Advancements in Genomic Analysis at Children’s Mercy Kansas City

Shane Corder
HPC Systems Engineer
Children’s Mercy Kansas City
Center for Pediatric Genomic Medicine
Children’s Mercy Research Institute
Advanced Research Computing Team

Survey

~8000 known genetic diseases

- affects 1 in 30 children in the US
- causes 1 in 6 children’s hospital admissions
- causes 1 in 5 deaths in the roughly 60,000 babies born in the Kansas City area

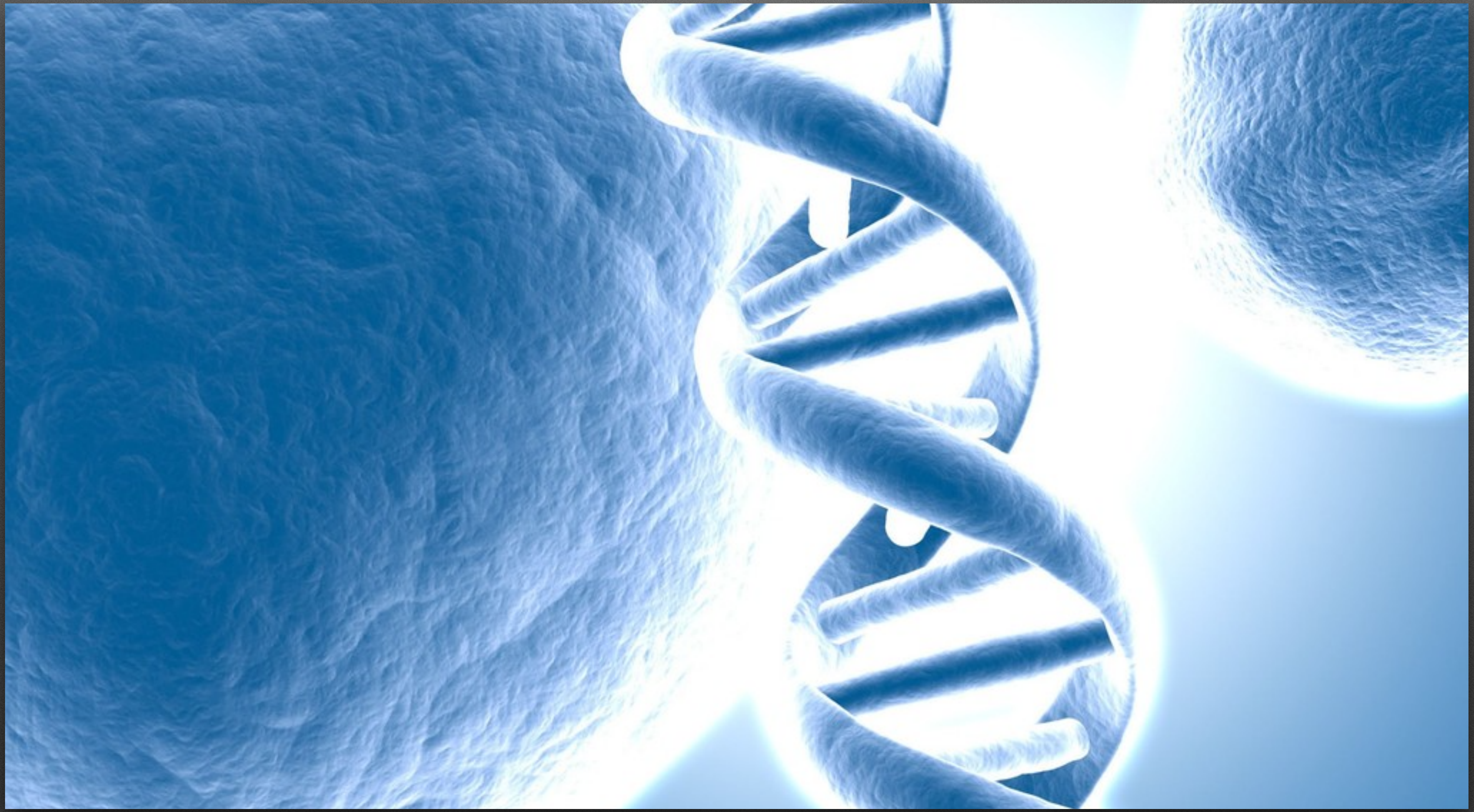
The hard facts

- We know the genetic cause of <5000 of these diseases
- Diagnosis often takes years
- Diagnosis often impacts treatment and always impacts families
- Imagine….

A critically ill newborn…

A sick 10 year old with muscle weakness…

A mother with no hope for an answer…

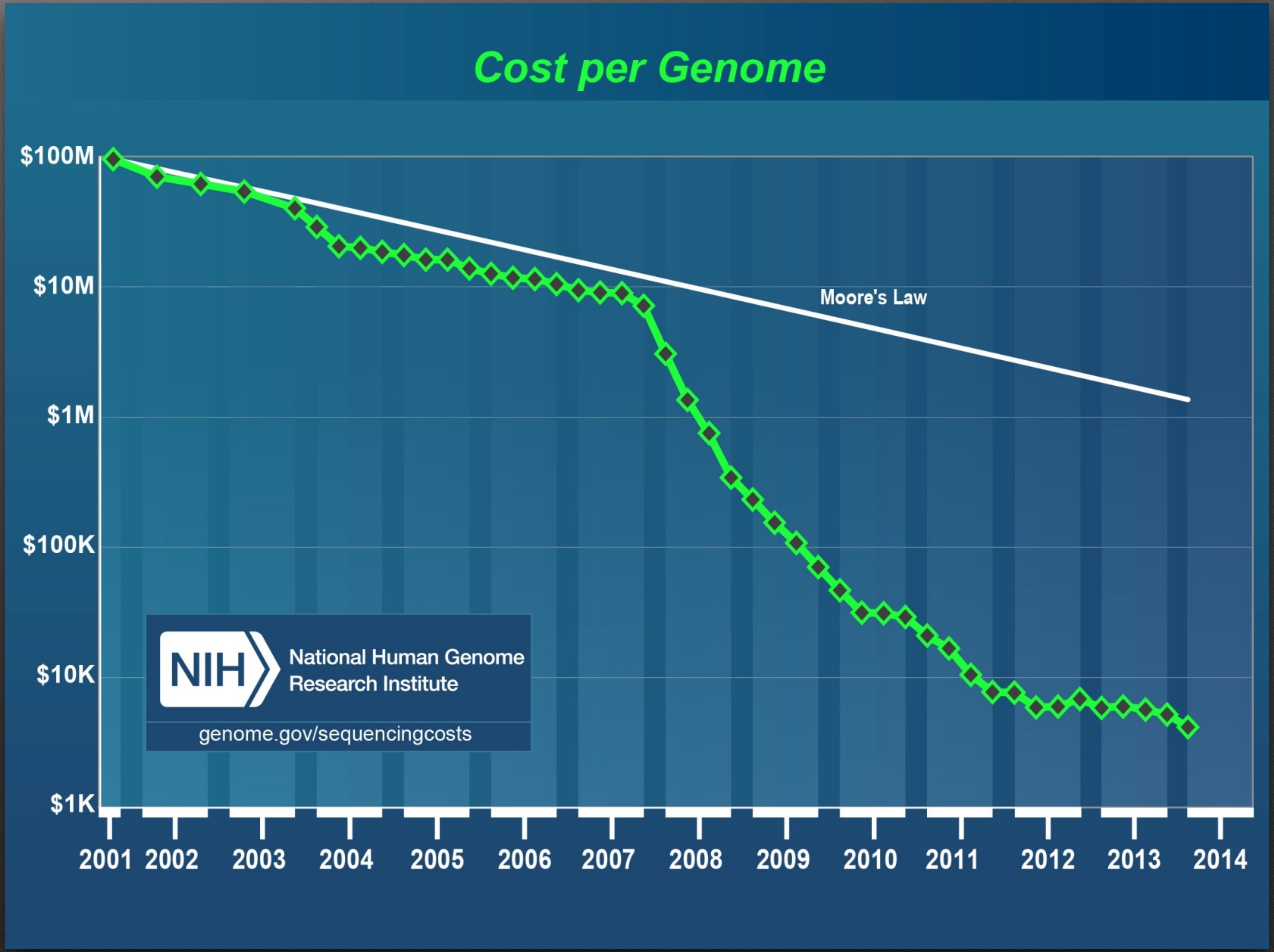


The Human Genome

6.4 billion letters in pairs
19,000 genes - coding for roughly 100,000 proteins

Decoding the Genome

- Human Genome Project generated first "draft" in 15 months
    - Generating the sequence draft cost $300 million
- Later released final sequence in 2003
    - Draft to final sequence cost an additional $150 million
- HGP was a 13 year project costing roughly $3 billion
- Today a HiSeq 4000 produces 16 human genomes in 3 days
    - Reagent costs of <$1,600 per genome



Data Deluge

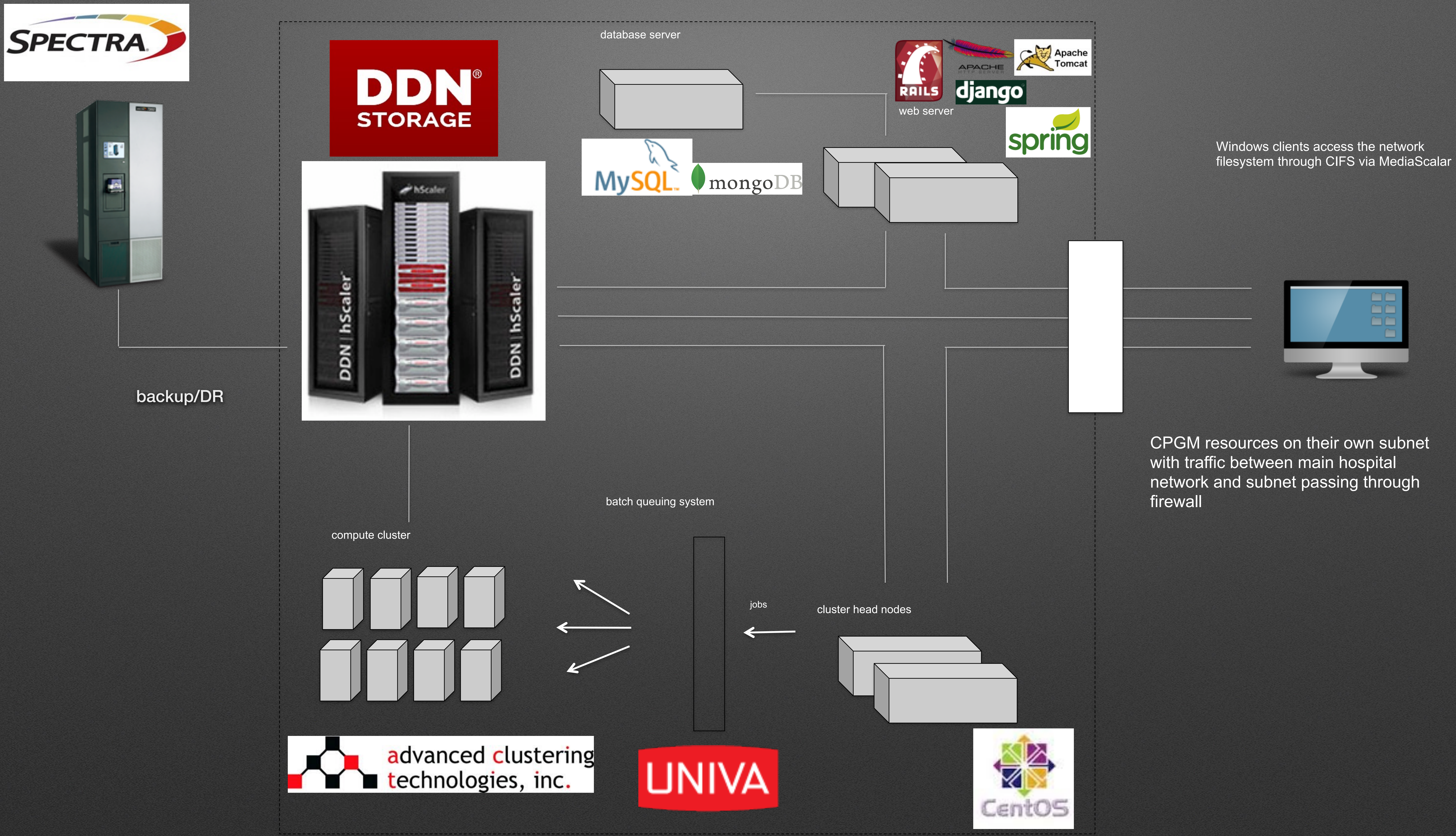
TRANSIENT DATA	1.62TB	primary analysis
	301GB	secondary data
PERMANENT DATA	104GB	Fastq
	71GB	BAM
	1.2GB	VCF - variants only
	825MB	annotated variant file
TOTAL	177GB	per genome

The center is capable of generating 64 whole genomes every 6 days.

- Compute
  - Pinnacle Flex blades
  - 900 core

- Storage
  - DDN GS7K
  - WOS (In the works)

Genome Center Network



database server

web server

Windows clients access the network filesystem through CIFS via MediaScalar

backup/DR

CPGM resources on their own subnet with traffic between main hospital network and subnet passing through firewall

batch queuing system

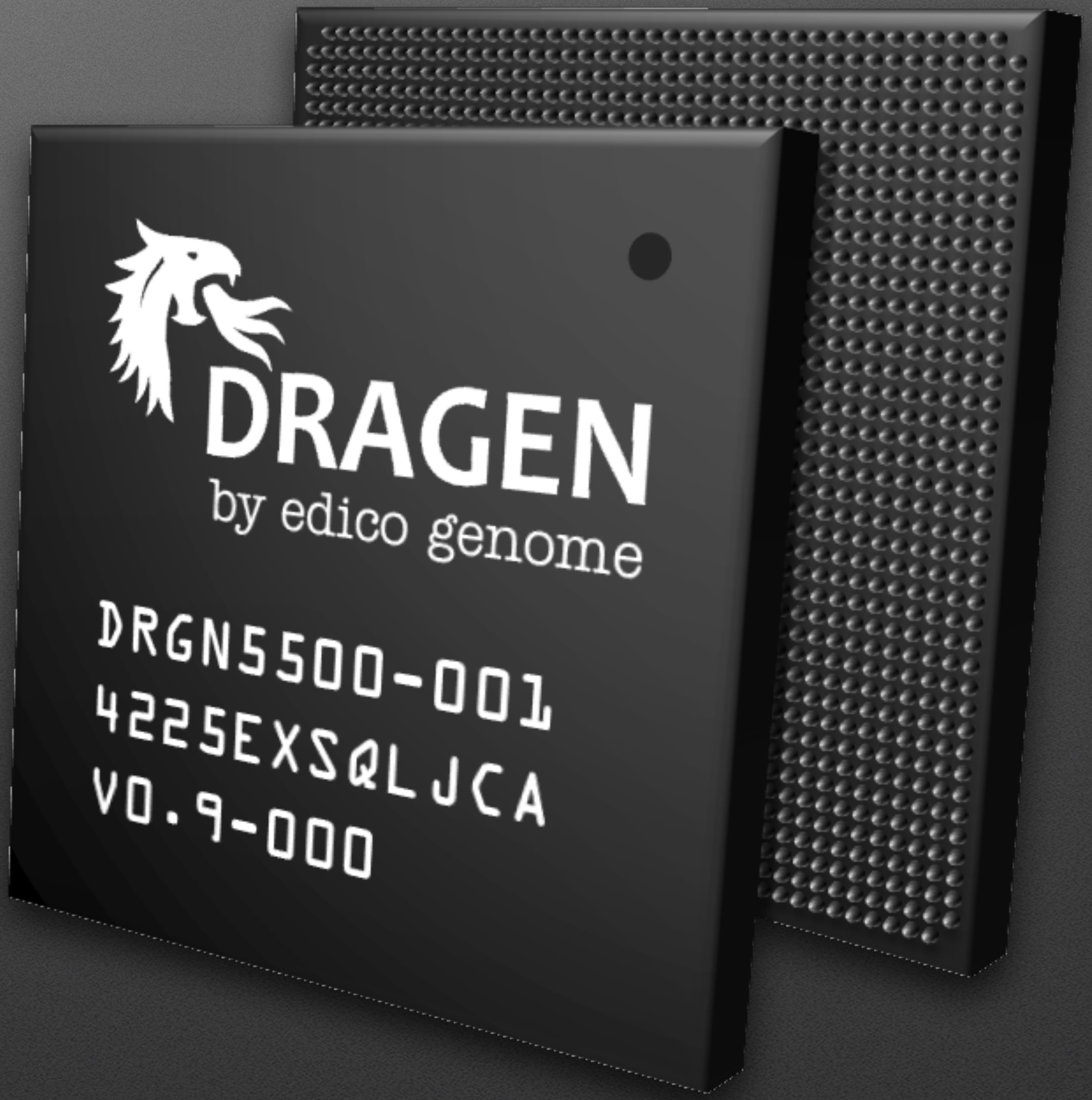
compute cluster

jobs

cluster head nodes

DRAGEN

- Intel E5-2690 v3 @ 2.60GHz
  - w/ HT for 48 core
- 128GB of RAM
- 120GB Intel SSD
- 2x 1.6TB NVME drives
  - RAID0 @ 3.2TB for staging
- 10GbE



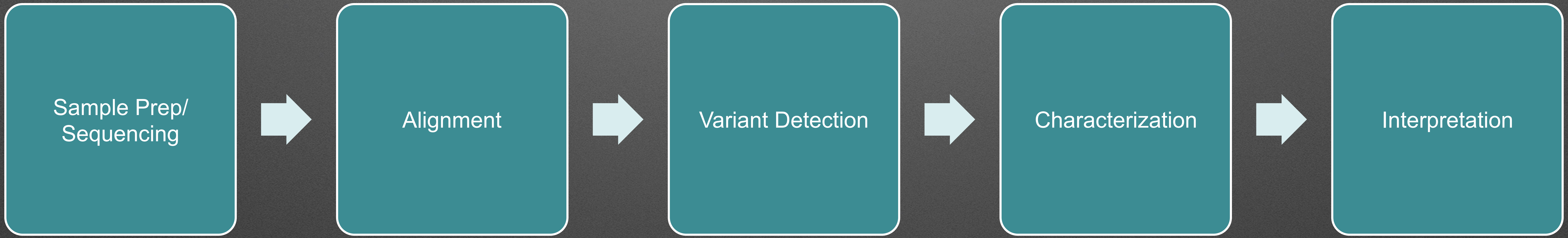
DRAGEN

Method	Sample	DNA Isolation, QC & Shearing	Library Prep	Library QC	SBS	Alignment	Variant Calling	RUNES Variant Annotation	VIKING Provisional Diagnosis	Total
Published WGS$_{50}$	Multiple[c]	2:30	3:15	1:30	25:30	14:40		2:30	0:05	50:00
SBS$_{18}$, GSNAP/GATK	5006-01	2:30	3:15	1:30	19:45	22:30		0:29	na	49:59
WGS$_{26\,,}$ SBS$_{18}$ & Dragen	UDT_173	2:30	3:02	1:30	17:58	0:15	0:15	0:34	0:04	26:08
WGS$_{26,}$ SBS$_{18}$ & Dragen	UDT_103	2:30	3:05	1:30	18:25	0:19	0:22	0:31	0:05	26:47
WGS$_{26}$, SBS$_{18}$ & Dragen	NA12878	2:30	3:15	1:30	18:00	0:19	0:22	0:33	n.a.	26:28
WGS$_{26}$, SBS$_{18}$ & Dragen	NA12878	2:30	3:15	1:30	18:36	0:10	0:11	0:35	na	26:47

Sample	Yield (GB)	Pipeline	Reads Aligned	Alignments with mapQ > 20	Variants Called	Analytic Sensitivity	Analytic Specificity
NA12878	133	DRAGEN	99.4%	95.48%	4,782,970	99.93%	99.87%
		GSNAP/GATK-1.6	98.5%	96.33%	5,343,988	99.54%	98.57%
NA12878	65[a]	DRAGEN	97.7%	91.31%	4,633,357	99.42%	99.46%
		GSNAP/GATK-3.2	96.2%	92.86%	4,571,157	97.29%	95.35%
UDT_173[b]	106	DRAGEN	99.5%	94.92%	4,742,150	96.13%	97.74%
		GSNAP/GATK-1.6	99.3%	96.88%	4,294,504	88.54%	98.06%

Stat-Seq - Rapid Medical Genome Sequencing

1. Identify candidate patient
2. Parental consent
3. DNA Sample



~~30~~ hours   ~~15~~ hours   ~~3~~ hours   ~~1.5~~ hours   .5 hours

23   .50   .25   .5

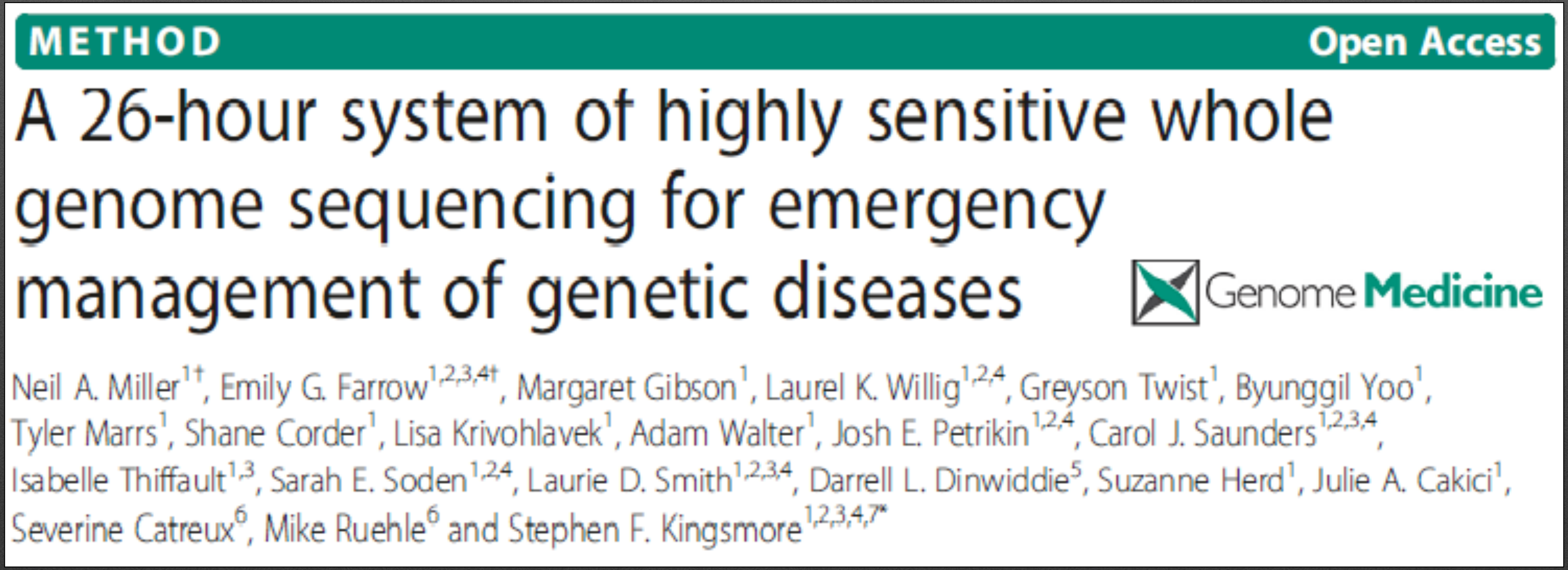
26

= ~~50~~ hours

Other benefits:

- Sure it’s fast and gives great accuracy…
- We’ll lessen our development load - mainly on our variant detection pipeline
- The Dragen can take BCL or Fastq files
- Our current max sequencing load can’t touch this thing
- Frees up our compute cluster to develop new things and for other compute heavy jobs to be scheduled

Does it make a difference?



METHOD Open Access

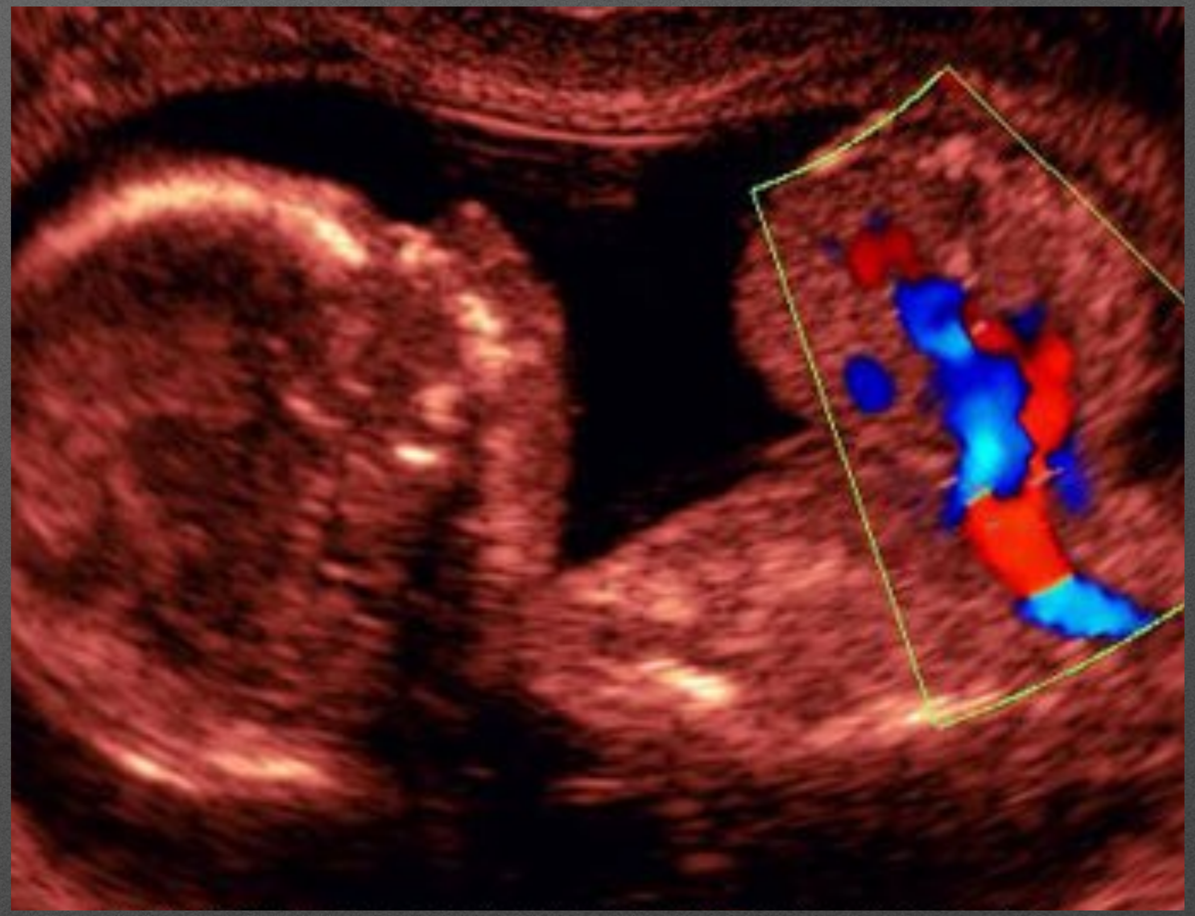
A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases

Genome Medicine

Neil A. Miller[1†], Emily G. Farrow[1,2,3,4†], Margaret Gibson[1], Laurel K. Willig[1,2,4], Greyson Twist[1], Byunggil Yoo[1], Tyler Marrs[1], Shane Corder[1], Lisa Krivohlavek[1], Adam Walter[1], Josh E. Petrikin[1,2,4], Carol J. Saunders[1,2,3,4], Isabelle Thiffault[1,3], Sarah E. Soden[1,2,4], Laurie D. Smith[1,2,3,4], Darrell L. Dinwiddie[5], Suzanne Herd[1], Julie A. Cakici[1], Severine Catreux[6], Mike Ruehle[6] and Stephen F. Kingsmore[1,2,3,4,7*]

Patient CMH000487

- Fetal MRI: Several congenital anomalies
- Delivery in the CMH materno-fetal health center
- Admitted to the NICU
- Acute liver failure @ 2 months of age
- Cause unknown despite extensive testing



Diagnosis and treatment change

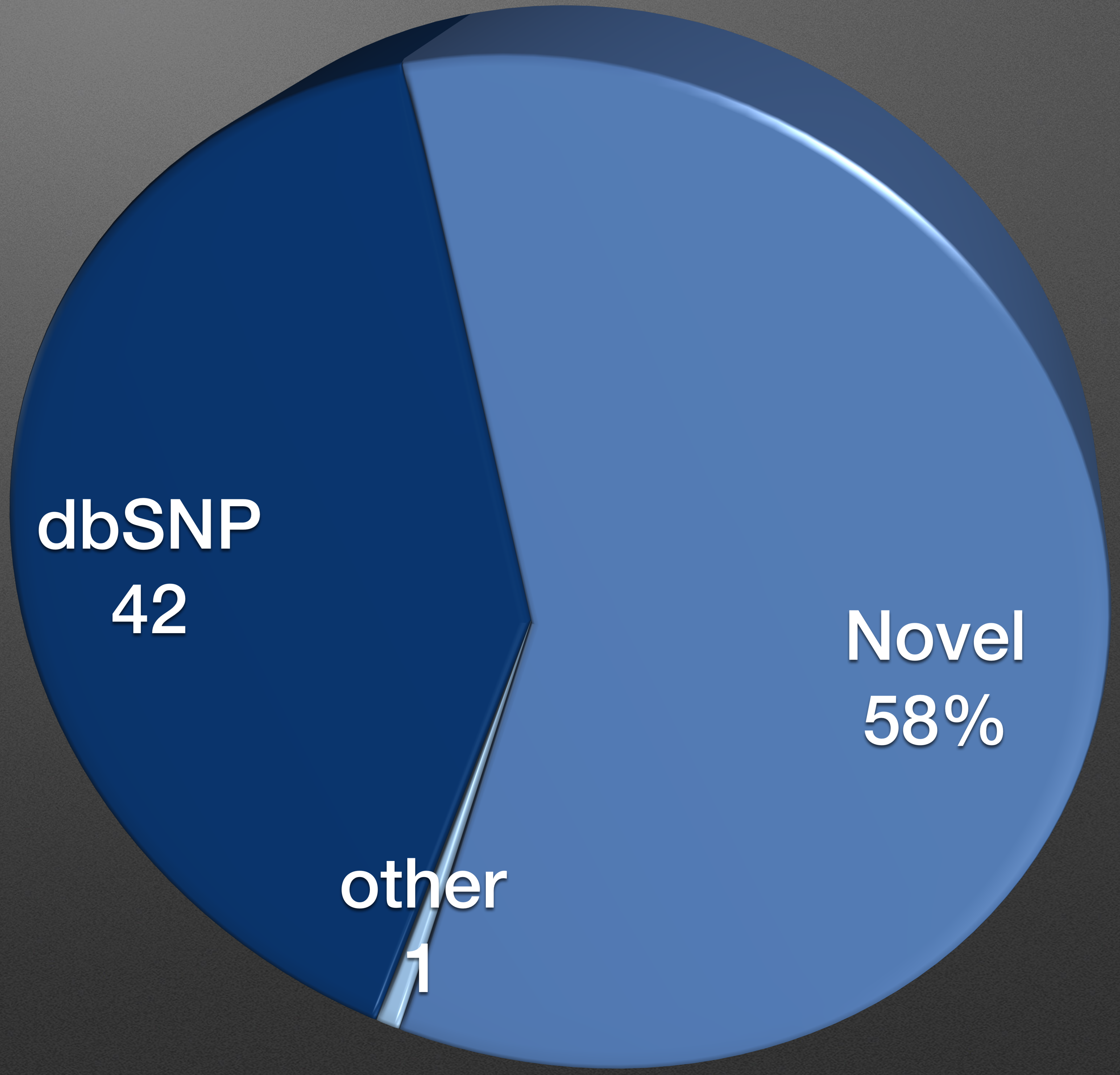
- Following testing and confirmation - IV corticosteroids & immunoglobin
- Liver function returned to normal and baby got to go home

What are we really accomplishing here?

- Rapidly diagnose critically ill patients
- Create a paradigm shift
- End the diagnostic odyssey
- Enable powerful treatment options
- Identify genes for which no other test exists
- Discover new disorders and disease genes

Majority of variants are novel

Majority of variants not observed in other public databases



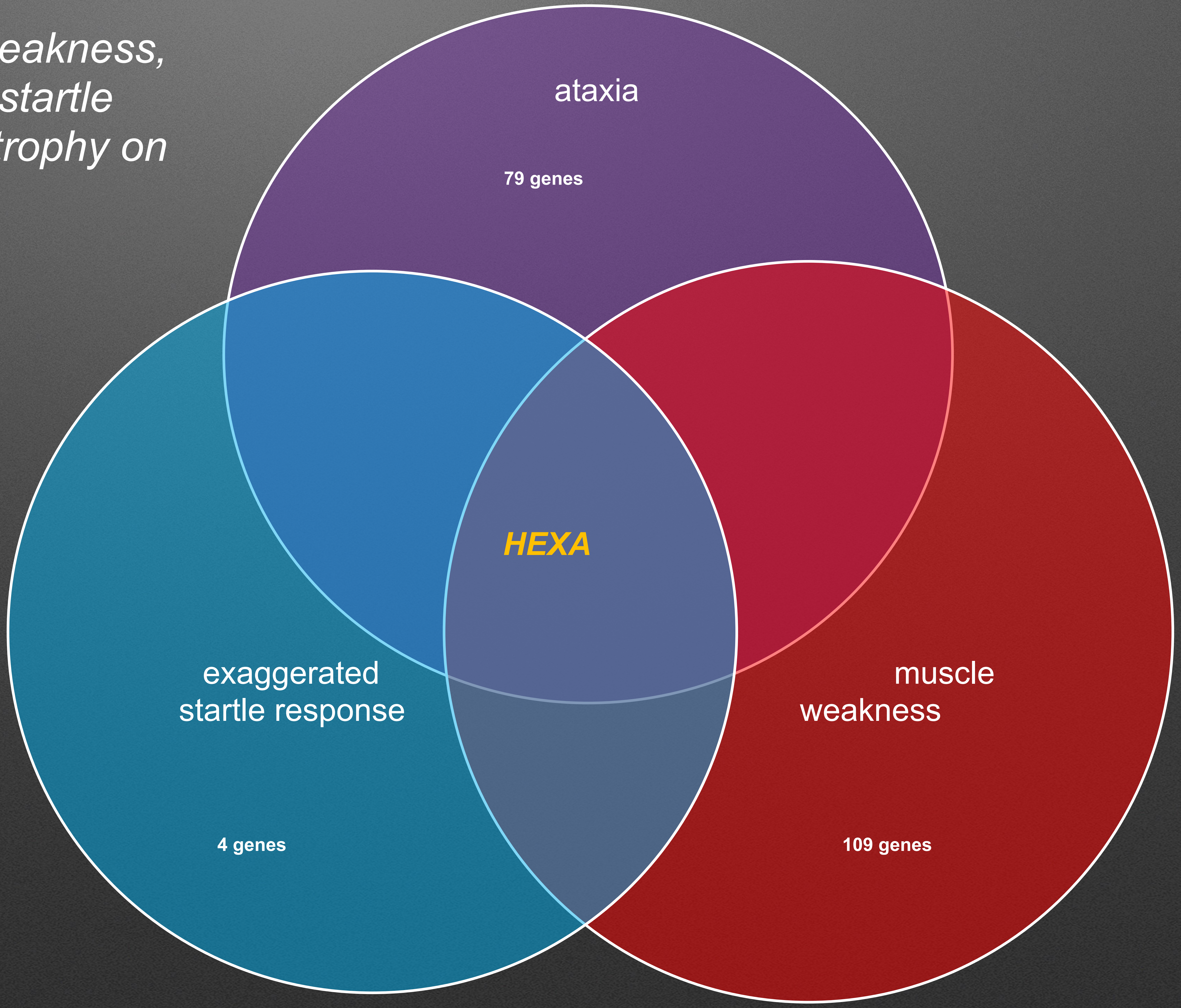
The nordic suite

SSAGA - clinical presentation

- Maps clinical symptoms to diseases to genes using standardized vocabularies
  - SNOMED, Human Phenotype Ontology
- Nominates superset of clinically relevant genes
- Shortens list of variants and candidate genes
- Standardizes clinical information to assist in interpretation of results
- Archival of patient clinical features

The nordic suite

13 mo. male with muscle weakness, developmental regression, startle reflex, ataxia and leukodystrophy on brain MRI…



The nordic suite

RUNES - variant characterization

What does that mean???

- Variant calls must be evaluated to determine their functional consequence
- Characterization done through prediction tools and cross-referencing with external databases
- Final ACMG variant score

annotations

affected gene(s)?
known disease gene?
transcript context?
cause loss of initiation
cause loss/gain of stop?
change amino acid?

disrupt translation frame?
disrupt splicing?
has variant been observed before?
known disease causing mutation?
known to be benign?
population allele frequency?

The nordic suite

The CMH Variant Warehouse

- database recording characterized results of every variant observed in the CPGM population
- 140M variants (as of 2016-03-28)
- 4584 patients and family members
  - 1803 TaGSCAN
  - 2315 exome
  - 447 whole genome
- Searchable by gene, category, allele frequency
- Curation tools based on ACMG recommendations for capturing in depth manual analysis

The nordic suite

VIKING - Interpretation

- Whole Genome Scale
- SSAGA and RUNES integration
- Trio/familial/set analysis
- Dynamic filtering
  - Relevant clinical features
  - Variant classification
  - Allele frequency

Free for Academic/Research Use

Variant Warehouse - https://warehouse.cmh.edu
RUNES/VIking - https://www.childrensmercy.org/genomesoftwareportal
SSAGA - https;//ssaga.cmh.edu

bioinformatics@cmh.edu

“The best way to find yourself is to lose yourself in the service of others.”

–Mahatma Gandhi

Thank you.