2 Next Generation Sequencing Data and Proteogenomics

Kelly V. Ruggles and David Fenyö

Abstract

The field of proteogenomics has been driven by combined advances in next-generation sequencing (NGS) and proteomic methods. NGS technologies are now both rapid and affordable, making it feasible to include sequencing in the clinic and academic research setting. Alongside the improvements in sequencing technologies, methods in high throughput proteomics have increased the depth of coverage and the speed of analysis. The integration of these data types using continuously evolving bioinformatics methods allows for improvements in gene and protein annotation, and a more comprehensive understanding of biological systems.

Keywords

Next generation sequencing • Proteogenomic integration • Bioinformatics • Peptide identification • Gene annotation

2.1 NGS Overview

NGS itself refers to a number of techniques, all of which perform massively parallel sequencing, in which millions of DNA fragments from a sample are sequenced at the same time (Muzzey et al. 2015). This produces a vast amount of data, in some cases adding up to 1 TB per run. With this level of data volume and faster data generation, bioinformatics has emerged as the true challenge in NGS data analysis and integration.

The most frequently used NGS methods at the DNA level are whole exome sequencing and whole genome sequencing (WGS). In whole

K.V. Ruggles
Department of Medicine, New York University
Medical Center, 550 First Avenue, New York,
NY 10016, USA

Center for Health Informatics and Bioinformatics,
New York University Medical Center, 227 East 30th
Street, New York, NY 10016, USA
e-mail: kelly.ruggles@nyumc.org

D. Fenyö (✉)
Institute for Systems Genetics, New York University
Medical Center, 430 East 29th Street, New York,
NY 10016, USA

Department of Biochemistry and Molecular
Pharmacology, New York University Medical Center,
550 First Avenue, New York, NY 10016, USA
e-mail: david@fenyolab.org

Á. Végvári (ed.), *Proteogenomics*, Advances in Experimental Medicine and Biology 926,
DOI 10.1007/978-3-319-42316-6_2

exome sequencing, only protein-coding regions of the genome are sequenced, removing the remaining ~99 % of the DNA and thereby significantly lowering the required time and cost. This method has been most often employed in studies of gene discovery and the identification of disease causing mutations. For WGS however, the entire genome is sequenced, which is useful for novel gene identification and for the analysis of non-coding regions including promoters and enhancers.

DNA NGS technologies have enabled researchers to detect differences between an experimental and a reference genome. These typically fall into two categories:

1. Large deletions/duplications (copy number variation (CNV))
2. Changes to the DNA sequence, also known as "variants", either as single nucleotide polymorphisms (SNPs) or short insertion/deletions (indels)

Both require alignment of NGS reads to a reference genome (Fig. 2.1).

NGS is also performed on RNA using RNA-Seq, a technique which is now frequently used in lieu of micro-arrays to assess gene expression. RNA-Seq enables researchers to investigate alternative splicing events, gene fusion events, SNPs and gene expression. The experimental procedure is similar to that of DNA sequencing, with an additional step of first deriving cDNA sequences from all RNA present in the sample.

Although different sequencing methods can produce different raw data types, these data are most often combined to create a FASTQ file, containing information on both sequence and quality. This data is first aligned to the reference genome and stored in a sequence alignment map (SAM) or binary alignment map (BAM) file using a sequence alignment algorithm (Li et al. 2009) (Fig. 2.1). A number of algorithms have been developed for this purpose, using Burrows-Wheeler Transformation (BWT) techniques (*e.g.*, Bowtie/Bowtie 2 (Langmead and Salzberg 2012), BWA/BWA-SW (Li and Durbin 2010)) and/or Smith-Waterman (SW) dynamic programing (*e.g.*, SHRiMP/SHRiMP2 (David et al. 2011; Rumble et al. 2009)).

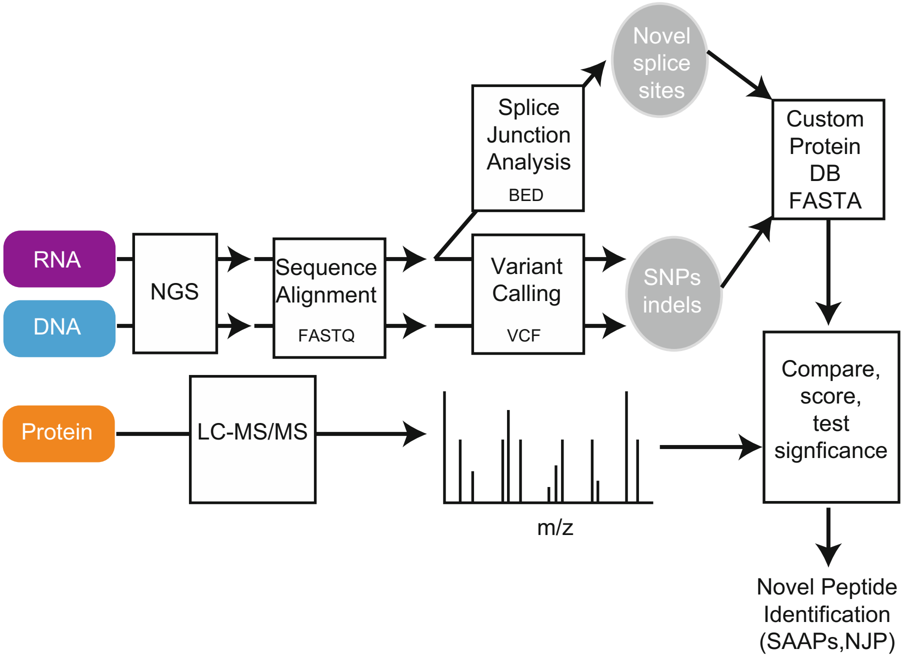

Fig. 2.1 Proteogenomic overview

2.2 Variant Identification Using Proteogenomics

2.2.1 Single Nucleotide Polymorphisms (SNPs)

Following alignment of the DNA or RNA sequence, subsequent variant calling, filtering and annotation can be completed. Variants are found through the identification of small deviations between the experimental and the reference genome (Figs. 2.1 and 2.2). These variants may be disease drivers, or mutations having little to no functional impact. Several programs have been developed specifically for the purpose of variant calling and each produces a list of variant positions stored in a Variant Call Format (VCF) file (Danecek et al. 2011).

A primary challenge with SNP variant calling is in identifying "true" variants and filtering out those due to errors in sequencing or alignment (Nielsen et al. 2011). Informatics packages have been developed for variant calling, including the popular Genome Analysis Toolkit (GATK) (McKenna et al. 2010) and VarScan (Koboldt et al. 2012). Indel mutation identification presents an additional set of complications, because it requires a more sophisticated approach to gapped alignment and paired-end sequence inference. Pattern growth approach software (*e.g.*, Pindel (Ye et al. 2009)), baysian-based algorithms (*e.g.*, Dindel (Albers et al. 2011)) and the variant calling algorithm GATK (McKenna et al. 2010) have all been refined for accurate indel identification (Neuman et al. 2013).

Following variant calling, filtering and annotation are common steps for isolating variants most likely to contribute to the pathology of interest. Although quality cutoffs for variant identification should always be employed, additional filtering becomes less important in proteogenomic analysis because proteomic data can be leveraged for variant validation.

2.2.2 Single Amino Acid Polymorphisms (SAAP)

Identifying variants that are expressed at the protein level presents a non-trivial informatics challenge in that mass spectrometric identification of peptide sequences is dependent upon the inclusion of that sequence in the protein database. Protein sequence database searching algorithms such as X!Tandem (Craig et al. 2005), Mascot (Perkins et al. 1999) and MSGF+ (Granholm et al. 2014) match the MS/MS spectra against a list of candidate peptide sequences and score the similarity of

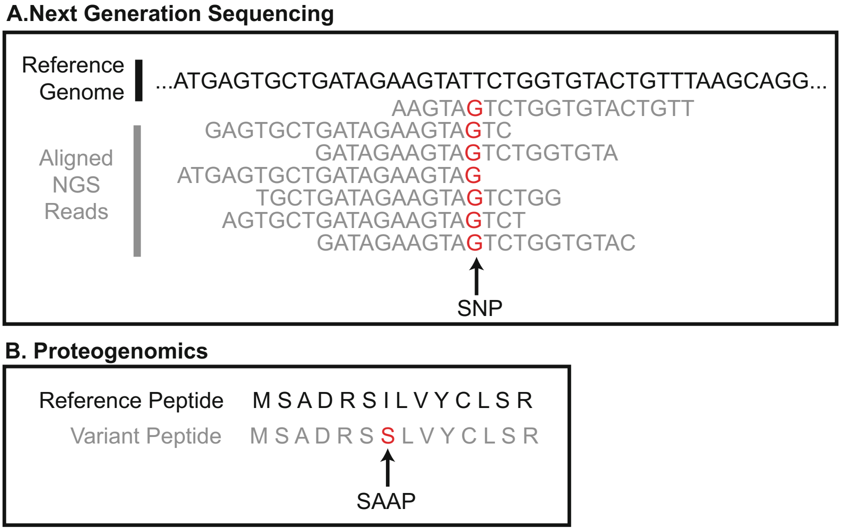

Fig. 2.2 Single nucleotide polymorphism and single amino acid polymorphism identification

a theoretical or library spectrum to the acquired spectrum based on mass. Databases with missing sequences will fail to identify these peptides in the MS/MS data and ideally, the protein database would contain all proteins present in the sample with minimal irrelevant sequences (Fig. 2.1).

Therefore, in order to identify single amino acid polymorphisms (SAAPs) occurring from non-synonymous genomic SNPs, one must create a protein sequence database that incorporates the sequencing data to contain corresponding SAAPs. These changes are integrated into the protein sequence data by first modifying the genomic reference sequence to include SNPs in the genome and/or transcriptome (Fig. 2.2a) and then completing an *in silico* protein translation of the modified sequences to attain a list of peptides containing SAAPs (Fig. 2.2b).

2.2.3 Bioinformatics Tools for Creating SAAP Protein Sequence Databases

Several tools have been developed to create NSG-integrated databases containing potential variant SAAPs. With the inclusions of these novel peptide sequences in the database, variant peptides can be identified from MS/MS data.

These tools include:

- **QUILTS**: Open source tool that incorporates SNPs from either DNA sequencing or RNA-Seq and allows for up to two variant VCF input files to accommodate cancer studies which require both germline and somatic (cancer specific) SNP options. QUILTS then creates a FASTA-formatted protein sequence database that can be used by common database searching algorithms (Ruggles et al. 2015). *quilts.fenyolab.org*
- **customproDB:** R package developed for customized protein database construction using SNPs and indels from RNA-Seq data. The output is also a FASTA-formatted sequence file (Wang and Zhang 2013). *www.bioconductor.org/packages/release/bioc/html/customProDB.html*

2.3 Alternative Splicing and Gene Annotation

Coding of novel gene regions and alternative splicing provides additional biological complexity. The advent of RNA-Seq has shown alternative splicing to occur in over 90 % of human genes (Pal et al. 2012), emphasizing the role of diverse protein isoforms in cellular function. RNA-Seq analysis provides information on splice junctions (intron / exon boundaries) present in a given sample, providing insight into both normal gene annotation and novel expression. Splice sites are identified following sequence alignment using splice-alignment software such as TopHat (Kim et al. 2013), BLAT (Fonseca et al. 2012) and MapSplice (Wang et al. 2010) (Fig. 2.1).

Comparing intron / exon boundaries identified through NGS to known junction boundaries can identify novel splice sites, including unannotated alternative splicing (two known exons) (Fig. 2.3a), partially novel splicing (one known exon) (Fig. 2.3b) and completely novel splicing (no known exons) (Fig. 2.3c) (Ruggles et al. 2015; Mertins et al. 2016). Hundreds of thousands of novel splice sites can be identified by one RNA-Seq experiment, but the fraction of functional versus "spurious" splicing requires additional information to be determined. Since *ab initio* methods for the identification of novel splice sites are limited (Barash and Garcia 2014), the validation of splice-junctions requires peptide evidence spanning these intron / exon boundaries.

2.3.1 Novel Splice Junction (NSJ) Peptides

As with SAAPs, NSJ peptide identification relies on the construction of a comprehensive protein database incorporating alternatively spliced isoforms and novel expression as coded in the transcriptome. These databases should contain all possible NSJ peptides in the sample to insure corresponding peptide identification from tandem MS analysis. Approximately one quarter of

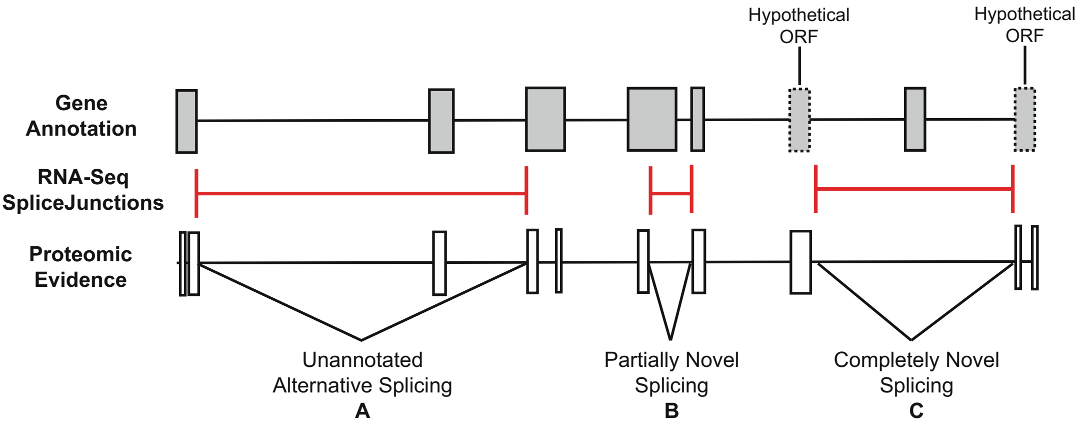

Fig. 2.3 Proteogenomic gene annotation

peptides cross a splice junction in humans, and these are particularly useful for intron / exon boundary and splicing verification. The identification of novel splice sites is most frequently used in:

- Improving gene annotation
- Cancer studies, where alternative splicing and novel expression have been reported to effect disease progression (Ning and Nesvizhskii 2010)

Gene annotation is the process of identifying genes and determining gene function. Prior to NSG, the identification of protein coding regions was done using comparative sequence analysis and gene prediction algorithms, both of which have inherent limitations. These limitations include difficulties in identifying gene start and stop sites and translational reading frames (Brent 2008), difficulty identifying splice boundaries (Reese et al. 2000), and issues in determining boundaries of short and overlapping genes (Warren et al. 2010). RNA-Seq has addressed most of these limitations, but studies have shown that many transcripts show no evidence of protein translation (Clamp et al. 2007; Eddy 2001). Proteogenomics is able to fill this gap by using MS-based proteomics in combination with RNA sequencing to verify gene coding regions and novel splice junctions.

2.3.2 Bioinformatics Tools for Identifying NSJ Peptides

A frequently applied method for proteogenomic gene annotation is the use of a six-frame translation of the DNA sequence of interest as the protein sequence database for the MS/MS peptide search, removing all bias based on the established genome annotation. This method is able to validate existing gene models and start / stop sites, and can also identify novel open reading frames (ORFs) (Fermin et al. 2006; Gupta et al. 2007; Kalume et al. 2005). Two limitations to this method are:

- The inclusion of a six-frame translation considerably increases the search space thereby reducing search sensitivity
- Splicing information cannot be determined, only intron / exon boundaries.

Tools for six-frame translation database searches include:

- **Peppy:** A Java-based software that searches a given six-frame translation database, returning peptide identifications at a user-specified false discovery rate (Risk et al. 2013). *http://geneffects.com/pepp*

- **PIUS (Peptide Identification by Unbiased Search):** Online tool that identifies peptides through a spectral match search of high-throughput MS/MS data using a six-frame translation database (Costa et al. 2013).

An alternative to a six-frame translation search is to use RNA-Seq derived splice junction data to identify novel alternative splicing in addition to unannotated ORFs. This requires more sophisticated informatics tools, which incorporate cases of unannotated alternative splicing (Fig. 2.3a), splicing at a novel intron / exon boundary (Fig. 2.3b), and splicing of novel, hypothetical open reading frames (Fig. 2.3c) to the genomic reference sequence. A Browser Extensible Data (BED) file, containing information on the location of these junctions, is created by most RNA-Seq alignment algorithms and used in the sequence modification step. The protein database is then created using an *in silico* protein translation of these modified sequences to obtain a full NSJ peptide list. Translation of these splice junctions can then be verified by the identification of peptide sequences bridging the transcribed intron / exon boundaries.

Tools that create NSJ protein sequence databases include:

- **QUILTS:** In addition to incorporating SNPs from NSG data to the protein sequence database, QUILTS accepts a Browser Extensible Data (BED) file containing RNA-Seq predicted splice junctions as input and creates FASTA files containing NSJ peptides corresponding to the transcriptome data (Ruggles et al. 2015). *quilts.fenyolab.org*
- **customproDB:** In addition to SNP-based protein database creation, customproDB creates FASTA database files using a putative junction BED file (Wang and Zhang 2013) *www.bioconductor.org/packages/release/bioc/html/customProDB.html*

2.4 Coordinated Gene and Protein Expression

In addition to facilitating the identification of SAAPs and NSJ peptides, proteogenomics can also support coordinated expression analysis based on genomic location. Copy number variation (CNV), defined as large (>1 kb) genomic deletions / duplications, can be derived from whole genome and exome sequencing. CNVs often result in gene dosage effects in multiple genes and have been shown to play a significant role in genetic variation and disease (Iafrate et al. 2004). Most methods for CNV detection can be categorized into two types: pair end mapping (PEM) methods and depth of coverage (DOC) methods. The more popular DOC algorithms such as SegSeq (Chiang et al. 2009) and CNV-seq (Xie and Tammi 2009) align reads on the genome and calculate read counts using sliding bins, which are further processed to determine a normalized copy number (Duan et al. 2013).

At the transcript level, differential gene expression is determined using methods that use read coverage to quantify transcript abundance. For example, RPKM (reads per kilo base per million mapped reads) and FPKM (fragments per kilobase per million) are commonly used methods to quantify normalized expression of a gene (Marioni et al. 2008) and many programs have been developed for the subsequent determination of differential gene expression, these include Cuffdiff (Trapnell et al. 2013), edgeR (Robinson et al. 2010), and DESeq (Anders and Huber 2010).

Proteogenomic tools have been developed that allow for coordinated expression analysis across data types, by converting proteomic location to genomic coordinates. This mapping allows researchers to analyze expression based on genomic location, for example in large areas of gene duplication / deletion, or at the exon level, rather than requiring gene-based analysis. This is particularly useful when displaying expression levels using genome browsers (*e.g.*, UCSF genome browser, Integrative Genomics Viewer (IGV)).

Bioinformatics tools for peptide mapping include:

- **PGx:** Open-source tool that maps peptides onto their putative genomic coordinates using a user-defined reference database. The software maps many peptides simultaneously, returning a BED (qualitative) and bedGraph (quantitative) file, which can be used to then be loaded into a genome browser for visualization (Askenazi et al. 2015). pgx.fenyolab.org
- **The Proteogenomic Mapping Tool:** Java-based software that searches peptides against a six-frame translated sequence database. Output includes a file containing the genomic location of each peptide match that can be visualized using a genome browser (Sanders et al. 2011). www.agbase.msstate.edu/tools/pgm/

2.5 Conclusions

Informatics-based proteogenomic methods help to determine which genomic variants and alternatively spliced gene forms are translated, revealing their biological potential. For example, mutations and novel splice junctions that are found at the peptide level have a higher likelihood of being a driver of disease. Additionally, integrating NGS and proteomic data using proteogenomic mapping tools allows for the simultaneous analysis of gene expression, which can help to better understand the complexities of gene regulation. We expect that as NGS and high throughput proteomic techniques continue to improve, the quantity and quality of associated data will continue to rise and will demand continuously evolving bioinformatics tools for proteogenomic integration and analysis.

Acknowledgement This work was supported by funding provided by the National Institutes of Health through contract S13-068 from Leidos.

References

Albers, C. A., Lunter, G., MacArthur, D. G., McVean, G., Ouwehand, W. H., & Durbin, R. (2011). Dindel: Accurate indel calls from short-read data. *Genome Research, 21*, 961–973. doi:10.1101/gr.112326.110.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology, 11*, R106. doi:10.1186/gb-2010-11-10-r106.

Askenazi, M., Ruggles, K. V., & Fenyö, D. (2015). PGx: Putting peptides to BED. *Journal of Proteome Research*. doi:10.1021/acs.jproteome.5b00870.

Barash, Y., & Garcia, J. V. (2014). Predicting alternative splicing. *Methods Molecular Biology, 1126*, 411–423. doi:10.1007/978-1-62703-980-2_28.

Brent, M. R. (2008). Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nature Reviews Genetics, 9*, 62–73. doi:10.1038/nrg2220.

Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J. T., Zhao, X., Carter, S. L., Russ, C., Nusbaum, C., Meyerson, M., & Lander, E. S. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods, 6*, 99–103. doi:10.1038/nmeth.1276.

Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M. F., Kellis, M., Lindblad-Toh, K., & Lander, E. S. (2007). Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 19428–19433. doi:10.1073/pnas.0709013104.

Costa, E. P., Menschaert, G., Luyten, W., De Grave, K., & Ramon, J. (2013). PIUS: Peptide identification by unbiased search. *Bioinformatics, 29*, 1913–1914. doi:10.1093/bioinformatics/btt298.

Craig, R., Cortens, J. P., & Beavis, R. C. (2005). The use of proteotypic peptide libraries for protein identification. *Rapid Communications in Mass Spectrometry, 19*, 1844–1850. doi:10.1002/rcm.1992.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics, 27*, 2156–2158. doi:10.1093/bioinformatics/btr330.

David, M., Dzamba, M., Lister, D., Ilie, L., & Brudno, M. (2011). SHRiMP2: Sensitive yet practical SHort read mapping. *Bioinformatics, 27*, 1011–1012. doi:10.1093/bioinformatics/btr046.

Duan, J., Zhang, J.-G., Deng, H.-W., & Wang, Y.-P. (2013). Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PloS One, 8*, e59128. doi:10.1371/journal.pone.0059128.

Eddy, S. R. (2001). Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics, 2*, 919–929. doi:10.1038/35103511.

Fermin, D., Allen, B. B., Blackwell, T. W., Menon, R., Adamski, M., Xu, Y., Ulintz, P., Omenn, G. S., & States, D. J. (2006). Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biology, 7*, R35. doi:10.1186/gb-2006-7-4-r35.

Fonseca, N. A., Rung, J., Brazma, A., & Marioni, J. C. (2012). Tools for mapping high-throughput sequencing data. *Bioinformatics, 28*, 3169–3177. doi:10.1093/bioinformatics/bts605.

Granholm, V., Kim, S., Navarro, J. C. F., Sjölund, E., Smith, R. D., & Käll, L. (2014). Fast and accurate database searches with MS-GF+Percolator. *Journal of Proteome Research, 13*, 890–897. doi:10.1021/pr400937n.

Gupta, N., Tanner, S., Jaitly, N., Adkins, J. N., Lipton, M., Edwards, R., Romine, M., Osterman, A., Bafna, V., Smith, R. D., & Pevzner, P. A. (2007). Whole proteome analysis of post-translational modifications: Applications of mass-spectrometry for proteogenomic annotation. *Genome Research, 17*, 1362–1377. doi:10.1101/gr.6427907.

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., & Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics, 36*, 949–951. doi:10.1038/ng1416.

Kalume, D. E., Peri, S., Reddy, R., Zhong, J., Okulate, M., Kumar, N., & Pandey, A. (2005). Genome annotation of Anopheles gambiae using mass spectrometry-derived data. *BMC Genomics, 6*, 128. doi:10.1186/1471-2164-6-128.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology, 14*, R36. doi:10.1186/gb-2013-14-4-r36.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., & Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research, 22*, 568–576. doi:10.1101/gr.129684.111.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*, 357–359. doi:10.1038/nmeth.1923.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics, 26*, 589–595. doi:10.1093/bioinformatics/btp698.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics, 25*, 2078–2079. doi:10.1093/bioinformatics/btp352.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-Seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research, 18*, 1509–1517. doi:10.1101/gr.079558.108.

Mertins, P., Mani, D. R., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., Wang, X., Qiao, J. W., Cao, S., Petralia, F., Mundt, F., Tu, Z., Lei, J. T., Gatza, M., Perou, C. M., Yellapantula, V., Lin, C., Ding, L., McLellan, M., Ping, Y., Davies, S. R., Townsend, R., Zhang, B., Rodriguez, H., Paulovich, A., Fenyo, D., Ellis, M., Carr, S. A., & The NCI CPTAC. (2016). Proteogenomic analysis of human breast cancer connects genetic alterations to phosphorylation networks. *Nature, 534*(7605), 55–62.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The genome analysis toolkit: A mapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20*, 1297–1303. doi:10.1101/gr.107524.110.

Muzzey, D., Evans, E. A., & Lieber, C. (2015). Understanding the basics of NGS: From mechanism to variant calling. *Current Genetic Medicine Reports, 3*, 158–165. doi:10.1007/s40142-015-0076-8.

Neuman, J. A., Isakov, O., & Shomron, N. (2013). Analysis of insertion-deletion from deep-sequencing data: Software evaluation for optimal detection. *Briefings in Bioinformatics, 14*, 46–55. doi:10.1093/bib/bbs013.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics, 12*, 443–451. doi:10.1038/nrg2986.

Ning, K., & Nesvizhskii, A. I. (2010). The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: A preliminary assessment. *BMC Bioinformatics, 11*(Suppl 11), S14. doi:10.1186/1471-2105-11-S11-S14

Pal, S., Gupta, R., & Davuluri, R. V. (2012). Alternative transcription and alternative splicing in cancer. *Pharmacolology & Therapeutics, 136*(3), 283–294. doi:10.1016/j.pharmthera.2012.08.005.

Perkins, D. N., Pappin, D. J., Creasy, D. M., & Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis, 20*, 3551–3567. doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2.

Reese, M. G., Hartzell, G., Harris, N. L., Ohler, U., Abril, J. F., & Lewis, S. E. (2000). Genome annotation assessment in Drosophila melanogaster. *Genome Research, 10*, 483–501.

Risk, B. A., Spitzer, W. J., & Giddings, M. C. (2013). Peppy: Proteogenomic search software. *Journal of Proteome Research, 12*, 3019–3025. doi:10.1021/pr400208w.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data.

Bioinformatics, 26, 139–140. doi:10.1093/bioinformatics/btp616.

Ruggles, K. V., Tang, Z., Wang, X., Grover, H., Askenazi, M., Teubl, J., Cao, S., McLellan, M. D., Clauser, K. R., Tabb, D. L., Mertins, P., Slebos, R., Erdmann-Gilmore, P., Li, S., Gunawardena, H. P., Xie, L., Liu, T., Zhou, J.-Y., Sun, S., Hoadley, K. A., Perou, C. M., Chen, X., Davies, S. R., Maher, C. A., Kinsinger, C. R., Rodland, K. D., Zhang, H., Zhang, Z., Ding, L., Townsend, R. R., Rodriguez, H., Chan, D., Smith, R. D., Liebler, D. C., Carr, S. A., Payne, S., Ellis, M. J., & Fenyo, D. (2015). An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. *Molecular Cellular Proteomics*. doi:10.1074/mcp.M115.056226.

Rumble, S. M., Lacroute, P., Dalca, A. V., Fiume, M., Sidow, A., & Brudno, M. (2009). SHRiMP: Accurate mapping of short color-space reads. *PLoS Computational Biology, 5*, e1000386. doi:10.1371/journal.pcbi.1000386.

Sanders, W. S., Wang, N., Bridges, S. M., Malone, B. M., Dandass, Y. S., McCarthy, F. M., Nanduri, B., Lawrence, M. L., & Burgess, S. C. (2011). The proteogenomic mapping tool. *BMC Bioinformatics, 12*, 115. doi:10.1186/1471-2105-12-115.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-Seq. *Nature Biotechnology, 31*, 46–53. doi:10.1038/nbt.2450.

Wang, X., & Zhang, B. (2013). customProDB: An R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics, 29*, 3235–3237. doi:10.1093/bioinformatics/btt543.

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., MacLeod, J. N., Chiang, D. Y., Prins, J. F., & Liu, J. (2010). MapSplice: Accurate mapping of RNA-Seq reads for splice junction discovery. *Nucleic Acids Research, 38*, e178. doi:10.1093/nar/gkq622.

Warren, A. S., Archuleta, J., Feng, W.-C., & Setubal, J. C. (2010). Missing genes in the annotation of prokaryotic genomes. *BMC Bioinformatics, 11*, 131. doi:10.1186/1471-2105-11-131.

Xie, C., & Tammi, M. T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics, 10*, 80. doi:10.1186/1471-2105-10-80.

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics, 25*, 2865–2871. doi:10.1093/bioinformatics/btp394.