Targeted resequencing

Sarah Calvo, Ph.D.
Computational Biologist
Vamsi Mootha laboratory

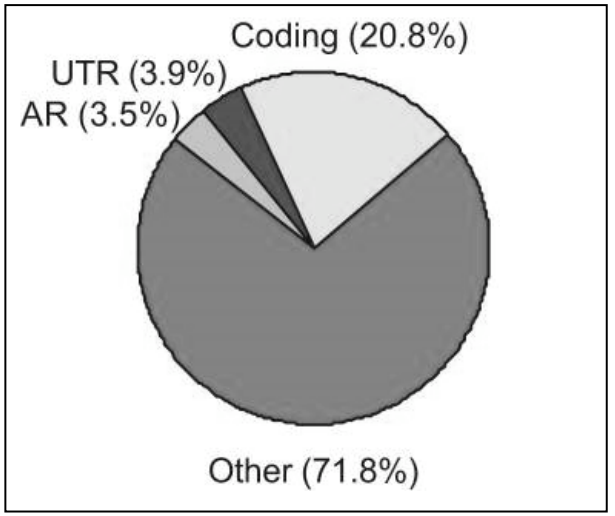
Snapshots of Genome Wide Analysis in Human Disease (MPG), 4/20/2010

Vamsi Mootha,
PI

How can I assess a small genomic target across many samples?

Limitations of whole exome sequencing

- expensive (1 Illumina lane/sample)
- coding only (~20% of all functional elements)



Multi-Species Conserved Seqs

Margulies et al. 2003

How can I assess a small genomic target across many samples?

Limitations of whole exome sequencing

- expensive (¥ 1 Illumina lane/sample)
- coding only (~20% of all functional elements)

Goal: sequence arbitrary small genomic regions, across many samples

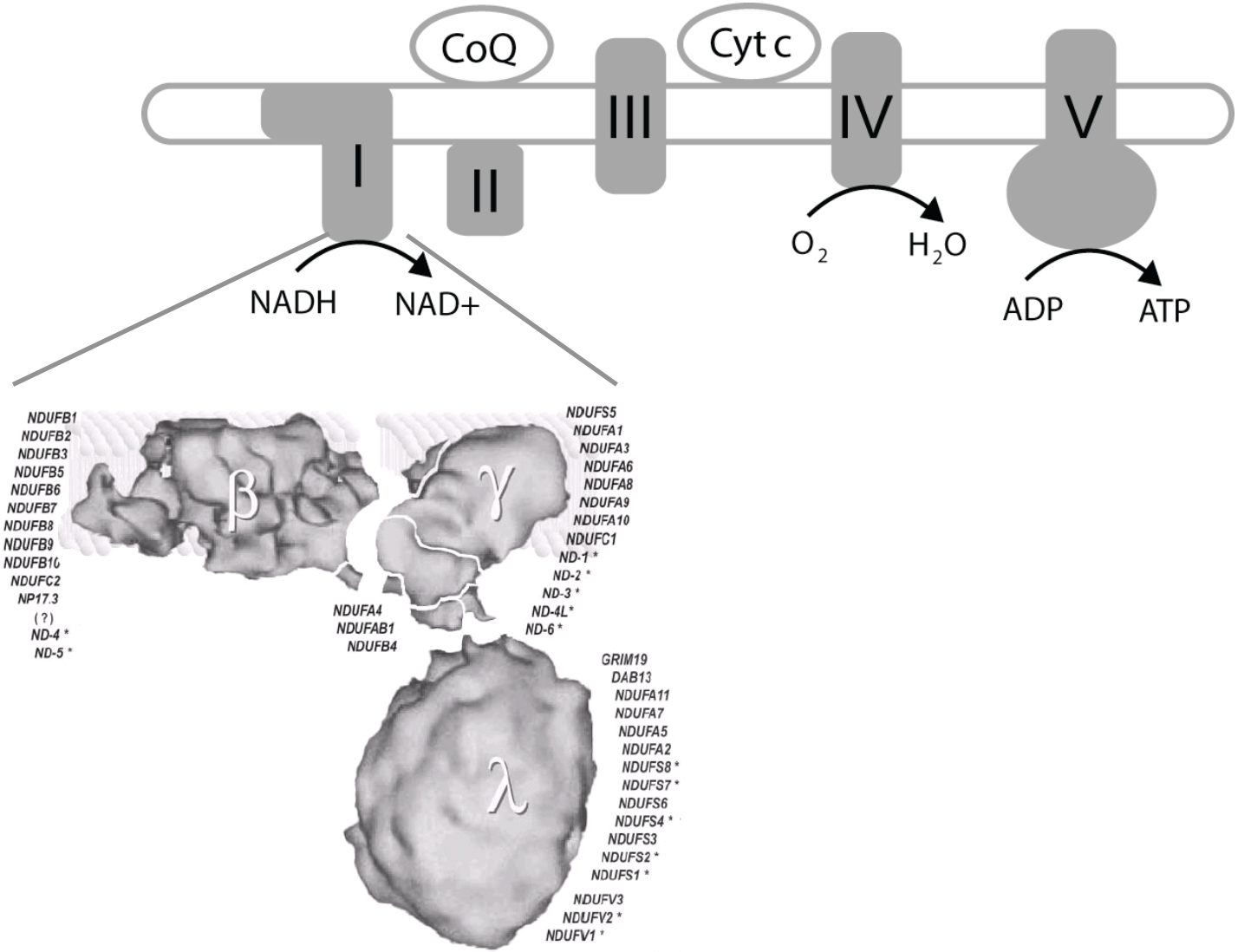
- genomic regions, eg GWAS
- non-coding genic, eg UTRs, regulatory, conserved seqs
- small number of candidate genes (100s)

Outline

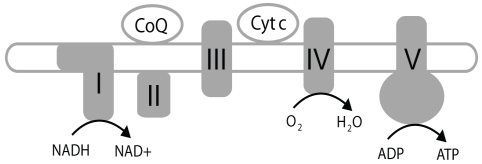
Snapshot of Mito10k (100 candidate genes x 100 patients)

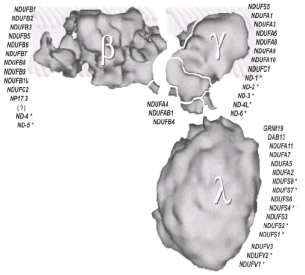
Lessons to share

Complex I of the mitochondrial respiratory chain



Complex I deficiency





Structure

- 46 subunits, 10 known assembly factors

Complex I Deficiency

- Most common OXPHOS disease
- Phenotypes
  - range from neonatal lethal, to adult-onset neurodegeneration
  - leukodystrophy, cardiomyopathy, seizures, Leigh syndrome
- 26 disease-related genes (19 subunits + 7 assembly factors)

Diagnosis difficult

No effective treatment

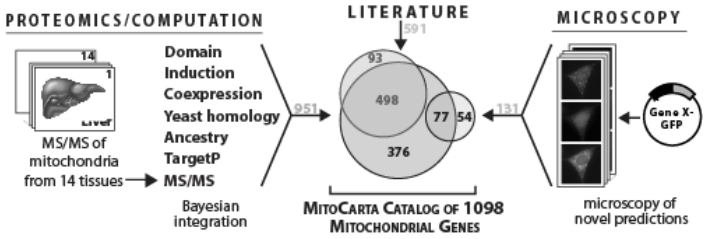
Goal: identify molecular basis for complex I deficiency

Method: sequence candidate genes in patient cohort

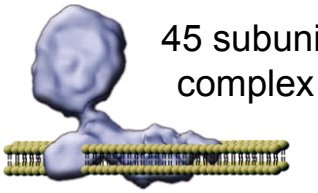
New candidates for complex I deficiency

Catalog of 1100 mitochondrial genes

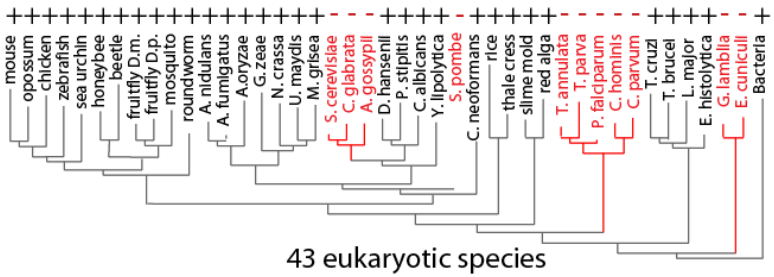
(Pagliarini, Calvo et al. Cell 2008)



Phylogenetic profiling



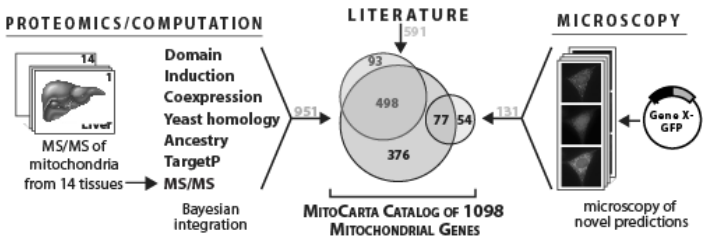
45 subunits complex I



New candidates for complex I deficiency

Catalog of 1100 mitochondrial genes

(Pagliarini, Calvo et al. Cell 2008)

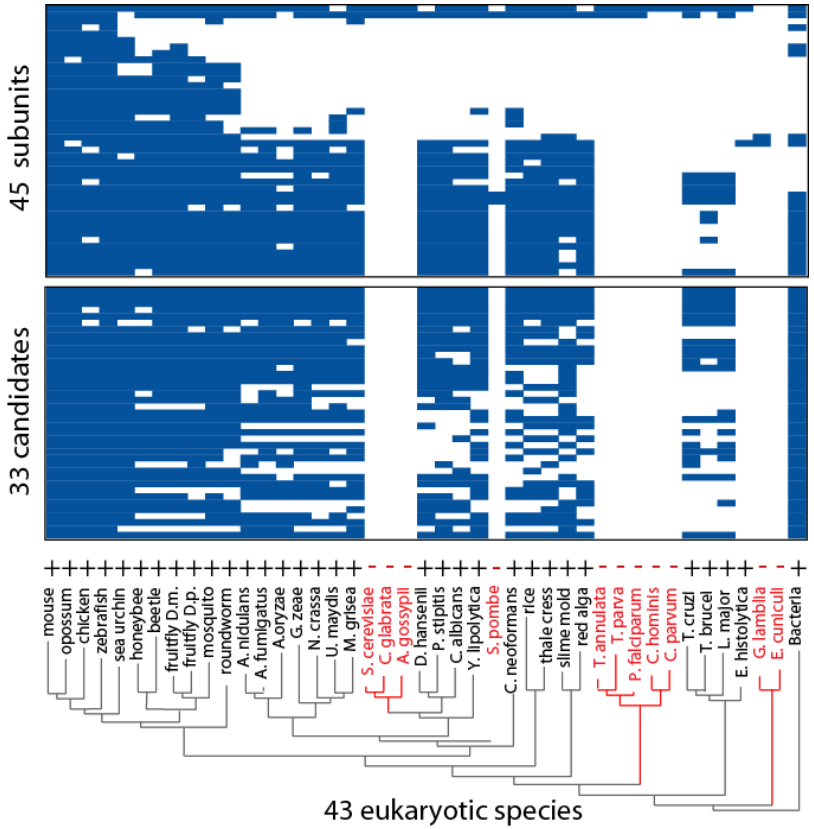


103 candidate genes

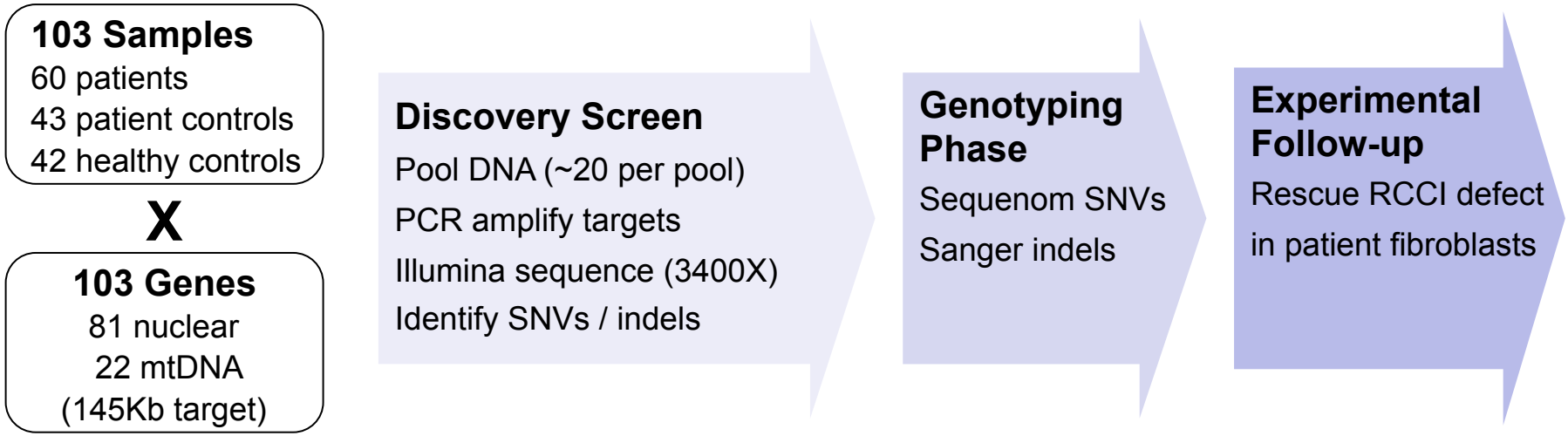
22 mtDNA

81 nuclear

Phylogenetic profiling



Mito10K: sequence 100 genes X 100 patients



Mito10K: sequence 100 genes X 100 patients

103 Samples
60 patients
43 patient controls
42 healthy controls

X

103 Genes
81 nuclear
22 mtDNA
(145Kb target)

Discovery Screen
Pool DNA (~20 per pool)
PCR amplify targets
Illumina sequence (3400X)
Identify SNVs / indels

Genotyping Phase
Sequenom SNVs
Sanger indels

Experimental Follow-up
Rescue RCCI defect in patient fibroblasts

Challenge: detect singleton SNVs, given high error rates

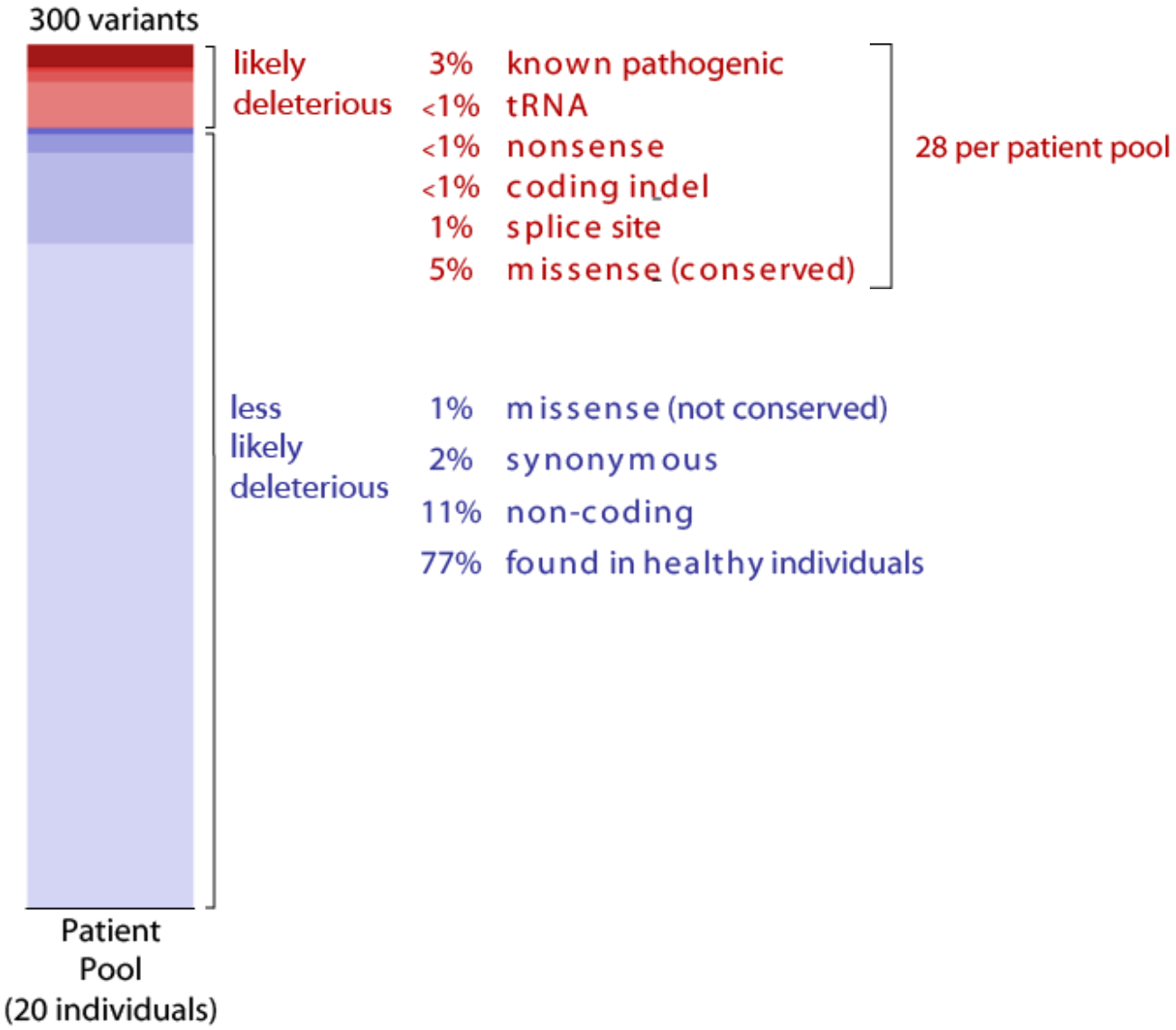
Method: Syzygy

- 92% sensitivity (76% singletons)
- 9% false positives

~300 SNVs detected per pool

Manuel Rivas

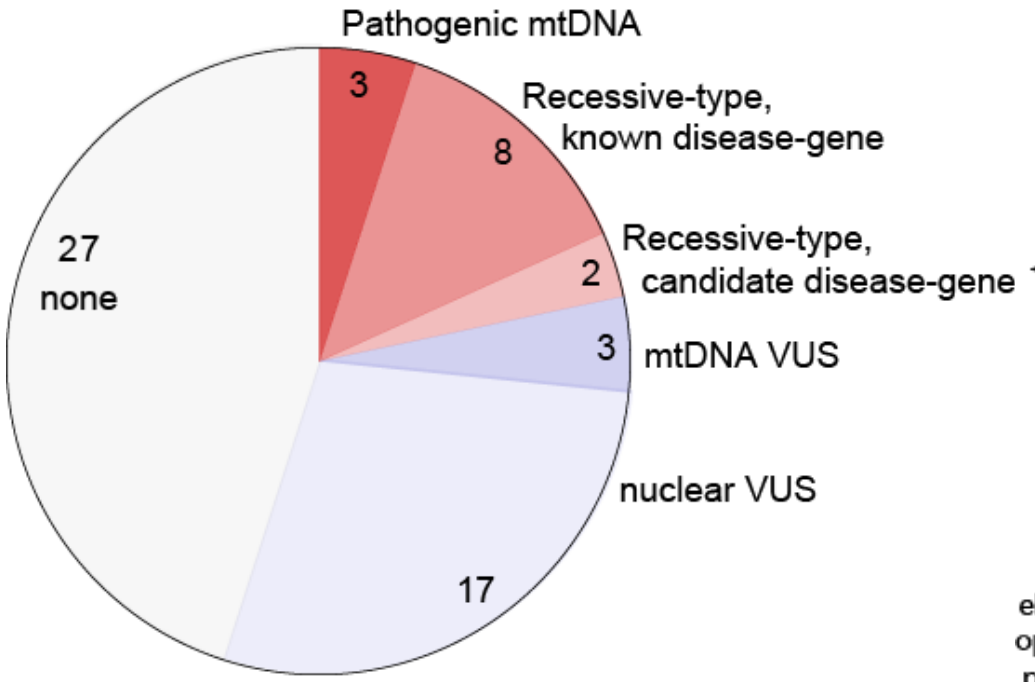
Can we identify likely deleterious variants?



Can we find causal mutations in patient cohort?

Expect: known pathogenic mtDNA variants

recessive-type variants in known nuclear genes

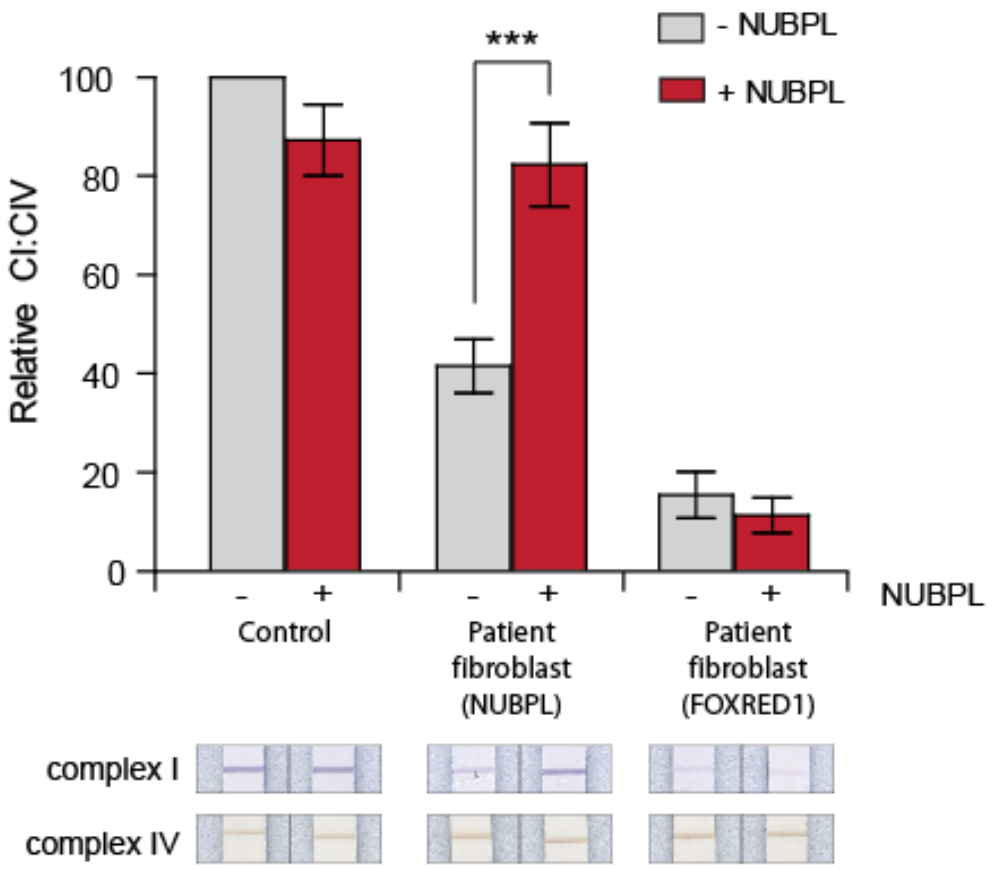


60 patients with complex I deficiency

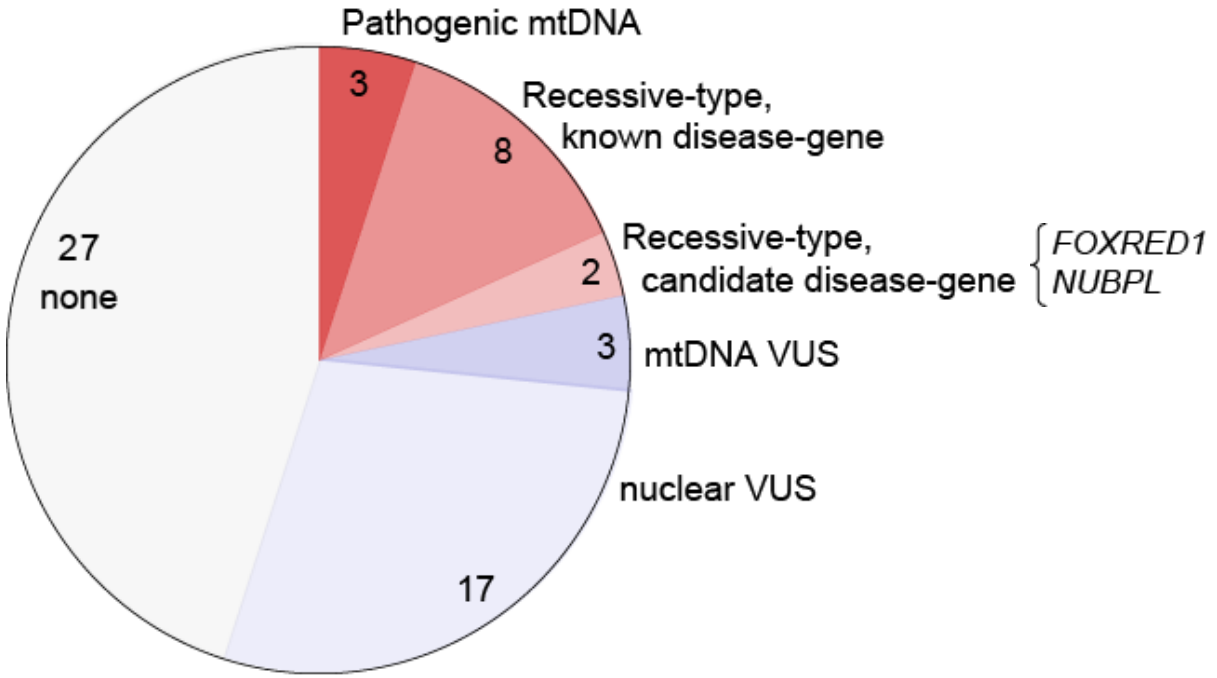
FOXRED1 (compound het)
NUBPL (homozygous)

How can we establish pathogenicity of novel genes?

Patient fibroblasts show complex I defect



Can we find causal mutations in patient cohort?



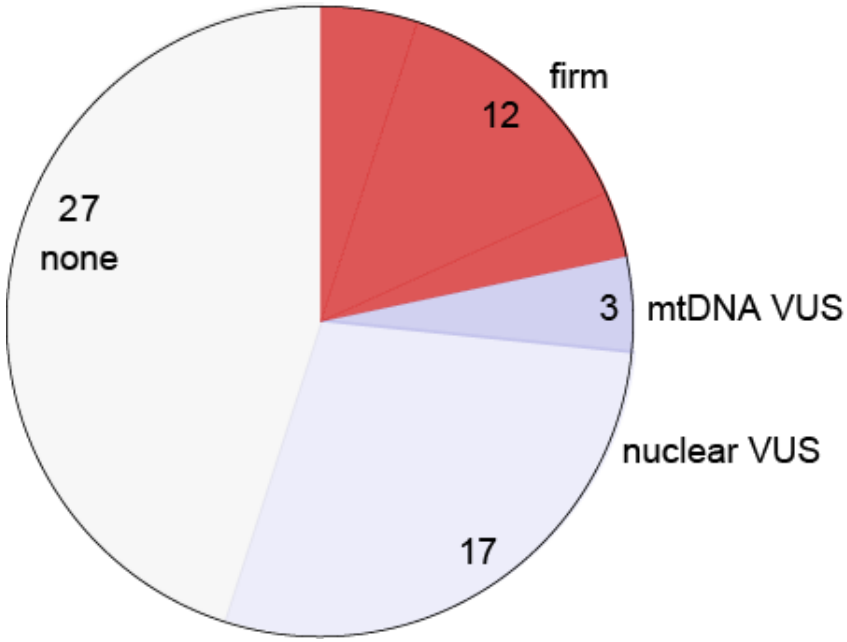
60 patients with complex I deficiency

Can we find causal mutations in patient cohort?

Diagnosis of 22% of patient cohort
2 novel disease-related genes
5 novel mutations in known disease genes

What about rest?

- non-targeted gene?
- non-coding region?
- poor coverage?
- less likely deleterious?
- dominant?
- synergistic variants?



60 patients with complex I deficiency

Mito10K summary

Mito10k: targeted resequencing of ~100 candidate genes x 100 patients

- cost-effective
- diagnosis in 22% of undiagnosed patients
- 2 novel disease-related genes
- Key: availability of cellular models

Available methods/tools

- Syzygy (http://www.broadinstitute.org/ftp/pub/mpg/syzygy/syzygy.zip)
- Definition of ‘likely deleterious’ (Appendix A in this presentation)
- IGV (http://www.broadinstitute.org/igv/)

Outline

Snapshot of Mito10k (100 candidate genes x 100 patients)

Lessons to share

1. sample preparation
2. validation of rare variants
3. cost
4. phasing

Sample preparation: what if I don't have enough DNA?

Issue: need 3ug DNA for sequencing

Method: whole genome amplification (WGA)

- pros: works well (straight-forward, inexpensive)
- cons: possible mutations
  - unequal amplification of regions
  - unequal amplification of alleles
  - rolling PCR artifact (if DNA slightly degraded)

Observations:

- Pooling: unequal amplification mtDNA → pools with 96% mtDNA from 1 person
- Single-sample: rolling PCR → exons with excessive coverage, false SNP/indel calls

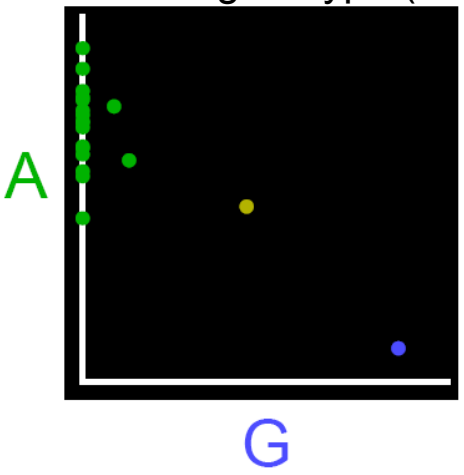
Conclusions: WGA works, however avoid if possible

How can I validate rare variants?

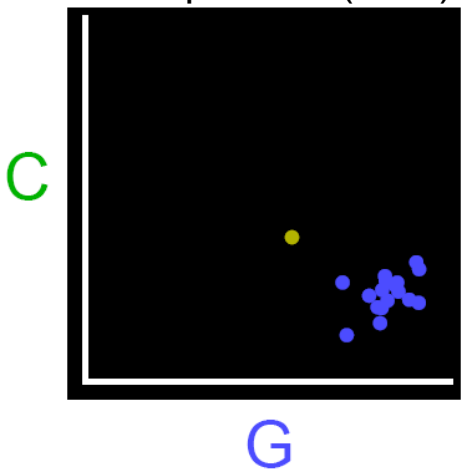
Method: genotype via Sequenom or BeadXPress

Assess Sequenom accuracy: Sanger validation of 82 DNA sites

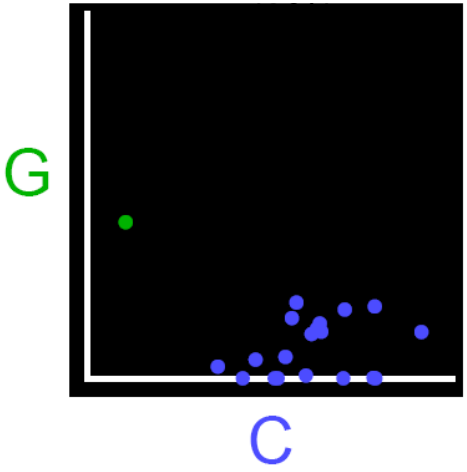
Accurate genotype (88%)



False positive (10%)



Hom/Het Miscall



Caveats (not necessary Sequenom error):

- WGA (mutation or unequal amplification of alleles)
- Sequenom design params (# samples iPLEXed)
- Sample mixups

Conclusions: care with design of rare variant validation

How should I minimize cost in project design?

Observation from Mito10K pilot:

- 7 Illumina lanes (~$10K)
- 40,000 Sequenom genotypes (~20K)
- Months data analysis

Conclusions: to minimize analysis costs: design fewer data tranches, larger # samples

How can I phase rare heterozygous variants?

Issue: If interested in mutations consistent with recessive inheritance,
need to determine if 2 observed rare hets are:
a) in-phase (same parental chromosome), or
b) compound heterozygous (different parental chromosomes)

Method:
common variants: computational methods (eg imputation)
rare variants: genotype familial DNA or experimental method to “phase” variants

Conclusions: obtain parental DNA during sample collection OR
consider experimental methods to phase

Acknowledgements





Mootha Laboratory
Vamsi Mootha
Olga Goldberger

Genome Analysis Platform
Stacey Gabriel
Noel Burtt
Candace Guiducci
Gabe Crawford
Rob Onofrio
Michelle Redman
Entire Genetic Analysis platform

Biological Samples Platform
Scott Mahan
Kristin Ardlie
Susan Flynn
Entire BSP platform

Sequencing Platform
Jen Baldwin
Jane Wilkinson
Lauren Ambrogio
Entire Seq platform

Genome Analysis Toolkit Team
Mark DePristo
Eric Banks
Andrey Sivachenko

Pooled Sequencing Analysis Team
Manuel Rivas
Jason Flannick
Mark Daly
James Pirruccello
Moran Cabili
Team PooledSeq

Murdoch Children's Research Institute
David Thorburn
Elena Tucker
Alison Compton

Patients!

Funding: NIGMS, NHGRI

Appendix A: definition of likely deleterious variants

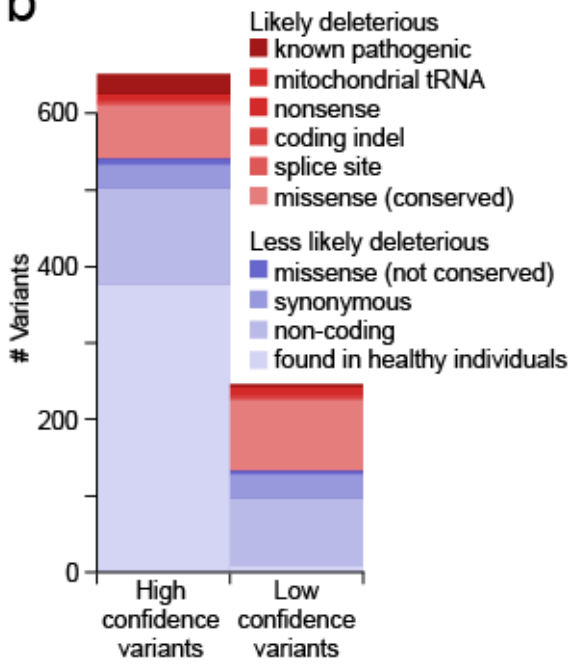
For the Mito10K project, we used the following algorithm to identify variants most likely to underlie a rare, devastating phenotype

1. Include all disease-associated variants (Human Gene Mutation Database & curation)
2. Next, exclude variants present in healthy individuals, based on
    1. dbSNP129
    2. 1000 genomes project, pilot 1
    3. 42 HapMap controls also sequenced
    4. mtDNA variants present in mtDB with minor allele frequency > 0.005
3. Next, include:
    1. coding indel
    2. nonsense variant
    3. missense variant at an amino acid conserved in ≥10 aligned vertebrate species, based on the multiz44way genome alignments (UCSC), or predicted as 'damaging' by PolyPhen-2.0 (HumVar training data)
    4. splice-site (splice acceptor sites -1,-2,-3, and splice donor sites -1,1,2,3,5 selected based on training data consisting of all 8189 HGMD disease-associated splice variants)
    5. tRNA variants in mitochondrial genome
    6. 5' UTR variants which create or disrupt an AUG alternate start codon

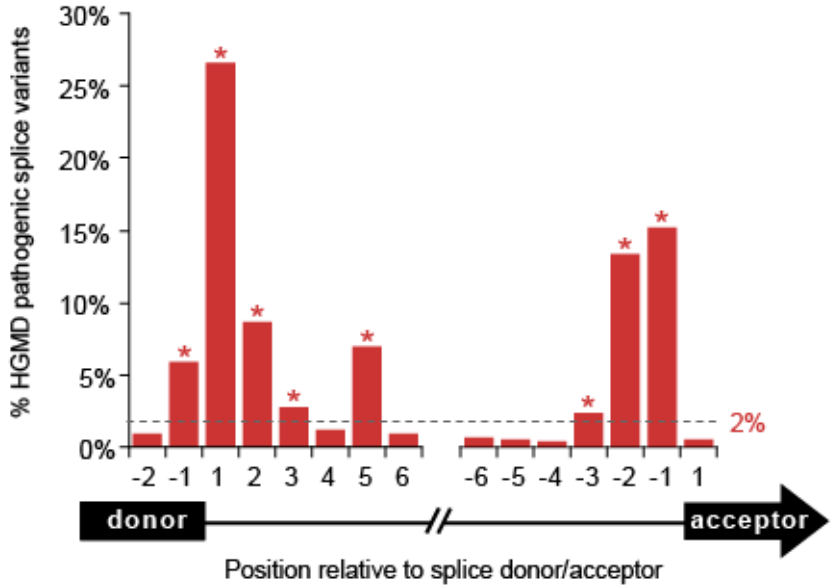
a

Variant type	High confidence variant calls			Low confidence variant calls		
	Discovered in patients	Likely deleterious	Validated	Discovered in patients	Likely deleterious	Validated
nDNA						
nonsense	3	2	1	5	5	1
missense	131	60	51	97	86	9
splice	78	28	22	40	16	2
synonymous	92	0	0	33	0	0
UTR	214	0	0	71	0	0
coding indel	3	3	3	0	0	0
mtDNA						
nonsense	0	0	0	0	0	0
missense	37	14	12	0	0	0
synonymous	85	0	0	0	0	0
noncoding	9	2	2	0	0	0
Total	**652**	**109**	**91**	**246**	**107**	**12**

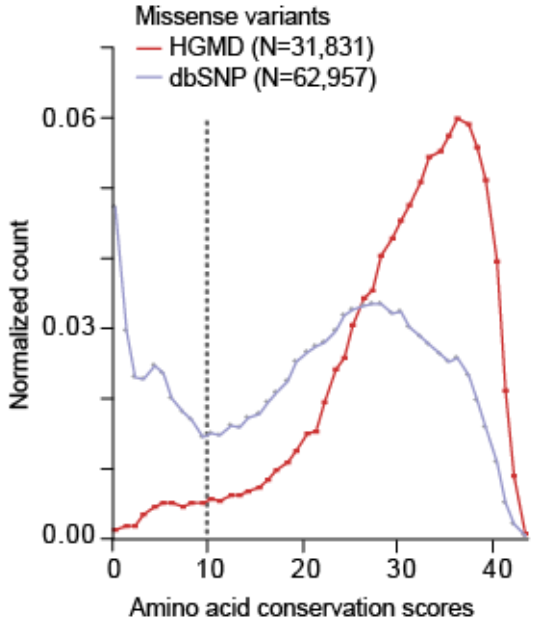
b



c



d



Genetic diagnosis of CI cohort (94 unrelated individuals)