Molecular Pathology Informatics

Alexis Carter, MD
Physician Informaticist
Department of Pathology and Laboratory Medicine

Children's℠
Healthcare of Atlanta
Dedicated to All Better

Notice of Faculty Disclosure

In accordance with ACCME guidelines, any individual in a position to influence and/or control the content of this ASCP CME activity has disclosed all relevant financial relationships within the past 12 months with commercial interests that provide products and/or services related to the content of this CME activity.

The individual below has responded that he/she has no relevant financial relationship(s) with commercial interest(s) to disclose:

Alexis Carter, MD

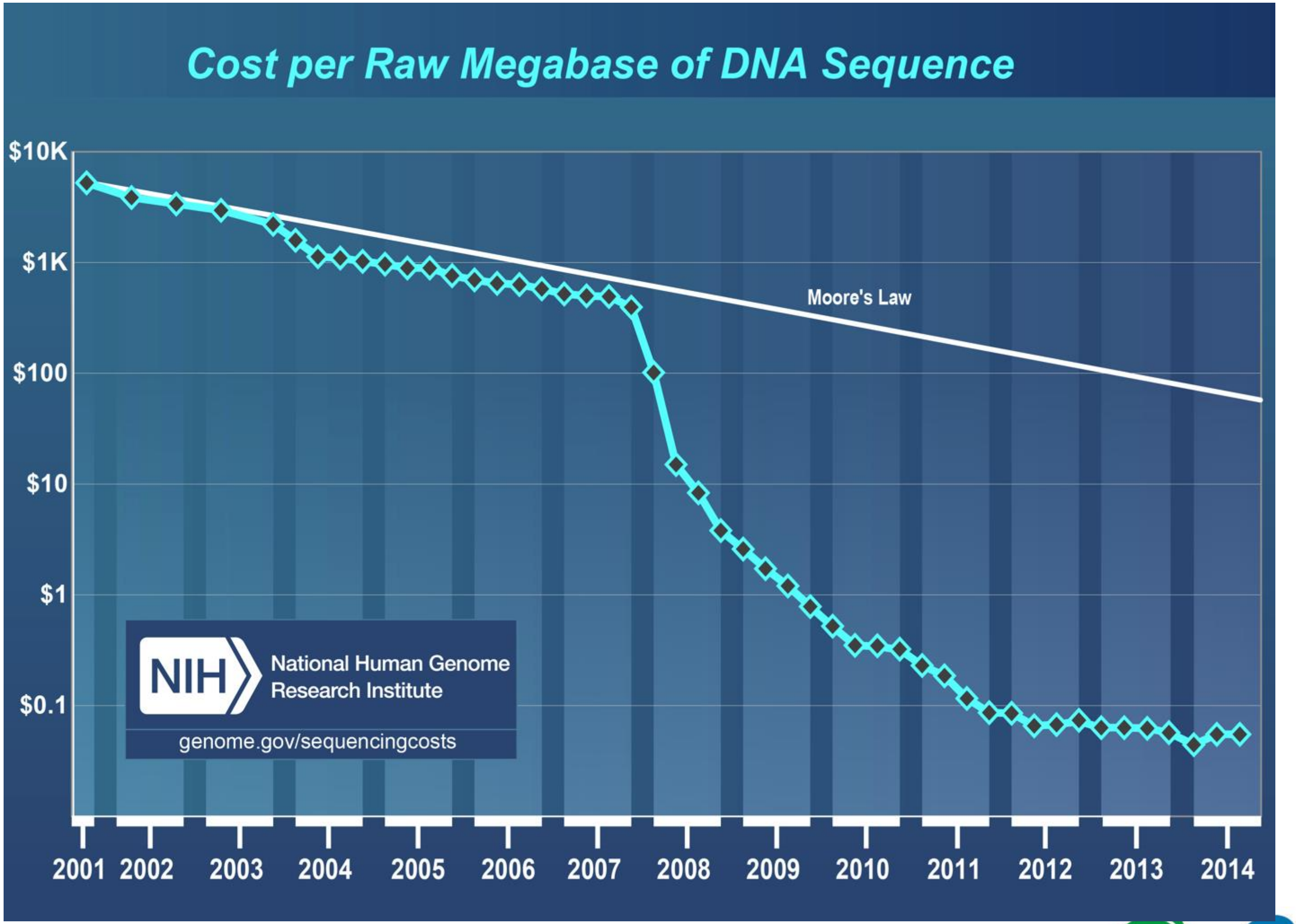
Objectives

- By the end of this presentation, the participant should be able to:
  - Understand the pre-analytic, analytic and post-analytic informatics challenges for molecular laboratories
  - Describe the major file types and quality metrics used in bioinformatics pipelines for next generation sequencing (NGS)
  - Explain the limitations of existing pipeline standards for file formats and validation
  - Understand big data and computational pathology
  - Understand the challenges and threats to NGS testing including LIS and EHR limitations for reporting

Focus on NGS Informatics – Why?

- Result generation heavily dependent on computational algorithms
- Original sequencing reaction results not decipherable without computational algorithms
  - no electropheresis gel or electropherogram to look at
- People are rarely talking about anything else
- Good example of big data analysis
- US Federal Government scrutiny

DNA Testing – More bases for $$$



http://core-genomics.blogspot.com/2015/01/the-not-so-rapid-decreasing-costs-of.html

DNA Testing – Today vs. Before

Previously

- Testing each gene required many tests
- Expensive to do more than one gene
- Could not test entire DNA

Today (Next-Generation)

- Can sequence many, many genes at one time
- Cost per amount of DNA has decreased a lot
- Can find more variants with less money

Big Data

- Many people use this term
  - Most cannot accurately define it
- Many people think big data refers to:
  - Next generation sequencing
  - Whole slide imaging

Big Data

- So what is big data? Why do we care?
- High quality **Computational Pathology** is rooted in sound principles of analyzing and using big data
- Characterized by **three** Vs:

Volume	Large amounts of data

Berman JJ. *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information.* Amsterdam: Morgan Kaufmann; 2013.

Big Data

- So what is big data? Why do we care?
- High quality **Computational Pathology** is rooted in sound principles of analyzing and using big data
- Characterized by **three** Vs:

Volume	Large amounts of data
Variety	Many different types of data
Velocity	Constantly accumulating new data

Berman JJ. *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information.* Amsterdam: Morgan Kaufmann; 2013.

Big Data

	Small Data Resource	Big Data Resource
Design	Answer **specific** questions or serve specific purpose	Provide answers to **protean** questions on variable topics, current and future, and to serve many different and flexible purposes
Location	Within **one** institution, server, computer or file	In **many** places
Structure	**Highly structured**; limited data types	**Unstructured data of many types** (e.g., free text, sound, images, video)
Preparation	**Few** prepare the data (usually the end-user)	**Many** prepare the data (usually **not** the end-user)
Longevity	**Short** (discarded when project is completed)	**Long** (data is kept in perpetuity)

Berman JJ. *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information.* Amsterdam: Morgan Kaufmann; 2013.

Big Data (cont.)

	Small Data Resource	Big Data Resource
Measurements	**One** set of standard units of measure for data; easy to verify data quality	**Many** different sets of units of measure; difficult to verify quality of data
Reproducibility	**Easy** to repeat a project with new data to verify quality of results	**Hard (to impossible)** to repeat a project with new data to verify quality of results
Stakes	**Small** costs; easy to recover from project failure	**Expensive**; failure can lead to bankruptcy
Introspection	**Highly organized** data (rows and columns)	**Loosely or unorganized** data (may be inscrutable)
Analysis	Analysis can occur **all together** and all at the **same time**	Analysis occurs in **incremental steps** (unless performed on grid/parallel/super computing resources)

Berman JJ. *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information.* Amsterdam: Morgan Kaufmann; 2013.

Big Data

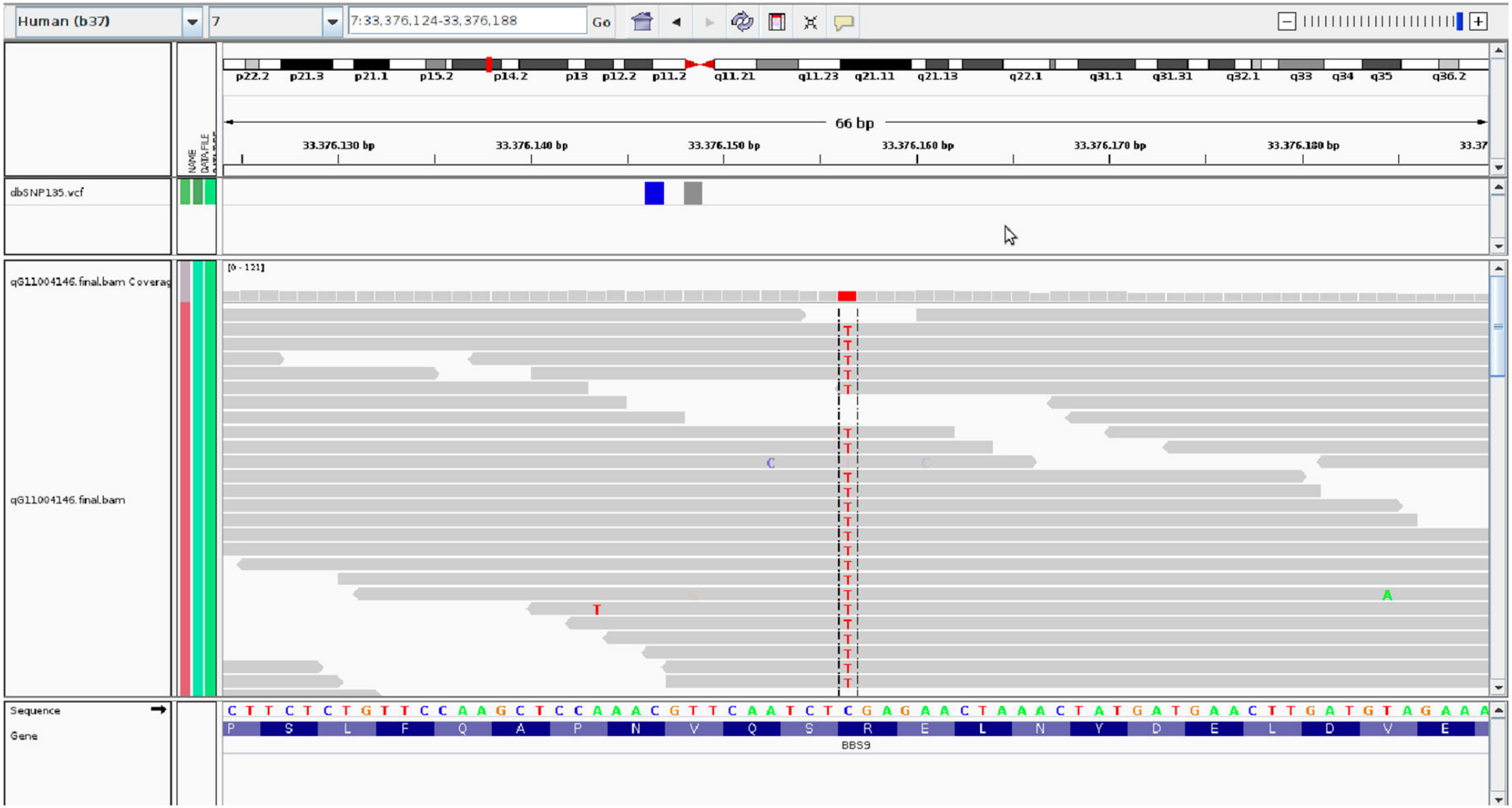
- Many people think big data refers to:
  - Next generation sequencing
    - FASTQ, BAM and VCF files have volume but **lack** velocity and variety unless…
      - Multi-patient exome/genome level sequences acquired on an ongoing basis from different analyzers
    - HOWEVER, definitely big data, regardless of input, when you are trying to **interpret** variants produced

NGS - Overview

- Next-generation sequencing
- Better term: massively parallel sequencing
- DNA is sequenced in short overlapping fragments then aligned to the reference and variants detected

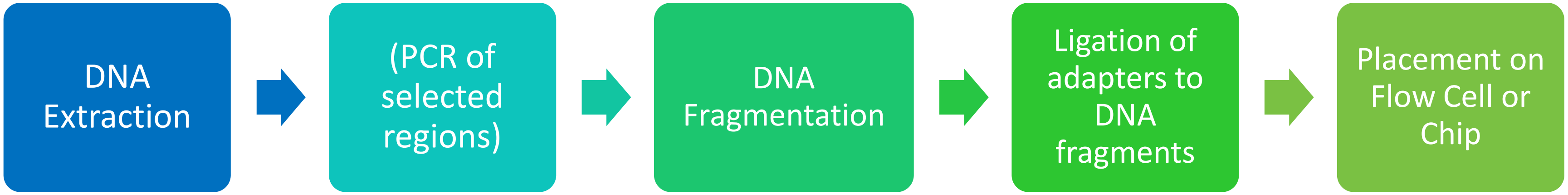
DNA Testing – Next Generation Sequencing

Integrated Genomics Viewer https://www.broadinstitute.org/software/igv/download


Human (b37)
7
7:33,376,124-33,376,188
Go
p22.2 p21.3 p21.1 p15.2 p14.2 p13 p12.2 p11.2 q11.21 q11.23 q21.11 q21.13 q22.1 q31.1 q31.31 q32.1 q33 q34 q35 q36.2
66 bp
33,376,130 bp
33,376,140 bp
33,376,150 bp
33,376,160 bp
33,376,170 bp
33,376,180 bp
NAME
DATA FILE
dbSNP135.vcf
qG11004146.final.bam Coverag
[0 - 121]
qG11004146.final.bam
Sequence
Gene
C T T C T C T G T T C C A A G C T C C A A A C G T T C A A T C T C G A G A A C T A A A C T A T G A T G A A C T T G A T G T A G A A A
P S L F Q A P N V Q S R E L N Y D E L D V E
BBS9

http://www.analesdepediatria.org/es/sindrome-bardet-biedl-aplicacion-diagnostica-secuenciacion/articulo/S1695403313003822/

NGS Analysis



DNA Extraction

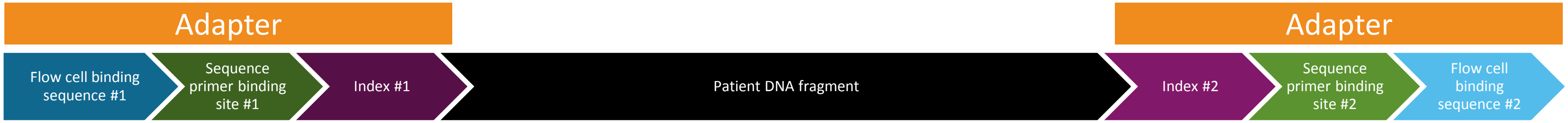
(PCR of selected regions)

DNA Fragmentation

Ligation of adapters to DNA fragments

Placement on Flow Cell or Chip

- Adapter placed on each end of single strand of DNA sequence
- Each adapter contains:
  - Sequence with known complementarity to binding site in chip bead or flow cell oligonucleotide
  - Unique index (**molecular barcode**) (8-12 bp for Illumina)
    - Allows **multiple patient samples on a single chip or flow cell**
  - Primer binding sites for sequencing reaction



Adapter

Flow cell binding sequence #1

Sequence primer binding site #1

Index #1

Patient DNA fragment

Index #2

Sequence primer binding site #2

Flow cell binding sequence #2

Adapter

NGS Analysis – Raw Sequencing Data

- Illumina technology
  - Flow cell has a lane coated with oligonucleotides complimentary to flow cell binding sequences #1 and #2
  - Sample flows down the lane and binds to the oligos
  - Sequencing by synthesis reaction follows
  - Different color for each nucleotide
  - Visual fluorescence recorded for each reaction and location
  - DNA strands then “bridge” fold and bind other end.
  - Reaction repeats in reverse direction.

NGS Analysis – Raw Sequencing Data

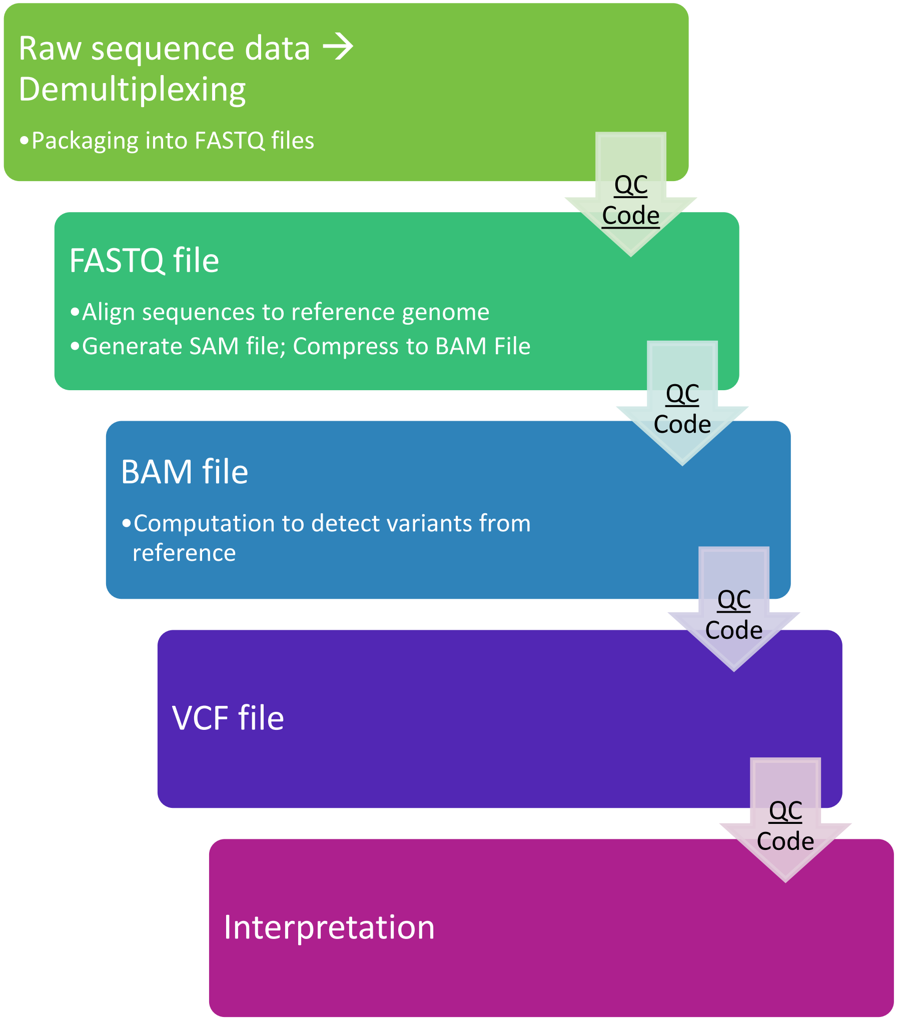
- Ion Torrent technology
  - Beads are coated with oligos complimentary to the binding sequence
  - DNA binds to the beads
  - Amplification reaction occurs to coat the bead with identical sequences
  - DNA-covered beads flow through semiconductor chip and bind to wells in chip (one bead per well)
  - Single nucleotide washed over all cells of chip x 15 s.
  - Cells which incorporate that base release hydrogen
  - pH is measured in the well and base incorporation (or not) recorded

Demultiplexing

- Sequences come off of the instrument all mixed together
- Before analysis of sequences can begin
- Patient samples separated based on their index sequences
- Computer code which does this must be robust and have integrity checks

NGS Bioinformatics Pipeline

- Bioinformatics pipeline
  - Multiple sets of one or more computational algorithms performed in series to analyze biological data
  - Not limited to NGS data
- Critical to collect and check quality metrics along the way
- Many, many software packages with variable quality



Raw sequence data → Demultiplexing

- Packaging into FASTQ files

QC Code

FASTQ file

- Align sequences to reference genome
- Generate SAM file; Compress to BAM File

QC Code

BAM file

- Computation to detect variants from reference

QC Code

VCF file

QC Code

Interpretation

FASTQ

- FASTA file format
  - Simple text file format for nucleic acid sequence
  - No well-defined or accepted standard
- FASTQ file format
  - FASTA file format with additional quality data for each base
  - Also no well-defined or accepted international standard
  - *De facto* standard for representing sequences in NGS
  - Developed around 2000 by Wellcome Sanger Trust Institute

Cock P, et al. Nucleic Acids Res. 2010 Apr; 38(6): 1767–1771.
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847217/

FASTQ

- Incorporates a Phred (Q) score for each base call
- Three versions – none very well defined
  - Sanger, Solexa, Illumina
- Sanger Phred (Q) score

$$Q_{PHRED} = -10\ x\ log_{10}(P_e)$$

  - P_e: Probability of error

Chance that wrong base is incorporated	Q score calculation	Q Score
1 in 10	$-10\ x\ log_{10}(0.1)$	10
1 in 100	$-10\ x\ log_{10}(0.01)$	20
1 in 1,000	$-10\ x\ log_{10}(0.001)$	30
1 in 10,000	$-10\ x\ log_{10}(0.0001)$	40

FASTQ

- For a single read:

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Line	Starts with	Contains
1	@	Sequence identifier and optional description (free text; not structured; **no requirement for sample identification**)
2	<none>	Raw sequence letters
3	+	May be blank; optionally repeats sequence identifier and description
4	<none>	Quality values for each base in line 2 (uses single character ASCII representation); **May also contain @ and + symbols**

FASTQ

ASCII DEC	Phred (Q) Score (subtract 32 from ASCII DEC)	Symbol
32		
33	1	!
34	2	"
35	3	#
36	4	$
37	5	%
38	6	&
39	7	'
40	8	(
41	9	)
42	10	*
43	11	+
44	12	,
45	13	-
46	14	.
47	15	/
48	16	0
49	17	1
50	18	2

ASCII DEC	Phred (Q) Score (subtract 32 from ASCII DEC)	Symbol
64	32	@
65	33	A
66	34	B
67	35	C
68	36	D
69	37	E
70	38	F
71	39	G
72	40	H
73	41	I
74	42	J

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

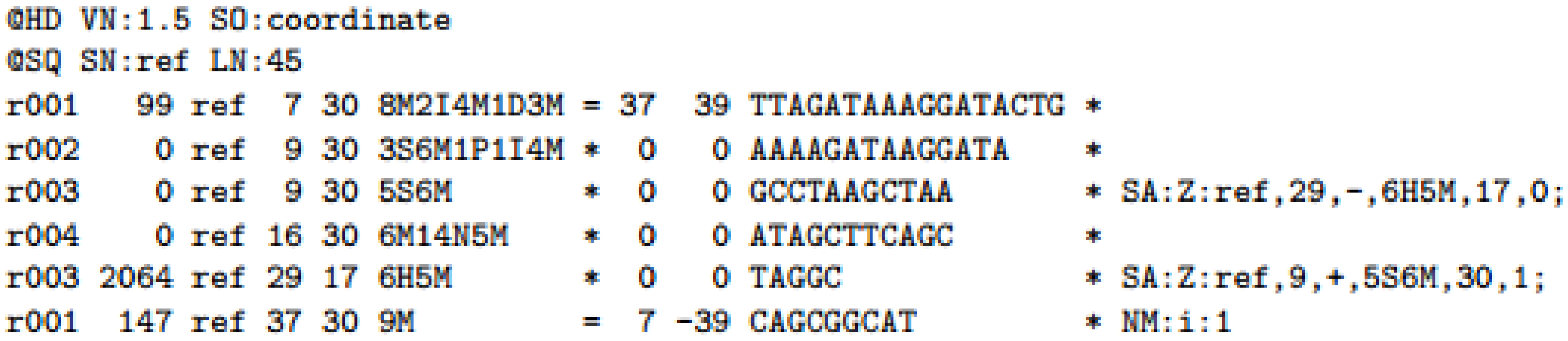
Q = 25

SAM and BAM

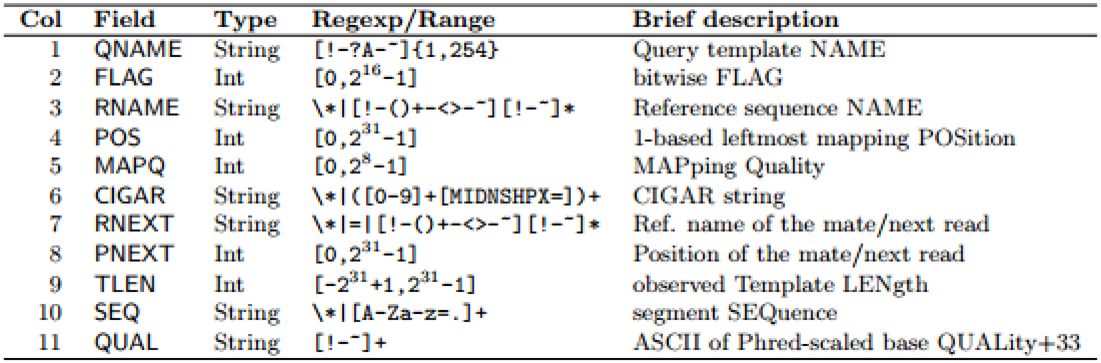
- SAM: Sequence Alignment/Map
  - Free text file
  - Contains data showing where the sequence in the FASTQ aligns to the “reference” sequence
    - **hg19 / GRCh37: 2009 (most commonly used; http://grch37.ensembl.org/index.html)**
    - **hg38 / GRCh38: 2013**
  - More structure than FASTQ
  - **No requirement or standard for sample identification**

SAM and BAM

- https://samtools.github.io/hts-specs/SAMv1.pdf



```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```



Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

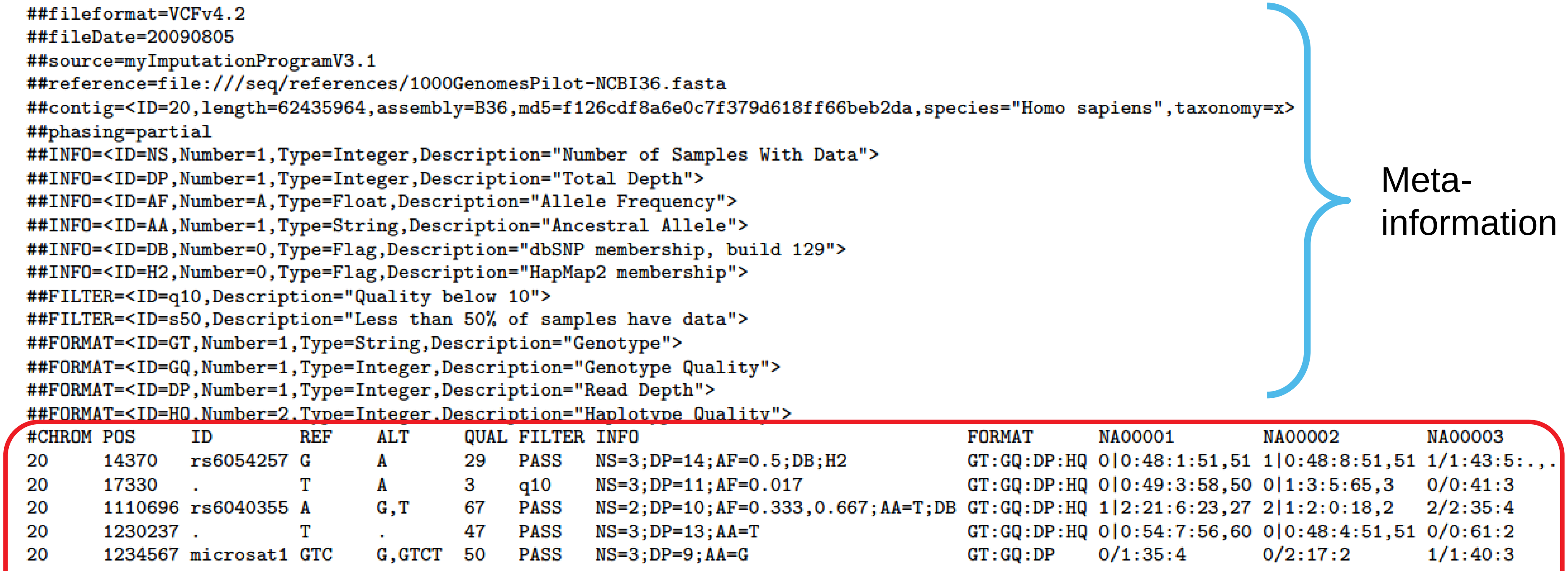
SAM and BAM

- BAM: Binary version of SAM (compressed)
  - Provides good compression of SAM while allowing efficient random indexed access to data
  - Most commonly used file format for alignment because lower file size

VCF

- VCF: Variant Call Format file
- Text file that contains a list of variants that the sample has compared to the reference genome
  - May include artifacts, benign, unknown and pathogenic variants
- Again, no requirement or stringency for sample identification
- Multiple versions of VCF in use
- https://samtools.github.io/hts-specs/VCFv4.2.pdf

VCF



```
##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=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##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=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
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 1/1:43:5:.,.
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   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
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 0/0:61:2
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       1/1:40:3
```

Meta-information

Variant data

VCF for Clinical Use

- Clinical Grade VCF format
- 2012: Centers for Disease Control and Prevention facilitated working group
- Goal: Identify and build consensus around the requirements for a clinical grade variant file format
- http://vcfclin.org/

NGS sequence data

Read	A single output sequence from an NGS sequencing reaction. A single sequencing reaction in a single flow cell or chip generates trillions of reads.
Depth of coverage	The number of reads which contain a specific nucleotide. The higher the depth of coverage, the more sensitive and accurate an assay will be to low percentages of variants.

- Germline (inherited) testing
  - Variant allele burden expected to be 50% or 100%
  - Depth of coverage OK to be lower (100x to 250x)
- Somatic (acquired) testing (e.g., cancer)
  - Variant allele burden quite variable
  - Depth of coverage needs to be high to catch low allele burdens (e.g., 500x or higher)

When it can all go wrong…

- Pipelines can be set up to filter data based on certain pre-defined criteria
- This filtering, if not properly designed and validated, can cause variants to be **hidden from view**

When it can all go wrong…example

- Lab notified of discrepant result
- Cancer sample analyzed **at another lab** had 15 bp insertion in *EGFR;* original lab NGS test was negative
- Data re-analyzed
- Original pipeline was built to exclude any variant if three or more unrelated variants occurred at the same location (regardless of percentages)
  - 15% alleles with 15 bp insertion in exon 19 of *EGFR* (confers increased sensitivity to EGFR TKIs)
  - <1% alleles with unrelated variant #1 at same location
  - <1% alleles with unrelated variant #2 at same location
- Entire variant hidden from view of pathologist

Interpretation and Annotation

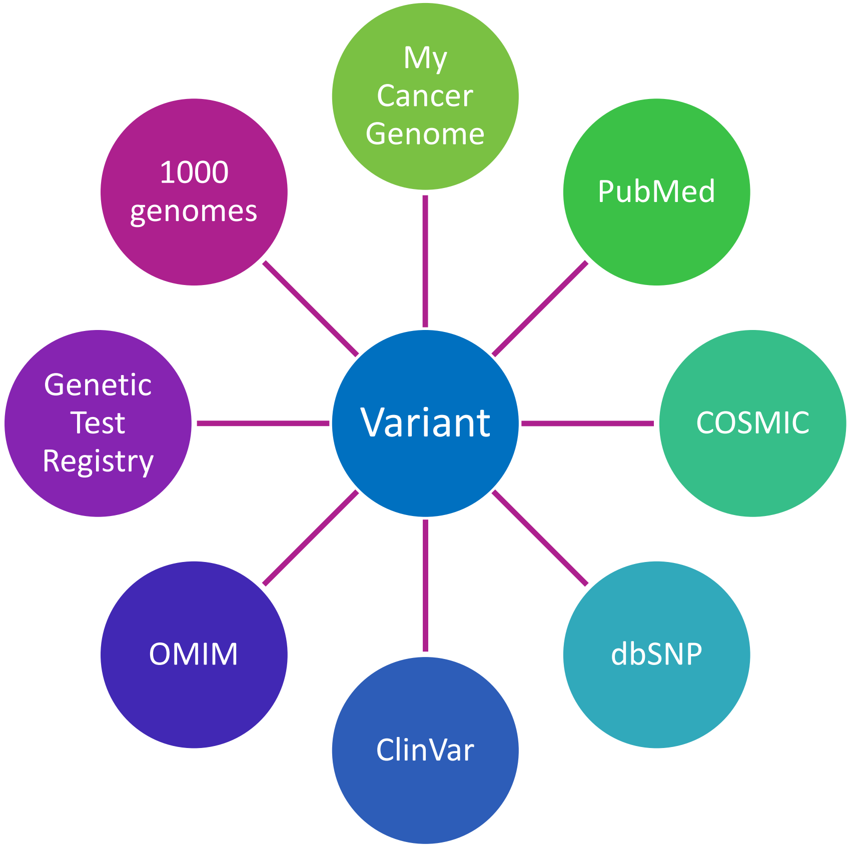
- **Interpretation** is the assignment of clinical significance to the variant
  - In most cases, interpretation must be made by an advanced laboratory professional
  - May occur with or without assistance of other validated tools
  - Basic variant interpretations:
    - Artifact (false positive generated by sequencing process)
    - Benign polymorphism
    - Non-coding and synonymous variants
    - Known pathogenic variant
    - Variant of unknown significance

Interpretation and Annotation

- **Annotation** is the labeling of a variant in the context of a particular clinical presentation (e.g., tumor type, tissue of origin, signs, symptoms) for future use in the analysis of other samples
  - Allows linkage of variant to online databases for that variant
  - Laboratories lack adequate tools to annotate variants and retrieve those annotations for future analysis

Annotation and Interpretation

- Only about 20% of variants have known significance
- Other 80% have to be researched

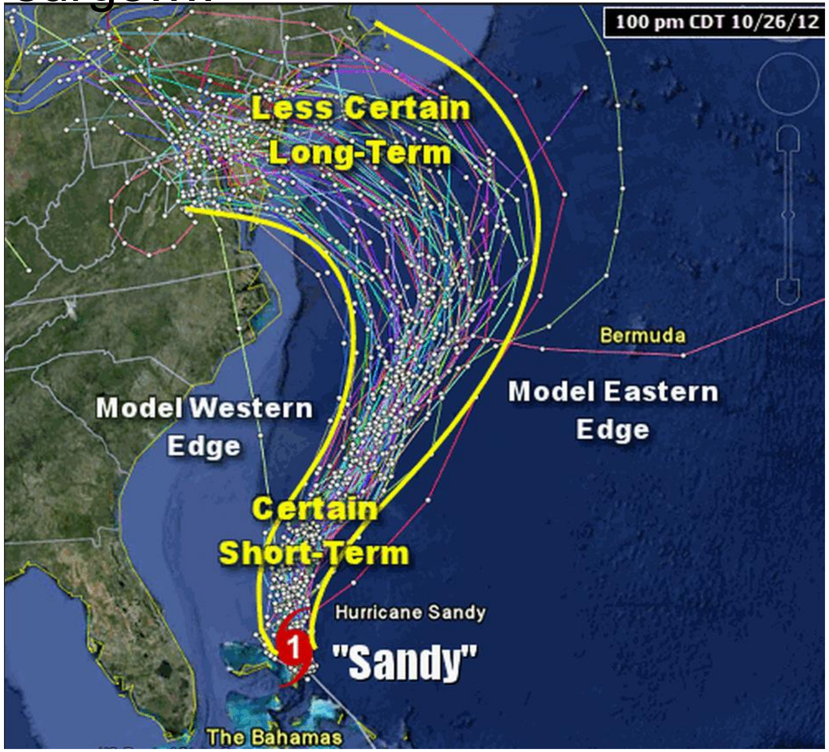


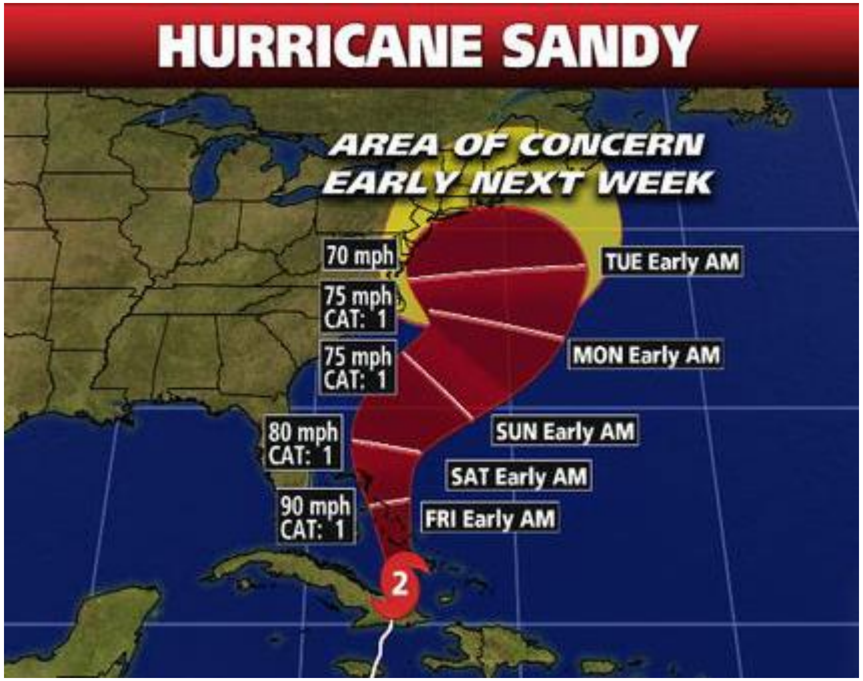
- Online genomic references to help determine significance of variants are
  - Are constantly being updated by multiple (often anonymous) sources
  - Data may be unstructured
  - Data **often** uncurated

This is **Big Data**

Mathematical models predict path, intensity, size, timing, storm surge….

…presented in a usable view





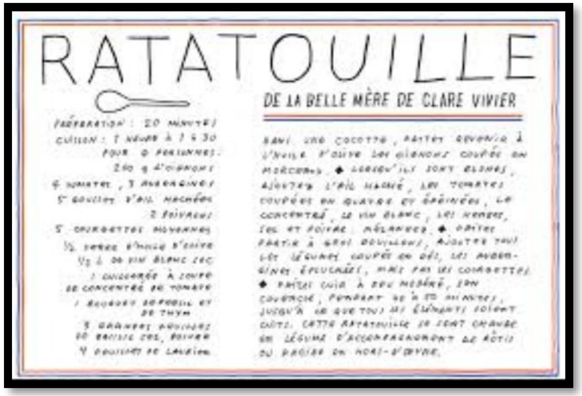
Analogy courtesy of **John R. Gilbertson, MD, PhD**

Images courtesy of Google

Small data

Structured data

Tools



RATATOUILLE

DE LA BELLE MÈRE DE CLARE VIVIER

PRÉPARATION : 20 MINUTES
CUISSON : 1 HEURE À 1 H 30
POUR 6 PERSONNES :
250 g D'OIGNONS
4 TOMATES, 3 AUBERGINES
5 GOUSSES D'AIL HACHÉES
2 POIVRONS
5 COURGETTES MOYENNES
½ VERRE D'HUILE D'OLIVE
½ L DE VIN BLANC SEC
1 CUILLERÉE À SOUPE DE CONCENTRÉ DE TOMATE
1 BOUQUET DE PERSIL ET DE THYM
3 BRANCHES BROYÉES DE BASILIC SEL, POIVRE
4 FEUILLES DE LAURIER

DANS UNE COCOTTE, FAITES REVENIR À L'HUILE D'OLIVE LES OIGNONS COUPÉS EN MORCEAUX. ◆ LORSQU'ILS SONT BLONDS, AJOUTEZ L'AIL HACHÉ, LES TOMATES COUPÉES EN QUATRE ET ÉPÉPINÉES, LE CONCENTRÉ, LE VIN BLANC, LES HERBES, SEL ET POIVRE. MÉLANGEZ. ◆ FAITES PARTIR À GROS BOUILLONS, AJOUTEZ TOUS LES LÉGUMES COUPÉS EN DÉS, LES AUBERGINES ÉPLUCHÉES, MAIS PAS LES COURGETTES. ◆ FAITES CUIRE À FEU MODÉRÉ, SANS COUVERCLE, PENDANT 45 À 50 MINUTES, JUSQU'À CE QUE TOUS LES ÉLÉMENTS SOIENT CUITS. CETTE RATATOUILLE SE SERT CHAUDE EN LÉGUME D'ACCOMPAGNEMENT DE RÔTIS OU FROIDE EN HORS-D'ŒUVRE.

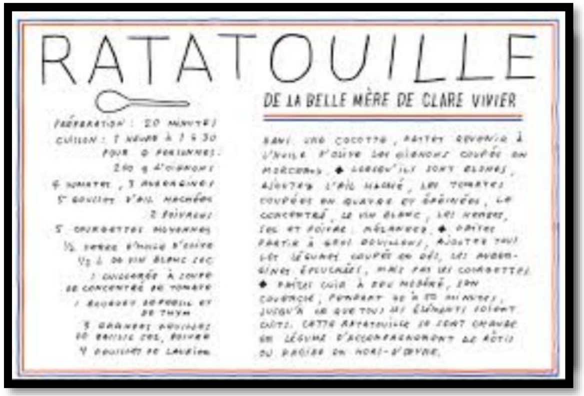
Knowledge

ACTIONABLE

BIG data

Unstructured BIG data

Tools



RATATOUILLE

DE LA BELLE MÈRE DE CLARE VIVIER

PRÉPARATION : 20 MINUTES
CUISSON : 1 HEURE À 1 H 30
POUR 8 PERSONNES :
250 g D'OIGNONS
4 TOMATES , 3 AUBERGINES
5 GOUSSES D'AIL HACHÉES
2 POIVRONS
5 COURGETTES MOYENNES
½ VERRE D'HUILE D'OLIVE
½ L DE VIN BLANC SEC
1 CUILLERÉE À SOUPE DE CONCENTRÉ DE TOMATE
1 BOUQUET DE PERSIL ET DE THYM
3 BRANCHES [illegible] DE BASILIC SEL, POIVRE
4 FEUILLES DE LAURIER

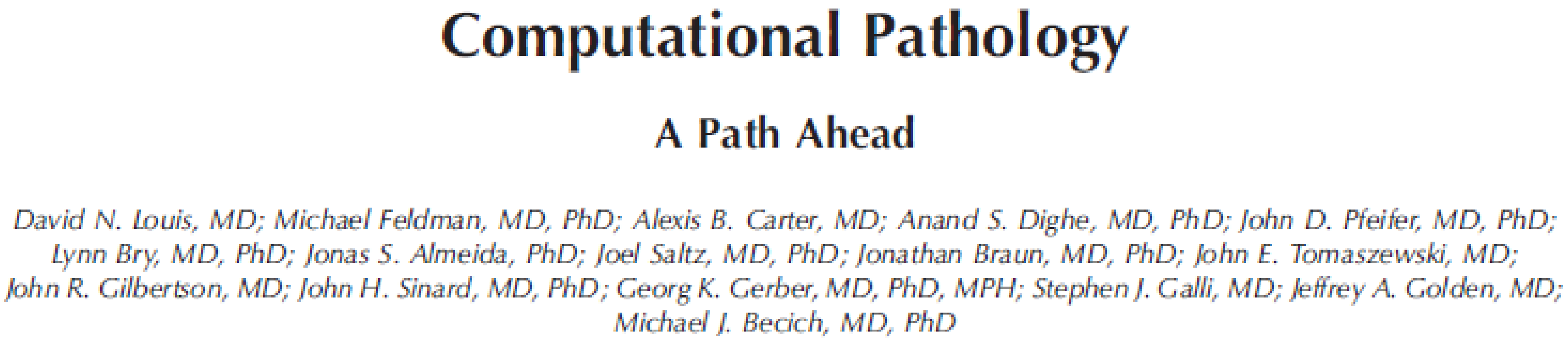
DANS UNE COCOTTE, FAITES REVENIR À L'HUILE D'OLIVE LES OIGNONS COUPÉS EN MORCEAUX. ◆ LORSQU'ILS SONT BLONDS, AJOUTEZ L'AIL HACHÉ, LES TOMATES COUPÉES EN QUATRE ET ÉPÉPINÉES, LE CONCENTRÉ, LE VIN BLANC, LES HERBES, SEL ET POIVRE. MÉLANGEZ. ◆ FAITES PARTIR À GROS BOUILLONS, AJOUTEZ TOUS LES LÉGUMES COUPÉS EN DÉS, LES AUBERGINES ÉPLUCHÉES, MAIS PAS LES COURGETTES. ◆ FAITES CUIRE À FEU MODÉRÉ, SANS COUVERCLE, PENDANT 40 À 50 MINUTES, JUSQU'À CE QUE TOUS LES ÉLÉMENTS SOIENT CUITS. CETTE RATATOUILLE SE SERT CHAUDE EN LÉGUME D'ACCOMPAGNEMENT DE RÔTIS OU FROIDE EN HORS-D'ŒUVRE.

Knowledge

Structured data

ACTIONABLE

2015



Computational Pathology

A Path Ahead

David N. Louis, MD; Michael Feldman, MD, PhD; Alexis B. Carter, MD; Anand S. Dighe, MD, PhD; John D. Pfeifer, MD, PhD; Lynn Bry, MD, PhD; Jonas S. Almeida, PhD; Joel Saltz, MD, PhD; Jonathan Braun, MD, PhD; John E. Tomaszewski, MD; John R. Gilbertson, MD; John H. Sinard, MD, PhD; Georg K. Gerber, MD, PhD, MPH; Stephen J. Galli, MD; Jeffrey A. Golden, MD; Michael J. Becich, MD, PhD

- Working group of pathology chairs and informatics experts
- Defined scope and future needs to develop discipline

Ack

From left to right (top row):

Michael Prystowsky, Einstein, Michael.Prystowsky@einstein.yu.edu
Steve Galli, Stanford, sgalli@stanford.edu
Jeff Golden, Brigham and Women's, jagolden@partners.org
Jonas Almeida, Univ of Alabama, jalmeida@uab.edu
John Tomaczevski, SUNY Buffalo, johntoma@buffalo.edu
Jeff Saffitz, Beth Israel Deaconess, Jsaffitz@bidmc.harvard.edu
Jon Morrow, Yale, jon.morrow@yale.edu
Jim Musser, Methodist in Houston, JMMusser@tmhs.org
Mike Becich, Univ of Pittsburgh, becich@pitt.edu
Metin Gurcan, Ohio State Univ, metin.gurcan@osumc.edu
Jonathan Braun, UCLA, jbraun@mednet.ucla.edu
Don Karcher, George Washington, dkarcher@mfa.gwu.edu
Joel Saltz, SUNY Stony Brook, joel.saltz@stonybrook.edu
Anand Dighe, MGH, asdighe@partners.org
Geoffrey Smith, Emory, ghsmith@emory.edu
Abul Abbas, Univ of San Francisco, abul.abbas@ucsf.edu
Tris Parslow, Emory, tparslo@emory.edu
Georg Gerber, Brigham and Women's, ggerber@partners.org
John Gilbertson, MGH, jrgilbertson@partners.org
Andrew Beck, Beth Israel Deaconess, abeck2@bidmc.harvard.edu
Bob McGonnagle, CAP Today, bmcgonn@cap.org
John Pfeifer, Washington Univ., pfeifer@path.wustl.edu

From left to right (bottom row):

David Louis, MGH, louis@helix.mgh.harvard.edu
Robert Daber, UPenn, robert.daber@upenn.edu
Rakesh Najarian, Washington Univ, rakesh@wustl.edu
Peter Jensen, Utah, peter.jensen@path.utah.edu
Kevin Roth, UAB, karoth@uab.edu
Ulysses Balis, Univ of Michigan, ulysses@med.umich.edu
Alexis Carter, Emory, abcart2@emory.edu
Mike Feldman, UPenn, feldmanm@upenn.edu
John Sinard, Yale, john.sinard@yale.edu
Lynn Bry, Brigham and Women's, lbry@partners.org
Bruce Beckwith, NorthShore Medical Ctr, bbeckwith@partners.org
James Versalovic, Texas Children's Hospital, jxversal@texaschildresnhospital.org

Not shown:

Brian Jackson, ARUP/Utah, brian.jackson@aruplab.com
Anant Madabhushi, Case Western, axm788@case.edu
David Roth, UPenn, david.roth2@uphs.penn.edu
Marianne Boswell, MGH, mboswell@partners.org
Rebecca Crowley, Univ of Pittsburgh, crowleyr@pitt.edu
Rajesh Dash, Duke, r.dash@duke.edu
Mahul Amin, Cedars Sinai
David Foran, Rutgers, djf@pleiad.umdnj.edu
Bruce Friedman, Univ of Michigan emeritus, bfriedma@med.umich.edu
Wade Rodgers, UPenn, rogersw@mail.med.upenn.edu

Conclusions of Working Group

- Strengths
  - Future is **ours to lose**
- Opportunities
  - Must be viewed as essential
  - We have **untapped pools of future experts**
    - We are missing out on our female population
      - 58% of new pathology residents are women (AAMC)
      - Only 15% of board certified informaticists are women
    - We are missing out on our minority populations as well
- Weaknesses
  - Lack of necessary number of trained experts
  - **Lack of computational culture**
- Threats
  - **FDA LDT draft guidance**
  - Someone else getting to it first

FDA LDT Draft Guidance

- FDA cited ‘‘high-tech instrumentation and software to generate results and interpretations’’ as reason for ‘‘increased risk” without oversight compared with the so-called traditional LDTs used prior to 1976
- FDA stated that in ‘‘considering whether to exercise enforcement discretion for Traditional LDTs,’’ several factors would be considered
  - one of which was whether the LDT was interpreted **with<u>OUT</u>** the use of automated instrumentation or software for interpretation

References

- Leggett RM, Ramirez-Gonzalez RH, Clavijo BJ, et al. Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. *Front Genet.* 2013; 4: 288*. doi: 10.3389/fgene.2013.00288*
- Louis DN, Feldman M, Carter AB, et al. Computational Pathology. *Arch Pathol Lab Med. 2015.*
- Louis DN, Gerber GK, Baron JM, et al. Computational pathology: an emerging definition. *Arch Pathol Lab Med. 2014;138(9):1133-1138.*
- Roy S, LaFramboise WA, Nikiforova MN, et al. Next-Generation Sequencing Informatics. Arch Pathol Lab Med. 2016 Feb 22. [Epub ahead of print]

Questions?