Mass General Brigham

What data is being generated? Implications for returning results

Matt Lebo, PhD, FACMG

Chief Laboratory Director - Lab for Molecular Medicine
Director of Bioinformatics – Mass General Brigham Personalized Medicine
Associate Professor of Pathology – Brigham and Women's Hospital and Harvard Medical School
Associate Member – Broad Institute of MIT and Harvard

Genomic assays

Genotyping array

- Typically target 500K – 2 million **sites** across the genome
- Can identify other potential variant sites via imputation

Targeted panel

- Typically target 10s-100s of **genes**, though some are in low 1000s
- Can be enhanced to target specific clinically relevant content (e.g., CNVs or deep intronic variants)

Exome

- Targets the ~2% of genome that is coding (i.e., all ~20,000-25,000 genes)

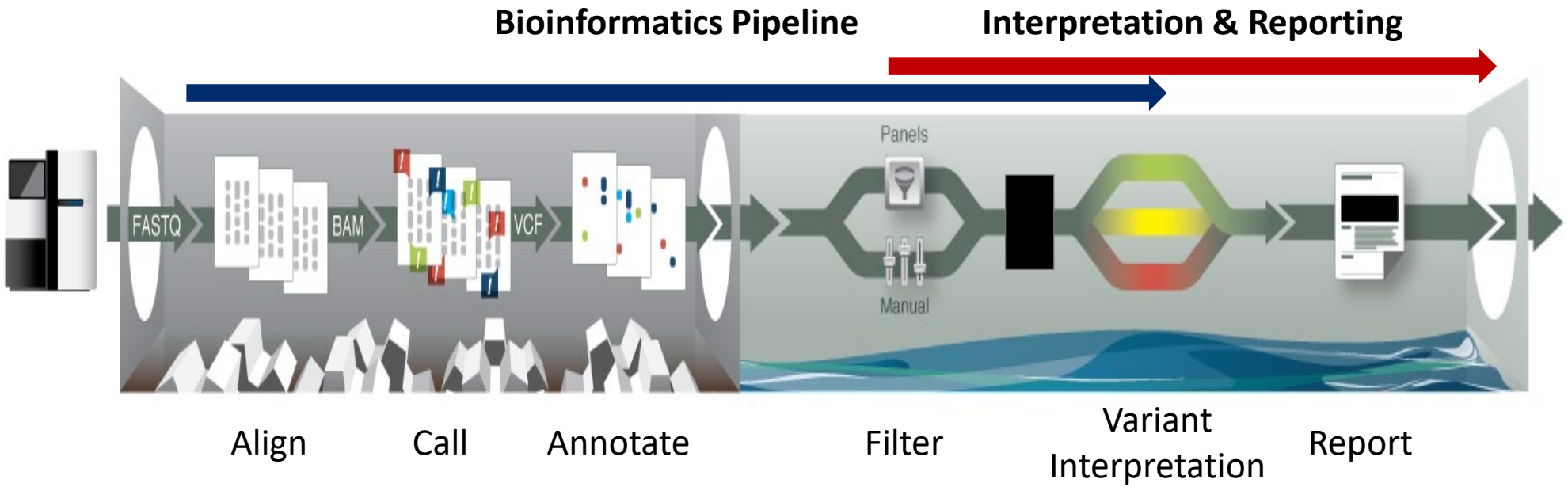
Genome

- Unbiased sequencing of entire genome, including mitochondrial genome

Low-pass genome

- Average coverage varies based upon costs and needs (0.25x to 5x)
- Requires imputation to identify haplotypes and variants

Genomic bioinformatics pipeline



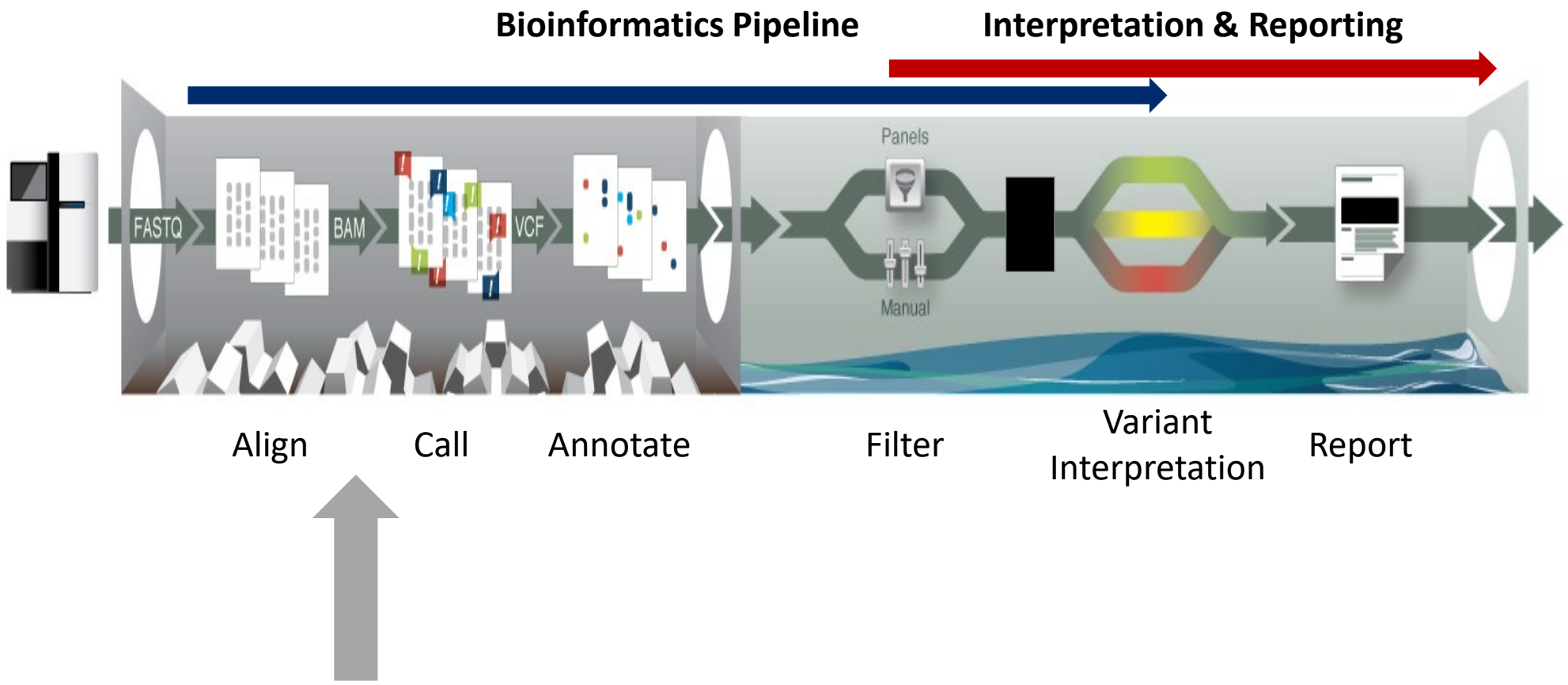
Raw data and storage

	IDAT	CRAM	VCF	Imputed VCF
Genotyping array	60M	NA	20M	750M
Targeted Panel	NA	500M	5M	NA
Exome	NA	5G	100M	NA
Genome	NA	30G	1G	NA

- File sizes may differ, particularly for VCF
  - Joint/merged VCF files may reduce average file size
  - Annotated VCF files will increase file sizes
  - VCF files and annotations may be represented efficiently in a database

*All sizes are estimates and vary on sequencing depth, panel size, and annotations from variant caller

Genomic bioinformatics pipeline



Alignment/Variant calling

Clustering (for arrays)

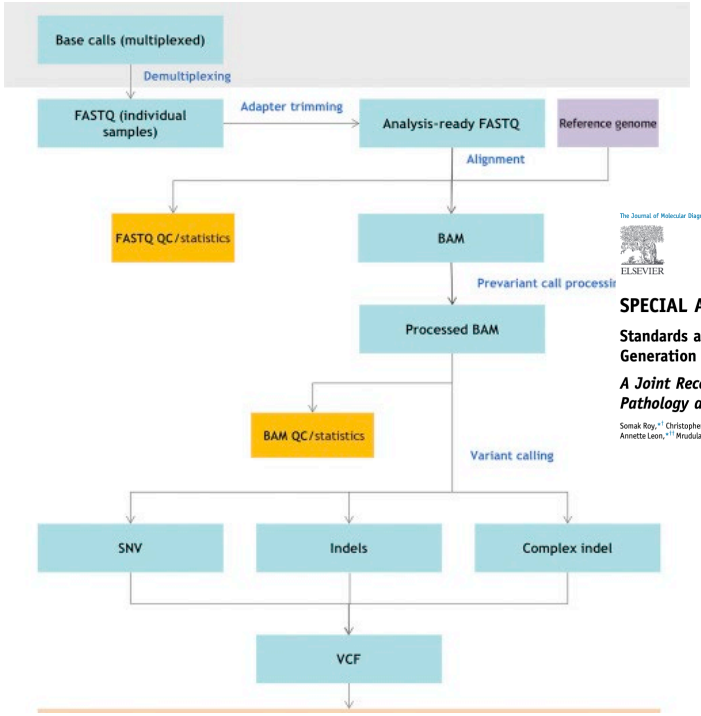
- Typically, not or can't be done for rare variants

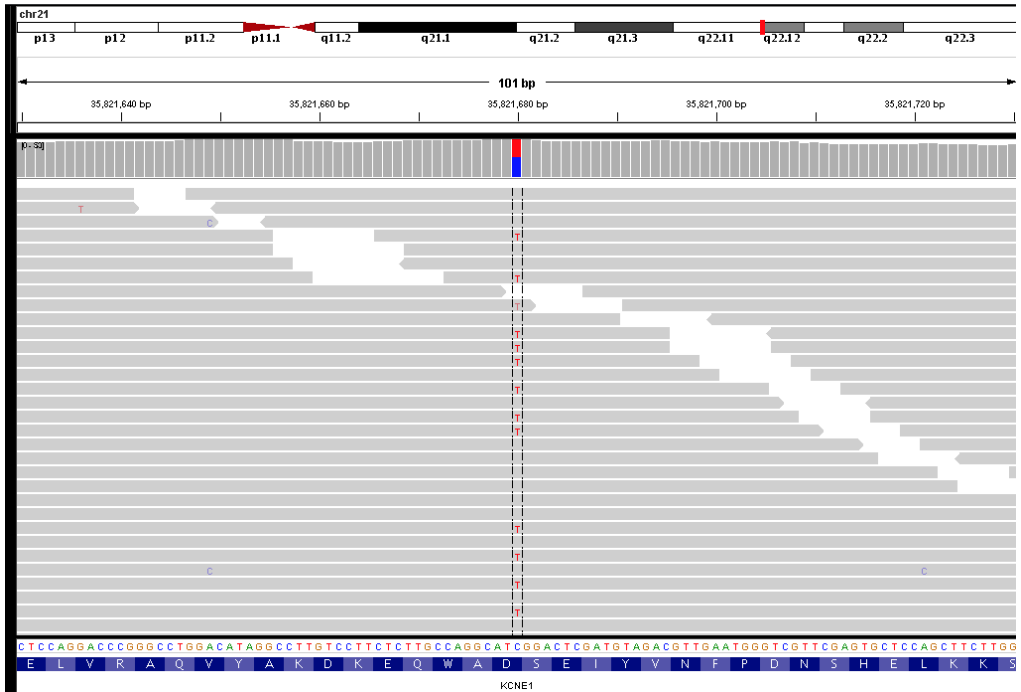
Alignment

- Choice of alignment tool affects accuracy
- Choice of reference genome may affect variant calling and downstream annotations

Variant calling

- Multiple variant callers may be needed to capture all relevant variant types
- Choice of caller affects accuracy





Genomic assays – Relative accuracy

	Common variants	Rare Variants	Copy Number Variants	Structural Variants
Genotyping array	Good, but limited	Poor and limited	Good, but generally for large events	Poor
Targeted Panel	Good	Good	Low-resolution, unless specifically assayed	None
Exome	Good	Good	OK, low-resolution	Poor
Genome	Good	Good	Good, but low PPV for small events	Emerging
Low-pass genome	Good	Poor	Good for larger events	Unknown

*All sequencing methodologies are using short reads

Difficult regions (i.e., missing data)

- Even with generally high accuracy, many critical variants are still challenging to detect
- Important to know where datasets may be incomplete

Clinical Area (total P/LP variants)	Challenging variants (%)	Distribution of variant types: Large indel	Small CNV	Complex Rearrangement	Low Complexity	Segmental Duplication	Putative Mosaic	Unique challenging variants (%)
Carrier (n=48,218)	20.4%							7.3%
Hereditary Cancer (n=44,818)	10.3%							12.5%
Pediatric Genetics (n=8,321)	11.2%							7.8%
Neurology (n=7,934)	19.4%							7.4%
Cardiology (n=7,249)	3.9%							5.0%
Metabolic/Newborn (n=4,332)	3.7%							5.1%
Preventive (n=3,166)	2.1%							9.8%
Immunology (n=2,880)	3.8%							8.4%
Other (n=792)	4.3%							8.3%
TOTAL (n=127,710)	**13.8%**							8.7%

Genetics inMedicine

www.nature.com/gim

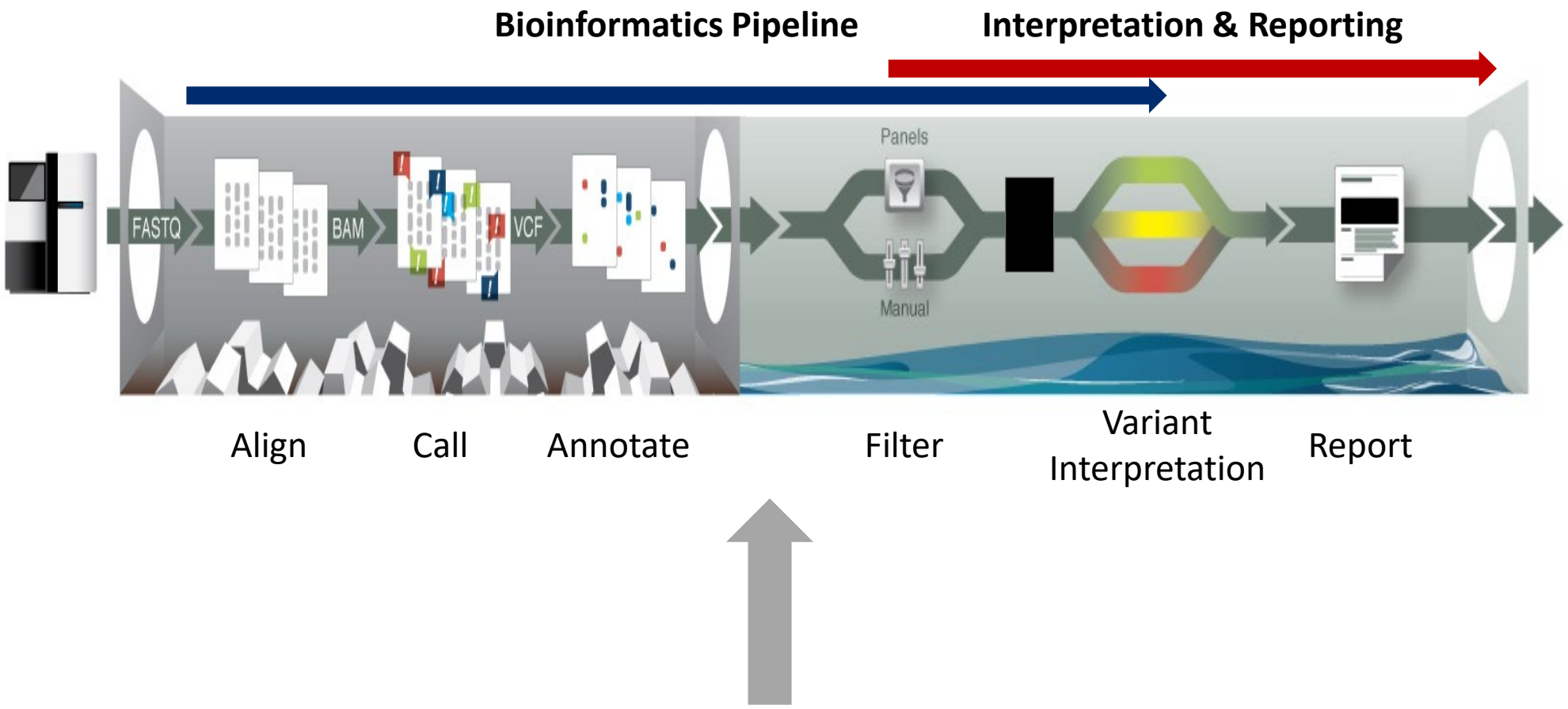
Check for updates

ARTICLE

One in seven pathogenic variants can be challenging to detect by NGS: an analysis of 450,000 patients with implications for clinical sensitivity and genetic test implementation

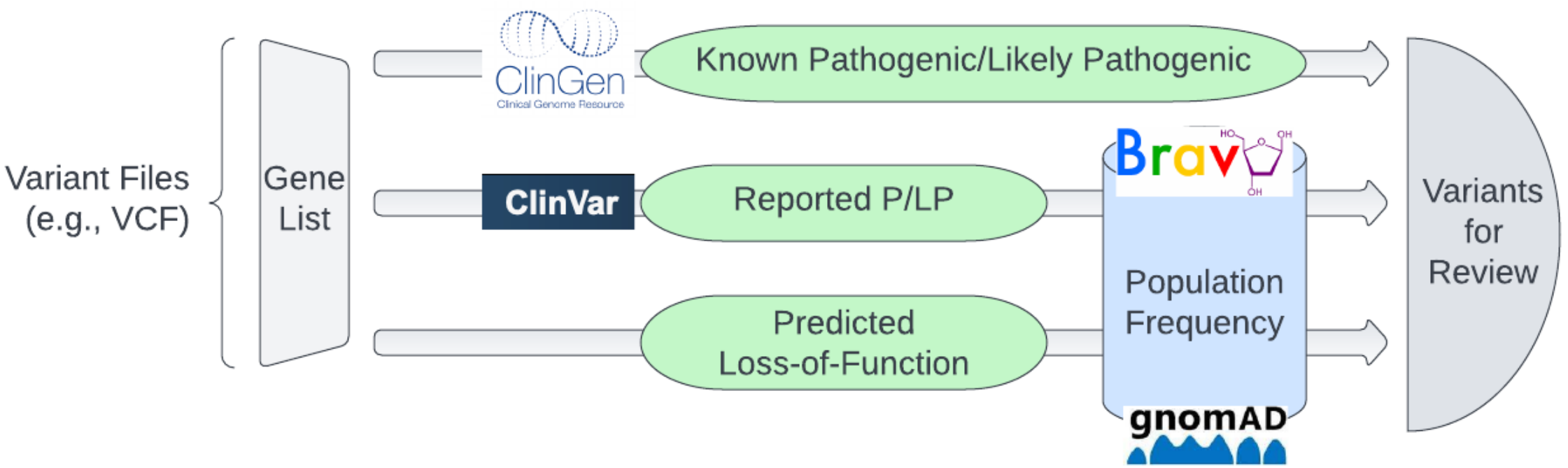
Stephen E. Lincoln[1], Tina Hambuch[1✉], Justin M. Zook[2], Sara L. Bristow[1], Kathryn Hatchell[1], Rebecca Truty[1], Michael Kennemer[1], Brian H. Shirts[3], Andrew Fellowes[4], Shimul Chowdhury[5], Eric W. Klee[6], Shazia Mahamdallie[7,10], Megan H. Cleveland[2], Peter M. Vallone[2], Yan Ding[5], Sheila Seal[7], Wasanthi DeSilva[4], Farol L. Tomson[8,11], Catherine Huang[8], Russell K. Garlick[8], Nazneen Rahman[7], Marc Salit[2,9], Stephen F. Kingsmore[5], Matthew J. Ferber[6], Swaroop Aradhya[1] and Robert L. Nussbaum[1]

Genomic bioinformatics pipeline



Filtration Methods

- Need to annotate and filter variants
- Goal: identification of returnable Pathogenic and Likely Pathogenic variants
  - No prioritization based upon patient phenotype
- Balance sensitivity with PPV



Standards in variant interpretation

Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology

Sue Richards, PhD[1], Nazneen Aziz, PhD[2,16], Sherri Bale, PhD[3], David Bick, MD[4], Soma Das, PhD[5], Julie Gastier-Foster, PhD[6,7,8], Wayne W. Grody, MD, PhD[9,10,11], Madhuri Hegde, PhD[12], Elaine Lyon, PhD[13], Elaine Spector, PhD[14], Karl Voelkerding, MD[13] and Heidi L. Rehm, PhD[15]; on behalf of the ACMG Laboratory Quality Assurance Committee

SCORE:

	Benign				Pathogenic							
	Strong		Supporting		Supporting		Moderate		Strong		Very Strong	
Population data	BA1	MAF too high					PM2	Absent (or rare) in pop db with coverage >20X				
Population data	BS1	MAF too high			PS4_P	Proband Count - Supporting	PS4_M	Proband Count - Moderate	PS4	Case-control OR Proband Count		
Population data	BS2	Observ in unaffected										
Computational and predictive data			BP1	Truncating disease causing; variant missense					PS1	Same AA change as establish pathogenic variant		
Computational and predictive data			BP3	In-frame indel in repeat region w/out known function			PM5	Diff pathogenic missense variant at codon	PM5_S	≥2 diff path missense variants at codon		
Computational and predictive data			BP4	Computational evidence suggests no impact	PP3	Computational evidence suggests impact	PVS1_M	Null variant - Moderate	PVS1_S	Null variant - Strong	PVS1	Null variant & LOF known mechanism
Computational and predictive data			BP7	Silent (or noncons splice) variant with no predicted splice impact			PM4	Protein length changing variant in non-repeat region				
Functional data					PP2	Missense in a gene with low rate of benign missense & path missense common	PM1	Mutation hotspot or fxnl domain				
Functional data	BS3	Estabished fxnl study shows no deleterious effect			PS3_P	Functional assay - Supporting	PS3_M	Functional assay - Moderate	PS3	Established fxnl study shows deleterious effect		
Seg Data	BS4	Lack of segregation in affected			PP1	Coseg with disease Dominant: 3 segs Recessive:	PP1_M	Coseg with disease Dominant: 5 segs Recessive:	PP1_S	Coseg with disease Dominant: 7 segs Recessive:		
De novo data							PM6	De novo (neither paternity or maternity confirmed)	PM6_S	≥2 independent occurences of PM6		
De novo data									PS2	De novo (paternity and maternity confirmed)	PS2_VS	≥2 independent occurences of PS2
Allelic data			BP2	Observed in trans with dominant variant OR observed in cis with path variant			PM3	Detected in trans with path variant (recessive disorders)	PM3_S	2 occurences of PM3	PM3_VS	≥3 occurences of PM3
Other			BP6	ClinVar expert panel = benign	PP5	ClinVar expert panel = pathogenic						
Other data\			BP5	Found in case with an alternative cause	PP4	Patient phenotype or FH high specific for gene						

Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen)

Erin Rooney Riggs, MS, CGC[1], Erica F. Andersen, PhD[2,3], Athena M. Cherry, PhD[4], Sibel Kantarci, PhD[5], Hutton Kearney, PhD[6], Ankita Patel, PhD[7], Gordana Raca, MD, PhD[8], Deborah I. Ritter, PhD[9], Sarah T. South, PhD[10], Erik C. Thorland, PhD[6], Daniel Pineda-Alvarez, MD[11], Swaroop Aradhya, PhD[4,11] and Christa Lese Martin, PhD[1]

Table 1 CNV interpretation scoring metric: copy-number loss

Evidence type	Evidence	Suggested points/case	Max score
Section 1: Initial assessment of genomic content			
Copy-number loss content	**1A**. Contains protein-coding or other known functionally important elements.	0 (Continue evaluation)	0
	1B. Does NOT contain protein-coding or any known functionally important elements.	−0.60	−0.60
Section 2: Overlap with established/predicted haploinsufficiency (HI) or established benign genes/genomic regions *(Skip to section 3 if your copy-number loss DOES NOT overlap these types of genes/regions)*			
Overlap with ESTABLISHED HI genes or genomic regions and consideration of reason for referral	**2A**. Complete overlap of an established HI gene/genomic region.	1.00	1.00
	2B. Partial overlap of an established HI genomic region • The observed CNV does NOT contain the known causative gene or critical region for this established HI genomic region OR • Unclear if known causative gene or critical region is affected OR • No specific causative gene or critical region has been established for this HI genomic region	0 (Continue evaluation)	0
	2C. Partial overlap with the 5′ end of an established HI gene (3′ end of the gene not involved)…	See categories below	
	2C-1. …and coding sequence is involved	0.90 *(range: 0.45 to 1.00)*	1.00
	2C-2. …and only the 5′ UTR is involved	0 *(range: 0 to 0.45)*	0.45
	2D. Partial overlap with the 3′ end of an established HI gene (5′ end of the gene not involved)…	See categories below	
	2D-1. …and only the 3′ untranslated region is involved.	0 (Continue evaluation)	0
	2D-2. …and only the last exon is involved. Other established pathogenic variants have been reported in this exon.	0.90 *(range: 0.45 to 0.90)*	0.90
	2D-3. …and only the last exon is involved. No other established pathogenic variants have been reported in this exon.	0.30 *(range: 0 to 0.45)*	0.45
	2D-4. …and it includes other exons in addition to the last exon. Nonsense-mediated decay is expected to occur.	0.90 *(range: 0.45 to 1.00)*	1.00
	2E. Both breakpoints are within the same gene (intragenic CNV; gene-level sequence variant).	See ClinGen SVI working group PVS1 specifications • PVS1 = 0.90 *(Range: 0.45 to 0.90)* • PVS1_Strong = 0.45 *(Range: 0.30 to 0.90)* • PVS1_Moderate *or* PM4 (in-frame indels) = 0.30 *(Range: 0.15 to 0.45)* • PVS1_Supporting = 0.15 *(Range: 0 to 0.30)*	See categories at left

ClinGen — Get Started · About Us · Curation Activities · Working Groups · Expert Panels · Documents · Tools

Clinical Domain Working Groups

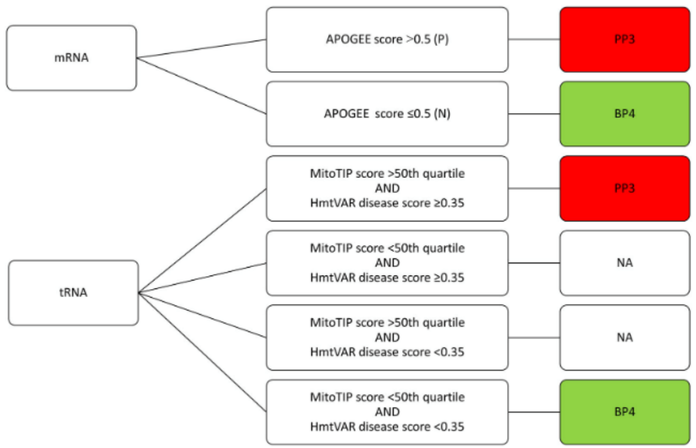
One specific goal of ClinGen is to develop teams of experts in different clinical domains to evaluate the clinical validity of gene-disease relationships and pathogenicity of individual genetic variants. These clinical domain working groups consist of expert clinicians, clinical laboratory diagnosticians, and researchers who will enlist representatives from community-organized efforts to implement standardized protocols for gene or sequence variant specific annotations of genes related to the specific disease domain.

Interested in becoming an Expert Panel? Learn more about the process?here.

Membership · Documents

Specifications of the ACMG/AMP standards and guidelines for mitochondrial DNA variant interpretation

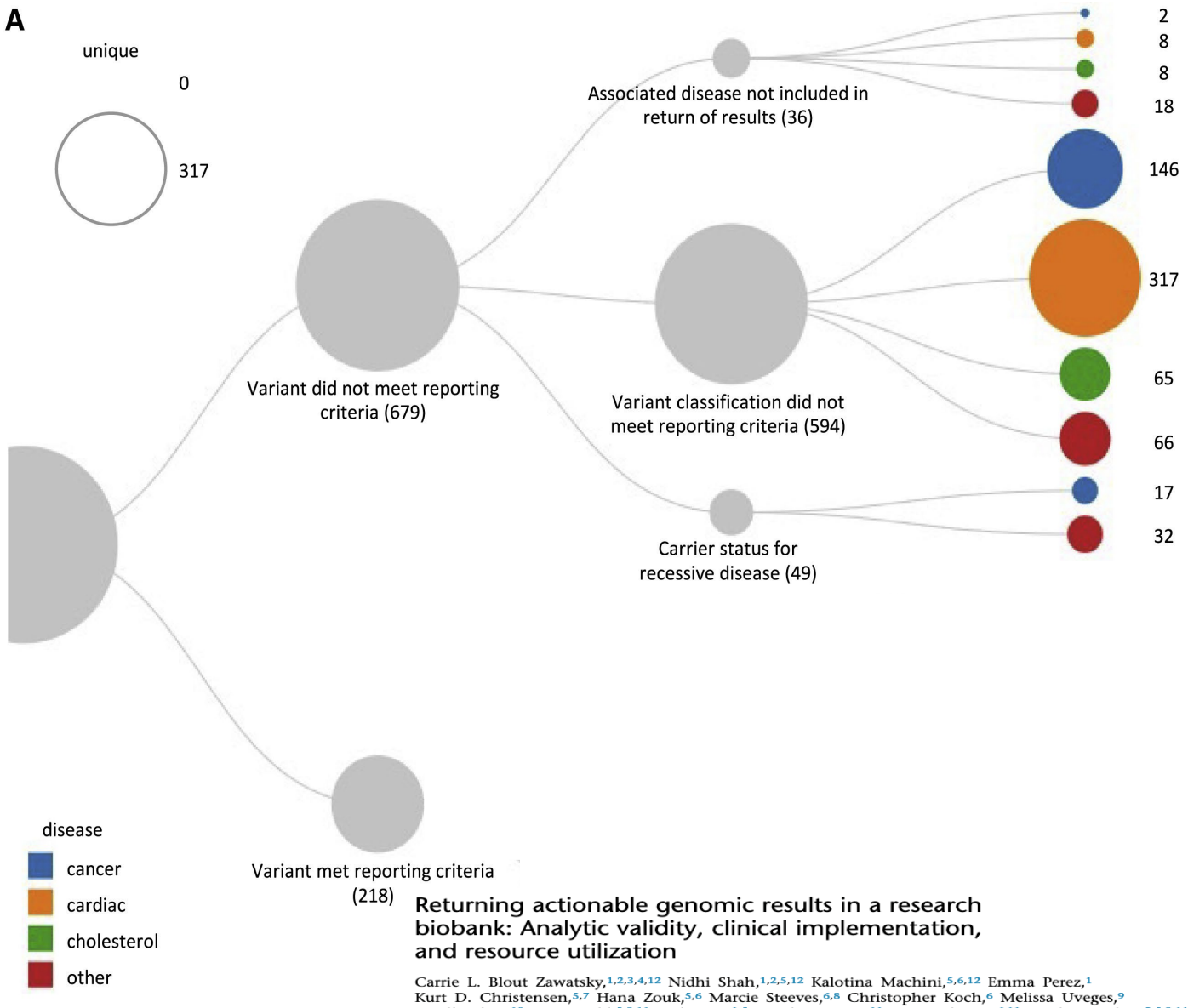
Elizabeth M. McCormick[1] | Marie T. Lott[2] | Matthew C. Dulik[3,4] | Lishuang Shen[5] | Marcella Attimonelli[6] | Ornella Vitale[6] | Amel Karaa[7] | Renkui Bai[8] | Daniel E. Pineda-Alvarez[9] | Larry N. Singh[2] | Christine M. Stanley[10,11] | Stacey Wong[9] | Anshu Bhardwaj[12] | Daria Merkurjev[5] | Rong Mao[13,14] | Neal Sondheimer[15] | Shiping Zhang[2,16] | Vincent Procaccio[17] | Douglas C. Wallace[2,3] | Xiaowu Gai[5,18] | Marni J. Falk[1,3]



Filtration Results

Results still require manual review for pathogenicity

- Not all previously reported variants meet criteria for P/LP
- Not all variants annotated as loss-of-function are actually LOF
- Not all diseases associated to a gene are returnable
- Variants may only be in carrier state for recessive disease

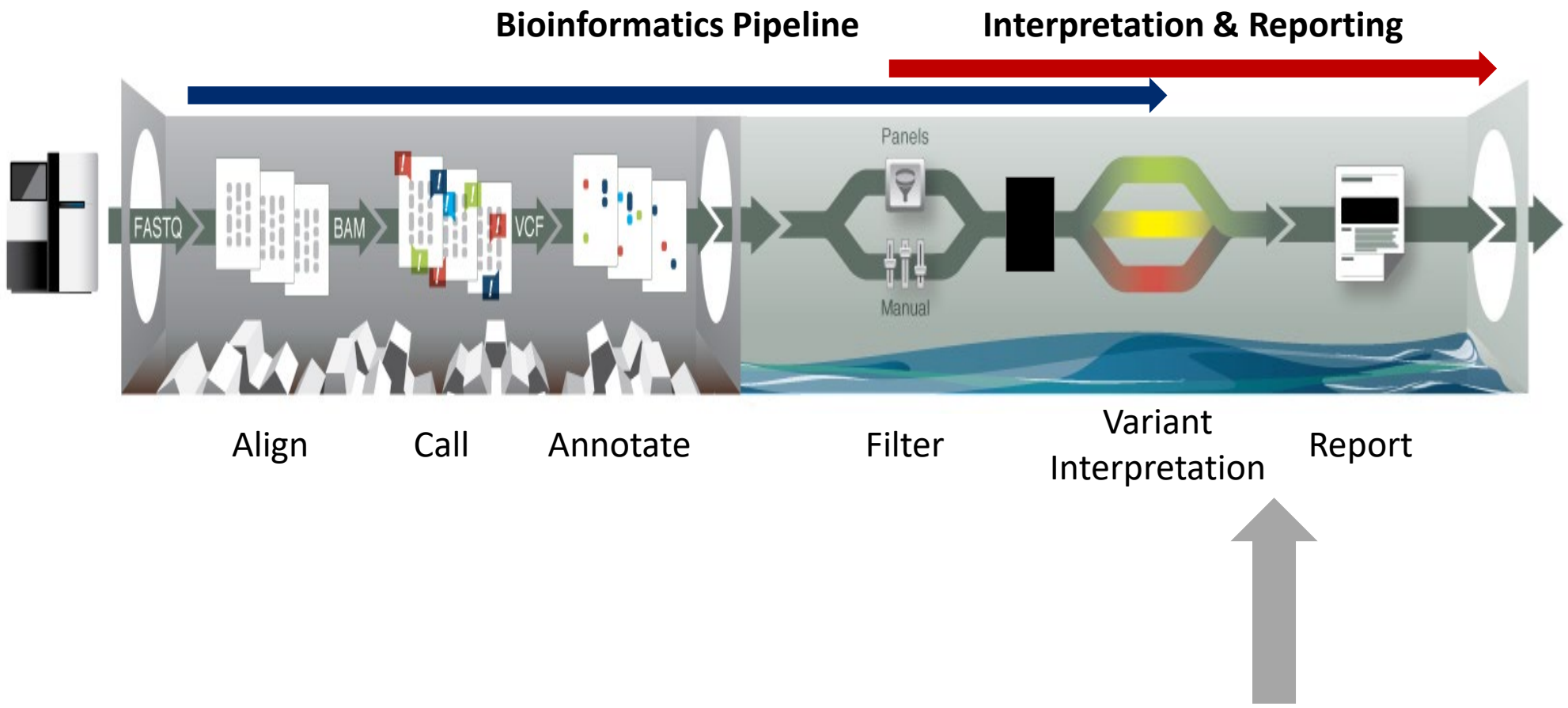


Returning actionable genomic results in a research biobank: Analytic validity, clinical implementation, and resource utilization

Carrie L. Blout Zawatsky,[1,2,3,4,12] Nidhi Shah,[1,2,5,12] Kalotina Machini,[5,6,12] Emma Perez,[1] Kurt D. Christensen,[5,7] Hana Zouk,[5,6] Marcie Steeves,[6,8] Christopher Koch,[6] Melissa Uveges,[9] Janelle Shea,[10] Nina Gold,[5,8,11] Joel Krier,[1,5] Natalie Boutin,[11] Lisa Mahanta,[6,11] Heidi L. Rehm,[2,5,8,11] Scott T. Weiss,[1,5,6,11] Elizabeth W. Karlson,[1,5,11] Jordan W. Smoller,[2,5,8,11] Matthew S. Lebo,[1,2,5,6,11,13] and Robert C. Green[1,2,3,5,11,13,*]

The American Journal of Human Genetics *108*, 2224–2237, December 2, 2021

Genomic bioinformatics pipeline



Returnable results: genotyping arrays vs. sequencing

Genotyping arrays

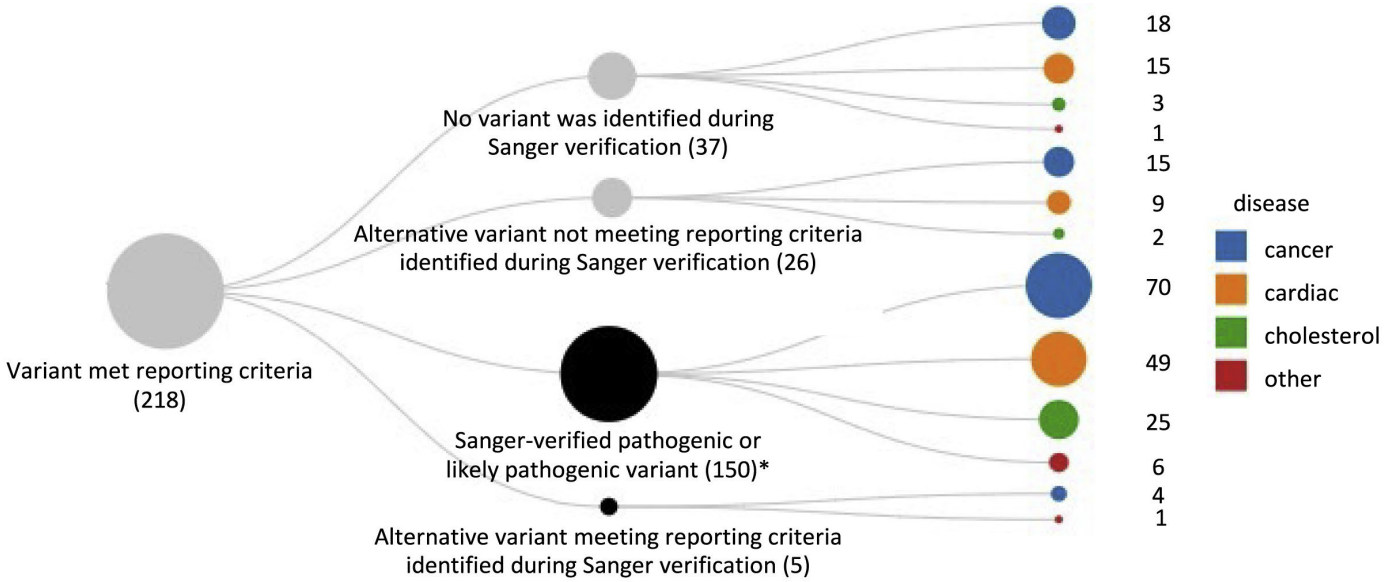
- Predefined list of variants interrogated

- Poor performance for rare variants
  - Inaccurate sites are often recurrent
  - Can have TP and FP for same site

Sequencing

- Interrogates whole coding sequence
  - Can get “novel” variants

- General high accuracy of variant calling
  - However, confirmation may still be necessary for to identify false positive calls

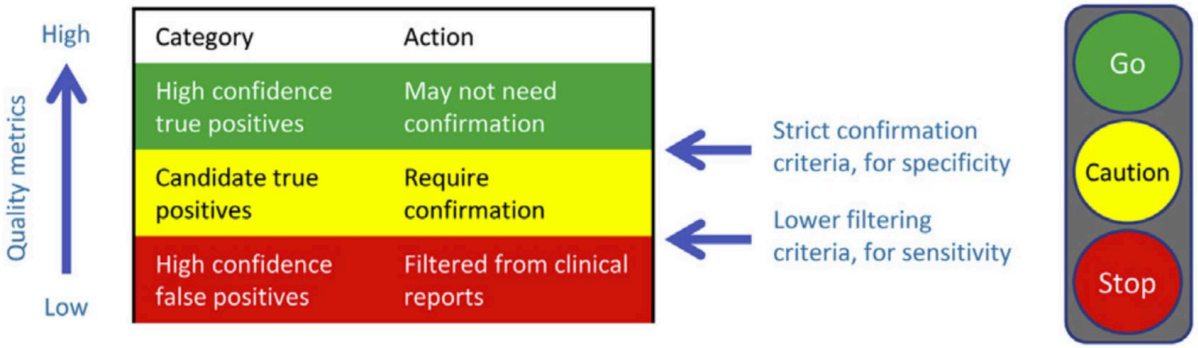


Returning actionable genomic results in a research biobank: Analytic validity, clinical implementation, and resource utilization

Carrie L. Blout Zawatsky,[1,2,3,4,12] Nidhi Shah,[1,2,5,12] Kalotina Machini,[5,6,12] Emma Perez,[1] Kurt D. Christensen,[5,7] Hana Zouk,[5,6] Marcie Steeves,[6,8] Christopher Koch,[6] Melissa Uveges,[9] Janelle Shea,[10] Nina Gold,[5,8,11] Joel Krier,[1,5] Natalie Boutin,[11] Lisa Mahanta,[6,11] Heidi L. Rehm,[2,5,8,11] Scott T. Weiss,[1,5,6,11] Elizabeth W. Karlson,[1,5,11] Jordan W. Smoller,[2,5,8,11] Matthew S. Lebo,[1,2,5,6,11,13] and Robert C. Green[1,2,3,5,11,13,*]

The American Journal of Human Genetics *108*, 2224–2237, December 2, 2021

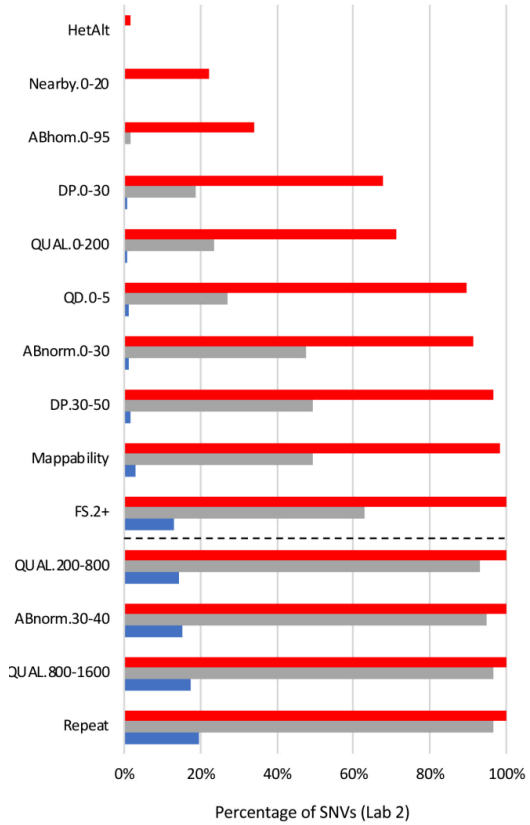
Confirmation of sequencing variants

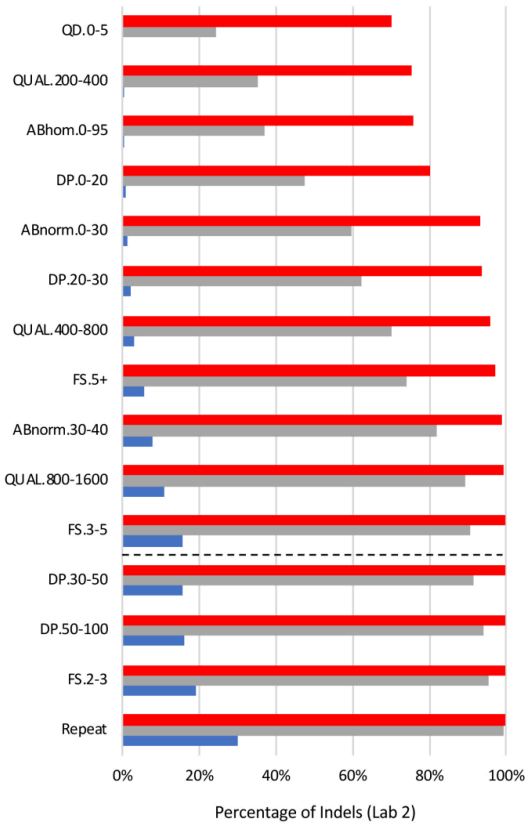


A Rigorous Interlaboratory Examination of the Need to Confirm Next-Generation Sequencing—Detected Variants with an Orthogonal Method in Clinical Genetic Testing

Check for updates

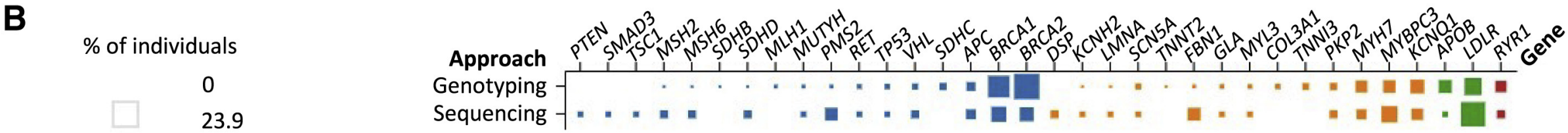
Stephen E. Lincoln,* Rebecca Truty,* Chiao-Feng Lin,†‡ Justin M. Zook,§ Joshua Paul,* Vincent H. Ramey,* Marc Salit,§¶ Heidi L. Rehm,†‡‖**†† Robert L. Nussbaum,*‡‡ and Matthew S. Lebo†‡**††



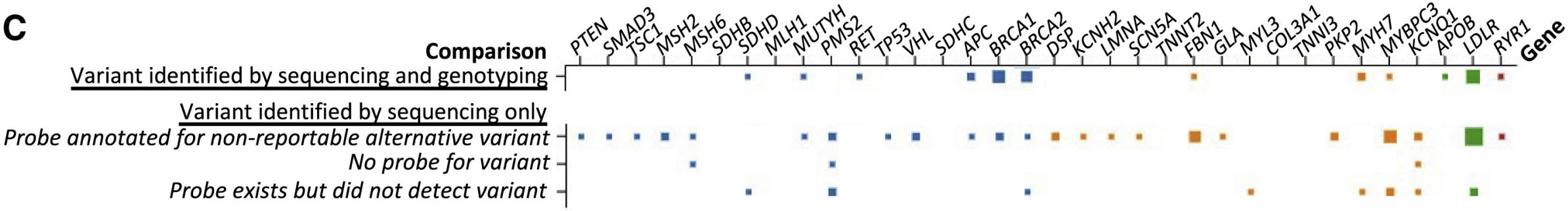


Returnable results: genotyping arrays vs. sequencing

Sequencing more accurately reflects true spectrum of variation



Missing variation from genotyping data often due to missing probes and poor performing probes

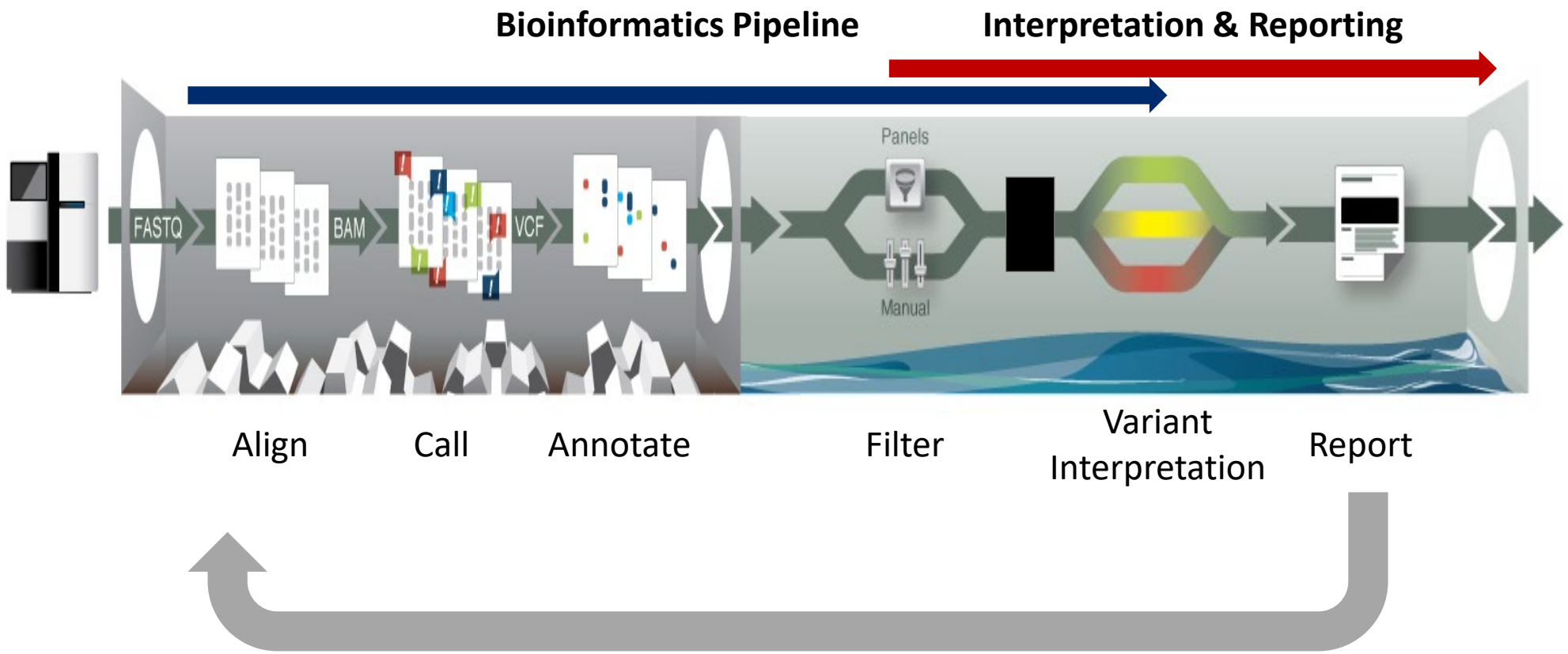


Returning actionable genomic results in a research biobank: Analytic validity, clinical implementation, and resource utilization

Carrie L. Blout Zawatsky,[1,2,3,4,12] Nidhi Shah,[1,2,5,12] Kalotina Machini,[5,6,12] Emma Perez,[1] Kurt D. Christensen,[5,7] Hana Zouk,[5,6] Marcie Steeves,[6,8] Christopher Koch,[6] Melissa Uveges,[9] Janelle Shea,[10] Nina Gold,[5,8,11] Joel Krier,[1,5] Natalie Boutin,[11] Lisa Mahanta,[6,11] Heidi L. Rehm,[2,5,8,11] Scott T. Weiss,[1,5,6,11] Elizabeth W. Karlson,[1,5,11] Jordan W. Smoller,[2,5,8,11] Matthew S. Lebo,[1,2,5,6,11,13] and Robert C. Green[1,2,3,5,11,13,*]

The American Journal of Human Genetics *108*, 2224–2237, December 2, 2021

Genomic bioinformatics pipeline



Revisiting raw data

Reassessment of variants can lead to additional returns

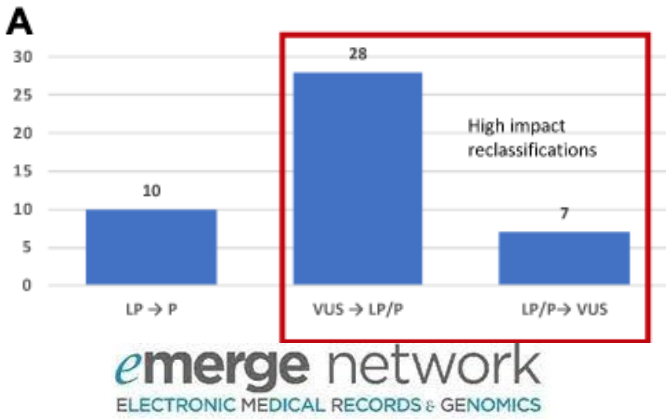
- New evidence available
- New guidelines for classification

Improved algorithms can more accurately call variants

- Especially true for CNVs/SVs
- Also for small variants, even with current high accuracy

Other improvements affecting bioinformatic pipelines

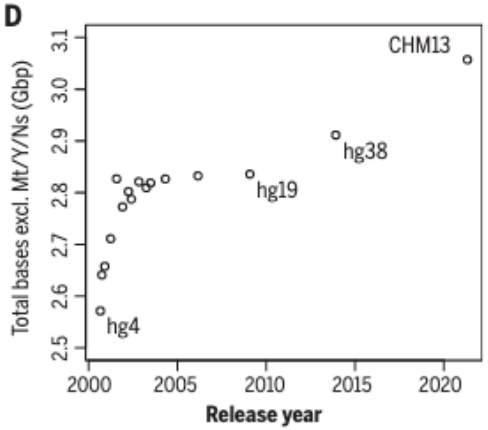
- Annotations (e.g., updated transcripts, LOF prediction)
- Improvements in reference genomes



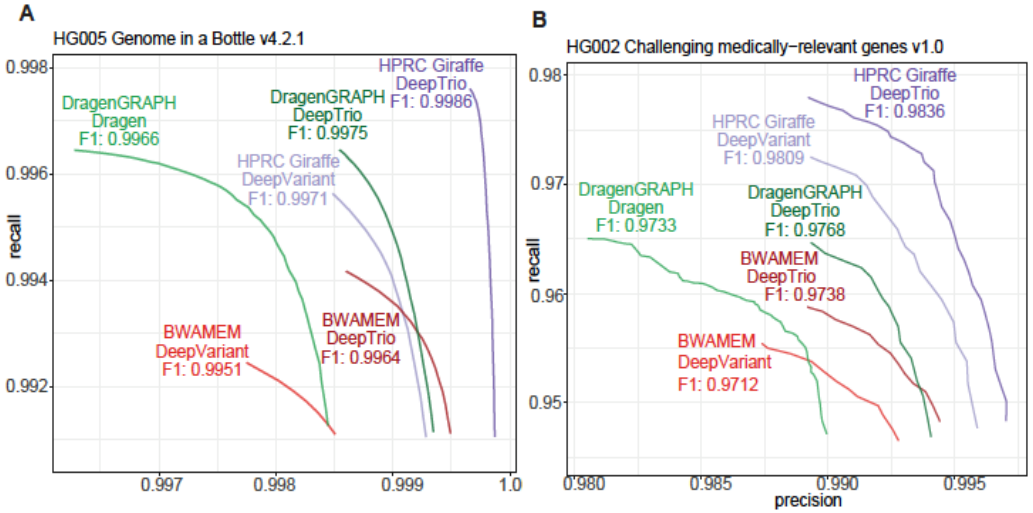
	Variant type	Sensitivity	PPV
Pipeline 1	SNV	0.9900	0.9941
	Indel	0.9897	0.9937
Pipeline 2	SNV	0.9967	0.9993
	Indel	0.9948	0.9967

Improvements in the human genome reference

- Telomere-To-Telomere
  - Complete haploid genome (CHM13)
  - Additional 200Mb of genomic content



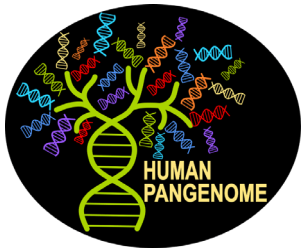
- Human Pangenome Reference
  - Captures genetic diversity of human species
  - Improved variant calling, especially in difficult regions
  - Improved SV/CNV calling

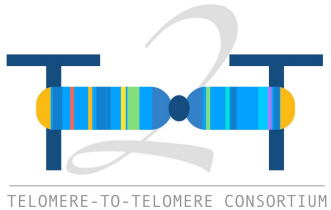


A Draft Human Pangenome Reference

Wen-Wei Liao, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Julian K. Lucas, Jean Monlong, Haley J. Abel, Silvia Buonaiuto, Xian H. Chang, Haoyu Cheng, Justin Chu, Vincenza Colonna, Jordan M. Eizenga, Xiaowen Feng, Christian Fischer, Robert S. Fulton, Shilpa Garg, Cristian Groza, Andrea Guarracino, William T Harvey, Simon Heumos, Kerstin Howe, Miten Jain, Tsung-Yu Lu, Charles Markello, Fergal J. Martin, Matthew W. Mitchell, Katherine M. Munson, Moses Njagi Mwaniki, Adam M. Novak, Hugh E. Olsen, Trevor Pesout, David Porubsky, Pjotr Prins, Jonas A. Sibbesen, Chad Tomlinson, Flavia Villani, Mitchell R. Vollger, Human Pangenome Reference Consortium, Guillaume Bourque, Mark JP Chaisson, Paul Flicek, Adam M. Phillippy, Justin M. Zook, Evan E. Eichler, David Haussler, Erich D. Jarvis, Karen H. Miga, Ting Wang, Erik Garrison, Tobias Marschall, Ira Hall, Heng Li, Benedict Paten

doi: https://doi.org/10.1101/2022.07.09.499321



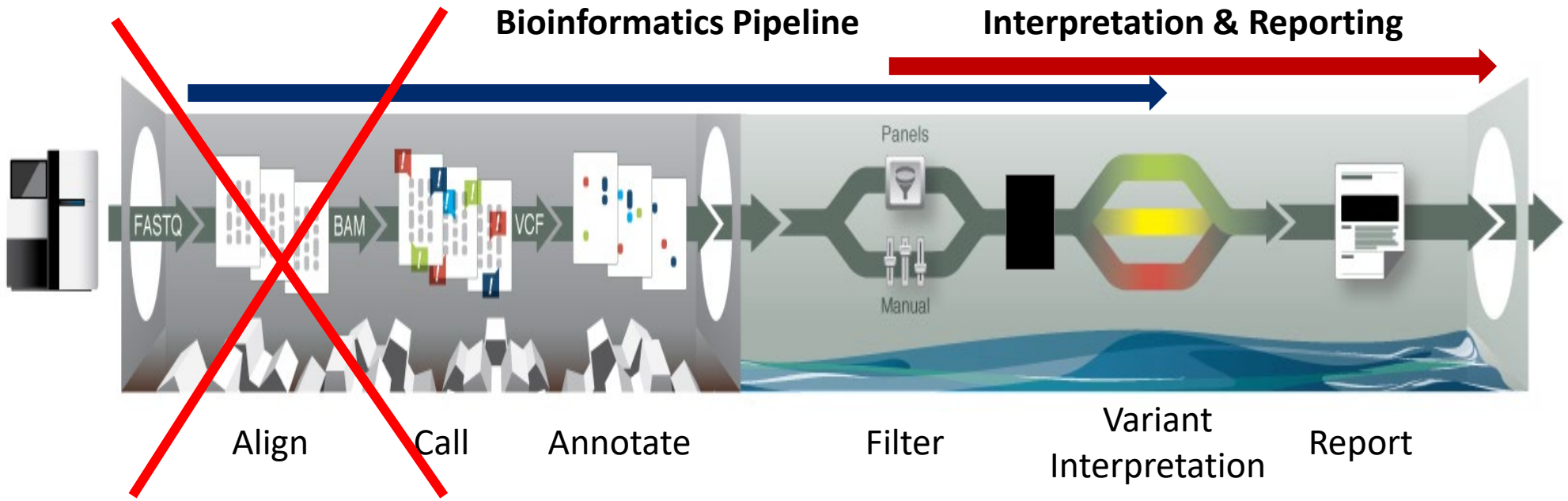


The complete sequence of a human genome

SERGEY NURK, SERGEY KOREN, ARANG RHIE, MIKKO RAUTIAINEN, ANDREY V. BZIKADZE, ALLA MIKHEENKO, MITCHELL R. VOLLGER, NICOLAS ALTEMOSE, LEV URALSKY, [...], AND ADAM M. PHILLIPPY +90 authors Authors Info & Affiliations

SCIENCE · 31 Mar 2022 · Vol 376, Issue 6588 · pp. 44-53 · DOI: 10.1126/science.abj6987

Genomic bioinformatics pipeline



Bioinformatic considerations for NOT reporting

Can you generate data without initiating return?

- Most cohorts have not masked/removed potentially returnable variants
  - Typically release unannotated VCF and/or aggregate data
  - Difficult to identify actionable variants without further information

Are there methods to mask rare variants?

- For genotyping arrays, you can remove them from the manifest
  - May appear in imputation if “common” enough

- Harder for sequencing-based methods
  - Due to identification of ”novel” variants
  - Possible to do it by population frequency (e.g., only keep variants >5% frequency)
    - Will remove more variants than may be wanted

Thank you!

Mass General Brigham