



NGS in Pathology Webinar

NGS Data Analysis

March 10 2016

Topics for today's presentation

Primary data analysis

Secondary data analysis

Cancer-specific considerations in NGS data analysis

Tertiary data analysis

Data visualization

IT infrastructure

Further information

Introduction

Next Generation Sequencing (NGS) is becoming a common and versatile tool for biological and medical research.

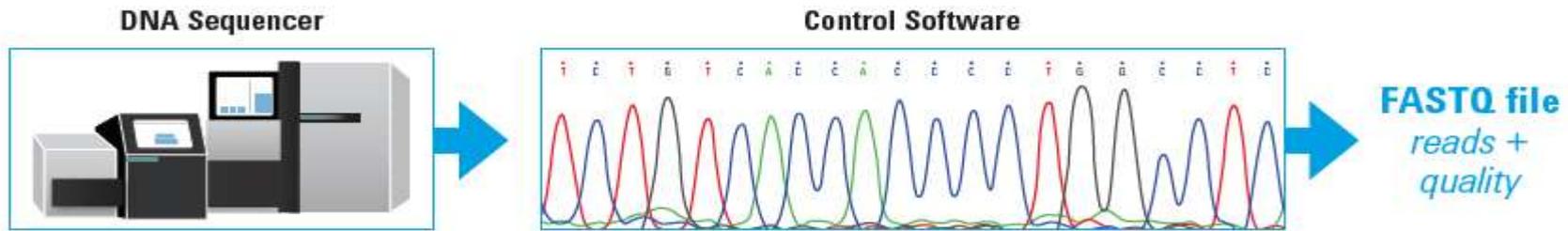
NGS data analysis remains one of the major bottlenecks preventing NGS from being implemented as a routine technique.

The average NGS data analysis software suite is composed of tools for **primary, secondary, and tertiary analysis**.

- This analysis includes image acquisition, quality control, base calling, alignment to a reference genome, variant calling, and biological interpretation tools. We also point out the cancer-specific considerations.

Primary data analysis - overview

In the primary data analysis, raw data are converted to sequence data.



First stage of NGS data analysis. Primary analysis: base calling.

In sequencing by synthesis, base pairs are typically identified after laser excitation and fluorescence detection, and an image is generated. Other technologies directly translate chemically encoded information (A, C, G, T) into digital information (0, 1) on a semiconductor chip.

Primary data analysis

Primary analysis is the responsibility of the sequencing platform and is highly integrated with the sequencing instruments. It is usually installed on the local hardware systems supporting the sequencing instruments.

This analysis often results in a FASTQ file, which is just a combination of the sequence data as a string of A, C, G, T and N (for no base call) characters and an associated Phred quality score for each of those bases.

Primary data analysis

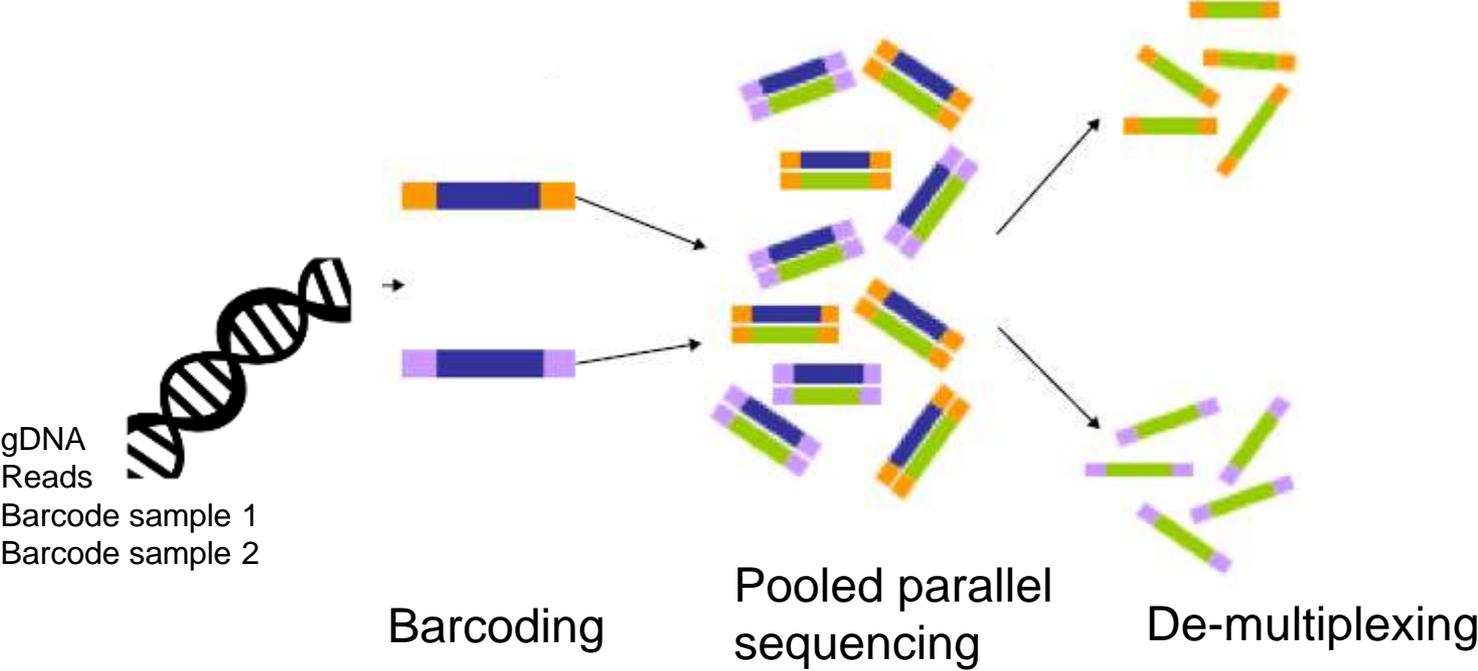
Primary analyses also provide quality values for each base that are used in subsequent phases of analysis, much like Phred quality scores Sanger sequencing. Phred quality scores are linked to error probabilities. Quality score = measure of confidence.

Phred quality score	Probability of incorrect base	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

Q30 is the acceptable score for most sequencing data. It indicates a 99.9% certainty that the base has been called correctly. This is considered high quality data and the standard value commonly used by sequencing facilities.

Primary data analysis

In many instances several samples are sequenced together. Every read from the same sample has the same barcode. After sequencing, the reads are sorted by barcode and grouped together (de-multiplexing).



Topics for today's presentation

Primary data analysis

Secondary data analysis

Cancer-specific considerations in NGS data analysis

Tertiary data analysis

Data visualization

IT infrastructure

Further information

Secondary data analysis – overview

Once the raw sequence data (usually in FASTQ format) is obtained from the NGS instrument, the computationally intensive step of read mapping or alignment is performed. Variants are then identified.



Second stage of NGS analysis. Secondary analysis: read alignment and variant calling.

Secondary data analysis

Mapping short reads **against a reference genome** is the most standard way to align data and should be as accurate as possible but also needs to complete in a reasonable amount of time.

BWA (Burrows-Wheeler Alignment) is the most popular mapping tool for short-read DNA sequences.

A **SAM or BAM alignment file** is generated. SAM stands for Sequence Alignment/Map format. BAM files are binary versions of SAM files, are smaller than SAM files and are non-human readable. These formats have become industry standards for reporting alignment/mapping information.

Secondary data analysis

Next, several **refinement steps** (such as adapter trimming, soft clipping, etc.) are performed. These steps routinely include flagging or filtering of duplicate reads likely to be PCR artifacts, and realignment, which leverages a collective view of reads around putative indels (insertions and deletions) to minimize erroneous alignment of read ends.

Quality scores allocated by the sequencing software will often be recalibrated on the basis of alignment data, before proceeding to the **variant calling stage**.

Topics for today's presentation

Primary data analysis

Secondary data analysis

Cancer-specific considerations in NGS data analysis

Tertiary data analysis

Data visualization

IT infrastructure

Further information

Cancer-specific considerations in NGS data analysis – overview

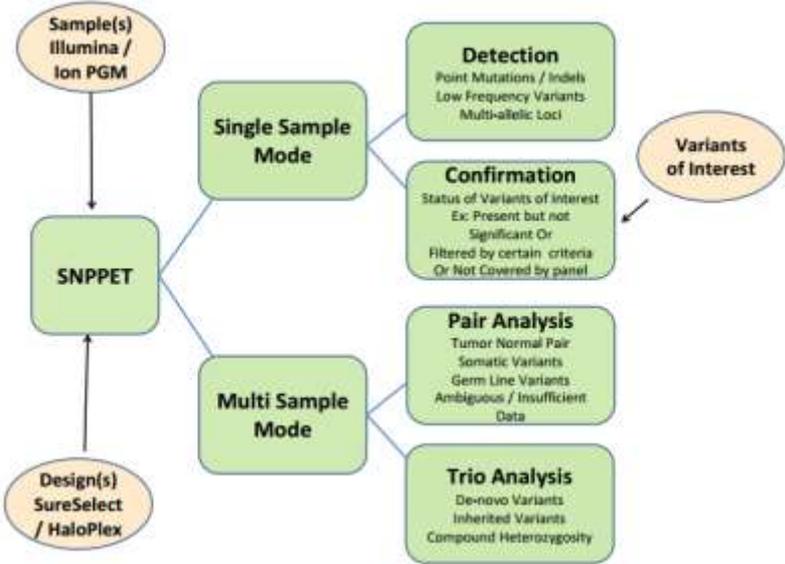
Most bioinformatic tools for processing NGS data have been designed for normal, i.e. **diploid**, genomes, and the assumptions behind their development are commonly invalid in tumor samples because of their *poor quality and limited amounts of DNA, paired-normal comparisons, intra-tumor clonal heterogeneity, polyploidy states, and the fact that many cancer genomes are highly rearranged.*

More recently, cancer-specific variant calling programs have been developed to process NGS data. Cancer-specific callers include JointSNVMix, Somatic Sniper, MuTect, and VarScan.

Cancer-specific considerations in NGS data analysis

Agilent developed its own algorithm called **SNPPET** that addresses the shortcomings of SAMtools and other algorithms that weren't designed to handle cancer samples. SNPPET, for example, detects low frequency variants reliably and handles mosaic samples well.

SNPPET configurations for single and multi-sample variant detection and confirmation.



Cancer-specific considerations in NGS data analysis

The identification of **indel** mutations compared to SNPs is more challenging. Several programs (for example BreakDancer, Dindel, and Pindel) are available to identify indels.

Larger **indels**, from tens to thousands of bases, and are common in cancer, need special methods for their identification. The use of paired-end reads, fragments sequenced from one end and from both ends, and then matched to each other, is one solution. This technique may also identify fusion genes, inversions, and translocations. And is often used to detect fusions and translocations in RNA.

Cancer-specific considerations in NGS data analysis

The roles of Copy Number Variants (CNVs) and Loss of Heterozygosity (LOH) are well known in tumorigenesis. At present, there are no consensus tools for performing these types of analyses. Two types of approaches can be taken and a combination of these may provide improved accuracy. The first approach regards **CNVs as very large indels** and looks for breakpoints within single reads, misalignment of paired reads, or overall difficulty in mapping. The second approach uses the **number of reads at any site as an indicator of copy number**, analogous to microarray technology.

Useful freeware sequence analysis tools for cancer NGS data

Mapping software programs		
BWA	http://bio-bwa.sourceforge.net	Burrows-Wheeler Alignment tool.
Stampy	http://www.well.ox.ac.uk/stampy	Maps short reads to a reference genome using Illumina reads. Particularly useful for indels. Can be used in combination with BWA.
Bowtie	http://bowtie-bio.sourceforge.net/index.shtml	Bowtie is an ultrafast, memory-efficient short read aligner.
Bowtia2	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml	Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences.
Variant callers		
GATK	http://www.broadinstitute.org/gsa/wiki/	Could be considered the gold standard for variant calling. Structured software library that has programs to analyze NGS data. Can be used for variant calling and identification of indels.
JointSNVMix	http://code.google.com/p/joint-snv-mix/	Analyses tumor and normal genome pairs simultaneously so that germline and somatic mutations can be distinguished.
MuTect	http://www.broadinstitute.org/cancer/cga/mutect	A variant caller to identify somatic point mutations from tumor normal paired sequencing data. The program can determine from the depth of coverage in tumor and normal whether there is sufficient sensitivity to call a somatic mutation.
SAMtools	http://samtools.sourceforge.net/	Tools for manipulating aligned data (including SNP finding).
Somatic Sniper	http://qmt.genome.wustl.edu/somatic-sniper/current/	The program compares tumor and normal data to produce a Phred-based probability score to determine the likelihood of the tumor and normal genotypes being different.
Varscan2	http://dkoboldt.github.io/varscan/	Can be used to identify somatic and germline variants and LOH events in tumor normal pairs. Has been used to identify CNVs in tumor normal exome data. It is a platform independent tool working on data with most NGS platforms including Ion Torrent.

Useful freeware sequence analysis tools for cancer NGS data

Freebayes

<https://arxiv.org/abs/1207.3907>

Bayesian genetic variant detector designed to find small polymorphisms, specifically SNPs, indels, MNPs (multi-nucleotide polymorphisms), and complex events (composite insertion and substitution events) smaller than the length of a short-read sequencing alignment.

SNPPET

<http://www.agilent.com/genomics/surecall>

Included in Agilent's data analysis software SureCall. See the SureCall* manual for a detailed description of SNPPET.

Indels and structural variant callers

BreakDancer

<http://breakdancer.sourceforge.net>

BreakDancer Max – can identify structural variants by noting paired-end reads, which are mapped at unexpected distances or are incorrectly orientated. Detects large insertions, deletions, inversions, inter/intra chromosomal translocations. BreakDancer Mini – used to detect small indels 10–100 bp.

Dindel

<https://sites.google.com/site/keesalbers/soft/dindel>

Small indel discovery. With deeper coverage the number of false positives can be reduced by filtering the data to ensure that each indel is present more than twice.

Genome STRiP

<http://www.broadinstitute.org/software/genomestrip/genome-strip>

Designed to detect structural variations shared by multiple individuals. Needs 20–30 genomes to achieve satisfactory results. Its current use is limited to uncovering and genotyping deletions relative to a reference sequence.

Pindel

<http://gmt.genome.wustl.edu/packages/pindel/>

Can be used to identify simple deletions and insertions. Uses paired-end reads to identify large breakpoints and medium size insertions. Can detect inversions and tandem duplications.



Topics for today's presentation

Primary data analysis

Secondary data analysis

Cancer-specific considerations in NGS data analysis

Tertiary data analysis

Data visualization

IT infrastructure

Further information

Tertiary data analysis - overview

Interpretation or tertiary data analysis, can be the most complex, experiment-specific, time-consuming and manual phase of the NGS data analysis pipeline.

To put things in perspective, a germline whole exome sequencing (WES) generates an average of about **30,000 SNPs** in the case of Agilent SureSelectXT2 Human All Exon V6+COSMIC that will need to be filtered.

Most of these SNPs are synonymous or benign changes. The remaining rare variants need to be further investigated and classified as **pathogenic, benign, or VOUS** (“**variant of unknown clinical significance**”).

Many tools and various databases are publicly or commercially available for variant interpretation. Software to predict the function of variants is also available, such as **SIFT and PolyPhen**. Several tools and resources that are useful in variant annotation and interpretation such as **CIViC**.

Agilent’s Cartagenia Bench Lab for Molecular Pathology contains many of these tools and databases and allows you to efficiently triage somatic variants and filter down to relevant candidates for review and assessment in their clinical context, including previous findings and information on biomedical relevance.

Resources for predicting mutant protein function and variant interpretation

Program	Description	URL
Polyphen-2	Mutation function prediction	http://genetics.bwh.harvard.edu/pph2
SIFT	Mutation function prediction	http://sift.jcvi.org
CHASM	Mutation function prediction	http://wiki.chasmsoftware.org
ANNOVAR	Annotation	http://www.openbioinformatics.org/annovar/
COSMIC	Catalog of Somatic Mutations in Cancer	http://www.sanger.ac.uk/genetics/CGP/cosmic
UCSC Cancer Genomics Browser	Web-based tools to visualize, integrate, and analyze cancer genomics and its associated clinical data	https://xenabrowser.net/heatmap/
Cancer Genome Workbench	Hosts mutation, copy number, expression, and methylation data from TCGA, TARGET, COSMIC, GSK, and NCI60 projects; tools for visualizing sample-level genomic and transcription alterations in various cancers	https://cqwbc.nci.nih.gov/
HGVS	Human Genome Variation Society; recommendations for the annotation of variants	http://varnomen.hgvs.org/
RefSeq Database	Derivation of coding sequence	www.ncbi.nlm.nih.gov/RefSeq
dbSNP	Single Nucleotide Polymorphism Database	http://www.ncbi.nlm.nih.gov/projects/SNP
HGMD	Human Gene Mutation Database	http://www.biobase-international.com/product/hgmd
ClinVar	Aggregates information about genomic variation and its relationship to human health	https://www.ncbi.nlm.nih.gov/clinvar/
CIViC	Clinical Interpretation of Variants in Cancer	https://civic.genome.wustl.edu/#/home

Topics for today's presentation

Primary data analysis

Secondary data analysis

Cancer-specific considerations in NGS data analysis

Tertiary data analysis

Data visualization

IT infrastructure

Further information

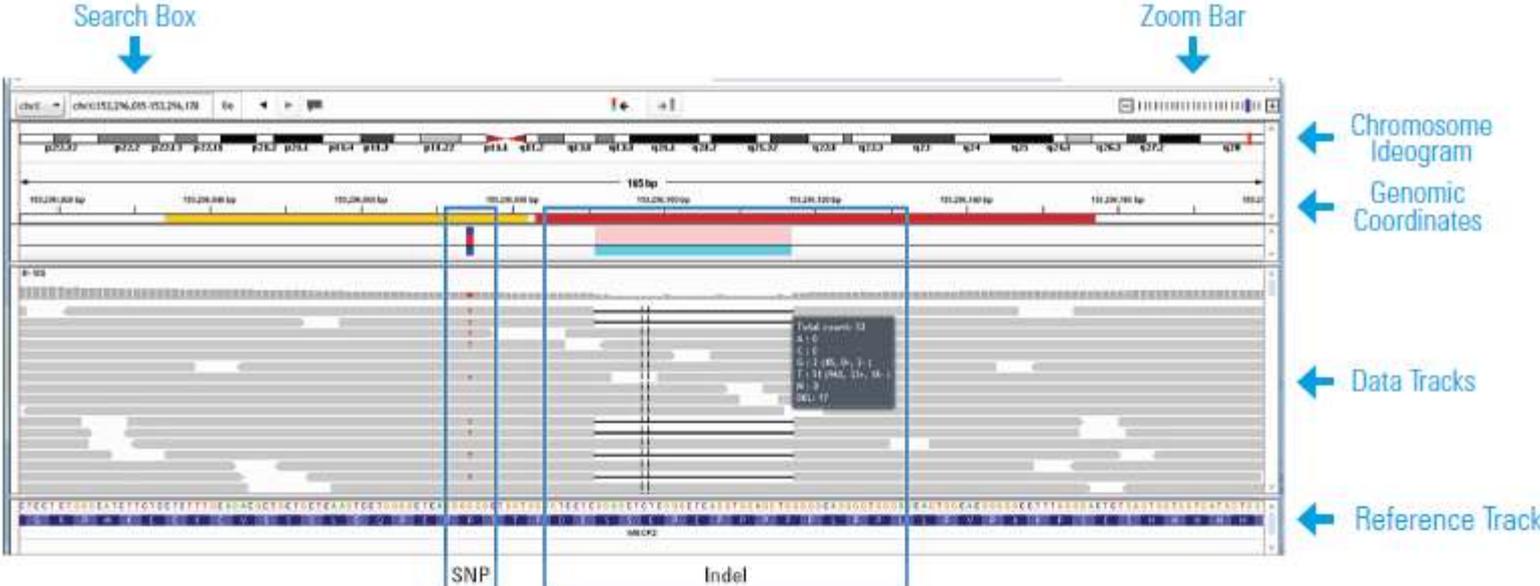
Data visualization – overview

In general, there is no need to view raw data. But indels or gene fusions can be problematic: **some indel reads** are discarded because they don't seem to align properly to the reference.

The **Integrative Genomics Viewer (IGV)** is a high-performance viewer that efficiently handles large data sets. For example, BAM files, VCF files and BED files can be visualized in IGV. IGV is a desktop application written in the Java programming language and runs on all major platforms (Windows, Mac and Linux). Its primary emphasis is to support researchers who wish to visualize and explore their own data sets or those from colleagues. IGV is freely available for download from <http://www.broadinstitute.org/igv>, under a GNU LGPL open-source license.

Data visualization

The top command bar with controls for selecting a reference genome, navigating and defining regions of interest. Just below the command bar is a header panel with an ideogram representation of the currently viewed chromosome, along with a genome coordinate ruler that indicates the size of the region in view. The remainder of the window is divided into one or more data panels and an attribute panel.



The IGV analysis window.

Topics for today's presentation

Primary data analysis

Secondary data analysis

Cancer-specific considerations in NGS data analysis

Tertiary data analysis

Data visualization

IT infrastructure

Further information

IT infrastructure - overview

The necessary IT infrastructure depends a lot on the size and volume of the sequencing data sets. Some data sets can be **enormous**. For example, whole-human-genome sequencing projects, including the raw sequences, alignments, and variant calls, can run into the hundreds of gigabytes per sample. Whereas, with targeted sequencing of samples, data sets are typically much smaller, probably just megabytes or a few gigabytes in size.

The other key resource, besides storage space, is **computing power**. These data sets can be so large, they often cannot reasonably be analyzed, especially aligned, using a desktop computer. On the low end, users typically need 16 to 24 gigabytes of memory to align sequencing data.

Many universities or research institutions offer **cluster resources**. A cluster can be viewed as a kind of ad-hoc supercomputer built by networking a number of smaller computers in parallel.

For researchers that do not have access to a cluster, several web- and cloud-based alternatives exist. One example of such a **cloud-based alternative** is Amazon Web Services (AWS). Systems like this offer tremendous flexibility by letting the cloud service provider do the heavy lifting on computationally intensive operations. Moreover, purchasing, maintaining, and upgrading IT infrastructure is no longer an issue with those solutions. For users who like a more end-to-end solution, there are commercial options like Agilent's Cartagena Bench Lab for Molecular Pathology.

Topics for today's presentation

Primary data analysis

Secondary data analysis

Cancer-specific considerations in NGS data analysis

Tertiary data analysis

Data visualization

IT infrastructure

Further information

Further information

For more information and resources

NGS Cancer Resource Center

<http://www.genomics.agilent.com/article.jsp?pagelid=8200008&CID=G012110>

Cartagenia Bench Lab for Molecular Pathology

<http://www.agilent.com/en-us/solutions/clinical-grade-variant-assessment/cartagenia-bench-lab-for-molecular-pathology>

File format glossary

Name	Description
FASTA	The FASTA file format is a simple text format that contains one or more records consisting of a start indicator (most commonly ">"), a comment, and then a sequence of nucleotides encoded as letters of the alphabet.
FASTQ	FASTQ is a file format initially developed by the Wellcome Trust Sanger Institute to bundle a FASTA sequence and its quality data.
SAM	SAM is a series of sequences and their alignment to a reference genome. The SAM (Sequence Alignment/Map) format is a very versatile and near-standardized format for storing many aligned nucleotide sequences.
BAM	BAM is also series of sequences and their alignment to a reference genome but is a more compact binary equivalent of the SAM format.
VCF	Variants between a sample and the reference genome standardized by the 1000 Genomes Project.
BED	A BED file (.bed) is a tab-delimited text file that defines a feature track.

References

- (1) Oliver *et al.* Bioinformatics for Clinical Next Generation Sequencing. *Clinical Chemistry* (2015) 61: 124-135.
- (2) Gullapalli *et al.* Next Generation Sequencing in Clinical Medicine: Challenges and Lessons for Pathology and Biomedical Informatics. *J Pathol Inform* (2012) 3:40.
- (3) Chang *et al.* Clinical Application of Amplicon-Based Next-Generation Sequencing in Cancer. *Cancer Genetics* (2013) 206: 413-419.
- (4) Thorvaldsdottir *et al.* Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration. *Briefings in bioinformatics* (2012) 14: 178-192.
- (5) Ulahannan *et al.* Technical and Implementation Issues in Using Next-Generation Sequencing of Cancers in Clinical Practice. *British Journal of Cancer* (2013) 109: 827-835.
- (6) Perkel *et al.* Sequence Analysis 101. *The Scientist* (2011) March 1.

Stay tuned for the next webinars in this series.....

WEBINAR

NGS in Cancer Pathology: NGS Data Interpretation and Reporting

Date: May 3
Time: 8 am PST, 5 pm Paris

Speaker: Melissa Rose, Application Scientist, Clinical Applications Division, Agilent Technologies

Cartagenia Bench Lab is marketed in the USA as exempt Class I Medical Device and in Europe and Canada as a Class I Medical Device.



Melissa Rose

[Register now](#)

WEBINAR

NGS in Cancer Pathology: Introduction to RNA-Seq

Date: May 10
Time: 8 am PST, 5 pm Paris

Speaker: Jennifer Carter Jones Ph.D., Field Application Scientist, Genomics Applications, Agilent Technologies

For Research Use Only. Not for use in diagnostic purposes.



Jennifer Carter Jones, Ph.D.

[Register now](#)

To register for these upcoming webinars, please look for the follow up email about today's webinar. It will contain a link to a recording of today's webinar and upcoming webinars in the series.