underwent exome sequencing

Multipoint LOD score 2.4 at 1q41

Pedigree MS239



*Individuals genotyped for whole genome scan
Circled individuals underwent exome sequencing

Multipoint LOD score 1.6 at 1q41

Exome Sequencing

- Family TAA288
  - Two affected individuals selected for exome sequencing
    - 16 variants shared by affected pedigree members
- Family MA239
  - Three affected and one unaffected individuals selected for exome sequencing
    - 5 variants shared by affected pedigree members
- In family TAA288 two variants within linkage region and in family MA239 only one variant
- Both families only had *TGFB2* variants in common

Identification of *TGFB2* variants

- Family TAA288
  - 5-bp deletion c.1021_1025del-TACAA in exon 6 which leads to a premature stop codon p.Try341Cysfs*25
    - Two-point LOD score 3.3
- Family MS239
  - Stop-gain variant in exon 4 p.Cys229*
  - Two-point LOD score 4.4
- Neither variant found in ExAC database
  - 61,486 “control” individuals
- In both pedigrees all affected individuals were heterozygotes for respective variants
- In both pedigrees there was reduced penetrance

Additional Screening of *TGFB2*

- French probands from a Marfan referral clinic
  - 62 familial cases
  - 74 sporadic cases
- USA probands with thoracic aortic disease
  - 214 familial cases
  - 57 sporadic cases
- In the French familial probands two variants were found
  - p.Glu102*
  - Frameshift duplication c.873_888dup leading to p.Asn297*
- Both probands had TAAD
- Neither variant was observed in ExAC
  - 60,706 “controls” individuals

Association Analysis for Mendelian Traits

Suzanne M. Leal
Center for Statistical Genetics
Baylor College of Medicine
sleal@bcm.edu

Association Analysis of Rare Variants

- Analysis of single rare variants are very poorly powered
- Many methods have been developed specifically to test for rare variant associations
  - To overcome the low power of testing for associations with individual rare variants
- Rare variant association methods are frequently referred to as
  - Aggregate
  - Burden
  - Collapsing

Association Analysis of Rare Variants

- Generally only Rare variants are analyzed, e.g. MAF< 0.5%
- Which are
  - Missense variants
  - Stop loss, gain variant
  - Spice site variants

Association Analysis - Mendelian

- If pedigree data are available
  - Linkage analysis and filtering approaches should be used for data analysis
- When only the proband is available for study
  - Or a limited number of family members
    - e.g. unaffected family members, a single affected sibling
- If the proband has a family history or
- It is suspected that the disease is due to a *de novo* variant
  - No parental data is available
- Association analysis can aid in finding genes which harbor pathogenic variants

Association Analysis - Mendelian

- Rare variant association analysis can be used in these situations
- Affected probands are compared to control individuals
- Care must be used in selecting controls
- Sequencing conditions should be the same for both cases and controls
  - Read depth
  - Capture array, etc

Controls

- If convenience controls are used
  - BAM files should be obtained and variants called for both cases and controls together
- Although frequencies for individual variants can be obtained from databases such as ExAC
  - These frequencies/counts should not be used to perform rare variant association analysis
    - Can lead to an increase in type I

Data Quality Control

- Unlike for filtering approached stringent data quality control should be performed
  - Removing variant sites which
    - Fail variant quality score recalibration [(VQSR) GATK]
    - High rates of missing variant calls, e.g. >10%
    - Fail Hardy Weinberg equilibrium , e.g. $p < 10^{-7}$
  - Removing variant genotypes with
    - Low read depth e.g < 10X
    - Low GQ scores e.g. < 20
- The quality control is data specific
- A balance must be met
  - between removal of data & false positive calls

Sample Size & Power

- For complex traits extremely large sample sizes are necessary
  - Tens of thousands of individuals
    - Due to low effect sizes of disease susceptibility variants
- For Mendelian diseases many fewer cases are necessary to detect an association
  - For some studies <50 cases may be necessary
  - To increase power large numbers of controls can be used
    - Although there is a diminishing return when the ratio of control to case is > 3:1

Influences on Power

- Mode of Inheritance
- Locus heterogeneity
  - Increasing locus heterogeneity leads to a decrease in power
- Allelic heterogeneity
  - Will not impact power
    - Unless benign variants are included in association test.

Types of Aggregate Analyses

- Frequency cut offs used to determine which variants to include in the analysis
  - Rare Variants (e.g. ≤1% frequency)
  - Rare and low (1-5%) frequency variants
- Maximization approaches
- Tests developed to detection associations when variants effects are bidirectional e.g. protective and detrimental
- Incorporate weights based upon – frequency or functionality

Misclassification

- When performing aggregate analysis
  - Misclassification of variants within a region can reduce power
- Exclusion of causal variants
  - Variants which are causal are erroneously not included in the analysis
- Inclusion of non-causal variants
  - Variants which are non-causal are included in the analysis

Caveats

- For exome data natural regions to aggregate rare variants are
  - Genes
  - Genes within pathways
- Analysis of genome sequence data outside of exonic regions is problematic
  - Unlikely a sliding window approach will work
    - Size of window unknown and will differ across the genome
  - A better understanding functionality outside the coding regions is necessary
    - Predicted functional regions, enhancer regions, transcription factors, DNase I hypersensitivity sites, etc.

A Few Rare Variant Association Tests

- Combined Multivariate Collapsing (CMC)
  - Li and Leal AJHG 2008
- Burden of Rare Variants (BRV)
  - Auer, Wang, Leal Genet Epidemiol 2013
- Weighted Sum Statistic (WSS)
  - Madsen and Browning PloS Genet 2009
- Kernel based adaptive cluster (KBAC)
  - Liu and Leal PloS Genet 2010
- Variable Threshold (VT)
  - Price et al. AJHG 2010
- Sequence Kernel Association Test (SKAT)
  - Wu et al. AJHG 2011

Fixed Effect Tests

Random Effect Test

A Few Rare Variant Association Tests

- Combined Multivariate Collapsing (CMC)
  - Li and Leal AJHG 2008
- Burden of Rare Variants (BRV)
  - Auer, Wang, Leal Genet Epidemiol 2013
- Weighted Sum Statistic (WSS)
  - Madsen and Browning PloS Genet 2009
- Kernel based adaptive cluster (KBAC)
  - Liu and Leal PloS Genet 2010
- Variable Threshold (VT)
  - Price et al. AJHG 2010

Fixed Effect Tests

Methods to Detect Rare Variant Associations Using Variant Frequency Cut-offs

- Combined multivariate & collapsing (CMC)
  - Li & Leal, AJHG 2008
- Collapsing scheme which can be used in the regression framework
  - Can use various criteria to determine which variants to collapse into subgroups
    - Variant frequency
    - Predicted functionality

CMC

- Define covariate X_j for individual j as

$$X_j = \begin{cases} 1 & \text{if rare variants present} \\ 0 & \text{otherwise} \end{cases}$$

- Compute Fisher exact test for 2x2 table

Number of cases for which one or more rare variants are observed e.g. nonsynonymous variants freq. ≤1%

	X=1	X=0
cases		
controls		

Number of cases without a rare variant

Number of controls for which one or more rare variants are observed

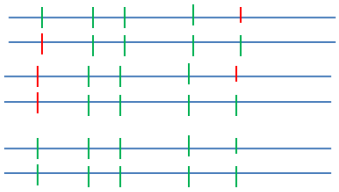
Number of controls without a rare variant

Can also use same coding in a regression framework

CMC

- Example of coding used in regression framework:
  - Binary coding $X_j = \begin{cases} 1 & \text{if rare variants present} \\ 0 & \text{otherwise} \end{cases}$

Gene region with 5 variant sites



Individual	Coding
1	1
2	1
3	0

Rare Variant Sites

Green bars: Major allele is observed in the study subject

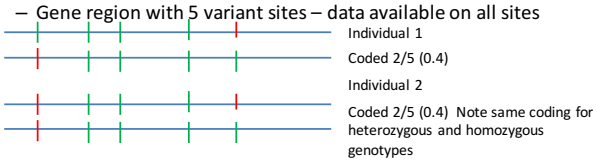
Red bars: Minor allele has been observed

Methods to Detect Rare Variant Associations Using Variant Frequency Cut-offs

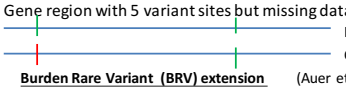
- Gene-or Region-based Analysis of Variants of Intermediate and Low frequency (GRANVIL)
  - Aggregate number of rare variants used as regressors in a linear regression model
  - Can be extended to case-control studies
    - Morris & Zeggini 2010 Genet. Epidemiol
  - Test also referred to as MZ

GRANVIL

- Example of coding used in regression framework
  - Gene region with 5 variant sites – data available on all sites



- Gene region with 5 variant sites but missing data on three variant sites



Burden Rare Variant (BRV) extension (Auer et al. 2013 Genet Epidemiol)
Individual 1: Coded 2
Individual 2: Coded 3
Individual 3: Coded 1

Methods to Detect Rare Variant Associations Weighted Approaches

- Group-wise association test for rare variants using the Weighted Sum Statistic (WSS)
  - Variants are weighted inversely by their frequency in controls (rare variants are up-weighted)
    - Madsen & Browning, PLoS Genet 2009
- Kernel based adaptive cluster (KBAC)
  - Adaptive weighting based on multilocus genotype
    - Liu & Leal, PLoS Genet 2010

Methods to Detect Rare Variant Associations Maximization Approaches

- Variable Threshold (VT) method
  - Uses variable allele frequency thresholds and maximizes the test statistic
  - Also can incorporate weighting based on functional information
    - Price et al. AJHG 2010
- RareCover
  - Maximizes the test statistic over all variants with a region using a greedy heuristic algorithm
    - Bhatia et al. 2010 PLoS Computational Biology

Significance Level for Rare Variant Association Tests

- For exome data where individual genes are analyzed usually a Bonferroni correction for the number of genes tested is used.
  - There is very little to no linkage disequilibrium between genes
- A Bonferroni correction for testing 20,000 genes is often used as the significance level cut-off
  - 2.5×10^{-6}

Rare Variant Aggregate Methods

- Ideally should be performed in a regression analysis framework
  - Logistic
  - Linear regression
- Almost all methods have been extended to be implemented within a regression framework
  - Can control for covariates which are potential confounders
  - Age
  - Sex
  - Population substructure/admixture

Rare Variant Aggregate Methods

- If the proportion of cases and controls sampled from each populations is different
  - Can occur due to
    - Disease frequency is different between populations
    - Sloppy sampling
- Population substructure\admixture can cause detection of differences in variant frequencies within a gene which is due to sampling and not disease status
  - False positive findings can be increased

Rare Variant Aggregate Methods

- Population substructure\admixture is often a confounder for genetic studies
  - A particular problem for rare variants
- Currently Principal Components Analysis (PCA) or Multidimensionality Scaling (MDS) is used to control for population substructure\admixture
  - For both studies of common & rare variants

Related Individuals

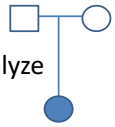
- Remove related individuals from the analysis
  - Only retain one member of a related pair/group in the analysis
- Perform analysis using mixed models
- Ignoring that related individuals are included in the analysis can increase type I error

Software to Perform Rare Variant Association Testing using NGS Data

- PLINK/SEQ
  - Developed by Shaun Purcell
    - https://atgu.mgh.harvard.edu/plinkseq/tutorial.shtml
- Variant Association Tools (VAT)
  - Reference Wang, Peng & Leal, 2014
    - http://varianttools.sourceforge.net/Association/HomePage

Testing for Associations using Trio Data

- Trio data are often sequenced to detect *de novo* events.
- However, transmitted as well *de novo* events can be analyzed
- The transmission disequilibrium test (TDT) is a natural choice to analyze trio data
- The TDT design can also be used to analyze Mendelian traits



Controlling for Population Admixture and Substructure Using the Trio Design

- The trio (two parents and an affected child) approach was developed to control for population substructure and admixture
  - Falk and Rubinstein 1987 Ann Hum Genet
  - Many additional trio methods have been described
- The Transmission Disequilibrium Test (TDT) is currently the mostly widely used trio method
  - Spielman et al. 1993 AJHG

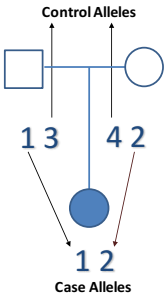
TDT

<u>Case Alleles</u>

•Transmitted parental alleles 1 and 2

<u>Control Alleles</u>

•Non-transmitted parental alleles 3 and 4



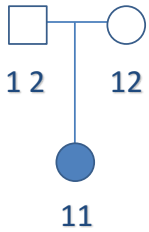
*Phenotype information from parents is not used in the analysis

TDT

Not Transmitted

Transmitted

	1 allele	2 allele
1 allele	a	b
2 allele	c	d



(McNemar's Test) $\frac{(b-c)^2}{b+c} \xrightarrow{d} \chi_1^2$

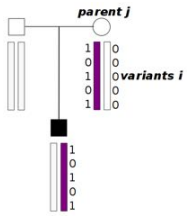
Only transmission events from heterozygous parents are informative, i.e. quadrants b & c

TDT-Aggregate Analysis

- The TDT was extended to incorporate rare variant association methods
  - Combined Multivariate Collapsing (CMC)
    - RV-TDT-CMC
  - Burden of Rare Variants (BRV)
    - RV-TDT-BRV
  - Weighted Sum Statistic (WSS)
    - RV-TDT-WSS
  - Variable Threshold (VT)
    - RV-TDT-VT

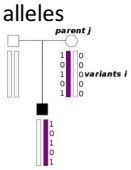
RV-TDT-CMC & RV-TDT-WSS

- RV-TDT-CMC
  - Collapse transmissions within a region
    - For parent j
      - b=0 and c=1
- RV-TDT-BRV
  - Aggregate transmissions within a region
    - For parent j
      - b=0 and c=3



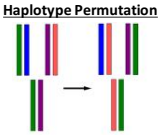
RV-TDT-WSS & RV-TDT-VT

- RV-TDT-WSS
  - Aggregate transmissions within a region weighted by the frequency of non-transmitted ("control") alleles
    - For parent j
      - b=0 and c= $\sum \omega_i$
- RV-TDT-VT
  - Maximizes the test statistic over minor allele frequencies using either CMC or BRV coding
    - For parent j
      - b=0 and c=1 or 3



Evaluating Significance

- Analytical
  - χ_1^2 (one-sided test)
  - CMC method only
- Empirical
  - Haplotype permutation
    - Shuffle parental haplotypes
  - All methods



What are the Necessary Sample Sizes for the Trio Design for a Mendelian Trait

- Assuming No Locus Heterogeneity
- Power 0.80

	Number of Trios	
Mode of Inheritance	Alpha	
	0.05	2.5×10^{-6}
Autosomal Recessive	4	15
Autosomal Dominant	13	59

References
He et al. 2014 AJHG
Krumm et al. 2015 Nature Genetic

RV-TDT Software
http://bioinformatics.org/rv-tdt/

The Collapsed Haplotype Pattern (CHP) Method for Performing Linkage Analysis Using Sequence Data

Suzanne M. Leal
sleal@bcm.edu
Center for Statistical Genetic
Baylor College of Medicine
https://www.bcm.edu/research/labs/center-for-statistical-genetics

Performing Linkage Analysis Using Exome and Genome Sequence Data

- As cost of performing sequencing falls
  - DNA samples from all informative pedigree members can undergoing sequencing
- Several studies have generated exome and genome sequence data on all informative family members
  - T2D-Genes (type 2 diabetes study)
    - Genome sequence data on 20 Mexican families (1,043 Individuals)
- Caveat performing linkage analysis on individual rare variants is not a powerful approach

Collapsed Haplotype Pattern (CHP) Method

- Motivated by rare variant aggregate association methods
  - Analysis of regions, usually genes
    - Instead of analyzing individual rare variants
- Rare variant aggregate associations methods are more powerfully than analyzing individual variants

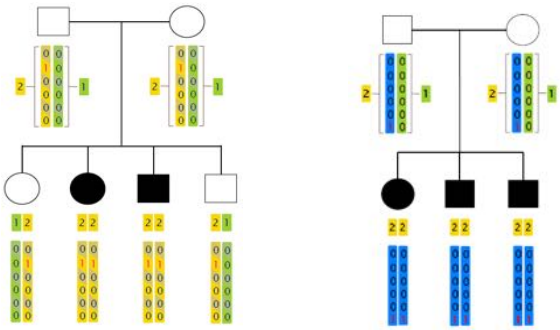
CHP Method

- Lander-Green algorithm is used for genetic phasing and reconstruction of haplotypes
- Missing genotypes are imputed
  - Conditional on family members genotypes and
  - Population allele frequencies
    - Obtained from founders if sample size is sufficiently large or
    - Frequencies are obtained from databases (e.g. ExAC)

CHP Method

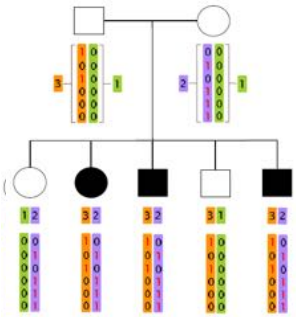
- For each pedigree variants on a regional haplotypes, e.g. LD blocks
  - Are assigned a single numeric value e.g.
    - 0 no minor alleles
    - 1 at least one minor allele
- Each regional haplotype within a family is uniquely represented

Example-CHP



- These two pedigrees both have rare variants in the same gene region
- Although they segregate a different rare variant and haplotype
  - The same coding can be used without a lost of information

Example-CHP



- A unique coding is provided for each haplotype
  - To avoid lost of information

CHP method

- Each pedigree is analyzed separately
  - Using allele frequencies that are correct for the haplotypes segregating in the pedigree
- Parametric LOD score results are summed across families by gene region
  - At the same Θ value, e.g. Θ=*0.0*

Evaluation of the CHP Method

- Data was generated for four nonsyndromic hearing impairment genes
  - Autosomal recessive mode of inheritance
    - *GJB2*, *SLC26A4*
  - Autosomal dominant mode of inheritance
    - *MYO7A* and *MYH9*
- All variants were generated based upon their frequency in European-Americans
  - Using data from Exome Sequencing Project
- Causal status of variants obtain from NCBI-ClinVar

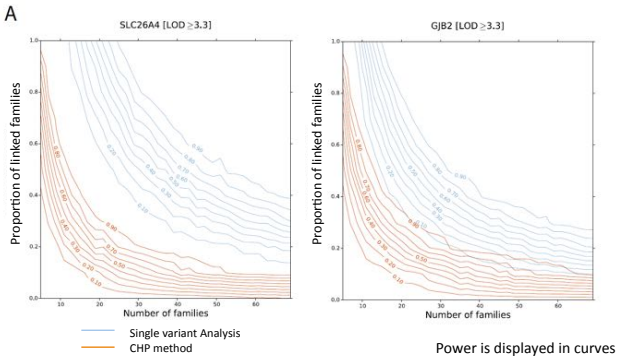
Evaluation of the CHP Method

- Families were generated with 3-8 children
  - Based on the number of children per family in the United States in 2012, rescaled to sum to 100%
    - 3 children: 69.34%
    - 4 children: 20.52%,
    - 5 children: 6.84
    - 6 children: 2.28%
    - 7 children 0.76%,
    - 8 children 0.26%
- RarePedSim was used to generate the pedigree data
  - http://bioinformatics.org/simped/rare/

Evaluation of the CHP Method

- Pedigrees were generated with varying degrees of locus heterogeneity
  - e.g. 20% of families linked to *GJB2* and 80% to *SLC26A4*
- Families with ≥2 affected children were "ascertained"
- Variants with a MAF≤0.01% analyzed
- Power was evaluated using 500 replicates
  - For a genome-wide α level of < 0.05
    - LOD > 3.3
    - HLOD > 3.6
- The CHP method was compared to single variant analysis

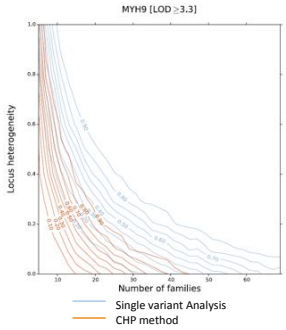
Results Autosomal Recessive Model Genes *SLC26A4* and *GJB2*



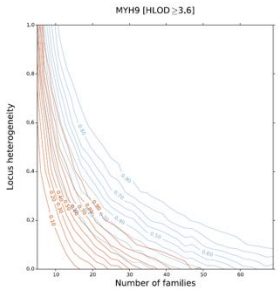
Power is displayed in curves

Results Autosomal Dominant Model Gene *MYH9*

Analysis under locus homogeneity



Analysis under locus heterogeneity



Power is displayed in curves

Linkage Analysis

- Unlike filtering approaches, linkage can provides statistical evidence of a variant's "involvement" in trait etiology
  - Caution should be used, variant may only be in LD with the pathological variant
- Because linkage incorporates mode of inheritance information and penetrance models
  - Less likely than filtering to exclude causal variants in the presents of phenocopies and/or reduced penetrance

References

Wang et al. 2015 EJHG

Ott, Wang, Leal 2015 NRG

Software

CHP incorporated in SEQLinkage

http://bioinformatics.org/seqlink

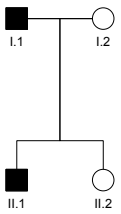
Evaluating Power Using Simulation Studies

Suzanne M. Leal
Center for Statistical Genetics
Baylor College of Medicine
sleal@bcm.edu

Simulation

- Simulation studies are used in many situations
  - Predict traffic jams
  - Flow from volcano eruptions, etc.
- For genetic studies simulation can be used for a variety of situations
  - Estimate the power to detect linkage for a given data set
  - Estimate empirical p-values
  - Compare various analysis methods

Example Generating Genotype Data



Generating Genotypes for a Pedigree

- A marker the following allele frequencies will be generated
  - 1=0.4
  - 2=0.1
  - 3=0.45
  - 4=0.05

Generating Genotypes for a Pedigree

- A random number generator is used
  - Random numbers between 0 and 1 are generated
- The numbers are generated according to a uniform distribution
  - Each number has equal probability of occurring
- Random number generators are in actuality pseudo-random number generators
  - If a simulation is carried our using the same starting seed the same results will be obtained.

Generating Genotypes

- If the random number selected is between <0.4
  - Then the 1 allele is chosen
- If the random number selected is between ≥0.4 and < 0.5
  - Then the 2 allele is chosen
- Etc

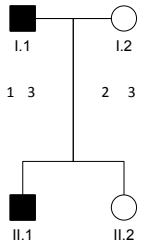
```
0          0.4      0.5                          0.95    1
|________1_______|__2__|____________3____________|_4__|
```

Generating Genotypes

- Since each individual needs two alleles to construct their genotype - two random numbers are generated
- Father
  - 0.84
    - Assign a 3 allele
  - 0.31
    - Assign a 1 allele
- Mother
  - 0.44
    - Assign a 2 allele
  - 0.63
    - Assign a 3 allele

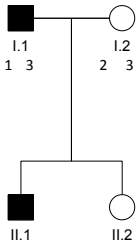
```
0              0.4    0.5                    0.95   1
|_______1______|__2__|__________3_________|_4__|
```

Generating Parental Genotypes



Generating Offspring Genotypes

Should the random number generator be used to generate two more genotypes for the children?



Generating Offspring Genotypes

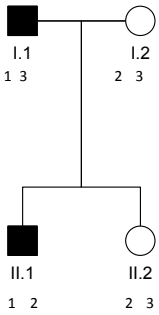
- No the alleles must segregate from the parents.
- It must be determined which of the two parental alleles each offspring "inherits"
  - With 50% probability

```
0  child receives first parental allele  0.5  child receives second parental allele   1
|_____________________________|_____________________________|
```

Random Numbers are Generated

- For child II.1
  - Random # 0.21
    - From father first allele
      - obtains a 1 allele
  - Random # 0.11
    - From mother first allele
      - obtains a 2 allele
- For child II.2
  - Random # 0.76
    - From father second allele
      - obtains a 3 allele
  - Random # 0.31
    - From mother first allele
      - obtains a 2 allele

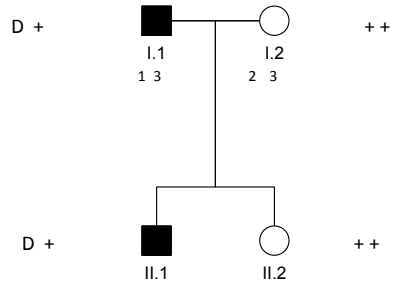
Generating Offspring Genotypes



Generating Genotypes for Pedigrees

- The genotypes in the previous example were generated unconditional (unlinked) to the disease phenotype
- Next marker will be generated linked to the disease locus
- Assumption
  - The disease phenotype is autosomal dominant
    - No phenocopies
    - No reduced penetrance
  - The marker and the disease locus are linked
    - Θ=0.04

Generating Genotypes Conditional on Disease Locus



Generating Genotypes Conditional on the Disease Locus

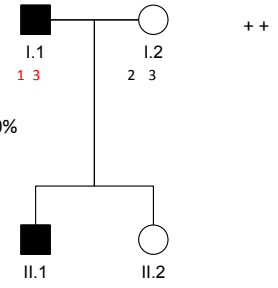
- Need to Generate offspring genotypes conditional on parental genotypes and underlying disease genotype
- Since the pedigree is phase unknown
  - Do not know grandparental genotypes
- Have to determine phase
- Assumption the disease the disease and marker loci are in linkage equilibrium
  - Each phase has 50% probability

Generating Genotypes Conditional on Disease Locus

Phase I
D +
1 3

Phase II
D +
3 1

Each with probability 50%



Father's Phase is Determined

- A random number generator is used to determine the phase for the father

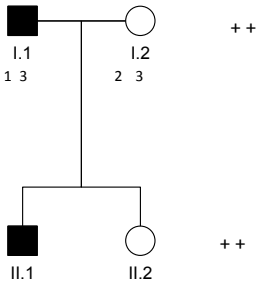
0 Phase I for the father 0.5 Phase II for the father 1

- The random # generated is 0.76
  - The father's phase is II

Generating Genotypes Conditional on the Disease Locus

D +

Phase II
D +
3 1



Generating the Offspring Genotypes

- The genotypes must be assigned conditional on the disease genotypes and whether or not a recombination event has occurred
- Whether or not a recombination has occurred will be determined by generating a random number

Determining Offspring Genotypes

- The first child II.1 is affected
- He either receives from his father
  - A 3 allele with probability 1-Θ
    - For this example 1-0.04
  - Or a 1 allele with probability Θ
    - For this example 0.04
- The second child II.2 is unaffected
- She either receives from her father
  - A 1 allele with probability 1-Θ
  - Or with probability Θ receives a 3 allele

0 Θ (0.04) 1

Determining Offspring Genotypes

- Child II.1 (affected)
  - Random # 0.88 is generated
    - He is assigned the 3 allele from his father
- Child II.2 (unaffected)
- Random # 0.01 is generated
  - She is assigned the 3 allele from her father

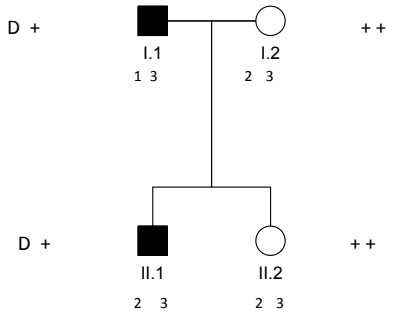
0 Θ (0.04) 1

Determining the Offspring Genotypes

- The mother provides no linkage information
  - Each child can be assigned either a 2 or a 3 allele
    - With probability 0.5
- Two Random numbers are generated
  - For Child II.1
    - The random # 0.32 is generated
      - The 2 allele is assigned from the mother
  - For Child II.2
    - The random # 0.43 is generated
      - The 2 allele is assigned from the mother

0 child receives first parental allele 0.5 child receives second parental allele 1

Generating Genotypes Conditional on the Disease Locus



Generating Haplotype Data

- Instead of generating and assigning individuals alleles
  - Haplotypes are generated
- When haplotypes are generated unconditional on disease phenotype or quantitative trait
- Based upon haplotype frequencies two haplotypes are assigned to each individual in the parental generation
  - Random numbers are used to determine which two haplotypes are assigned

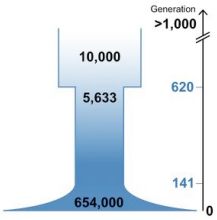
Generating Haplotype Data

- For each offspring recombination events between the two parental haplotypes are determined by genetic maps
  - Positions of recombination events are determined by random numbers
- One paternal and one maternal "new" haplotypes is assigned to the offspring from each of their parents
  - Each of the two parental haplotype has equal probability of beginning assigned to the offspring
    - Which haplotypes are assigned is determined by random numbers

Generating Sequence Data for Pedigrees

- Haplotype data can be generated using population demographic models

- Data generated on 16,568 genes
- Simulating variant data using reference sequence data a European population demographic model
  - Gazave et al. 2013
  - Haplotype pool generated for each gene
    - Each pool contains 1,308,000 haplotypes



Generating Sequence Data for Pedigrees

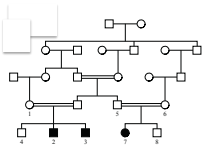
- Variant data frequencies can also be used from databases
  - e.g. ExAC
- Caution should be used that a sufficient large sample sizes is used to obtain variant frequencies
  - Otherwise very rare variants will be under-represented
    - Too few singletons, doubletons etc.
- Determine which variants are pathogenic using clinical databases
  - e.g. ClinVar

Generating Sequence data for Pedigrees

- To then generate the variant data conditional on the disease phenotypes
  - To generate data under the alternative
    - A penetrance model is used
    - The penetrance model should mimic the mode of inheritance in the pedigree
      - Autosomal dominant, Autosomal Recessive or X-linked
      - Fully penetrant or
      - Reduced penetrance and phenocopies

Generating Sequence data for Pedigrees

- Variants can be generated unconditional on the disease phenotype
  - To generate data under the null
- Variants are only generated for pedigree members which are available for study



First Step -Generating Pedigree Data

- Empirical p-values
  - Data is generated under the null hypothesis
    - Markers and disease are unlinked
  - Not necessary to know the underlying genetic model
    - Can be used for Mendelian and non-Mendelian traits
- Power
  - Marker(s) are generated linked ($\Theta<0.5$) to the disease locus
  - Must know underlying genetic model
    - For pedigree data can only be used for Mendelian traits

Second Step Analyzing Data Power, ELOD & EMLOD

- Must know underlying disease model
- Simulated data is analyzed using the same model as was used for data generation
  - Allele frequencies (marker and disease)
  - Penetrances
- Can evaluate the informativeness of pedigree data using several measures
  - Power
  - ELOD (Expected LOD Score)
  - EMLOD (Expected Maximum LOD Score)
  - Maximum LOD score

Second Step Analyzing Data Power, ELOD & EMLOD

- Power
  - The proportion of replicates where the null hypothesis of no linkage is rejected based upon a LOD score criterion (e.g. LOD score $\geq$3.3)
- ELOD
  - Is estimated by the average LOD score across at the recombination fraction the data was generated at across all replicates
- EMLOD
  - Is estimated by the average of the maximum LOD score across all replicates
- Maximum LOD score
  - Largest LOD score observed for all replicates
    - Only valid for fully penetrant disease without phenocopies

How Many Replicates Should be Generated?

- Depends on how accurate of an estimate is necessary.
- When estimating empirical p-values will be dependent on how small of a p-value is being estimated.
  - The smaller the p-values the more replicates
- For example if the p-value is in the range of 0.00001 need to generate many more than 1,000 replicates
  - Since by chanced under the null my never observe a p-value of $\leq$0.00001
- If only interested in estimating if an empirical p-value is <0.05
  - ~5,000 replicates may be sufficient

How Many Replicates Should be Generated?

- Power
- Usually need fewer replicates
- ~500 replicates
- But is some instances there can be great variability and many more replicates are necessary for accurate power estimates

Exercises

- Simulate pedigree data using SLINK
  - Generate marker data
- Analyze data with using MSIM
  - Perform parametric two-point linkage analysis
- Simulate rare variant data using RareSimPed
  - Simulates sequence data
    - Generates a VCF file
- Analysis data using SEQLinkage
  - Performs the Collapsed Haplotype Pattern (CHP) method

Simulation Programs

- SLINK
  - Generates genotype and haplotype data conditional or unconditional on affection status or quantitative trait
  - Generates phenotype data
    - Quantitative
    - Qualitative
  - Large and complex pedigree structures
  - Small number of marker loci ~$\leq$ 7 can be generated
- SIMULATE
  - Generates genotype data unconditional on affection status
  - Large and complex pedigree structures
  - Large number of marker loci can be generated

- SIMLINK
  - Generates genotype data conditional and unconditional on affection status or quantitative trait
  - Large pedigree structure
  - Small number of marker loci can be generated
    - One disease and one marker locus
  - Must modify the program in order to have it supply generated pedigree structures
- MERLIN
  - Generates genotypes data unconditional on affection status or quantitative trait
  - Large and complex pedigree structures
  - Large number of marker loci can be generated
- SOLAR
  - Generate genotypes unconditional on affection status or quantitative trait
  - Large and complex pedigree structures
  - Large number of marker loci can be generated

- GASP
  - Generates quantitative and qualitative phenotype data
    - Gene-gene and gene-environmental interaction
  - Generates genotype data conditional on generated phenotype data
  - Limited in size and structure of pedigrees
    - At most three generations
  - Can generate up to 400 marker loci
- SimPed
  - Pedigrees of virtually any size or complexity
  - Generation of >10,000 diallelic or multiallelic marker loci
    - Generates data for the autosomes and X chromosome
      - Haplotype data
        - » Markers in linkage disequilibrium
      - Genotype data
        - » Markers in linkage equilibrium

- SIMLA
  - Generates qualitative phenotype data
    - Gene-gene and gene-environmental interaction
    - Assigns affection status to pedigree members
  - Limited in pedigree structures that can be generated
    - user cannot provide pedigree structure
  - Large number of marker loci can be generated
  - Can also generate sequence data
- RareSimPed
  - Generates sequence data for Mendelian and Complex traits (qualitative and quantitative) regardless of pedigree structure
    - Using population based frequencies or demographic models
  - Generates genotype data conditional and unconditional on the phenotype
  - Generates phenotype data conditional on the generated genotype data

Functional Studies for Two Autosomal Recessive Nonsyndromic Hearing Impairment Genes:

lysyl-tRNA synthetase (*KARS*) & Adenylate cyclase 1 (*ADCY1*)

Suzanne M. Leal
sleal@bcm.edu
Baylor College of Medicine

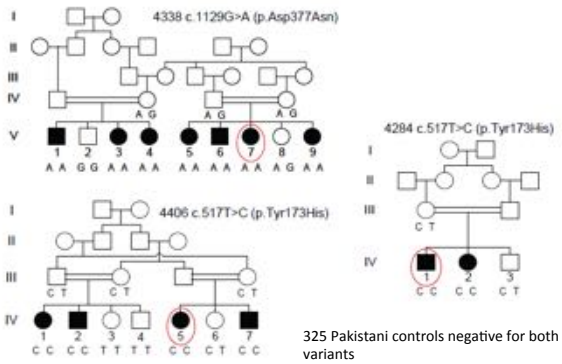
REPORT American Journal of Human Genetics

Mutations in *KARS*, Encoding Lysyl-tRNA Synthetase, Cause Autosomal-Recessive Nonsyndromic Hearing Impairment DFNB89

Regie Lyn P. Santos-Cortez,[1,8] Kwanghyuk Lee,[1,8] Zahid Azeem,[2,3] Patrick J. Antonellis,[4,5] Lana M. Pollock,[4,6] Saadullah Khan,[2] Irfanullah,[2] Paula B. Andrade-Elizondo,[1] Ilene Chiu,[1] Mark D. Adams,[6] Sulman Basit,[2] Joshua D. Smith,[7] University of Washington Center for Mendelian Genomics, Deborah A. Nickerson,[7] Brian M. McDermott, Jr.,[4,5,6] Wasim Ahmad,[2] and Suzanne M. Leal[1,*]

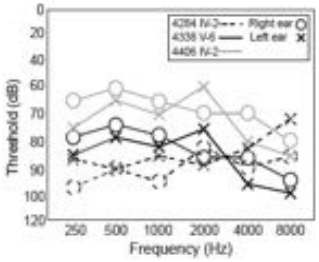
Describes the identification of *KARS* for nonsyndromic hearing impairment and functional studies which were performed

KARS Variants Segregate with HI in DFNB89 Families



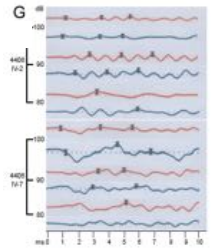
325 Pakistani controls negative for both variants

DFNB89 Hearing Impairment



Bilateral symmetric moderate-to-profound hearing impairment across all frequencies

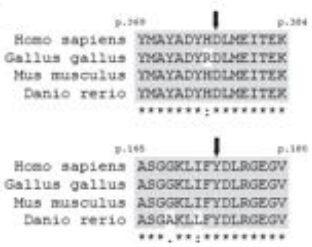
For individual V-6 all motor and sensory action potentials were normal



(+) ABR waveforms
Absent OAE
Normal EMG-NCV

Ruling out auditory neuropathy and supporting occurrence of cochlear pathology, particularly of the outer hair cells

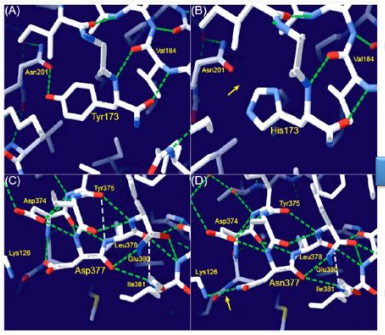
KARS Variants at Conserved Residues



p.Asp377 identical in 165 species
p.Tyr173 conserved in 162 species

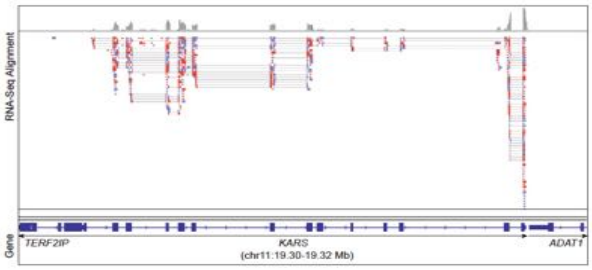
From primates to fungi

KARS Variants Predicted to Lower Catalytic Activity



Loss of β-2 strand
↓
Interfere with tRNA binding
↓
Decreased catalytic activity
↑
Affect tetramer binding
↑
Loss of α-helix 9

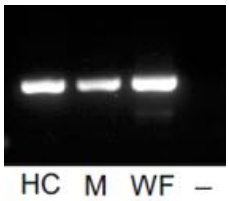
KARS is Expressed in Chicken Hair Cells



Expression of KARS in purified chicken hair cells was detected by RNA-seq

KARS is Expressed in Zebrafish and Mouse Hair Cells and Maculae

Zebrafish

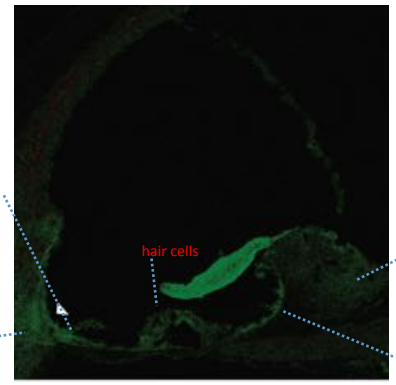


HC = Hair Cell
M = Macula
WF = Whole Fish
(-) = Control

Mouse

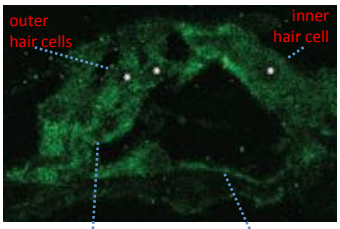
KARS Localized to Cochlear Duct

Mouse

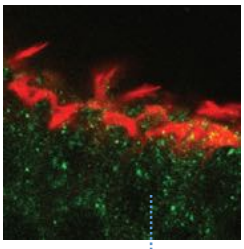


KARS Localized to Cochlear and Vestibular Hair Cells in the Mouse

Cochlea



Vestibule



* = hair cell nuclei

Labeling with KARS polyclonal antibody (green) & phalloidin (red)

Conclusions KARS

- *KARS* mutations define both a novel NSHI gene and a novel phenotype for *KARS*
- *KARS* is expressed in inner ears and hair cells of chicken, zebrafish and mouse
- KARS strongly localizes to otic fibrocytes, hair cells and cochlear supporting cells

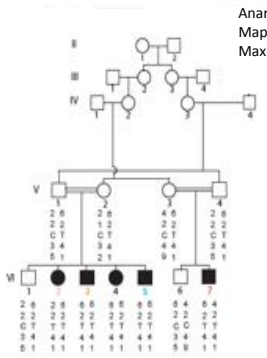
Human Molecular Genetics, 2014, Vol. 23, No. 12 **3289–3298**
doi:10.1093/hmg/ddu042
Advance Access published on January 29, 2014

Adenylate cyclase 1 (*ADCY1*) mutations cause recessive hearing impairment in humans and defects in hair cell function and hearing in zebrafish

Regie Lyn P. Santos-Cortez[1], Kwanghyuk Lee[1], Arnaud P. Giese[3,4], Muhammad Ansar[1,5], Muhammad Amin-Ud-Din[6], Kira Rehn[4], Xin Wang[1], Abdul Aziz[5], Ilene Chiu[2], Raja Hussain Ali[5], Joshua D. Smith[7], University of Washington Center for Mendelian Genomics, Jay Shendure[7], Michael Bamshad[7], Deborah A. Nickerson[7], Zubair M. Ahmed[3], Wasim Ahmad[5], Saima Riazuddin[4] and Suzanne M. Leal[1,*]

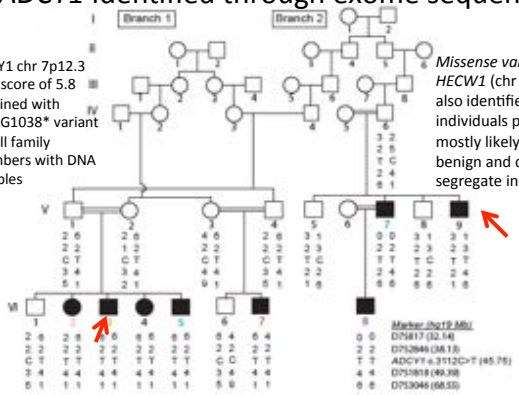
Describes the identification of *ADCY1* for nonsyndromic hearing impairment and functional studies which were performed

Family 4009 - DFNB44 Mapped to 7p14.1-q11.22



Anar et al. 2004
Mapped with STR Markers
Maximum Multipoint LOD Score 5.0

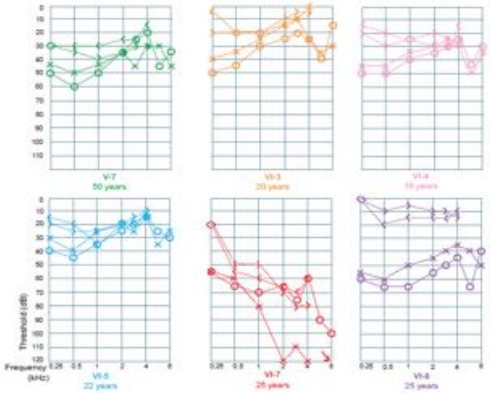
Nonsense Variant c.311C>T p.ARG1038 in ADCY1* Identified through exome sequencing



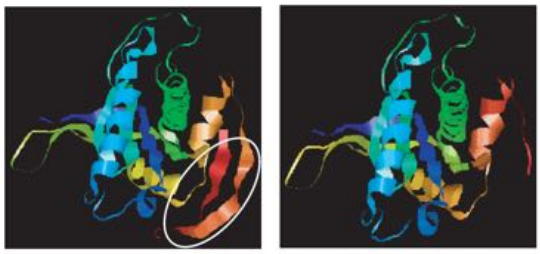
ADCY1 chr 7p12.3 LOD score of 5.8 obtained with p.ARG1038* variant for all family members with DNA samples

Missense variant in HECW1 (chr 7p14-13) also identified in both individuals predicted mostly likely to be benign and did not segregate in pedigree

Bilateral symmetric mild-to-moderate mixed hearing impairment in 5 of 6 family members of 4009



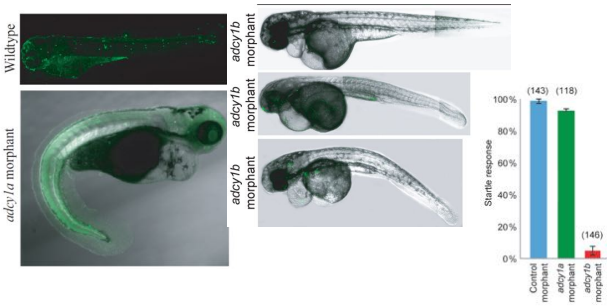
Predicted loss of two terminal beta-sheets due to ADCY1 p.Arg1038*



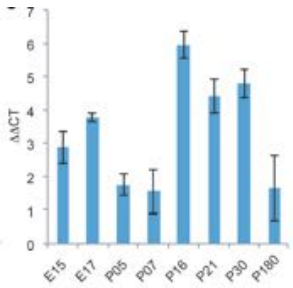
Wildtype ADCY1 ADCY1 p.Arg1038*

Predicted to cause loss of 82 amino acids from the cytoplasmic carboxyl tail and include highly conserved residues of the C_2 domain

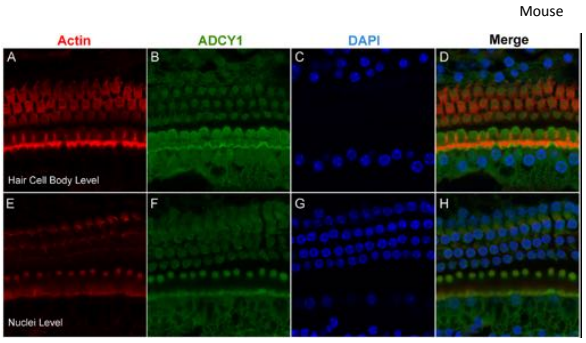
Failure of FM1-43 dye uptake and lack of startle response in *adcy1b* but not *adcy1a* morphant zebrafish



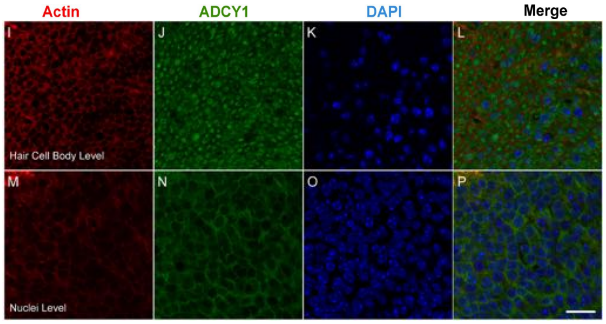
ADCY1 is expressed in mouse inner ear at various developmental stages, with highest expression at P16



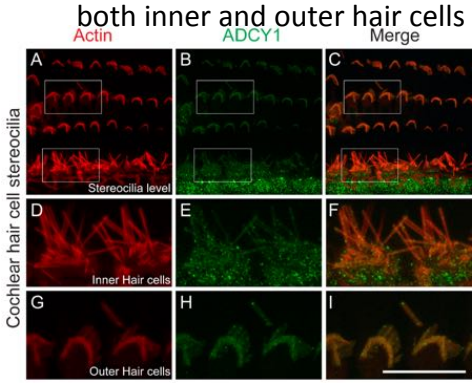
ADCY1 is localized to cochlear outer and inner hair cell bodies and nuclei with weaker staining in supporting cells



ADCY1 localizes to the vestibular hair cell bodies and also in supporting cells but no nuclei labeling was observed



ADCY1 is localized to the adult rat inner hair cell bodies and along the length of the stereocilia of both inner and outer hair cells



Conclusions-*ADCY1*

- *ADCY1* p.Arg1038* causes bilateral mild-to-moderate mixed hearing impairment in humans
- This mutation is predicted to decrease enzymatic efficiency and localization of ADCY1 to stereocilia
- ADCY1 has an evolutionarily conserved role in hearing

Conclusions –*ADCY1*

- *ADCY1* is expressed throughout inner ear development and maturation
- ADCY1 is localized to cytoplasm of inner ear hair cells and supporting cells and also to nuclei and stereocilia of cochlear hair cell
- Zebrafish *adcy1b* morphants had hair cell dysfunction and gross hearing impairment

Conclusions- overall

- With fast pace of NGS gene discovery, functional studies can be the rate-limiting step to publication
- Design of functional study depends on hypothesis for gene's role in target organ
- For inner ear, expression and localization within various cell types in rodent inner ear is usually performed as initial study
- If hair cells are involved zebrafish morphants can be studied

Conclusions - Overall Expression and Functional Studies

- Can aid in implicating a variant/gene in disease etiology
  - Particularly important if the variant/gene is found in a single family
    - Identified variant may be in LD with functional mutation
- Brings about a better understanding of disease etiology and the role the identified gene plays

Variant Annotation

Michael Nothnagel, michael.nothnagel@uni-koeln.de, 2015

Outline

- Forms of variant annotation
- Databases for annotation
- Software for annotation
- Notes of caution

Forms of variant annotation

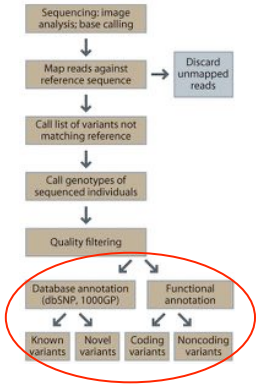
Technical information
- Sequencing instrument
- Quality metrics for filtering

Database annotation
- What is already known about the variant?
- Retrieval of information from databases

Functional annotation
- Prediction of functionality, pathogenicity etc. of variant
- Inference based on various

Multiple layers annotation
- Overlap with others sorts of genomic information, e.g. expression levels, transcription factors

Bioinformatic workflow



Jobling, et al. (2014)

Database annotation

- A wealth of information is already available from public databases for many variants
  - RefSeq numbers and other identifiers
  - Population frequencies (both global and population-specific)
  - Type of variant for coding regions (missense, stop, etc.)
  - Implication in human Mendelian diseases
  - Implication in human inherited diseases
  - Implication in human diseases and traits (GWAS?)
  - Literature
- Database annotation involves scripted or web-based analyses for
  - querying of public databases
  - storing retrieved information

Potential effects of small-scale gene mutations

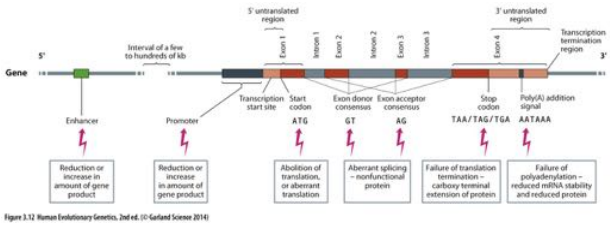


Figure 3.12 Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

Jobling, et al. (2014)

(Human) genomic databases (I)

www.ncbi.nlm.nih.gov/SNP

dbSNP, dbVar

- SNPs, indels, SV
- Global and population-specific frequencies

www.1000genomes.org

1000 Genomes

- SNPs, indels, SV
- Global and population-specific frequencies

www.hapmap.org

HapMap

- SNPs, indels
- Global and population-specific frequencies

www.nlgenome.nl

GoNL

- SNPs, indels
- Dutch population frequencies

(Human) genomic databases (II)

www.ebi.ac.uk/dgva

Genomic Variants Archive

- SNPs, indels

genome.ucsc.edu

UCSC Genome Browser

- the reference sequence and working draft assemblies for a large collection of genomes
- portal to ENCODE data at UCSC (2003-12) and to the Neanderthal project

www.ensembl.org

Ensemble

- European Bioinformatics Institute
- SNPs, indels

exac.broadinstitute.org

ExAC

- Exome Aggregation Consortium
- Exome data (including variants) for >60,000 unrelated individuals

Human disease databases (I)

www.ncbi.nlm.nih.gov/omim

OMIM

- Online Mendelian Inheritance in Men
- Catalog of human genes/disorders/traits
- Focus on molecular relationship between genetic variation and phenotypic expression

www.hgmd.cf.ac.uk

HGMD

- Human Gene Mutation Database
- Collate known (published) gene lesions responsible for human inherited disease

www.ebi.ac.uk/gwas

GWAS catalog

- QC-ed, manually curated, literature-derived collection of all published GWAS assaying >100,000 SNPs; all SNPs with $p<10^{-5}$

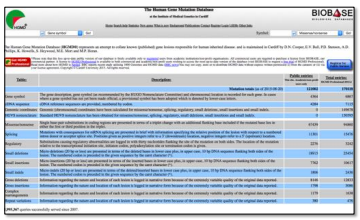
www.ncbi.nlm.nih.gov/pubmed

PubMed

- >24 million citations for biomedical literature from MEDLINE, journals, and online books

HGMD

Cooper & Krawczak (1993), Cooper, et al. (1998), Krawczak, et al. (2000), Stenson, et al. (2003), Stenson, et al. (2014)



Manually curated collection of published gene lesions responsible for human inherited disease; includes the first example of all mutations causing or associated with human inherited disease plus functional studies

Free access to mutations included >=3 years ago for registered academic users; otherwise professional version for up-to-date access

Human disease databases (II)

www.ncbi.nlm.nih.gov/clinvar

ClinVar

- Public archive of reports on relationships among human variations and phenotypes
- Supporting evidence and submitter visible
- Focus in medical genetics

cancergenome.nih.gov

The Cancer Genome Atlas (TCGA)

- Public catalog of genomic changes in tumors
- Search, download, and analysis of data sets generated by TCGA

cancer.sanger.ac.uk/cosmic

COSMIC

- Public catalog of genomic changes in tumors
- Search, download, and analysis of data sets generated by TCGA

There are (many) more databases.

Number of genic SNPs per genome

TABLE 3.2: ESTIMATED NUMBERS OF POTENTIAL CODING AND LOSS-OF-FUNCTION VARIANTS WITHIN PROTEIN-CODING GENES

Type of variant	Average number per genome
Synonymous	10,572–12,126[a]
Nonsynonymous (missense)	9966–10,819[a]
Generation of stop codon (nonsense)	26.2 (5.2)[b]
Splice site variant	11.2 (1.9)[b]
Small indel causing frameshift	38.2 (9.2)[b]
Large deletion	28.3 (6.2)[b]
Total number of LoF variants	103.9 (22.5)[b]

Data from the low-coverage dataset of 1000 Genomes Project Consortium (2010) *Nature* 467, 1061.
[a] Interquartile range of the number of variants per individual across the CEU, CHB, JPT, and YRI HapMap samples (see Box 3.6 for the three-letter abbreviations of the populations).
[b] Average number of variants in the CEU sample, with average number in homozygous state in parentheses; from MacArthur DG et al. (2012) *Science* 335, 823.
LoF, loss of function.

Table 3.2 Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

Jobling, et al. (2014)

Motivation for functional annotation

- Each individual carries multiple deleterious mutations:
  - On average 2% of all individuals carry a missense mutation in any given gene. (Andrews, et al., 2013, Trends Immunol)
  - For a given disease, multiple missense mutations will by chance be present in those genes that seemingly relate to the disease in an affected individual.
- Do these missense mutations
  - actually alter gene function or, more precise,
  - actually cause the disease/phenotype at hand?
- Functional assays or model organism experiments are
  - too costly for all observed missense mutations
  - too time-consuming
  - may raise ethical issues
- Computational inference may address some of these issues.
  - Trade-off between sensitivity and specificity

Approaches to functional annotation

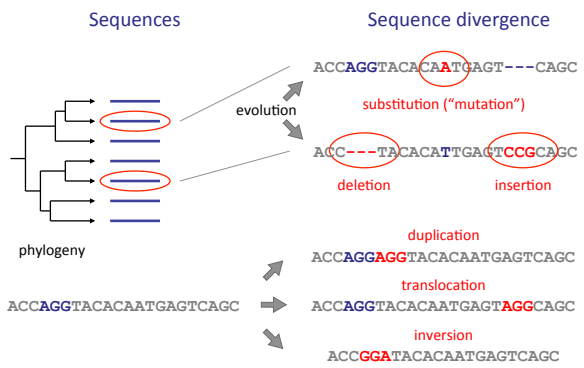
Variant is located in …

Coding sequence	Non-coding sequence

• Focus on proteins
• Primarily applied with exome data
• Possible analyses
  • Protein sequence conservation
  • Protein features
  • Gene relationships, interaction
  • Pathway analysis
  • …
	• Focus on regulation
	• Primarily applied with whole-genome data
	• Possible analyses:
	  • Methylation, epigenetic alterations
	  • Transcription-factor (TF) binding sites
	  • RNA interference
	  • Expression
	  • …

- Selection, reduced nucleotide diversity, etc.

Homology: sequence divergence by evolution



Physicochemical properties of proteins

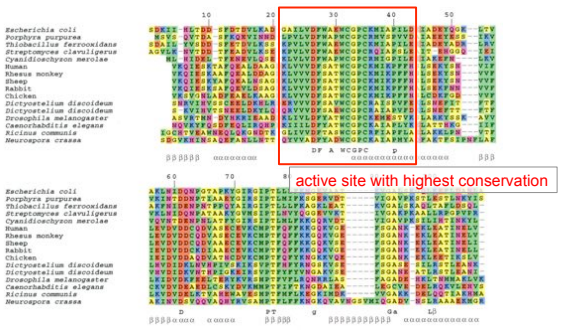
- Chemical composition and biophysical characteristics
  - Solubility in water, polarity, charge,
  - Cyclic, sulfur-containing
  - Molecular volume
- Structure
  - Primary: amino acid chain
  - Secondary: chain folding (mainly a and b helixes)
  - Tertiary: spatial arrangement through folding and coiling, sometimes including molecular chaperones
  - Quaternary: complex of two more polypeptide chains
- Sequence conservation
- …

⇨ Fixation of mutations with dissimilar amino acids is rare.
(Grantham, 1974, Science)

Cozzone (2002) Encyclopedia of Life Sciences; www.els.net

Conservation through selection

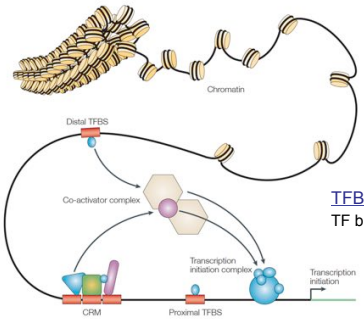
Thioredoxin of *E. coli* and 15 homologs



Lesk (2014)

Transcriptional regulation

Places of action for transcription factors (TF)

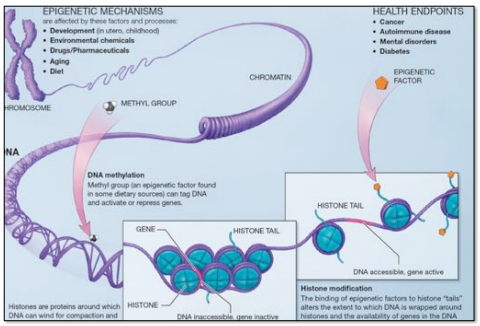


TFBS:
TF binding site

CRM:
cis-regulatory module (set of TFs)

Wasserman & Sandelin (2004) Nat Rev Genet

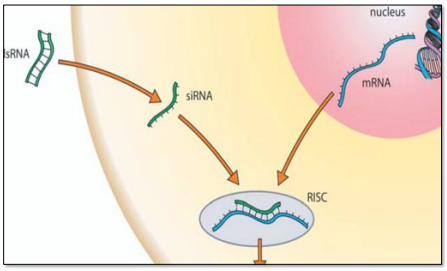
Epigenetic changes



http://commonfund.nih.gov/epigenomics/

RNA interference

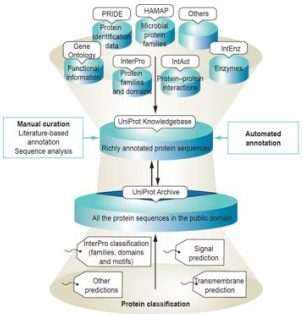
Silencing of gene expression by targeted degradation of mRNA



Robinson (2004) PLoS Biol

Data basis for functional annotation

UniProt database



- Universal Protein Resource
- http://www.uniprot.org/
- Comprehensive catalog of
  - protein sequence
  - functional information (annotation data)
- Several databases:
  - UniProtKB: Knowledgebase (annotation)
  - UniRef: Reference Clusters (sequences for UniProtKB)
  - UniParc: Archive (all sequences)
  - UniMES: metagenomes
- Merger of previous Swiss-Prot and TrEMBL databases

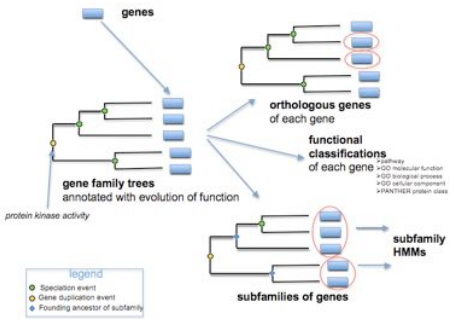
PANTHER

Thomas, et al. (2003) Genome Res 13:2129-41;
Thomas & Kejariwal (2004) PNAS 101:15398–15403.

- Protein ANalysis THrough Evolutionary Relationships
- http://www.pantherdb.org/
- Classification system of proteins and their genes
- Classification by:
  - Family (evolutionarily related proteins) and subfamily (related proteins that have the same function)
  - Molecular protein function (e.g. kinase)
  - Biological protein function (e.g. mitosis)
  - Pathway relationships
- Compilation by human curation as well as bioinformatic algorithms
- >11,900 protein families, >83,000 subfamilies in 2015

PANTHER

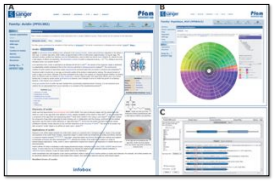
Information types stored in PANTHER



http://www.pantherdb.org

PFAM database

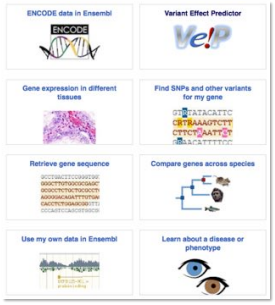
Punta, et al. (2012) Nucleic Acids Res 40:D290-D301;
Finn, et al. (2014) Nucleic Acids Res 42:D222-D230



- Protein families
- http://pfam.xfam.org/
- Database of protein families (>16,200 in 2015)
- Contains, for each family, multiple sequence alignments and Hidden Markov models (HMMs) for seed alignment
- Contains information about protein domains
- Grouping of families into clans

Ensemble database

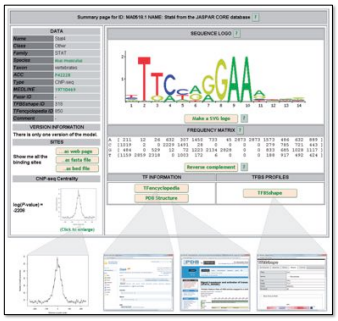
Cunningham, et al. (2015) Nucleic Acids Res 43:D662-D669



- http://www.ensembl.org
- Genomic interpretation system
- Annotations, querying tools, access methods for chordates and key model organisms
- Annotation includes:
  - Gene annotation (GENCODE gene set)
  - Regulatory region / epigenetic annotation
  - Variation annotation (germline & somatic), also including 1000G, HapMap, EVS and other data
  - Comparative annotation (mutation age, multiple sequence alignment, secondary protein structures, …)
- Web-based queries and API

JASPAR database

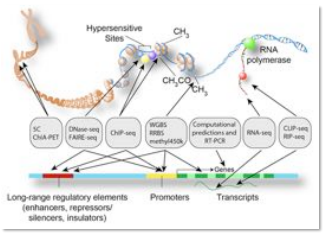
Portales-Casamar, et al. (2010) Nucleic Acids Res 38:D105–D110;
Mathelier, et al. (2014) Nucleic Acids Res 42:D142-D147



- http://jaspar.genereg.net/
- Collection of databases:
  - JASPAR CORE: database of transcription factor binding motifs
  - JASPAR COLLECTIONS: databases for splice forms, meta-models, and others

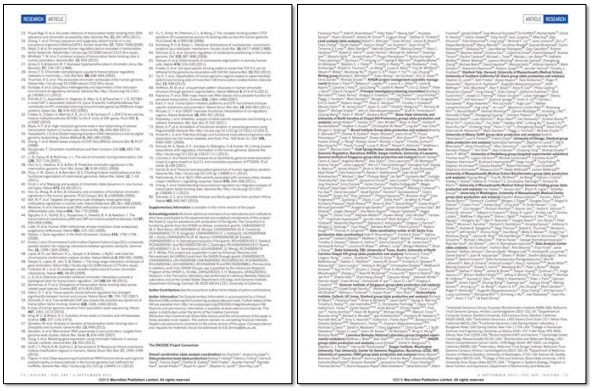
ENCODE database

The ENCODE Project Consortium (2012) Nature 489:57–74



- Encyclopedia of DNA Elements
- https://www.encodeproject.org/
- Projects aims to build a comprehensive catalog of all functional elements in the human genome
- International collaboration funded by the National Human Genome Research Institute (NHGRI

ENCODE: author list



The ENCODE Project Consortium (2012) Nature

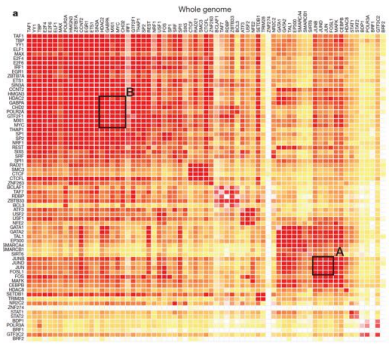
ENCODE: annotations

- Candidate enhancers and promoters for DNase hypersensitivity
- Gene expression over ~60 cell types
- Transcription start sites (TSS)
- Peaks (sites of transcription factor binding or DNase hypersensitivity)
- Amount of RNA for different types of RNA and in various cell lines
- Promoter regions
- Predicted enhancers
- Semi-automated genome annotation (SAGA); summarization of chromatin accessibility, patterns of histone modifications, transcription factor binding, …
- High Occupancy of Target (HOT) regions (regions in which a large number of different transcription-related factors bind)
- Connectivity of transcription factors
- Motifs (DNA binding sites) for transcription related factors
- and more …

https://www.encodeproject.org/

ENCODE

Co-association between transcription factors



The ENCODE Project Consortium (2012) Nature

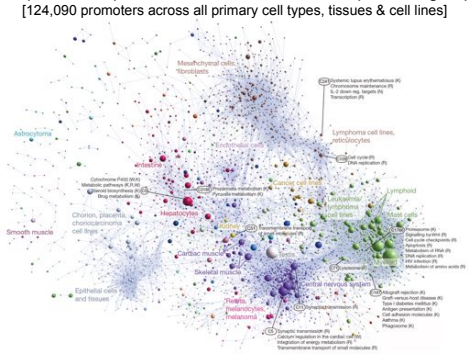
FANTOM5 database

FANTOM Cons., RIKEN PMI & CLST, et al. (2014) Nature 507:462-70

- Functional Annotation of the Mammalian Genome
- http://fantom.gsc.riken.jp/5/
- Annotation of regulation, expression and function of mammalian genes
- Promotor atlas, cell-type-specific TF
- Tools for visualization and exploration
- Based on systematic sampling of the distinct mammalian cell types (975 human and 399 mouse samples, including primary cells), tissues and cancer cell lines
- RIKEN-led consortium

FANTOM5

Collapsed co-expression network of 4882 co-expression groups
[124,090 promoters across all primary cell types, tissues & cell lines]



FANTOM Consortium, et al. (2014) Nature

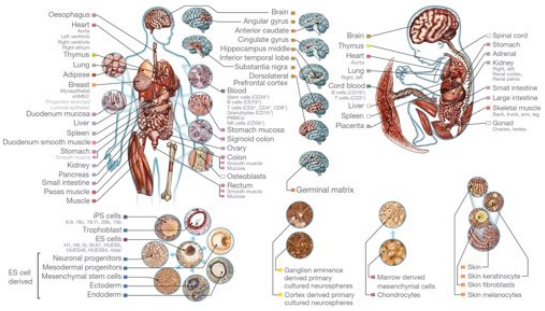
Human epigenome

Roadmap Epigenomics Consortium, et al. (2015) Nature 518:317-30

- http://www.roadmapepigenomics.org/
- Map of
  - DNA methylation
  - Histone modifications
  - Chromatin accessibility
  - Small RNA transcripts
- Considered locations:
  - Stem cells
  - Primary ex vivo tissues
- Sites selected to represent the normal counterparts of tissues and organ systems frequently involved in human disease
  → Convenience control for such studies

Human epigenome

Profiled tissues and cell types



Roadmap Epigenomics Consortium (2015) Nature

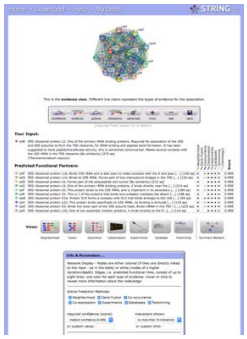
Human epigenome

Chromatin state annotation in 127 epigenomes

Roadmap Epigenomics Consortium (2015) Nature

STRING database

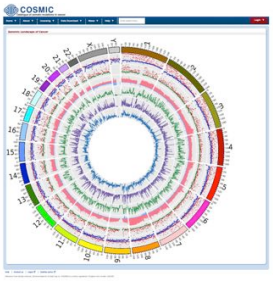
Franceschini, et al. (2012) Nucleic Acids Res 41:D808-D815



- http://string-db.org/
- Database of known and predicted protein-protein interactions (both direct [physical] and indirect [functional] associations)
- Based on:
  - Genomic context
  - High-throughput experiments
  - Co-expression
  - Previous knowledge
- Builds upon numerous other databases
- >9,600,000 proteins from >2000 organisms in 2015

COSMIC database

Forbes, et al. (2015) Nucleic Acids Res 43: D805-D811



- http://cancer.sanger.ac.uk/cosmic
- Catalog of somatic mutations in cancer
- Two types of data:
  - Manual curation data from peer reviewed publications by COSMIC expert curators (aka non-systematic/targeted screen data)
  - Systematic screen data: uploads from large scale genome screening publications and from other databases (TCGA, ICGC); unbiased molecular profiling of diseases

GenomeRNAi database

Horn, et al. (2007) Nucleic Acids Res 35:D492-7;
Gilsdorf, et al. (2010) Nucleic Acids Res 38:D448-52;
Schmidt, et al. (2013) Nucleic Acids Res 41:D1021-6

GenomeRNAi

- http://www.genomernai.org/GenomeRNAi/
- Database containing phenotypes from RNA interference screens in Drosophila and Homo sapiens
- Provision of RNAi reagents and their predicted quality.

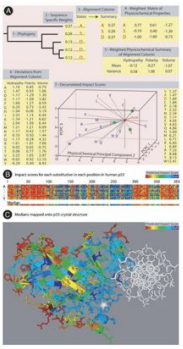
There are more databases…

Software

Schiffahrtsmuseum Brake, Germany

MAPP

Stone & Sidow (2005) Genome Res 15:978-86



- Multivariate Analysis of Protein Polymorphism
- http://mendel.stanford.edu/sidowlab/downloads/MAPP
- Steps:
  1. Multiple alignment of homologous sequences, phylogeny-weighted scores
  2. Interpretation of scores by quantified physicochemical properties, yielding constraints on these properties for each variant
  3. Create new feature space by PCA of all physicochemical properties
  4. MAPP score: distance to the origin of the new feature space

GERP/GERP++

Cooper, et al. (2005) Genome Res 15:901-13;
Davydov, et al. (2010) PLoS Comp Biol 6:e1001025.

- Genomic Evolutionary Rate Profiling
- http://mendel.stanford.edu/sidowlab/downloads/gerp
- Identification of constrained elements by a deficit of substitution events due to purifying selection
- Comparison of estimated evolutionary rates between
  - individual alignment column (residue/variant) and
  - a tree describing neutral substitution rates (ML-based phylogenic inference)
- Constraint regions exhibit fewer than expected changes
- RS score (metric of constraint): rejected substitutions
- GERP++: additional aggregation of constrained sites into constrained sequences

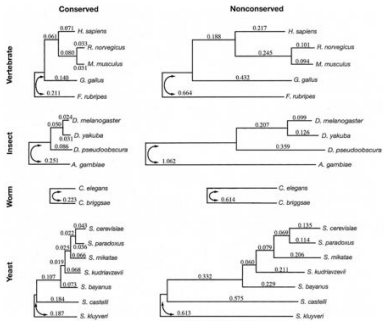
PhastCons

Siepel, et al. (2005) Genome Res 15:1034-1050

- Part of the PHAST (Phylogenetic Analysis with Space /Time Models) package:
  - http://compgen.bscb.cornell.edu/phast/
  - Engine behind the Conservation tracks in the UCSC Genome Browser
- Aims at conservation scoring and identification of conserved elements from multiple sequence alignment
- Predicting sequences as being conserved / not conserved
  - using a phylogenetic Hidden Markov Model (HMM)
  - different values for branch length scaling parameter (average substitution rate) in phylogenetic tree between both types
  - Unsupervised learning without use of external information
- Calculation of conservation score

PhastCons

Assumed tree topologies and branch length



Siepel, et al. (2005) Genome Res

PhastCons: conservation score

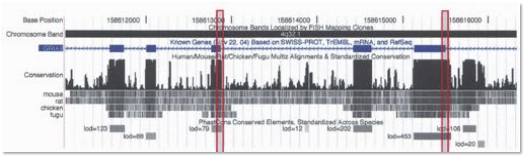
Conservation score: posterior probability that each site was generated from a conserved state in the phylo-HMM

LOD score: log-ratio of the likelihoods for a region under the conserved phylogenetic model compared to the nonconserved model

Note 1: LOD – logarithm of the odds

Note 2: This is not the LOD score from linkage analysis (although scaled in a similar way).

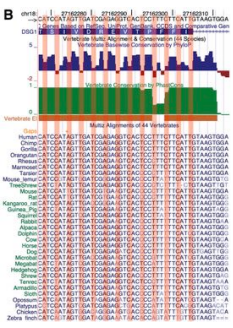
Conservation track in UCSC Genome Browser:



Siepel, et al. (2005) Genome Res

PhyloP

Pollard, et al. (2010) Genome Res 20:110-121



- phylogenetic P-values
- http://compgen.bscb.cornell.edu/phast/
- Aims at detecting deviations from the neutral rate of substitutions
  - Conservation: less than under drift
  - Acceleration: more than under drift
- Additionally allows for clade-specific differences in the phylogeny
- Software implementation of four tests, including likelihood-ratio and score tests, a number-of-substitutions test (SPH), and GERP
- The conservation track of the UCSC genome browser contains PhyloP scores (SPH p-values for deviation from drift).

LogRE

Clifford, et al. (2004) Bioinformatics 20:1006-14

- http://lpgws.nci.nih.gov/cgi-bin/GeneViewer.cgi
- Prediction whether amino acid (AA) changes in conserved domains are likely to affect protein function
- Based on output of the HMMER/2/3 software (multiple sequence alignment using HMMs and profiles) and Pfam profiles (conservation in protein families)
- E-value in sequence alignment: expected number of sequences with an alignment score equal to or even more extreme than that of the observed sequence
- LogRE value:

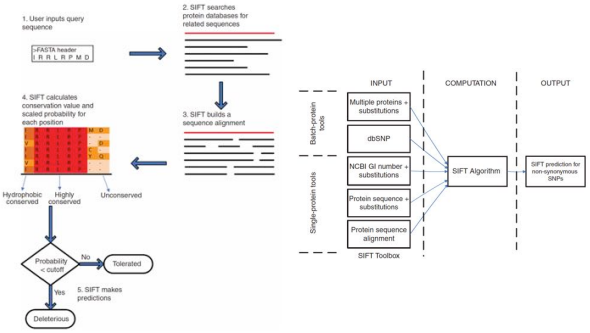
  $\log_{10}$ of ratio E(deviant AA) / E(canonical AA)

SIFT

Kumar, et al. (2009) Nat Protoc 4:1073-81, and others

- Sorting Tolerant From Intolerant
- http://sift.jcvi.org/
- Protein function prediction due to an AA substitution (nsSNP)
- Based on
  - Multiple sequence alignment
  - Conservation with respect to functionally related protein sequences
  - Similarity between the alternate amino acids
  - No incorporation of protein structure
- Output
  - Score: probability of substitution for being tolerated (i.e. values near 0 imply high probability for being deleterious)
  - Qualitative prediction of being 'tolerated' or 'deleterious' by thresholding

SIFT: workflow



High sequence conservation in functionally related protein sequences

→ nsSNP unlikely to be tolerated

Kumar, et al. (2009) Nat Protoc

SIFT score

Multiple sequence alignment of homologous amino acid (AA) sequences

For a given position, calculation of the relative frequencies of the 20 AA at this position in the alignment, normalized by the maximum relative frequency

SIFT score:
normalized probability of the observed AA
(i.e. frequency of the observed AA relative to the most common AA at this position in the alignment)

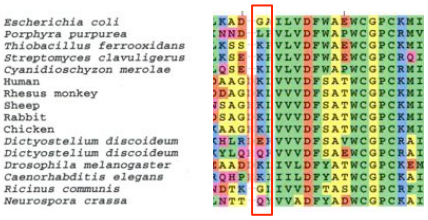
SIFT score close to 0:

The observed AA almost never occurs at this position in the homologous sequences, indicating high conservation and a probably deleterious effect.

Kumar, et al. (2009) Nat Protoc

SIFT score

Thioredoxin of *E. coli* and 15 homologs



Relative frequency	f(G)=2/16	f(L)=1/16	f(K)=10/16	f(E)=1/16	f(Q)=2/16
SIFT	S(G)=2/10	S(L)=1/10	S(K)=10/10	S(E)=1/10	S(Q)=2/10

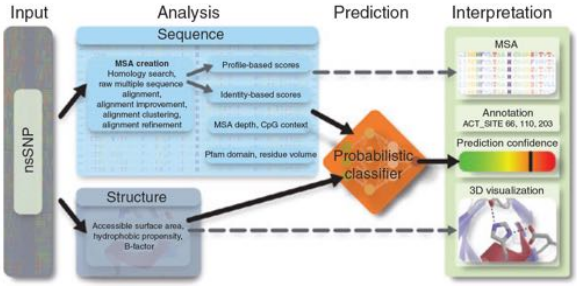
Lesk (2014)

PolyPhen-2

Adzhubei, et al. (2010) Nat Methods 7(4):248-249

- http://genetics.bwh.harvard.edu/pph2/
- Prediction of the functional effects of an amino acid substitution on the structure and function of a protein
- Naïve Bayes classifier based on
  - Sequence conservation
  - Chemical properties of amino acids
  - Protein structure
  - Sequence context
- Output
  - Score: probability of substitution for being deleterious (i.e. values near 1 imply high probability)
  - Qualitative prediction of being 'probably damaging', 'possibly damaging', 'benign' or 'unknown'

PolyPhen-2: workflow

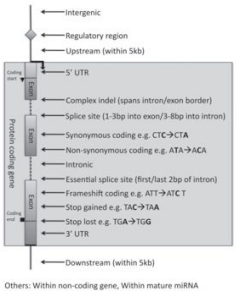


Adzhubei, et al. (2010) Nat Methods

SNP Effect Predictor (SEP)

McLaren, et al. (2010) Bioinformatics 26:2069-70.

Predicted consequences



- Annotation of SNVs in transcripts (i.e. coding sequence)
- Part of Ensemble; annotation based on Ensemble databases
- Web-based tool and Application Programme Interface (API, written in Perl) available
- http://www.ensembl.org/info/docs/api/

ANNOVAR (I)

Wang, et al. (2010) Nucleic Acids Res 38:e164.

- http://annovar.openbioinformatics.org/
- Widely used tool; builds upon numerous databases and many other tools
- Annotation of SNVs, InDels and CNVs
- Conversion utilities for numerous file types (including VCF)
- Perl command line tool
- Web-based access to some functionality via wANNOVAR (http://wannovar.usc.edu/)
- Gene-based annotation:
  - Identification of protein-coding changes
  - Flexible use of many gene definition systems (RefSeq, UCSC, ENSEMBL, GENCODE, AceView, and others)

ANNOVAR (II)

- Region-based annotation:
  - Identification of conserved regions among 44 species,
  - Prediction of transcription factor binding sites, segmental duplication regions, GWAS hits, database of genomic variants, ENCODE sites, ChIP-Seq peaks, RNA-Seq peaks, …
- Filter-based annotation:
  - Presence (and reported frequency) in specific databases (dbSNP, 1000 Genome, NHLBI-ESP 6500 exomes, ExAC, and others)
  - Calculation of scores (e.g. SIFT, PolyPhen-2, LRT, MutationTaster, MutationAssessor, FATHMM, MetaSVM, MetaLR, GERP++)
- Other functionalities:
  - Retrieval of nucleotide sequence in any user-specific genomic positions in batch
  - Candidate gene list for Mendelian diseases from exome data
  - and more

SnpEff

Cingolani, et al. (2012) Fly 6:1-3

- SNP Effect
- http://SnpEff.sourceforge.net/
- Annotation of SNVs, InDels, MNP (multiple nucleotide polymorphism) in coding sequence
- Multiple input file formats (VCF, mpileup, text)
- Gene annotation has similar scope as in ANNOVAR
- Integration with computational biology platform Galaxy (http://gmod.org/wiki/Galaxy) and GATK
- Superseded ANNOVAR when integrated in GATK
- Tool SnpSift for VCF file manipulaitn and filtering

Condel

González-Pérez & López-Bigas (2011) Am J Hum Genet 88:400-9

- Consensus deleteriousness score of missense mutations
- http://bg.upf.edu/condel
- Mulitple sequence alignment of homolous sequences
- Weighted combination of five predictors: Logre, MAPP, Mutation Assessor, PolyPhen-2 and SIFT
- Definition of different simple and averaged scores for the 0/1 prediction and the normalized scores of each of the five predictors
- Combinations these derived scores used for classification of a variant being deleterious or neutral

FATHMM, FATHMM-MKL

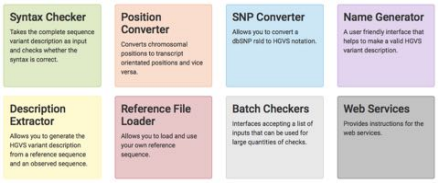
Shihab, et al. (2013) Hum Mutat 34:57-65; Shihab, et al. (2015) Bioinform.

- Functional Analysis through Hidden Markov Models
- http://fathmm.biocompute.org.uk/
- Prediction of functional consequences for both coding and non-coding SNVs
- Web service
- Based on
  - conservation of homologous sequences, protein domain functionality and pathogenicity (inferred from relative frequencies of disease-associated variants)
  - SVM using functional annotation from numerous ENCODE tracks
- Incorporates numerous databases, e.g. HGMD, UniProt, VariBench and SwissVar

Mutalyzer

Wideman, et al. (2008) Hum Mutat 29:6-13

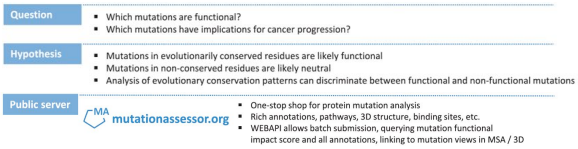
- https://mutalyzer.nl/
- Checking sequence variant nomenclature according to the guidelines of HGVSt (Human Genome Variation Society)
- Some automated extraction of variant annotation
- Web-based service



Mutation Assessor

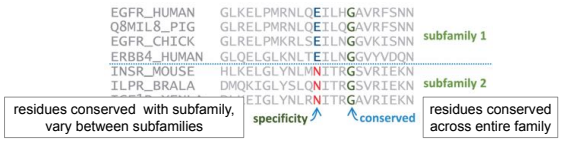
Reva, et al. (2011) Nucleic Acids Res 39:e118

- http://mutationassessor.org/
- Functionality predicted from inter-species conservation and known 3D structures
- Somatic cancer mutations are additionally evaluated for recurrence, multiplicity and annotation based on the COSMIC database



Mutation Assessor: functional impact score (FIS)

Multiple sequence alignment of a large number of homologs for both protein families and subfamilies



Strength of residue conservation: distributional entropy of alignment column

Conservation score: effect of mutation described as difference in residue conservation

Specificity score: conservation score within data-defined sequence subfamily

Functional Impact Score = conservation score + specificity score

Reva, et al. (2011) Nucleic Acids Res

MutationTaster / MutationTaster2

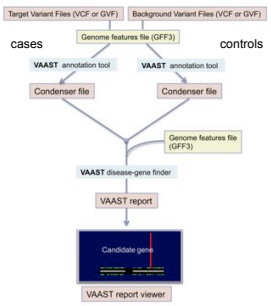
Schwarz, et al. (2010) Nat Methods 7:575-6;
Schwarz, et al. (2014) Nat Methods 11:361-2

- http://www.mutationtaster.org/
- Web-based service, upload of VCF files
- Prediction of functional consequences for amino acid substitutions (nsSNVs), intronic and synonymous SNVs and InDels and exon-intron border variants
- Bayes classifier trained on 1000G and HGMD Professional
- Integration of: 1000G, HapMap, ClinVar, HGMD Public, ENCODE, JASPAR, PhyloP/PhastCons [conservation], NNSplice [splicing], …

VAAST

Yandell, et al. (2011) Genome Res 21:1529-42

Search procedure



- Variant Annotation, Analysis, and Search Tool
- Annotation of amino acid substitutions (coding sequence) and non-coding
- Likelihood-ratio test for disease association; aggregation of rare variants (similar to CMC approach)
- Severity of SNVs assessed by comparison to OMIM
- Scoring of non-coding and synonymous variants by use of sequence conservation, OMIM, 1000 Genomes, ENCODE,

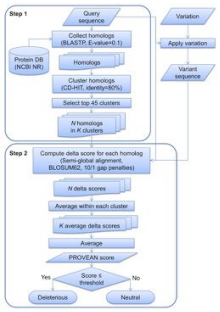
VAT

Wang, et al. (2014) Am J Hum Genet 94:770-83

- Variant Analysis Tools
- http://varianttools.sourceforge.net/
- Different gene set references
- Presence in dbSNP, ExAC, 1000G, HapMap, database of genomic variants, catalog of somatic mutations in cancer
- Prediction scores from dbNSFP database (SIFT, PolyPhen, MutationTaster, and others)
- Conserved or duplicated regions
- Automatic annotation using ANNOVAR and SnpEff
- Many more tasks possible; coded in Python

PROVEAN

Choi, et al. (2012) PLoS ONE 7: e46688



- Protein Variation Effect Analyzer
- http://provean.jcvi.org/
- Annotation of the functional impact based on conservation of homologous protein sequences
- Focus on InDels, multiple substitutions
- Impact measured by Delta Score Δ:
  - defined as the difference in the alignment scores for the given protein and a homologous sequence, average over many homologous sequences
  - Thresholding Δ for prediction

CADD

Kircher, et al. (2014) Nat Genet 46:310-5

- Combined Annotation - Dependent Depletion
- http://cadd.gs.washington.edu/
- Annotation of SNVs and InDels
- Based on 63 partially different annotations (VEP, ENCODE, GERP, phyloP, TF binding, SIFT, PolyPhen, …)
- Integration of numerous annotations into a single C score
- Assessment of the "deleteriousness" of a variant by simulation
  - Genome-wide simulation of de-novo germline variation without selection
  - Comparison against fixed or nearly fixed derived alleles in humans (as compared to chimpanzee) with respect to annotation

CADD: C score

Support-vector machine (SVM) for distinguishing nearly fixed variation from simulated neutral variation (14.7×10^6 vs. 14.7×10^6)

SVM trained on 63 annotations and some selected interaction terms (but 949 features in the model due to dummy coding of categorical variables!)

Application to all 8.6 billion possible substitutions in GRCh37, yielding the distribution of the combined score from the SVM (C-score) for variants in the human reference genome

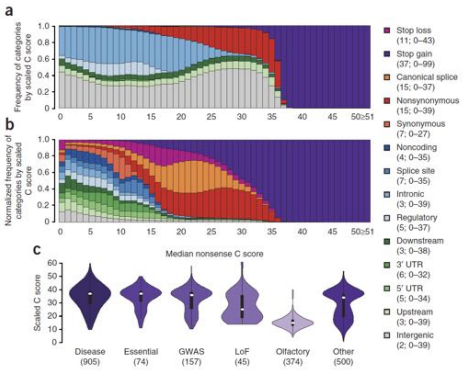
Phred-scaling of the rank of the score (scaled C-sore):
$-10\log_{10}$ (rank/total number of substitutions).

Comparison of the scaled C-score of a variant at hand against this distribution

Example: A variant with a scaled C-score of 20 indicates that it is rank at 1% of the most deleterious substitutions in the human genome

Kircher, et al. (2014) Nat Genet

CADD: typical C scores for SNVs



Kircher, et al. (2014) Nat Genet

GWAVA

Ritchie, et al. (2014) Nat Methods 11:294-6

- Genome-wide annotation of variants
- https://www.sanger.ac.uk/resources/software/gwava/
- Functional annotation of non-coding sequence variants
- Integration of genomic and epigenomic annotations (1000G frequencies and ancestral allele calls, GERP scores [conservation], several ENCODE tracks, TF binding motifs)
- Classification of variants having a pathogenic effect or not via random forest, trained on HGMD and 1000G
- Validation by application to the COSMIC database

SuRFR

Ryan, et al. (2014) Genome Med 6:79

- SNP Ranking by Function R package
- http://www.cgem.ed.ac.uk/resources/
- Annotation of non-coding variants
- Incorporation of 1000G, ENCODE, FANTOM5, Epigenome Roadmap
- Prioritization of variants by a rank-of-ranks approach:

$$R = rank_i\left(\sum\left(r_{ij}.w_j\right)\right)$$

r_{ij} – ranks within annotation category, w_j – weight for category, R – overall rank

- Three pre-trained weighting schemes available
- Implemented as package for the R statistical language

There are more annotation tools…

http://omictools.com/variant-annotation-c104-p1.html

Performance of prediction

Comparison of methods

Comparison of 1,100 common polymorphisms (1000G) and 1,100 known disease mutations (HGMD)

Table 1 | Comparison between MutationTaster2 and other prediction tools

Tool	*n*	NPV	PPV	Sensitivity	Specificity	Accuracy
PPH2-var	2,200	0.808	0.875	0.789	0.887	0.838
PPH2-div	2,200	0.853	0.827	0.858	0.821	0.840
PROVEAN	2,200	0.798	0.865	0.778	0.878	0.828
SIFT	2,200	0.832	0.854	0.827	0.858	0.843
MutationTaster1	2,200	0.850	0.870	0.846	0.874	0.860
MutationTaster2	2,200	0.886	0.875	0.887	0.874	0.880

Details about the methods and further statistics are presented in **Supplementary Methods** and at http://www.mutationtaster.org/info/statistics.html. *n*, number of cases; NPV, negative prediction value; PPV, positive prediction value; PPH2-div, PolyPhen-2 with HumDiv classifier; PPH2-var, PolyPhen-2 with HumVar classifier.

Schwarz, et al. (2014) Nat Methods

Comparison of methods

Table 1 Characteristics of three selected in silico prediction tools

Characteristic	SIFT	PolyPhen-2
Target	nsSNV	nsSNV
Algorithm	Sequence alignment	Bayes classifier
Features	Amino acid sequence	Amino acid sequence, secondary and tertiary structure
Input	Amino acid sequence or SwissProt ID or rs number or location, amino acid substitution	Amino acid sequence or SwissProt ID or rs number or location, amino acid substitution
Classification	Tolerated, damaging	Probably damaging, possibly damaging, benign, unknown
Additional output	Number and median conservation of aligned sequences	False and true positive rate, protein structure
URL	http://sift.jcvi.org/	http://genetics.bwh.harvard.edu/pph2/

Table 2 Performance of three selected in silico prediction tools

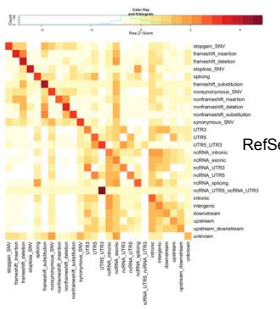
Tool	SIFT	PolyPhen-2[a]
Sensitivity	0.68	0.73 (0.86)
Specificity	0.62	0.70 (0.51)
Matthews correlation coefficient	0.30	0.43 (0.39)

Knecht & Krawczak (2014) Hum Genet

Comparison of methods

Same software (ANNOVAR), different annotation databases

[~81 million variant calls from 276 samples of immune disease & cancer cases; from the WGS500 project (University of Oxford)]



$\log_{10}$ number of variants

RefSeq annotation

Ensembl annotation

McCarthy, et al. (2014) Genome Med

Comparison of methods

Same annotation database (Ensembl), different annotation software

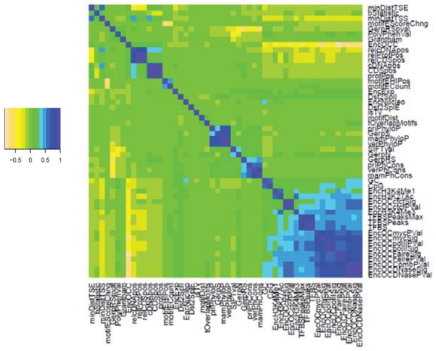
[~81 million variant calls from 276 samples of immune disease & cancer cases; from the WGS500 project (University of Oxford)]

	ANV+VEP	ANV	VEP	Exact match	Category match	ANV match rate (%)	VEP match rate (%)	Overall category match rate (%)	Overall exact match rate (%)
LOF total	104,915	77,527	96,761	68,284	69,373	88.08	70.57	66.12	65.09
Frameshift	19,021	15,822	16,685	13,486	-	85.24	80.83	-	70.90
Stop gained	16,758	14,960	16,146	14,348	-	95.91	88.86	-	85.62
Stop lost	1,113	906	1,077	870	-	96.03	80.78	-	78.17
All splicing	69,112	45,839	62,853	39,580	-	86.35	62.97	-	57.27
MISSENSE total	350,806	324,242	347,752	318,056	321,188	98.09	91.46	91.56	90.66
Inframe indel	9,455	8,650	6,600	5,795	-	66.99	87.80	-	61.29
Missense	343,284	315,592	339,953	312,261	-	98.94	91.85	-	90.96
Initiator codon	1,199	0	1,199	0	-	-	0.00	-	0.00
SYNONYMOUS and OTHER CODING total	182,120	172,463	175,483	165,643	165,826	96.05	94.39	91.05	90.95
Synonymous	181,873	172,463	175,053	165,643	-	96.05	94.62	-	91.08
Stop retained	203	0	203	0	-	-	0.00	-	0.00
Other coding	227	0	227	0	-	-	0.00	-	0.00
ALL LOF	104,915	77,527	96,761	68,284	69,373	88.08	70.57	66.12	65.09
ALL LOF and MISSENSE	455,721	401,769	444,513	386,340	390,561	96.16	86.91	85.70	84.78
ALL EXONIC	637,841	574,232	619,996	551,983	556,387	96.13	89.03	87.23	86.54

McCarthy, et al. (2014) Genome Med

Correlation between different annotations

14.7 millions human-derived alleles with ≥95% population frequency



Kircher, et al. (2014) Nat Genet

Limits of in-silico functional prediction

Back to square one: motivation for in-silico prediction

Do these missense mutations actually cause the disease/ phenotype at hand?

Model organism experiments are too costly for all observed missense mutations.

Comparison of predicted and actual consequences of missense mutations

Lisa A. Miosge[a,1], Matthew A. Field[b,1], Yovina Sontani[a], Vicky Cho[a,b], Simon Johnson[a,b], Anna Palkova[a,b], Bhavani Balakishnan[b], Rong Liang[b], Yafei Zhang[b], Stephen Lyon[c], Bruce Beutler[c], Belinda Whittle[b], Edward M. Bertram[b], Anselm Enders[d], Christopher C. Goodnow[a,e,2,3], and T. Daniel Andrews[a,2,3]

(A) Generation of random mutations in mouse pedigrees using ENU; Breeding to homozygosity and phenotyping of mice with 1 of 33 potentially disruptive de-novo points mutations in 23 essential immune system genes [already known to produce a fully penetrant detectable phenotype]

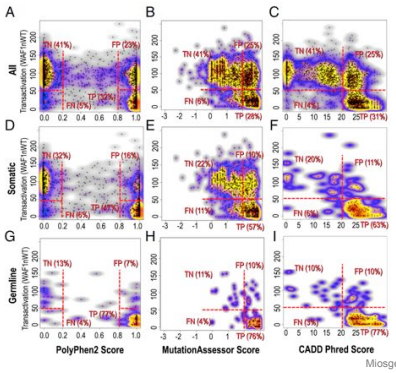
(B) In vitro phenotyping (translational activity) of all possible TP53 mutations

Prediction by PolyPhen-2, CADD, SIFT, GERP, MutAssessor &PANTHER

Miosge et al. (2015) PNAS

Limits of in-silico functional prediction

Predicted damage vs. experimentally measured activity for TP53



Miosge et al. (2015) PNAS

Limits of in-silico functional prediction

"The discordance between the predicted and actual effect of missense mutations revealed here creates the potential for many FP conclusions in clinical whole genome sequencing.
…
Hence, for interpretation of a clinical genome sequence at present, it is essential to measure experimentally the consequence of any missense mutation thought to be causal."

… We conclude that for de novo or low-frequency missense mutations found by genome sequencing, half those inferred as deleterious correspond to nearly neutral mutations that have little impact on the clinical phenotype of individual cases but will nevertheless become subject to purifying selection.

Miosge et al. (2015) PNAS

That's it!