MERCURY: NEXT-GEN DATA ANALYSIS AND ANNOTATION PIPELINE

David Sexton SFAF June 2012

Great Interest in Exome Interpretation

ORIGINAL ARTICLE

Clinical application of exome sequencing in undiagnosed genetic conditions

Anna C Need,[1] Vandana Shashi,[2] Yuki Hitomi,[1] Kelly Schoch,[2]
Kevin V Shianna,[1] Marie T McDonald,[2] Miriam H Meisler,[3] David B Goldstein[1,4]

Bioinformatics for personal genome interpretation

Emidio Capriotti, Nathan L. Nehrt*, Maricel G. Kann* and Yana Bromberg**

Submitted: 19th September 2011; Received (in revised form): 8th November 2011

An integrative variant analysis suite for whole exome next-generation sequencing data

Danny Challis[1†], Jin Yu[1†], Uday S Evani[1], Andrew R Jackson[2], Sameer Paithankar[2], Cristian Coarfa[2], Aleksandar Milosavljevic[2*], Richard A Gibbs[1,2*] and Fuli Yu[1,2*]

Whole-Exome Sequencing Identifies Compound Heterozygous Mutations in *WDR62* in Siblings With Recurrent Polymicrogyria

**David R. Murdock,[1] Gary D. Clark,[2,3] Matthew N. Bainbridge,[1] Irene Newsham,[1] Yua
Donna M. Muzny,[1] Sau Wai Cheung,[4] Richard A. Gibbs,[1] and Melissa B. Ramocki[2,3,**

Whole-Genome Sequencing in a Patient with Charcot–Marie–Tooth Neuropathy

James R. Lupski, M.D., Ph.D., Jeffrey G. Reid, Ph.D., Claudia Gonzaga-Jauregui, B.S., David Rio Deiros, B.S., David C.Y. Chen, M.Sc., Lynne Nazareth, Ph.D., Matthew Bainbridge, M.Sc., Huyen Dinh, B.S., Chyn Jing, M.Sc., David A. Wheeler, Ph.D., Amy L. McGuire, J.D., Ph.D., Feng Zhang, Ph.D., Pawel Stankiewicz, M.D., Ph.D., John J. Halperin, M.D., Chengyong Yang, Ph.D., Curtis Gehman, Ph.D., Danwei Guo, M.Sc., Rola K. Irikat, B.S., Warren Tom, B.S., Nick J. Fantin, B.S., Donna M. Muzny, M.Sc., and Richard A. Gibbs, Ph.D.

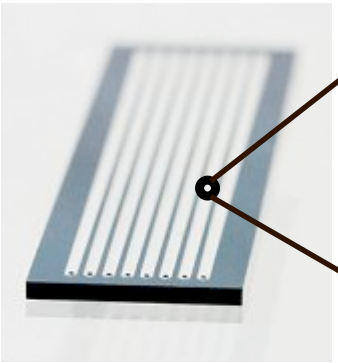
Rational for Automated Pipeline

- Filtering of data to biologically significant targets
- Novel/rare variant detection
- HGSC produces 16-18 Tbases/month
  - WGL can sequence up to 60 Exomes/month
- Quality assurance/control
- Complexity of pipeline
- Manpower

BCM Exome Sequencing Overview

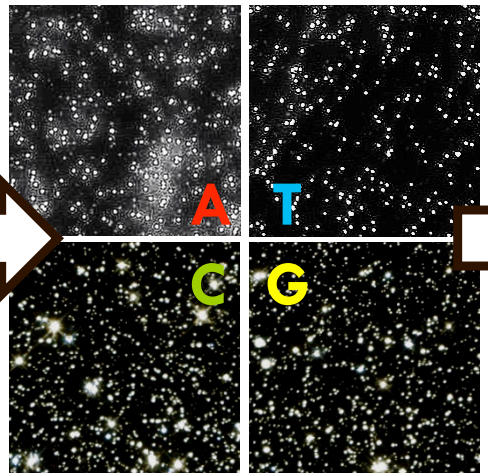
- HGSC Exome Pipeline
  - Now in place at the WGL and HGSC
- Prep Sample DNA
- Enrich for exon sequence – VCrome2.1
- Sequence sample "exome" – HiSeq2000
- Identify variants
- Annotate variants for biological significance
- Provide clinical interpretation through the WGL

Primary Sequence Data Production



…Amplify to form Clusters

Sequencing Instrument



Four-channel Images

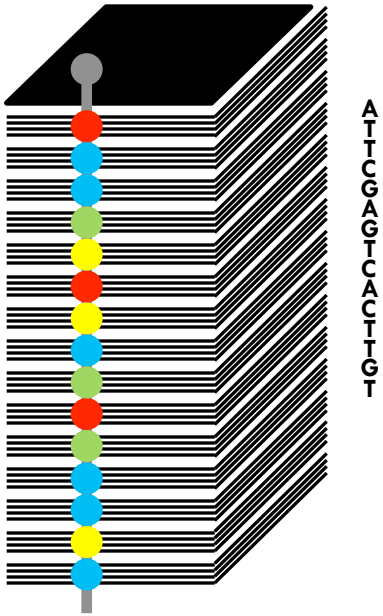
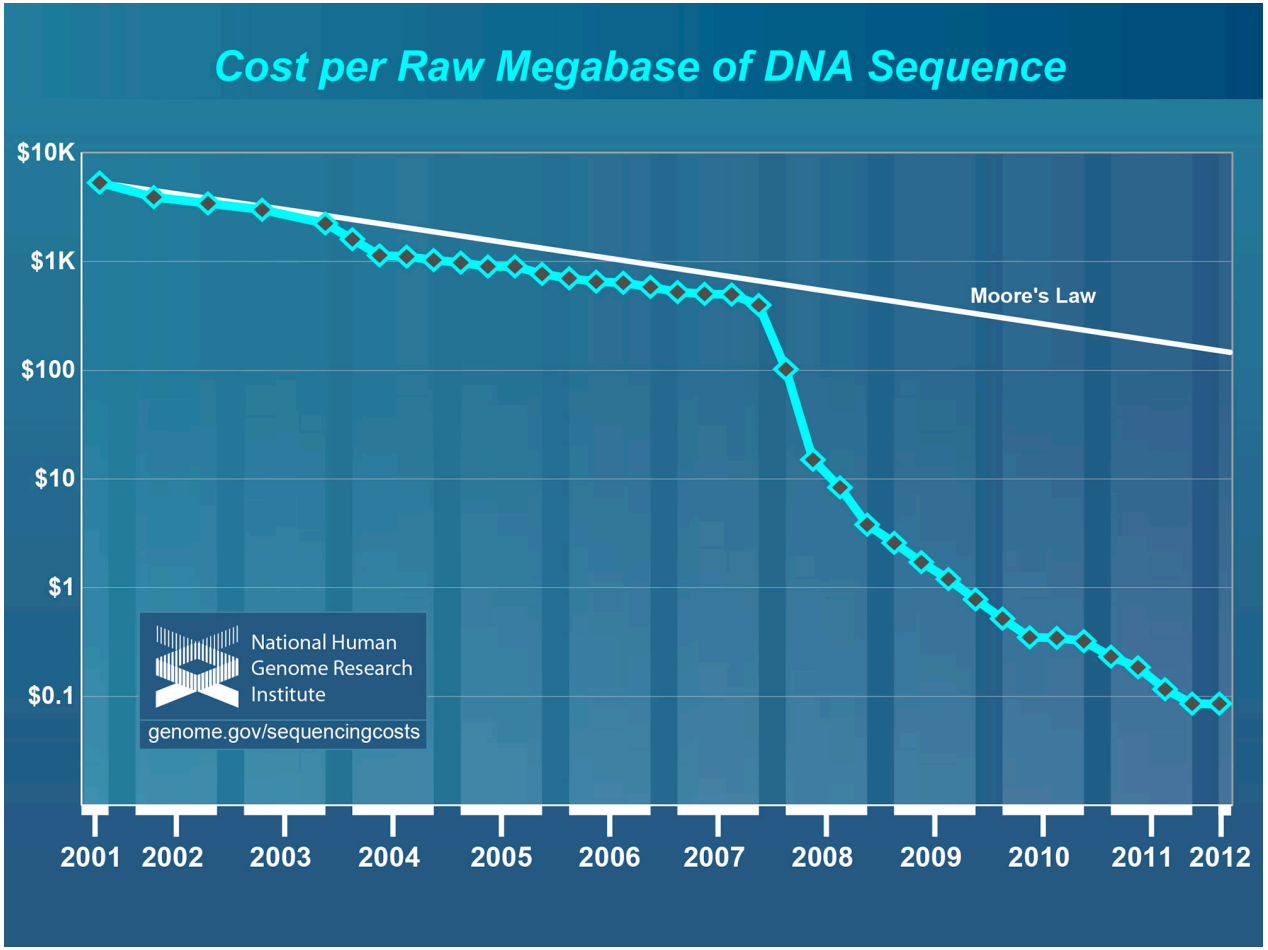


Image Processing & Basecalling

```
@HWI-ST115_0142:1:1:9815:1977#0/1
ATTCGAGTCACTTGTACGGCCTTCACTGACAAAGAGCAGAGTGTAGTGTTGGAT
+HWI-ST115_0142:1:1:9815:1977#0/1
BIIQNPMMOMYZZY[]\Y\]]___[_`^^`^``Z``^_^_UVVMURTPPP_ZZZ
```

Sequence Reads & Qualities (Basecall Confidence)

Obligatory Cost Per Megabase



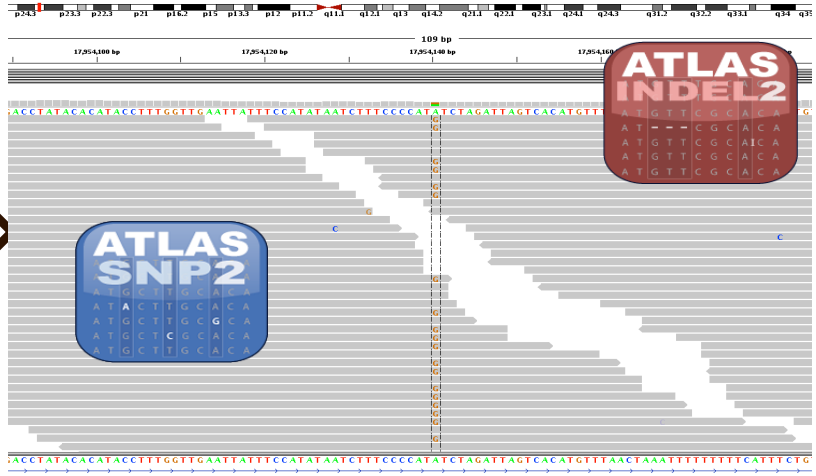
Pipeline Design Requirements

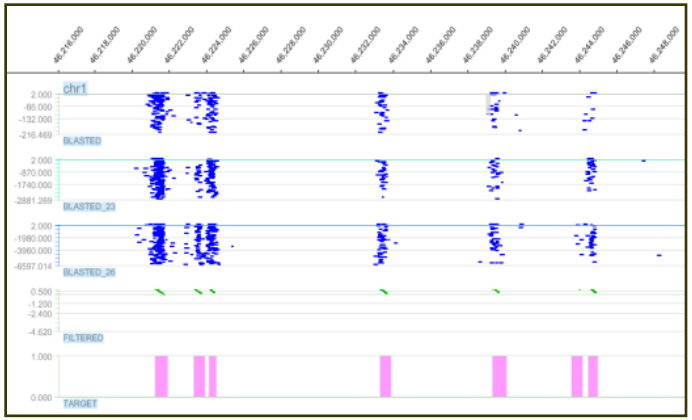
- Modularity
  - Allow components to be substituted at will
- LIMS compatibility
  - Need to communicate with LIMS
- Automated
  - As little human intervention as possible
- Hardened
  - Graceful error catching with alerts
- Code reusability - Ruby

Sequence Data Analysis

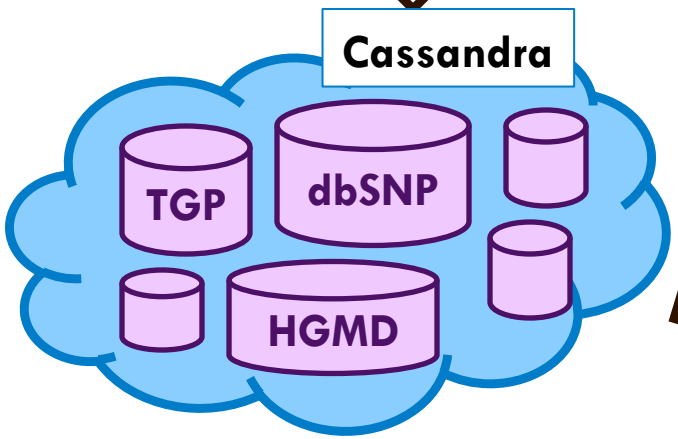
~10Gbp of Reads

Production &
Data QA/QC





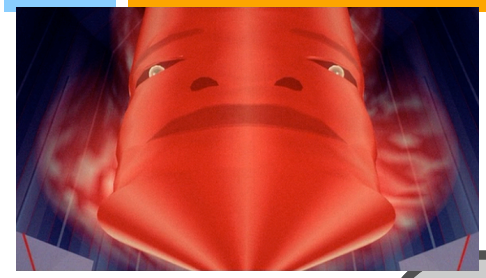
Align to Reference Genome

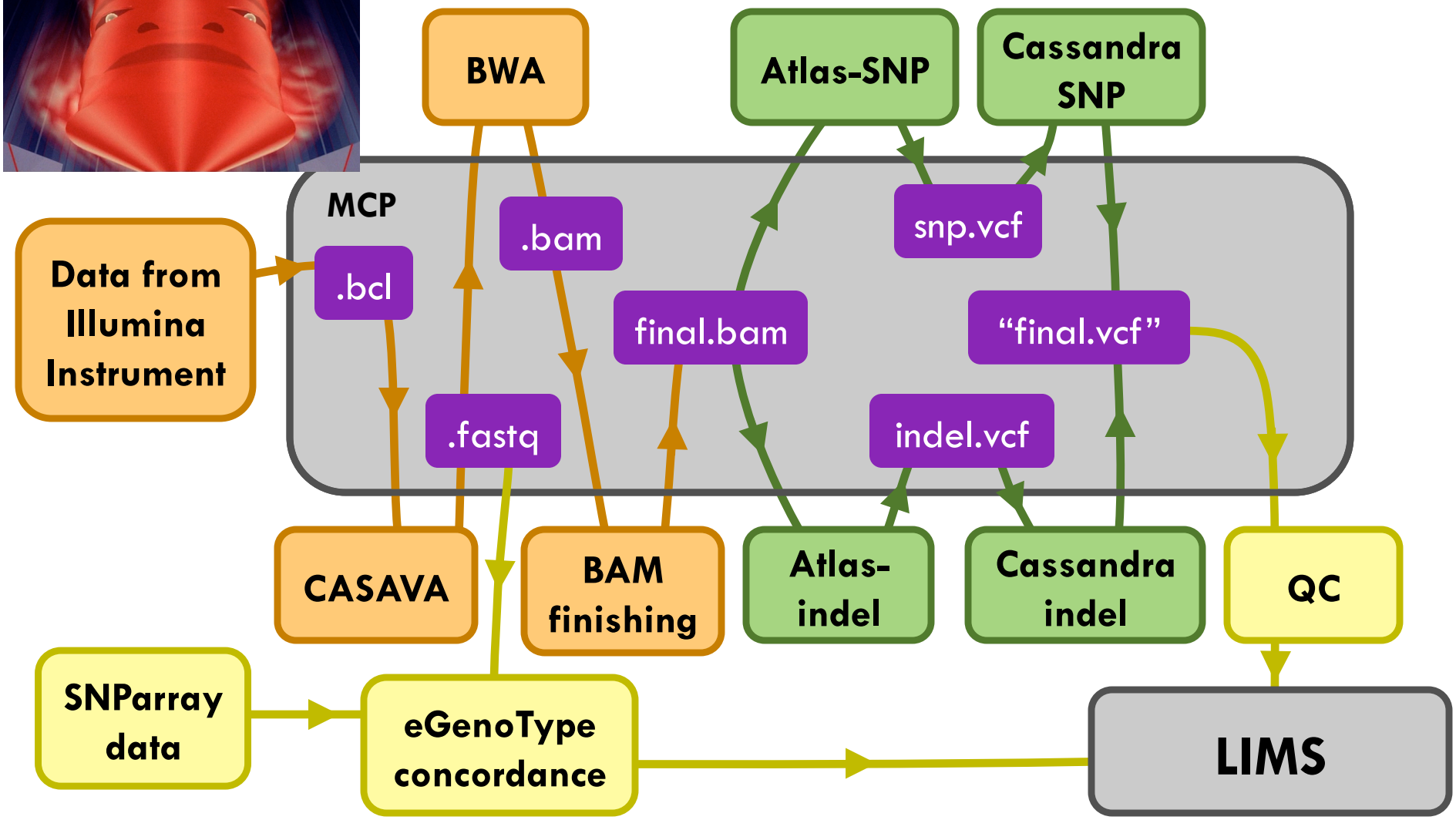


Annotate Variants

Call Variants and
Estimate Quality

Mercury Pipeline





Variant Calls & Annotation

- Variant Calls with **Atlas** Suite (**Fuli Yu**)
  - Logistic regression & Bayesian framework for variant quality
  - Adjustable (very light in WGL) heuristic filtering
  - Genotyping
  - Available on HGSC website & Genboree workbench
- Annotation with **Cassandra** (**Matthew Bainbridge**)
  - Queries databases of known variation
  - Provides 'Pileup String', Atlas QC measures
  - Identifies effect on genes & key regions
  - Includes gene function information to aid interpretation

Databases Used in Cassandra

SwissProt	SwissProt data, funcational description, Expression disease association
Phase1MAF EXOME	MAF data from Thousand genomes exome
Phase1MAF WGS	MAF data from Thousand genomes low cov whole genome
UW MAF	MAF data from the UW Variant Exome Server
CG MAF	MAF data from the complete genomics 75 whole genomes
dbNSFP	deleteriousness data from dbNSFP
Mappability	The mapability of the genome from UCSC
HGMD snp	SNV data from HGMD
dbSNP clinical	Clinically implicated variants from dbSNP
dbSNP	dbSNP variants
ESE indel	Indels from the HGSC database ESE
HGMD indel	Indels from HGMD
dbSNP indel	Indels from dbSNP
TG indel exome	Indels from Thousand Genomes exome project

Technical Replication

Sample #	Test1	Test1 (%)	Test1 and 2	Test1 and 2(%)	Test2	Test2(%)
HS-1011	133	0.561%	23,320	98.409%	244	1.029%
HS-1015	155	0.542%	28,312	98.910%	157	0.548%
HS-1016	155	0.644%	23,752	98.732%	150	0.624%
HS-1017	105	0.456%	22,767	98.768%	179	0.777%
HS-1018	165	0.693%	23,531	98.795%	122	0.512%
HS-1019	162	0.682%	23,441	98.686%	150	0.631%
HS-1020	493	2.041%	23,518	97.355%	146	0.604%
HS-1021	161	0.681%	23,188	98.125%	282	1.193%
Average	**191**	**0.718%**	**23,979**	**98.473%**	**179**	**0.613%**

Possible Sources of Error

- ~1% of sequenced bases are errors
- Repeats, pseudogenes, etc. lead to misplaced reads (mismapping)
- Uneven sampling (statistical sampling errors) lead to undercalled alleles
- Annotation databases imperfect & incomplete
- Imperfect phenotyping can be misleading

Computational Requirements

	Casava		BWA		GATK/Picard/SAMtools			Atlas		Cassandra
	bcl to fastq	final fastq	BWA align	build BAM	Merge	recal/ relgn	pileup	SNP call	in/del call	Annotation
nodes	1	0.25	2	1	1	0.5	0.5	1	1	1
RAM(GB)	28	8	28	28	28	28	8	8	8	16
hours	8	1	12	24	3	11	1	10	4	1
node*hrs	8	0.25	24	24	3	5.5	0.5	10	4	1

HGSC Compute Cluster

- 370 Nodes
- 3040 processors
- Moab/Torque
- 4PB Storage

We Need a Bridge

Example Report

Table 1. Deleterious Mutations in Disease Genes Related to Clinical Phenotype

Disease	Inheritance Pattern	Gene	Isoform	Nucleotide	Amino Acid	Zygosity	References /Comments	Relevance of the Variant to Patient's Current Symptoms
Congenital Variant of Rett Syndrome	Autosomal Dominant	FOXG1	MN9999.9	c.111G>T	p.G37X	Heterozygous	PMID:488888	Likely to be relevant

Table 2. Variants of Unknown Clinical Significance in Disease Genes Related to Clinical Phenotype

Disease	Inheritance Pattern	Gene	Isoform	Nucleotide	Amino Acid	Zygosity	References /Comments	Relevance of the Variant to Patient's Current Symptoms
None Detected								

Table 3. Medically Actionable Deleterious Mutations in Disease Genes Unrelated to Clinical Phenotype

Disease	Inheritance Pattern	Gene	Isoform	Nucleotide	Amino Acid	Zygosity	References /Comments	Variant Classification/Notes
Neurofibromatosis Type 1	Autosomal Recessive	NF1	NM_000492.3	c.3846G>A	p.W334X	Heterozygous	PMID:2236053	Deleterious
Lynch Syndrome	Autosomal Dominant	PMS2	NM_000535.5	c.99G>T	p.R33X	Heterozygous	PMID:4444444	Deleterious

Table 4. Carrier Status for Recessive Mendelian Disorders

Disease	Inheritance Pattern	Gene	Isoform	Location	Nucleotide	Amino Acid	Zygosity	References /Comments
Cystic Fibrosis, Congenital Absence of the Vas Deferens, CFTR-Related Hereditary Pancreatitis	Autosomal Recessive	CFTR	NM_000492.3	Exon 3	c.3846G>A	p. W1282X	Heterozygous	PMID:2236053

Table 5. Pharmacogenetic Profile Variant Alleles (optional)

Drug	Allele	Zygosity	References/Comments
Plavix	CYP2C19*17	Heterozygous	PMID 16413245; 17625515; 20083681

Note: An expanded report of the Whole Exome Sequencing Test is available. The Expanded Report will give additional information on mutations and variants in genes which cause disease unrelated to the indication for testing and predicted deleterious mutations in genes with no known current association with disease. If this information is requested by the physician, please complete Requisition Form and Patient Consent for Expanded Report.

Future Directions

- Cancer Pipeline
  - Tumor/Normal, Tumor/Normal/Normal
- More Instrument Integration
- Variant calls for list of VIPs
- Addition of More Annotation Databases
- More Filtering Tools
- Automation of Physician Assisted Diagnosis
  - NLP, Heuristics

Acknowledgements

- Code will be open source.

- Art Beaudet & the whole WGL team
- Jeff Reid, Richard Gibbs, & Donna Muzny
- Eric Boerwinkle, Fuli Yu, & Matthew Bainbridge
- Peter Pham & Mark Wang
- Alicia Hawes, Ziad Khan, & Mike Dahdouli